ONE LOOP

B2B SaaS Funnel Book

Quantitative growth from first touch to renewal



B2B SaaS Funnel Book

One Loop: Quantitative growth from first touch to renewal

Manuscript edition · 30 October 2026

This edition contains the current manuscript, including its illustrative company cases, glossary, appendices and source-verification notes. The company cases are invented teaching examples.

Generated from the working manuscript. The final index follows author review.

Contents

B2B SaaS Funnel Book

01 · Choose the growth objective

02 · Value customers and cash

03 · Model recurring growth

04 · Choose markets and accounts

05 · Choose acquisition channels

06 · Improve offers and conversion

07 · Allocate sales attention

08 · Remove buying friction

09 · Set prices and packages

10 · Forecast conversion and timing

11 · Set targets and staff capacity

12 · Activate customers

13 · Retain customers profitably

14 · Expand accounts and generate referrals

15 · Find the most valuable constraint

16 · Decide what to measure

17 · Decide what the evidence can resolve

18 · Choose whether to test or act

19 · Commit with uncertainty

20 · Fund the next best action

21 · Decide when a change needs action

22 · Diagnose what changed

23 · Learn from outcomes

24 · Run the growth system

Glossary

Appendix A: One possible implementation

Appendix B: Parameters and start values

Appendix C: The three companies: facts and timeline

Appendix D: Formula sheet

Appendix E: Sources and further reading

B2B SaaS Funnel Book

One Loop: Quantitative growth from first touch to renewal

You choose a market, acquire demand, help accounts buy and earn their renewal. Each choice changes the customer population, the data you observe and the profit you can expect. This book teaches that work as one discipline: quantitative growth. Commercial judgement, statistical modelling and experimentation form one continuous argument.

The reader can enter from marketing, analytics or data science. The practice has one standard. State the action, explain the mechanism, define the comparison and carry its uncertainty through to customer profit. You do not need to derive every equation to follow that reasoning. You do need to understand the assumptions that determine what the equation means.

Read through decisions

The book follows the commercial choices of a recurring-revenue business. The foundation sets the objective, value basis and revenue model. The next parts choose customers, acquire demand, convert it and create retained growth. The final parts develop the shared evidence and allocation methods, then show how the company learns and operates.

Part	The decisions	Chapters
Define valuable growth	Choose an objective, value customer paths and model revenue	Choose the growth objective; Value customers and cash; Model recurring growth
Choose markets and customers	Choose reachable markets and valuable accounts	Choose markets and accounts
Acquire demand profitably	Choose channels, offers and conversion changes	Choose acquisition channels; Improve offers and conversion
Convert demand with scarce capacity	Allocate attention, remove friction, price and plan capacity	Allocate sales attention; Remove buying friction; Set prices and packages; Forecast conversion and timing; Set targets and staff capacity
Create retention and expansion	Activate, retain, expand and generate referrals	Activate customers; Retain customers profitably; Expand accounts and generate referrals
Allocate resources under uncertainty	Find constraints, gather evidence and fund actions	Find the most valuable constraint; Decide what to measure; Decide what the evidence can resolve; Choose whether to test or act; Commit with uncertainty; Fund the next best action
Learn and operate	Detect, diagnose, evaluate and operate	Decide when a change needs action; Diagnose what changed; Learn from outcomes; Run the growth system

Start with *Choose the growth objective* and *Value customers and cash*. Then follow the decision you face. If spend is the question, read *Choose acquisition channels*. If seller time is scarce, read *Allocate sales attention*. If customers leave before obtaining value, read *Activate customers*.

The minimum model to carry into every decision

Define the unit before calculating: a person, account, site, opportunity or contract. People inside the same buying organisation may share an outcome, so they are not automatically independent observations. Record the time origin and the date through which outcomes are observed.

A cohort is a group that entered at a common time. Compare cohorts at equal age. An unfinished deal is not a lost deal; a customer who departs before activation is not simply an unfinished activation. These distinctions affect the denominator and what an outcome means. *Forecast conversion and timing* and *Activate customers* develop the time models.

Compare an action with an explicit alternative. Attribution assigns credit under a reporting convention. Incrementality asks what the action changes. A forecast of who will buy does not by itself show whom a sales call will persuade.

Value the additional customer path after conversion, service costs, retention and delay. An average value describes the existing mix; a marginal value concerns the next change. Keep capacity in the model. Demand that cannot be contacted or delivered does not immediately earn the value of served demand.

Uncertainty belongs to the recommendation. It includes sampling noise, measurement gaps and assumptions about behaviour. A model's precision can exceed its evidence. When a method appears before its full development, the chapter explains the immediate assumption and links to its canonical home.

How a chapter develops a decision

A chapter starts with a commercial choice and its alternative. It explains the mechanism and the observations needed to evaluate it. The maths connects those observations to economic consequences. Assumptions and failure conditions remain in the main argument. Formula callouts provide exact relationships; the surrounding prose explains what changes the decision. Deeper method sections can be revisited when needed.

The shared instruments have one full home. The technical reference chapters develop measurement, noise, experiments, decision records and monitoring through the operational choices they support. They belong to the same practice as acquisition and retention. Links let you use them without reading every reference chapter first.

Volume determines what the evidence can resolve. The standard operating window is 28 days. A year holds 13 such windows, or 364 days; the company simulation also has calendar months, whose dates remain explicit. A quarter is 91 days, or 3.25 windows. Low, medium and high volume describe the smallest changes that can be read in a useful period, not simply company size. *Decide what the evidence can resolve* gives the method.

Follow the three companies

Alder, Birch and Cedar are invented companies with one shared set of reproducible facts. All company examples are Illustration only. They report in euro and sell in Europe.

Company	Motion	Continuing decision problem
Alder	High-touch sales to manufacturing organisations	Commit sensibly when individual accounts matter and rate comparisons are weak
Birch	Inbound demand with sales assistance for finance teams	Connect acquisition quality, seller capacity and customer economics
Cedar	Self-serve product adoption with assistance for larger accounts	Connect fast experiments with slower retention and value outcomes

The chapters follow decisions, so the episodes are not in calendar order. Every original episode keeps its date and a complete narrative home. A later result described in an episode is a retrospective check, not evidence available at its earlier decision. *The three companies: facts and timeline* provides the chronological view from Year 1 January to Year 3 June.

Some episodes span more than one decision. A discovery and packaging change can share an account story; a scoring change can coincide with a capacity commitment. Keep the complete episode in view through the links rather than treating each metric as an independent experiment. Cedar's self-serve and assisted motions also have different volumes and evidence limits.

Read the evidence labels

Label	Meaning
Arithmetic	The result follows from the stated quantities and assumptions
Standard practice	A cited source supports the method or convention under its conditions
Plausible but untested	A starting assumption that needs checking in the relevant population
Assertion	A claim without supporting evidence
Illustration only	An invented case or quantity used to explain a mechanism

A mathematically correct calculation can depend on an uncertain behavioural estimate. The label does not remove that uncertainty. Keep a source's population and design beside the claim. Practitioner frameworks help formulate hypotheses; they do not measure a causal effect in your company.

Use the working parts

The glossary provides definitions and their development chapters. *Formula sheet* locates the principal calculations. *Parameters and start values* distinguishes parameters from measured facts. *Sources and further reading* identifies the sources and unresolved bibliographic details. *One possible implementation* connects the practice to a possible technical stack.

The companion Python and SQL implement the principal instruments. Code computes reproducible quantities. People decide the objective, assumptions and commitments. Agents can assist inside the controls explained in *Run the growth system*. The final index follows the author's review.

This is a handbook for decisions with imperfect evidence. Its purpose is to let you explain what you will do, why it can create value, what you still do not know and when the outcome will change your next choice.

Ø1 · Choose the growth objective

A growth decision commits money or capacity to change customer behaviour. The same decision determines which data you need, which comparison is credible and what an outcome is worth. We practise these as one discipline. We call it quantitative growth: the work of choosing, measuring and improving actions across the customer system.

The objective is expected discounted gross profit, subject to cash, capacity and risk constraints. A sign-up, a qualified account and a renewal are useful because they help us make that choice. Their value depends on who enters, what happens next and which action caused a difference.

Before using any result, name its unit, time origin, observation date and alternative action. A cohort is a group that entered at a common time. A young cohort has not had the same chance to convert or renew as an older one. An attributed customer is credited by a reporting rule; an incremental customer exists because of an action. Neither distinction can be repaired by a more elaborate dashboard. All company examples in this book are Illustration only.

One discipline, two kinds of return

A practitioner starts with a commercial question and carries it through the calculation. Choosing a channel entails assumptions about response, conversion, service cost and retention. Choosing a statistical comparison entails assumptions about which customers and actions the resulting estimate represents. Both choices determine the same investment.

The operating loop returns evidence to the next decision. A customer growth loop returns an output to a subsequent growth input. A customer can refer another account; product use can create material that attracts another buyer; retained contribution can fund another acquisition cycle. Each mechanism has a delay, a cost and a limiting resource. Learning can improve it, but learning is not itself another acquired customer.

Reforge's growth-loop argument asks how outputs return to inputs (Balfour, Winters, Kwok and Chen, 2018, *Growth Loops are the New Funnels*). Balfour's process argument asks how a business repeatedly finds and improves growth mechanisms (Balfour, *The First Step To Building A Growth Machine*). We use both questions. A loop compounds only under the conditions of its model. Saturation, declining marginal response and limited delivery capacity can change its behaviour.

Before a decision, name the path by which it changes value. When you cannot estimate that path, write a belief and the observation that could change it. The purpose of a quantitative argument is to make the commitment better, including the decision to wait or stop.

What a B2B SaaS company sells

SaaS means software as a service: the provider hosts the software, and the customer rents it over the internet. B2B means that the customer is an organisation, in which several people usually share the buying decision. The customer signs a contract for a term, often one year, and pays for that term. When the term ends, the customer decides again, and renews only while the software keeps doing its job.

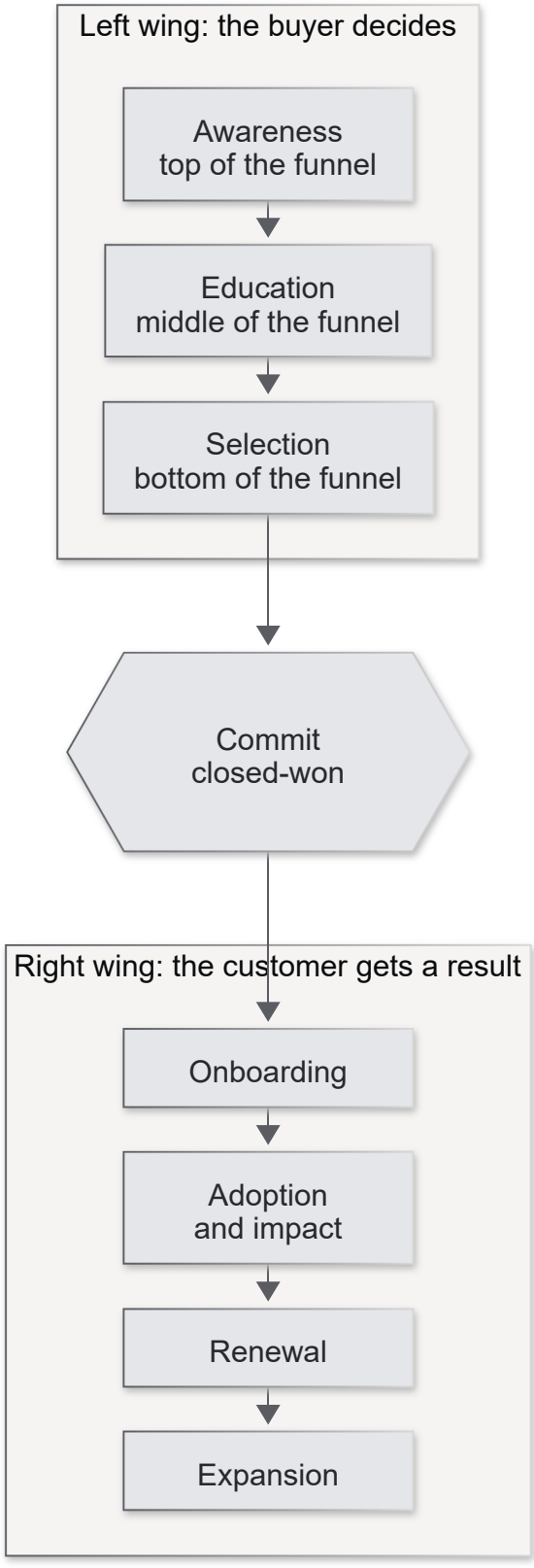
So the company sells a continued result, not a product handed over once. We measure that revenue as annual recurring revenue (ARR): the yearly value of all active contracts on one date. The chapter *Model recurring growth* defines it in full.

Two facts follow. First, every renewal is a new sale. Alder, one of our three companies, faces about 150 renewal decisions a year and wins about 43 new customers. Birch faces about 2,250 renewal decisions and wins about 590. In both, renewals outnumber new wins by more than 3 to 1. Second, the company spends money to win a customer before the customer pays it back. The chapter *Value customers and cash* prices that investment.

The classic **funnel** shows only the first half of this story. It is the sequence of stages that a stranger passes to become a customer, drawn as a funnel because fewer people reach each stage. It ends at the sale, where SaaS revenue only begins.

The bowtie: one map from first touch to expansion

The **bowtie** is that map. Winning by Design drew it for recurring-revenue companies (Winning by Design, 2023, *Revenue Architecture*). It has two wings that meet at a knot. The left wing is the buyer's path to a decision. The knot is the signature. The right wing is the customer's path to a result and to the next decision.



The bowtie: two wings and a knot. The funnel is the left wing only.

We use the term **go-to-market** for the whole bowtie read as one system: the market, the offer, the price, the channels and the sales motion (the way a company sells and serves: self-serve, sales-assisted or sales-led). Each stage is named from the buyer's side: it says what the buyer or customer has decided or can now do, not what the company did.

The left wing: awareness, education and selection

In **awareness**, a buyer recognises a problem and learns that your company may solve it. In **education**, the buyer learns how to solve the problem, what good looks like and whether the problem has priority. In **selection**, the buyer compares options and picks one.

Companies map these three stages to the **top of the funnel** (TOFU), the **middle of the funnel** (MOFU) and the **bottom of the funnel** (BOFU). Typical measures are impressions (times an ad is shown), clicks, visits and leads at the top, qualified leads in the middle, and opportunities and signed deals at the bottom.

The knot: the commit

At the knot, the buyer signs. We call this **commit**, short for mutual commit. The company now owes a result by a date. The customer owes time, data and a named person, and keeps paying while the result arrives.

The right wing: onboarding, adoption and impact, renewal and expansion

In **onboarding**, the company connects data, invites users and configures the product. The first moment of value, which we call **activation**, ends onboarding. We define it as one event that the customer can see, for example the first report built from the customer's own data. In **adoption**, the product becomes part of the routine: more of the intended users use it, and use it more deeply.

Impact is the measurable result that the customer gets. For a tool that shortens the monthly financial close, the impact is days saved at the close, not logins. In **renewal**, the customer decides whether to extend the contract. In **expansion**, the recurring revenue of a customer that stays grows, through more seats, more usage, more products or a higher price.

Stage names in common use

Inside the left wing, companies use finer names. Each name marks one event with a date.

- A **lead** is a person or a company that gave contact details, or that the company identified as a prospect. It is not yet judged worth a sales touch.
- A **marketing-qualified lead (MQL)** is a lead that meets marketing's written definition of worth a sales touch, by score or by behaviour. A typical definition: a lead that matches the ideal customer profile and viewed the pricing page twice. The **ideal customer profile (ICP)** is the written description of the companies that gain most from the product and are worth most to the company.
- A **sales-accepted lead (SAL)** is an MQL that sales has agreed to work. Some companies skip this stage, and some use the word for the SQL, so write down your meaning. Some companies put acceptance after qualification. Birch does, as *Choose markets and accounts* shows.
- A **sales-qualified lead (SQL)** is a lead that sales has contacted and judged to have a real problem and a fit.
- An **opportunity** is a potential sale with an expected value and a stage in the sales process. Date it on the day it was opened.
- **Closed-won** is an opportunity that ended in a signed contract. **Closed-lost** ended with no sale. Log both, or every rate of won deals is wrong.

After the signature, three more names matter: activated, renewed and expanded. The table puts all names on the bowtie.

Bowtie stage	Names in common use	Event that moves an item in	Owner in many companies
Awareness (top of the funnel)	visit, engaged account (a target company, or account, whose people took a defined action), lead	a person gives contact details, or the company identifies a prospect	Marketing
Education (middle)	MQL, SAL	the lead meets the written MQL definition; sales accepts it	Marketing, then sales development
Selection (bottom)	SQL, opportunity	sales judges the fit; an opportunity is opened	Sales
Commit (the knot)	closed-won, closed-lost	the contract is signed or the deal is lost	Sales, with finance
Onboarding	implementation, activated	the customer reaches the first value event	Onboarding team or customer success
Adoption and impact	active, value realised	the customer reaches the agreed result	Customer success
Renewal	renewed, churned	the term ends and the customer decides	Customer success
Expansion	expanded	the recurring revenue of a staying customer rises	Customer success or account management

The chapter *Choose markets and accounts* shows how to define each stage by one date.

Pirate metrics on the bowtie

Many teams also use the **pirate metrics**, named by Dave McClure in his 2007 talk *Startup Metrics for Pirates*. The initials of his five stages, AARRR, read like a pirate's cry. We use six stages, written AAARRR: awareness, acquisition, activation, retention, revenue and referral. We add awareness because the buyer meets your company first.

Pirate stage	Meaning in B2B SaaS	Bowtie stage	Funnel names	One example metric
Awareness	Target accounts learn that the company exists	Awareness	Top of the funnel; visit, engaged account	engaged accounts in a window (28 days; see the section on volume)
Acquisition	A stranger becomes a known lead, then an MQL	Awareness and education	Top and middle; lead, MQL	MQLs in a window
Revenue	The deal closes and the first payment is billed	Commit, then expansion	Bottom; closed-won, new ARR (the yearly value of contracts won in the window)	new ARR in a window
Activation	The customer reaches first value	Onboarding	Right wing; activated	share of new customers that reach the first value event within the activation window
Retention	The customer keeps using the product and renews	Adoption and renewal	Right wing; renewed	gross revenue retention (the share of existing ARR kept, before any growth)
Referral	Customers bring or vouch for new customers	Right wing, advocacy	Right wing; referred lead	referred leads in a window

Three points matter when you read this table.

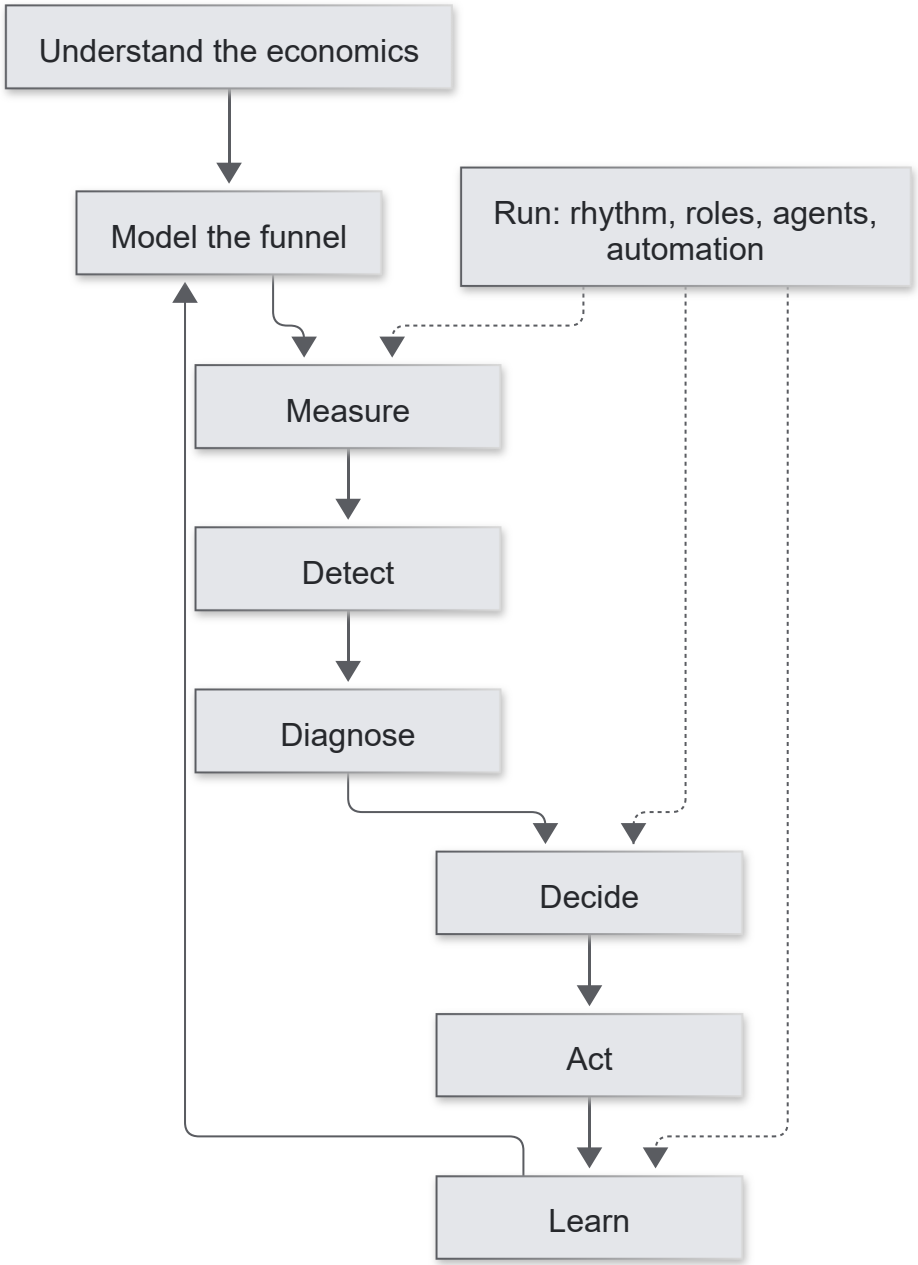
First, the order of events depends on the contract. In a product-led trial, activation comes before revenue: the user reaches value and then pays. Cedar works this way. When the contract is billed in advance, revenue comes first: the customer pays at the signature and reaches value later. Birch works this way, and the table follows that order. At Birch, the 72% of new customers who reach their first result within 30 days renew for the first time at 90%, and the others at 72% (Illustration only).

Second, counts at different stages are in different units, so they cannot be compared. Cedar counts about 800,000 visitors (people), 40,000 sign-ups (workspaces, the team spaces that people create in the product), 14,400 activated workspaces and 1,368 paying accounts in a window. A fall from 800,000 to 1,368 is not a leak. Read each rate against its own history and its own published range. Cedar's sign-up to paid rate is 3.4%. The published range for freemium and trial products is 3% to 6% (ChartMogul and ProductLed, 2026, *The SaaS Conversion Report*; a self-reported sample of 200 products).

Third, a stage with no metric has no owner. Many companies do not measure referral. Keep the stage on the map and mark the gap.

One loop of eight moves

The book runs the funnel as one **loop** of eight moves. A loop is a path by which the result of an action returns to influence the next action.



The loop of eight moves. Learning changes the model, and the loop starts again. Run surrounds the loop.

Move	The question it answers	Where the book answers it
Understand the economics	What is a customer worth, and what does it cost to win and keep one?	*Model recurring growth*, *Value customers and cash*
Model the funnel	How do volumes, rates and time make revenue?	*Choose markets and accounts*, *Forecast conversion and timing*, *Set targets and staff capacity*, *Find the most valuable constraint*
Measure	What is true, and how much noise does it carry?	*Decide what the evidence can resolve*, *Decide what to measure*
Detect	Did something change beyond noise?	*Decide when a change needs action*
Diagnose	What caused the change?	*Diagnose what changed*
Decide	What do we do, and what evidence do we need first?	*Commit with uncertainty*, *Choose whether to test or act*, *Choose acquisition channels*, *Fund the next best action*

Move	The question it answers	Where the book answers it
Act	Which lever (one change that a team can make) do we pull, and where on the bowtie?	*Improve offers and conversion*, *Remove buying friction*, *Retain customers profitably*
Learn	Were we right, and how do we get better?	*Learn from outcomes*
Run (around all eight)	Who does what, at which rhythm, with which automation?	*Run the growth system*

Three rules hold in every move. Code computes every number. AI models read text and draft prose. People decide, asynchronously by default: a written brief goes to one owner, who decides by a deadline and logs the decision. A meeting happens only when a decision needs one, for example when two owners truly conflict.

Open and closed loops

In an **open loop**, nobody compares the output with the intention. A campaign with no date to read the result is an open loop. In a **closed loop**, also called a feedback loop, someone measures the output, compares it with the intention and changes the input. A thermostat is the standard picture: the sensor reads the temperature, and the gap to the target switches the heating on or off.

Four questions test any loop. What is the output? What is it compared with? Which input does the difference change? Who has the authority to change that input? If one answer is missing, the loop is open.

Every loop also has a **loop delay**: the time between an action and the moment you can read its result honestly. A loop cannot run faster than its delay (Standard practice in control engineering). For a change to the selling process, the delay is at least the median time to the outcome plus one reading window of 28 days. At Alder that is 120 plus 28, so 148 days. At Birch it is 34 plus 28, so 62 days, and for Cedar's self-serve funnel 9 plus 28, so 37 days.

Correcting and compounding loops

A **correcting loop** shrinks a gap between output and intention, as the thermostat does. A **compounding loop** feeds its own growth. The referral loop is the usual example: a customer brings other customers, and they bring more.

Follow a compounding loop step by step. Say that each customer brings g new customers. A first round of customers comes from other sources. Each later round is g times the round before, and the rounds add up to a finite total only when g is below 1.

> **Calculation**
>
> total customers = first round ÷ (1 - g)

With g at 0.10, a first round of 100 customers gives 100, then 10, then 1, then 0.1, and the total is about 111. With g at 0.25, the total is about 133. At g of 1 or more, the rounds never shrink, and the formula does not apply. The same logic runs backwards: each customer that leaves also removes the customers it would have brought.

The argument holds under three conditions. You measured g on your own data. The value of g is stable over time. And the referred customers are not already counted in paid acquisition: if a referred customer first clicked a paid advertisement, you count that customer twice. In this book, g stays unknown until a company measures it.

Why siloed work fails

In many companies, marketing owns the left wing up to the MQL, sales owns the rest of the left wing, customer success owns the right wing, and finance counts the money. Each team has its own metric and tools, and each can do well by its own measure while the whole company fails.

Two mechanisms explain how. The first is local optimisation. Marketing raises the number of MQLs by loosening the definition. Sales raises its share of won deals by refusing the hard leads. A discount-led quarter brings in customers outside the ICP. Many of them never reach their first result, and the loss shows up a year later as churn, the loss of customers who leave. Every team's number rose, and the company lost (Plausible but untested: the size of the harm differs by company).

The second mechanism is the hand-off. A silo has two weak points that no stage owner (the person who answers for one stage) holds: marketing to sales, and sales to customer success. Records change hands there, definitions differ, and nobody answers for the gap.

The cure has three parts. The first is shared definitions: one written meaning and one date for each stage, in every system. The second is one unit of value for every team: expected gross profit over a horizon, discounted. In plain words, it is the profit that a customer should bring over five years, after the direct cost of serving it, with later money counted for less. The default discount rate of 12% a year (Gupta, Lehmann and Stuart, 2004, *Valuing Customers*) is a start value that each company sets for itself (Plausible but untested). We call this one unit of value the **value basis**. The third part is one named owner for the whole bowtie, who reads both wings together and may stop a change that helps one wing and harms the other. We call this person the **bowtie owner**. The chapter *Run the growth system* sets up that role.

From data-informed to data-centric

In a **data-informed** company, people decide, and data is one input next to experience and judgement. In a **data-centric** company, the process itself is built around data. Every stage event (the record that an item entered a stage) has a date, code computes the numbers, and each decision is logged with its expected effect and graded later. Data-centric does not mean automatic. A rule that decides alone can trap a company at a local best, because data refines what exists and rarely proposes what is new (Croll and Yoskovitz, *Lean Analytics*).

The right position depends on volume and on the cost of being wrong. When volume is low and noise is high, the decision stays data-informed: judgement leads, and data limits it. When volume is high and the decision repeats, the process becomes data-centric.

Quantitative methods also come in three kinds. A **top-down** method starts from a total, such as an ARR target or a budget, and splits it into parts, for example stage rates or channels. A **bottom-up** method starts from the units, such as accounts, deals or cohorts (groups of customers that started in the same period), and adds them up. A **hybrid** method does both and compares the answers, and the gap between them is information. The book picks one kind per question. A budget split is top-down, a cohort forecast is bottom-up, and next year's commitment is hybrid: the board target is checked against the sum of the unit forecasts.

Volume is a parameter

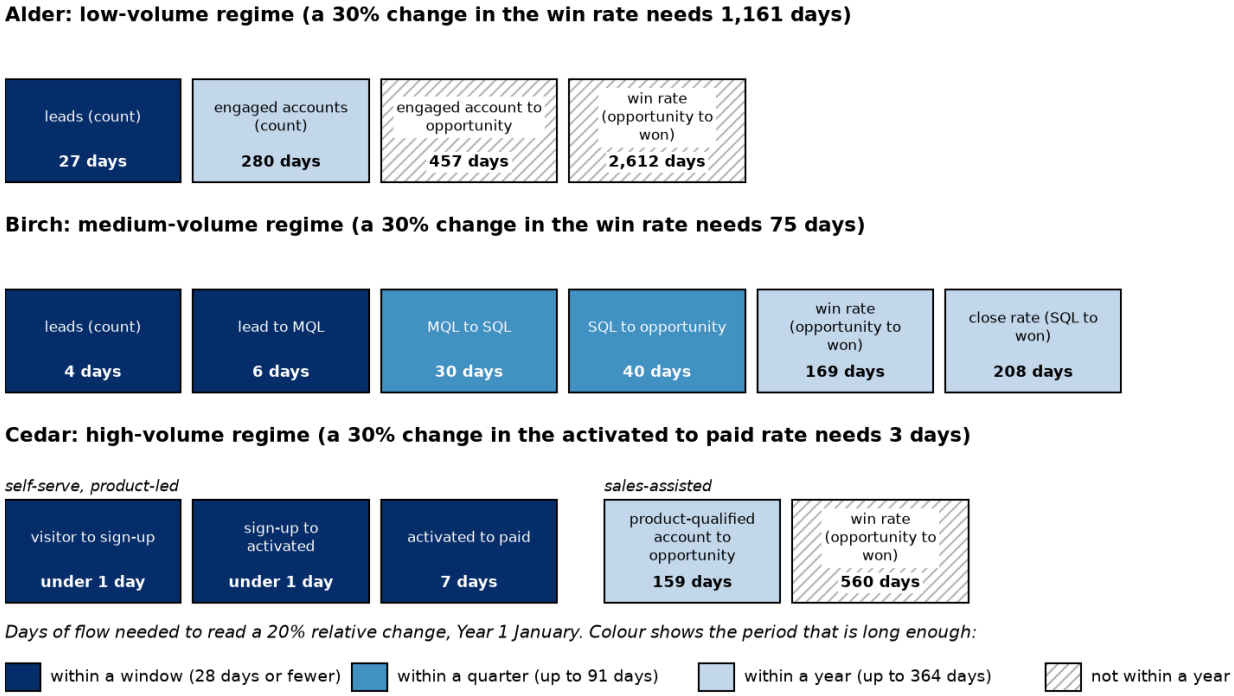
Every rate in a funnel is a sample. A rate of 22% in one period is not exactly the rate of the next period, even when nothing changed, because chance moves it. So the question is not "what is the rate?" but "what change can this volume show?" The chapter *Decide what the evidence can resolve* gives the full method.

The basic period of the book is the **28-day window**. Each window is compared with its expectation, as *Decide when a change needs action* shows. Four whole weeks hold the same mix of weekdays each time, so the weekday pattern cancels. A year holds 13 windows (364 days). A quarter of 13 weeks (91 days) holds 3.25 windows, not 3. The 28-day window is a convention of this book (Arithmetic). The facts of the three companies count calendar months. A weekly chart is allowed, but it is only a **display grain**: a setting for how a chart or table looks, never a second comparison window.

A rule of thumb tells how many items must enter a stage to read a change in its rate. For a rate p and a relative change r, for example 0.20 for a 20% change, each of two equal periods needs n items (Standard practice: van Belle, 2008, *Statistical Rules of Thumb*; Kohavi and others, 2014, *Seven Rules of Thumb for Web Site Experimenters*, KDD).

> **Calculation**
>
> n per period = 16 × (1 − p) ÷ (p × r × r)
>
> days to read = 28 × n per period ÷ items entering the stage per window

We call this the **readability rule**. Half the change needs four times the items. Take Birch, which has about 1,128 MQLs a window. Its MQL to SQL rate is 25%. A 20% change needs 1,200 MQLs per period, so Birch reads it in about 30 days. Its win rate, the share of opportunities that end in a signed contract, is 32%. A 20% change needs 850 opportunities per period. Birch has about 141 a window, so it needs about 169 days, which is about 6 months. Alder, with a win rate of 22% and about 15 opportunities a window, needs about 2,612 days, which is about 7 years. So a stage reads a change within a window, within a quarter, within a year, or not within a year.



The barometer strip: how long each stage of each company needs to read a 20% change at Year 1 January.

A company also needs one label for its volume: the **volume regime**. We count the days of flow that the last stage rate before revenue in the main motion needs to read a 30% change. For a sales-led company that rate is the win rate; for a self-serve product (the customer buys and starts alone, without a seller) it is activated to paid. High volume is 28 days or fewer. Medium volume is more than 28 and up to 91 days, one quarter. Low volume is more than 91 days. For a win rate of 25%, high volume starts at about 533 opportunities a window, and medium volume at about 164. By this rule, Alder is low (1,161 days), Birch is medium (75 days) and Cedar is high (3 days, through its self-serve motion). The 30% change and the two limits are a convention (Plausible but untested), and each company may set its own.

Each example in this book opens with one sentence of at most 30 words called the **barometer**. It names the company and the date, the new customers per window, where that count sits against a published reference, the volume regime, and the deepest stage that reads a 20% change within the period that fits the regime. The period is a window at high volume, a quarter at medium volume and a year at low volume. One short second sentence may add detail. For a self-serve company such as Cedar, a new customer is a new paying account. The reference is the **derived typical** number of new customers. Published medians show growth of 27% a year and retention losses of about 9%. So a company adds about 36% of its opening ARR as gross new ARR. About 65% of that, or 23.4% of opening ARR, comes from new customers (Benchmarkit and Pavilion, 2024, *2024 B2B SaaS Performance Metrics Benchmark Report*; SaaS Capital, 2023, *SaaS Retention Benchmarks for Private B2B Companies*). The medians come from different samples, so the result is only an order of magnitude.

> **Calculation**
>
> derived typical new customers per window = 0.234 × ARR ÷ (13 × yearly contract value)

Volume is therefore a parameter. Every conclusion in this book that depends on it says at which volume it holds.

How the book labels evidence

Five plain labels tell you how far to trust a claim. **Arithmetic** means that it follows from the numbers. **Standard practice** means that many practitioners use the method. **Plausible but untested** means that it is reasonable but not tested on your data. **Assertion** means that it is stated without support. **Illustration only** means that it is a made-up number that shows a mechanism. All company numbers in this book carry the last label.

Some arguments have no source, such as the compounding gain above. They carry no label. Instead, the text shows each step, shows how each step reduces uncertainty, and states the conditions under which the argument holds.

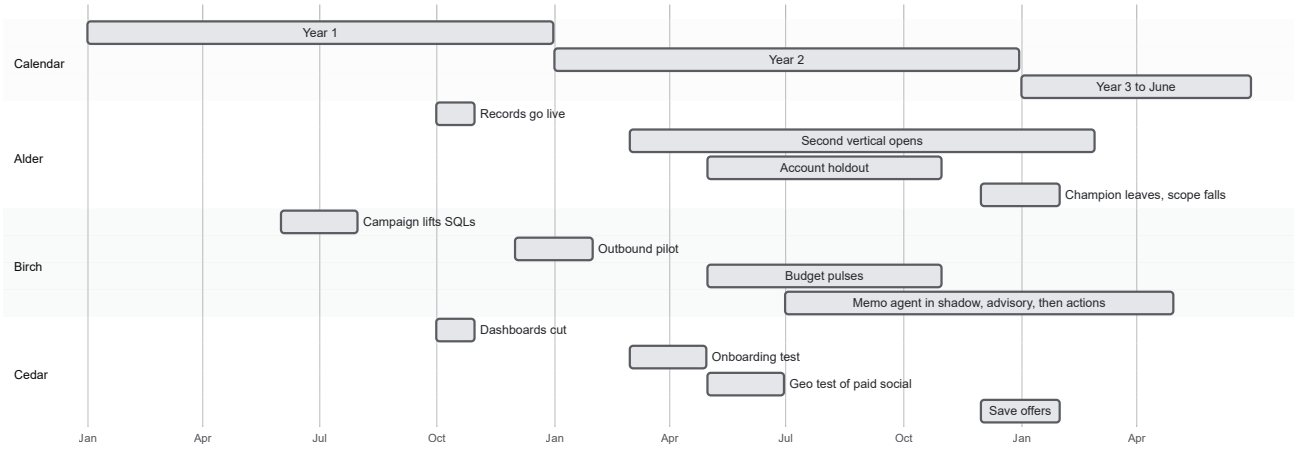
Meet the three companies

Three invented companies run the loop beside you for 30 months, from Year 1 January to Year 3 June. All report in euro and sell in the EU. Code computes every number about them from one facts file, with rates inside published ranges. The most quoted funnel tables repeat one vendor's data, so we treat every benchmark as a range, never as a target.

	Alder	Birch	Cedar
Volume	Low	Medium	High
What it sells	Quality-management software for manufacturing plants	Finance software for the monthly close and spend control	Team scheduling and collaboration for small and medium businesses
Motion	Sales-led, high touch; account-based marketing (marketing aimed at a chosen list of target accounts), events, partners	Sales-assisted inbound: demo request, sales development, account executive	Product-led: free plan plus trial and self-serve plans; a sales-assist team for larger accounts
Contract and billing	Annual and three-year contracts, billed yearly in advance	Annual, billed in advance, renewing automatically	Monthly and annual plans paid by card; assisted deals are annual
ARR at the start (million euro)	12	22	37
Paying customers	190	2,250	26,640
Yearly value of a new contract (euro)	60,000	9,600	821 self-serve; 16,000 assisted
New customers per window	3.3	45.1	1,383 (1,368 self-serve, 15 assisted)
Days to read a 30% change in the last rate before revenue	1,161	75	3
Arc of the story	Learning to decide well when the numbers are too few to prove anything	From reports and arguments to a priced decision loop	High volume is not high certainty

Cedar has two motions. Its company regime follows the motion with the most new customers: self-serve. At the trend level, its assisted motion has 60 opportunities a window and needs about 249 days to read a 30% change in its win rate. A high-volume company can face a low-volume decision.

The examples follow one shared calendar, with dates in words, such as Year 2 March. Alder and Birch slow down in December. Cedar has a peak in January and a dip in summer.



The shared 30-month calendar with the main events of each company.

What changes with volume

The chapter makes four conclusions that depend on volume.

Conclusion	Where it holds
The first gain of the loop is shared definitions and explicit judgement, not rate dashboards.	Low volume: no rate below the lead reads within a year (Alder).
A fight between teams can be settled by a rate read over a quarter, once the definition is shared.	Medium volume: MQL to SQL reads within a quarter at Birch, but the close rate (SQL to won) needs about 7 months.
Local optimisation is easy, because every test reads in days. The risk moves to blindness about revenue.	High volume: trial to paid reads in about 8 days at Cedar.
The right wing holds more decisions than the left wing.	Wherever renewals outnumber new wins, as at Alder and Birch.

Alder, Year 1 January: three customer counts for one company

Alder, Year 1 January: about 3.2 new customers a window, 0.9 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.2 years to read even a 30% change, and the lead count reads a 20% change in 27 days.

A new chief executive asks for a data-driven funnel and puts one question to the three teams: how many customers do we have? Marketing says 214. The sales team's customer relationship management system (CRM), where sales records its deals, shows 203. Finance, which counts the contracts it bills, says 190. The marketing count is 24 above billing, which is 12.6% higher. The CRM count is 13 above. Then the three teams argue about pipeline (the open deals that sales may still win), and each team argues from its own base.

Such gaps usually come from definitions that nobody wrote down. One team counts every account that ever signed. Another counts signed deals that are not yet live. Finance counts what it bills. Only one count ties to Alder's 12 million euro of ARR: the 190 paying customers in billing. It becomes the reference. The other two counts get a written definition and a reconciliation (a check that two sources give the same total). The chapter *Decide what to measure* shows how.

On the bowtie, marketing works the awareness end. Sales works the bottom of the funnel and wins about 3 customers a window. Three customer success managers work the right wing. Nobody owns the hand-offs or reads the two wings together.

The loop shows where the decisions are. Alder faces about 150 renewal decisions a year and wins about 43 new customers, which is 3.5 renewals for each new customer. Over Year 1, expansion adds about 1.96 million euro of ARR, and new customers add about 2.62 million. Expansion is therefore 43% of the ARR that Alder adds, and the pipeline argument is about the other 57%. The right wing holds more decisions than the left wing, and the three-way argument ignores it.

With about 3 new customers a window, one deal more or less moves the count of wins by about 30%. Only the lead count reads a 20% change in a window, in 27 days. The engaged-account count needs 280 days, and no rate reads within a year. A dashboard of conversion rates would show numbers that move with chance, and the team would chase them. So the first gain of the loop is not a rate. It is four plain things:

- one written definition of a customer, tied to billing;
- one named owner for the whole bowtie, for now the chief executive;
- one written monthly read of both wings, with judgement stated openly;
- stage dates recorded in the CRM, which has no stage history yet. They go live in Year 1 October, as the chapter *Decide what to measure* shows.

Here is what stated judgement looks like at this volume. Alder expects about 11 new customers in the first quarter, with an 80% range of 7 to 15 (Arithmetic on the company facts). The range is wide, and its width is the message: 7 wins and 14 wins would both be normal. A monthly read that says so is more honest, and more useful, than a conversion rate with a decimal point.

At low volume, the first gain of the loop is shared definitions and explicit judgement, not rate dashboards.

Birch, Year 1 January: "sales is slow" and "leads are bad"

Birch, Year 1 January: about 47 new customers a window, 1.1 times the derived typical; medium volume: a quarter reads a 20% change down to the SQL to opportunity rate. The win rate reads a 30% change in 75 days, and the close rate reads a 20% change in about 7 months.

Marketing says that sales is slow to follow up leads. Sales says that the leads are bad. The numbers disagree from the start, because two systems define the MQL differently. Marketing automation counts 1,128 MQLs a window. The CRM counts 902. The CRM figure is 80% of the marketing figure, and the 226 MQLs in between belong to nobody.

The table shows Birch's chain at the trend level, with the days that each rate needs to read a 20% change.

Stage	Per window	Rate into the stage	Days to read a 20% change
Leads	2,821		4 (a count)
MQLs	1,128	40%	6
SQLs	282	25%	30
Opportunities	141	50%	40
New customers	45	32% (win rate)	169

Stage	Per window	Rate into the stage	Days to read a 20% change
Close rate, SQL to won	282 SQLs	16%	208

The MQL to SQL rate of 25% is inside the published range of 13% to 51%. The SQL to opportunity rate of 50% is at the top of the range of 35% to 50% (First Page Sage, 2025, *B2B SaaS funnel conversion benchmarks*, and other vendor pages). Published stage benchmarks conflict and rest on vendor data, so neither team can win by pointing at a benchmark.

Can rates settle the argument? The MQL to SQL rate reads a 20% change in about 30 days. So a claim that the leads got worse can be checked within a quarter, once both systems use one MQL definition. The close rate needs about 208 days, which is about 7 months. A claim about it cannot be checked for two quarters or more.

Now look at "sales is slow". Birch has 10 account executives, and each takes about 30 SQLs a window. Capacity is 300 SQLs a window. In January, about 296 SQLs arrive. The seller load is 98.7%, and it is 94.0% at the trend level. At that load, new SQLs wait. So "sales is slow" may be true for a reason unrelated to lead quality. "Leads are bad" may be true as well, since the definition is unclear. Both claims may be partly right, and the rates cannot decide between them.

One value basis can. On the basis of the chapter *Value customers and cash*, a new Birch customer is worth 31,538 euro in expected gross profit over five years, discounted at 12% (Illustration only). With a close rate of 16%, an SQL is worth about 5,046 euro, and an MQL is worth about 1,262 euro. Now marketing can say what an extra MQL is worth, and sales can say what a waiting SQL costs. At 98.7% load, an extra MQL does not add 1,262 euro, because an SQL that nobody contacts cannot close. The chapter *Find the most valuable constraint* turns this into a calculation.

The first pass of the loop at Birch makes four things. It writes one MQL definition and applies it in both systems. It puts a date on each stage, so the hand-off can be timed. It gives both teams one number for seller load. And it prices an MQL and an SQL in euro. The two teams then argue about the same facts. In Year 1 May, Birch adds a sales-accepted stage to the chain, as the chapter *Choose markets and accounts* shows.

A silo argument cannot be settled by rates that the volume cannot read. One loop and one value basis can settle it.

Cedar, Year 1 January: sign-ups up, net revenue flat

Cedar, Year 1 January: about 1,600 new paying accounts a window, 1.7 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 32 days; the assisted motion, with 72 opportunities a window, needs about 8 months to read a 30% change.

Cedar's growth team of 12 runs about 15 experiments a month. The team judges each one by sign-ups per window, and the scoreboard shows lift after lift. Finance watches revenue and sees net new ARR, the change in ARR after all gains and losses, that is flat month by month. Nobody has put the two views side by side. Nine dashboards show 340 metrics, and the test register is poorly kept.

Apply the four loop questions to the tests. The output is sign-ups per window. The intention is a lift. The inputs are the page, the copy and the offer. The team has the authority to change all three. Every answer is clear, so the loop is closed, but on the wrong output. Sign-ups sit several rates away from revenue. On the pirate map, the tests work on acquisition and activation, and nobody tests revenue or retention.

The January month shows what finance sees.

Movement of ARR in Year 1 January	Million euro
New ARR	1.59
Expansion	+0.52
Contraction	-0.12
Churn	-0.76
Net new ARR	1.23

Read the table from the right wing. Contraction is ARR lost from customers that stay but pay less; churn is ARR lost from customers that leave. Of every euro of new ARR, 55 cents leave the existing book in the same month. Cedar wins 1,647 new customers in the month and loses 744. A year later, in Year 2 January, sign-ups are 15% higher, but net new ARR is only 3.2% higher. Losses from the growing book rise 30.6%.

A sign-up test works on the first line only. Suppose that a test lifts sign-ups by 10%. That adds 4,800 sign-ups a window. If they pay at the average rate of 3.4%, that is about 164 more paying accounts a window. In January it would raise self-serve new ARR by about 134,000 euro a month, which is 11% of net new ARR. If finance cannot find that gain, the extra sign-ups do not pay, or they leave early, or other changes cancel it, or it hides in the seasonal swing. The scoreboard cannot say which.

The facts show how early customers leave. 70% of new self-serve plans are monthly. Monthly plans lose 11% of their customers in the first month, 8% in the second and 6% in the third. After three months, only 77% of a monthly cohort remains. A test reads in days on sign-ups and takes months to show in retention. The loop delay is short on the left wing and long on the right wing. A team that reads only the short one will improve only the short one.

Volume makes the blindness easy. Visitor to sign-up reads a 20% change in less than one day of flow, so every sign-up test gives a quick answer. The January peak also flatters: January sign-ups run 20% above the trend level, and the chapter *Set targets and staff capacity* explains why.

One value basis shows the cost of the blindness. On the basis of the chapter *Value customers and cash*, a paying account is worth about 1,683 euro in expected gross profit over five years. At a paid rate of 3.4%, an average sign-up is worth about 58 euro. A test that adds 4,800 sign-ups looks like a gain of 10%, and at the average value it is worth about 276,000 euro. Suppose the extra sign-ups come from low-intent visitors and are worth a tenth of the average. The same test is then worth about 28,000 euro (Illustration only). The scoreboard shows the same lift in both cases.

The first pass of the loop at Cedar gives each test a second read that links to money, such as paying accounts still active after three months. That read arrives later, so each test record names its date. A named owner then reads the left and the right wing together. In later chapters Cedar reads weekly cohorts and cuts nine dashboards to three.

High volume makes local optimisation easy and system blindness likely.

The decision to carry forward

Choose one economic objective and state the constraints that can override it. Follow every local improvement through acquisition, delivery and retained value. The operating loop updates the beliefs behind those choices; a growth mechanism must also explain how an output becomes a later input.

02 · Value customers and cash

Should you acquire another customer, retain an existing one or conserve cash? You cannot compare those options with revenue alone. Their service costs, timing and survival differ. This chapter builds the economic quantity that every later decision uses: expected gross profit over a stated horizon, discounted to a common date.

Start with an action and its alternative. A customer value estimate describes the cash-generating path of a customer; an action value describes the difference between paths caused by a decision. They are related quantities, and they answer different questions. Buying access to a likely customer does not earn their whole value if they would have bought anyway.

You will connect acquisition and delivery costs to survival, cash recovery and uncertainty. The aim is to find the assumptions that change your choice. The three company episodes show why an attractive average can conceal a poor investment.

Gross margin and gross profit

Gross profit is the euro that remains from revenue after the cost of revenue. Gross margin is the same thing as a percentage of revenue. The book keeps them apart: all money in the book is gross profit, and margin is the rate that turns revenue into it.

> **Calculation**
>
> gross profit = revenue − cost of revenue
>
> gross margin = gross profit ÷ revenue

The cost of revenue is the cost of delivering the product to customers who already pay. Hosting, software built into the product, customer support, and onboarding and delivery work all sit there. Practice differs for customer success, and no accounting rule fixes it. In a survey of private software companies, the average firm put 36% of its customer success cost in the cost of revenue and 57% in sales and marketing. The same survey shows that 84% of support cost sits in the cost of revenue (KeyBanc Capital Markets, 2021, *Private SaaS Company Survey*; Standard practice for the split, no binding rule).

The book splits by function. Work that helps customers use the product goes into the cost of revenue. Work that owns a renewal or an upsell (a sale of more seats or a higher plan to a customer) goes into sales and marketing. If one team does both, split its cost by time and write the rule down. Cost that serves existing customers but sits outside the cost of revenue is the *cost to serve*: the renewal work and the expansion work.

Cost line	Where it sits	Why
Hosting, software in the product, support	Cost of revenue	Grows with customers and keeps the service running
Onboarding and implementation delivery	Cost of revenue	Delivers what the contract promised
Customer success, adoption work	Cost of revenue	Helps customers get value
Customer success, renewal and upsell work	Cost to serve	Owns a commercial transaction

Cost line	Where it sits	Why
Account managers for expansion	Cost to serve	Owns a commercial transaction

Two cautions apply. First, count each euro in one place only: customer success that sits in the margin is not deducted again from the value of a customer. Second, a cut to customer success lifts the margin at once, while the harm shows up at renewal, months later. Never manage to the margin alone.

At Year 1 April, the three companies have these margins, under the rule above: Alder 79.9%, Birch 80.5% and Cedar 81.4%. A published benchmark puts the median subscription gross margin at 79% (Benchmarkit and Pavilion, 2024, *2024 B2B SaaS Performance Metrics Benchmark Report*, in US dollars). State the rule beside every margin you print. Without the rule, a margin of 80% can mean two different things.

Customer acquisition cost

Customer acquisition cost (CAC) is the cost of winning one customer. The one word hides three different figures, and mixing them is a common error.

Figure	Cost in the top	Customers in the bottom	Use
Paid CAC	Paid media spend	Customers won from paid sources	Media efficiency
Fully loaded CAC	Paid media plus the people, programmes, tools and overhead of marketing and sales that win new customers	All new customers	Unit economics, payback
Cohort paid CAC	Paid media spend of one entry period	Customers that this period's paid leads produce, however long they take	Channel decisions

The rule that holds them together is the *population rule*: the cost and the customers must come from the same population. If the top counts paid spend, the bottom counts customers that the paid spend produced. If the top counts all new-customer cost, the bottom counts all new customers. A cost per lead on paid leads only must never sit beside a CAC that divides by all customers.

Fully loaded CAC needs one more rule: it counts the cost of winning new customers, not of keeping or growing existing ones. Each cost line therefore has a new-customer share. An account executive who closes new deals is 100% new-customer cost. An account manager who works only on expansion is 0%. A sales leader who does both carries a share of each, for example 85% and 15%. Finance keeps the shares in a table and changes them only with a note.

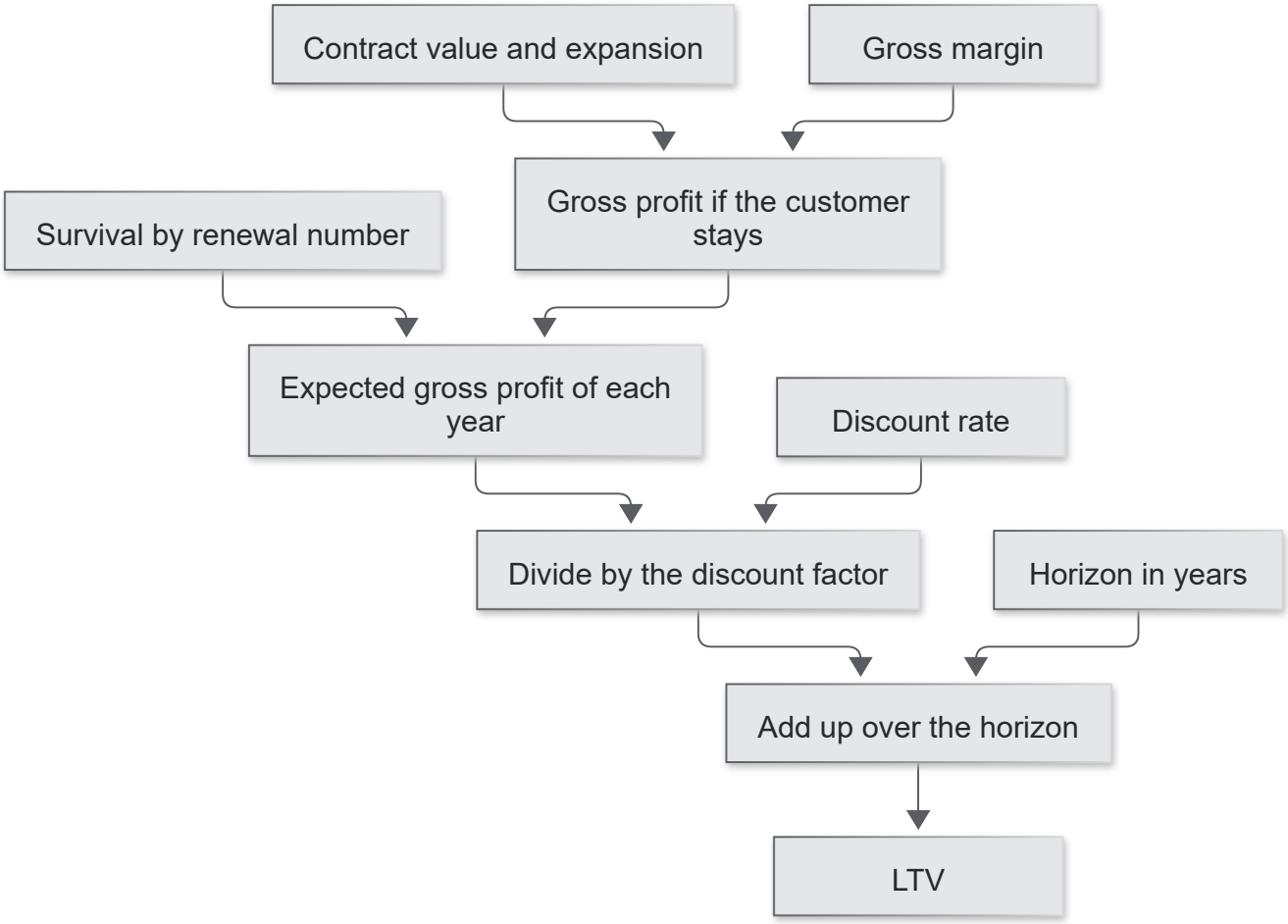
Divide by the new-customer ARR instead of the count of customers, and you have the *new-customer CAC ratio*: the cost of each euro of new annual recurring revenue. It compares companies with different contract sizes. At Year 1 April the ratio is 1.33 for Alder, 1.27 for Birch and 0.72 for Cedar. The published median is 1.76 (Benchmarkit and Pavilion, 2024, in US dollars, with no currency conversion), so all three sit below it.

Cohort paid CAC answers a different question. A customer who signs in June may have clicked an advertisement in March. Divide June's spend by June's customers and you pair the spend of one crowd with the customers of another. The cohort figure pairs March's paid spend with the customers that March's paid leads produce. It matures slowly, so you read it with an age and an interval. *Forecast conversion and timing* explains the ageing.

Lifetime value as expected discounted gross profit

Lifetime value (LTV) is what one new customer is worth in gross profit over the years that it stays. It has three ingredients: the chance that the customer is still there in each year, the gross profit of a year in which it is there, and a discount. A euro next year is worth less than a euro today.

The chance of being there comes from survival by renewal number, as *Model recurring growth* defines it. At Birch, the renewal rate is 85% at the first renewal, 90% at the second, then 92%, then 94%. Survival is the product of those rates: 100% in the first year, 85% in the second, 76.5% in the third, 70.4% in the fourth, 66.2% in the fifth.



Calculation

LTV = sum over years k = 0 to H − 1 of [S(k) × GP(k) ÷ (1 + r)^k]

S(k) is the share of customers still active in contract year k, with S(0) = 1.

GP(k) is the gross profit in year k of a customer that is active.

r is the discount rate and H the horizon in years.

The formula takes each year's gross profit at the start of the year, which fits contracts billed annually in advance, as Alder and Birch bill. Each year is weighted by the chance that it happens.

The table below works the formula for one Birch customer with a contract value of 9,600 euro and a gross margin of 80.5%. The column "with net expansion" lets the contract of a customer who stays change at each renewal. It falls by the 3% contraction at renewal and grows by the 12.8% a year of expansion that Birch measures. Together the two lift the contract by 9.4% a year.

Contract year	Share still a customer	Expected gross profit, contract flat	Expected gross profit, with net expansion	Discount factor	Present value, with net expansion
1	100.0%	7,724	7,724	1.000	7,724
2	85.0%	6,565	7,184	0.893	6,415
3	76.5%	5,909	7,076	0.797	5,641
4	70.4%	5,436	7,124	0.712	5,070
5	66.2%	5,110	7,328	0.636	4,657
Total					29,506

With the contract held flat, the present values add up to 25,412 euro. Net expansion adds 4,094 euro, which is 13.9% of the total. Show expansion on its own line: it comes from a different team and cost, and it is the least certain part.

The sum of 29,506 euro is a yearly sketch. A monthly model with all of Birch's contract rules gives 31,538 euro, 6.9% more. Most of the gap has one cause: the customers who leave hold smaller contracts than average. With leavers of average size, the same model gives 29,690 euro, only 0.6% above the sketch. The book uses 31,538 euro from here on.

The quick check

A shortcut replaces the survival curve with one constant yearly retention, as a check. Birch's gross revenue retention of 89.9%, applied as a constant, gives 26,100 euro for the flat contract, 2.7% above the survival-based 25,412 euro. The two agree because Birch's renewal rates rise only a little with age. The shortcut fails when the curve is far from flat: Cedar loses 11% of its monthly customers in the first month and 2.2% a month at the end. Do not turn gross revenue retention into a lifetime of 9.9 years either; *Model recurring growth* explains why.

One value basis for every decision

The basis

Every euro in this book is expected gross profit over a horizon, discounted at the company's rate. This is the *value basis*. It has two parameters: inputs that each company sets and the data do not fix. The book gives a start value for each, and you tune it on your own data.

Parameter	Start value	Range	Evidence label
Discount rate	12% a year	8% to 20%	Standard practice for the method; Plausible but untested for the number 12

Parameter	Start value	Range	Evidence label
Horizon	5 years	3 to 5, as a sensitivity	Plausible but untested; five years stops where cash flows become guesses

The rate stands for the cost of capital and for the risk that the survival curve is wrong. It does not stand for churn: survival already carries churn, so a high rate on top of it counts the risk of leaving twice.

One study took 12% as the middle of the 8% to 16% that finance textbooks give (Gupta, Lehmann and Stuart, 2004, *Valuing Customers*, Journal of Marketing Research). Damodaran's data for software firms put the cost of capital at 9.3% in January 2026, in US dollars (Damodaran, 2026, *Cost of Equity and Capital (US)*). Practitioners use 15% for private firms that grow in a predictable way and 20% for early-stage private firms (Skok and Reiss, undated, *Discount rate for DCF*, Matrix Partners). Do not use venture-capital target returns, because survival already counts failure. No paper tests which rate is right for a software firm, so the label for 12% is Plausible but untested.

To set your own rate, ask finance for the rate that the company uses for investment decisions. If none exists, start from the sector cost of capital and add 0 points for a mature profitable firm, 3 to 5 points for a private firm that is scaling, and 8 to 10 points before scale. Round it, fix it for 12 months and never tune it until a project passes. Print the sensitivity at four points below and four points above.

Birch, value of a new customer	8%	12%	16%	20%
Five-year horizon	34,295	31,538	29,164	27,106
Change against 12%	+8.7%	0	−7.5%	−14.1%

The whole plausible range moves the value by 21.0% of the value at 8%. A horizon of three years instead of five, at 12%, gives 20,626 euro, which is 34.6% lower. The horizon moves the value more than the whole rate range does. A year cut off at the end removes a whole year of expected gross profit, while a higher rate only shrinks each year a little. For Cedar the same cut is 27.0%, smaller, because its customers leave early and the late years carry less of the value.

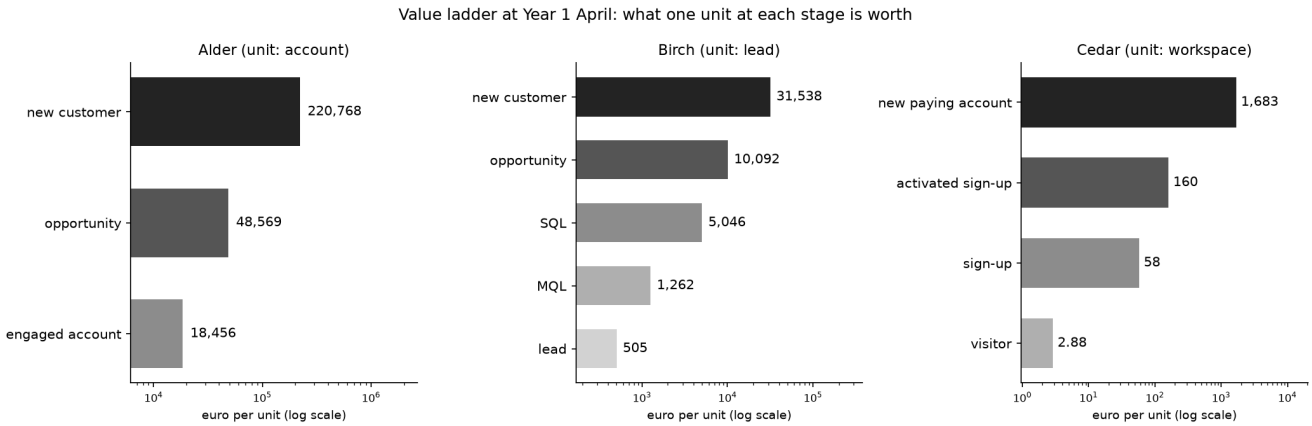
One point of retention moves the value a little more than one point of discount rate: 1.76% against 1.43% at a retention of 90% (Arithmetic). Measure retention well.

The value ladder

The *value ladder* gives the euro value of one unit at each stage of the funnel.

> **Calculation**
>
> value of a unit at a stage = chance that the unit becomes a customer × value of one customer
>
> chance = product of the step rates from that stage to the customer

At Birch, the step rates (the share of units that pass from one stage to the next) at Year 1 April are: lead to MQL 40%, MQL to SQL 25%, SQL to opportunity 50% and opportunity to won 32%. One new customer is worth 31,538 euro. One opportunity is worth 32% of that, 10,092 euro. One SQL is worth half of the opportunity, 5,046 euro. One MQL is worth a quarter of the SQL, 1,262 euro. One lead is worth 40% of the MQL, 505 euro.



Value of one unit at each stage for the three companies at Year 1 April, on a log scale.

Suppose a change lifts the MQL-to-SQL rate from 25% to 26%. Birch has 1,150 MQLs per window, so the change adds 11.5 SQLs. At 5,046 euro an SQL, that is worth 58,037 euro per window. A rise of one point looks small, but it is a relative gain of 4.0%. With a price on each unit, a marketer and the finance owner can discuss the change in one currency.

Three limits apply. First, the ladder assumes that later rates stay the same. If sellers are at capacity, extra SQLs wait and do not become customers; *Allocate sales attention* prices this. Second, do not compare rungs across units: an account at Alder and a lead at Birch are different things, as *Choose markets and accounts* shows. Third, time passes between stages. A cycle median of 34 days at Birch reduces the value by about 1% at 12%, which you can leave out. You cannot leave it out for a cycle of 150 days.

Cedar's ladder is for its self-serve motion: a new paying account is worth 1,683 euro, an activated sign-up 160 euro, a sign-up 58 euro and a visitor 2.88 euro. The assisted motion has its own ladder, with its own contract value and rates.

Payback and ratios

Accrual payback

Payback is the time that the gross profit of a customer needs to repay the fully loaded CAC. *Accrual payback* counts gross profit as it is earned, month by month.

> **Calculation**
>
> accrual payback in months = fully loaded CAC ÷ (contract value × gross margin ÷ 12)
>
> equivalent form: 12 × new-customer CAC ratio ÷ gross margin

At Year 1 April the accrual payback is 20.0 months for Alder, 18.9 for Birch and 10.6 for Cedar. For Cedar's self-serve motion alone it is 12.0 months, and for its assisted motion 4.1 months. The assisted figure leaves out the product-led acquisition behind its product-qualified accounts, so do not read it alone.

Payback needs a limit to be useful, and the limit depends on the segment. The standard targets are 12 months for small-business sales, 18 months for mid-market and 24 months for enterprise (Bessemer Venture Partners, 2021, *Scaling to $100 Million*; Standard practice). The book calls this limit the *payback ceiling*. Alder, which sells to manufacturing plants, uses 24 months, Birch 18 and Cedar's self-serve motion 12.

Do not compare your payback with a "median payback" from a report. Published medians of 16 to 18 months come from different samples and definitions: the formula applied to the published median CAC ratio of 1.76 and a margin of 79% gives 26.7 months (Arithmetic). Use the segment ceilings as limits and your own history as the baseline.

Cash payback

Contracts billed in advance bring cash before the service is delivered. *Cash payback* counts the cash that arrives, not the profit that is earned. We use the invoice times the gross margin as the cash gross profit of each invoice, as a simple convention. A customer pays the first invoice when the contract starts, and a renewing customer pays the second invoice twelve months later.

The first invoice carries 47,965 euro of gross profit, which is 60% of its fully loaded CAC of 79,782 euro. A quarter of Alder's new deals are three-year contracts that cannot leave at the first renewal, so the chance of a second payment is 0.75 × 0.85 + 0.25 × 1.00 = 88.8%. The expected cash after the second invoice is 90,534 euro, which exceeds the CAC. Expected cash payback is therefore 12 months, against 20.0 on the accrual basis. Birch gets the same result: 7,724 euro from the first invoice and 14,289 euro expected after the second, against a CAC of 12,160 euro.

Compare the accrual figure with the segment ceiling. The cash figure tells the finance owner how long the company funds growth itself.

Payback is a limit, not a goal

Payback is a liquidity constraint: a company that funds growth from its own cash needs payback inside a ceiling. But a project can pay back fast and add little value, and another can pay back slowly and add much. The objective is expected gross profit on the value basis, and payback is a feasibility test next to it. The decision card (one page that prices one decision and its options) carries this test, as *Commit with uncertainty* shows. The chance of breaching the ceiling must stay below a limit. The start value is 20% (Plausible but untested), and you tune it on your own history.

The LTV to CAC ratio and the magic number

The LTV to CAC ratio divides the value of a customer by the CAC of the same customer. Always name the CAC in the denominator. On the value basis, with the fully loaded CAC of Year 1 April, the ratio is 2.77 for Alder, 2.59 for Birch and 2.65 for Cedar's self-serve motion. An old rule says to aim above 3 (Skok, undated, *SaaS Metrics 2.0*, a practitioner essay). That rule was set for undiscounted values, so a discounted value gives a lower ratio for the same company. A ratio is not a decision rule.

The SaaS magic number is a quick efficiency check: the growth of recurring revenue in a quarter against the sales and marketing spend of the quarter before. With net new ARR and the January run-rate as a stand-in for the previous quarter, Birch scores 0.75 for the first quarter of Year 1, right on an older alarm floor of 0.75. The number blends new, expansion and lost revenue and ignores the sales cycle, so use it as a coarse alarm, never as a basis.

Every number carries its noise: Monte Carlo simulation

Take the figure that the board prints: payback of 20.0 months for Alder. It is an average. Alder wins 3.4 customers in an average window, and a window with two customers gives a different figure from a window with five. A model fed with average inputs returns the output of the average case, which is not the average output. Savage called this the flaw of averages (Evans, 2017, *Business Analytics: Methods, Models, and Decisions*, chapter on risk analysis).

Monte Carlo simulation is the remedy. It computes the output for many random sets of inputs, and the spread of the outputs shows how uncertain the result is. The method has five steps.

1. List the inputs of the calculation.
2. Give each input a distribution: the set of values it may take, with the chance of each.
3. Draw 10,000 sets of inputs from the distributions. With this count, the value that you read as the 10th percentile truly sits between about the 9.4th and the 10.6th (Arithmetic).
4. Compute the output for each set.
5. Read the spread: the 10th, 50th and 90th percentiles, and the chance that the output crosses a limit.

P10, P50 and P90 are the values below which 10%, 50% and 90% of the simulated outcomes fall. P50 is the median, and P10 to P90 is a range that holds 80% of the outcomes. Fix the seed of the random numbers and log it, so that anyone can repeat the run. Round the outputs, because the draws are a sample too.

Inputs as distributions

In the parameter table, each input gets a distribution: a family, a centre and range, a source and an owner. The table below is for one Alder window of new-customer acquisition. Owners are roles.

Input	Family	Centre and range	Source	Owner
New customers in a window	Poisson count	mean 3.38	funnel model	sales owner
Contract value of a new deal	lognormal	mean 60,000; log spread 0.25	Illustration only	sales owner
Cost of acquiring in the window	fixed	269,421	finance ledger	finance owner
Value per euro of new ARR	lognormal	mean 3.68; log spread 0.10	value of a new customer ÷ 60,000; spread from the renewal data	analytics lead

A Poisson count is the family for a count of independent events in a fixed time. Its spread is about the square root of its mean, so a mean of 3.4 customers typically moves by about 1.8. *Decide what the evidence can resolve* explains the family in full. A lognormal input is never negative and has a long right tail, which suits sizes and values.

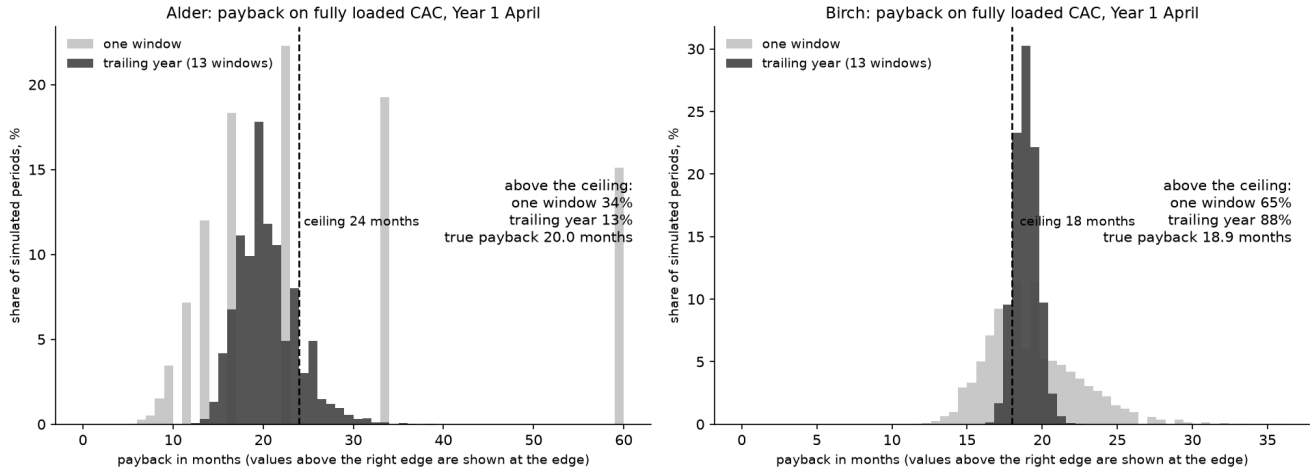
Two rules keep the table honest. First, a distribution is a belief, unless it comes from data. Where it comes from data, give the standard error (the typical size of the random error in an estimate). The spread of 0.10 for the value per euro of ARR comes from the renewal data. With about 44 first-renewal decisions a year, the first-renewal rate has a standard error of 5.4 points. Moving every renewal rate by that amount moves the value by +9.8% and −9.2%. Where it comes from an owner's belief, label it and ask for the P10 and the P90, not for one number.

Second, the cost of people and committed programmes is fixed within a window. A bad window does not shrink the salary bill. If cost per customer stayed fixed instead, the spread would come out too small.

Draws and outputs

The table runs the payback calculation for Alder and Birch at Year 1 April, with a Poisson number of customers and a fixed cost.

Period	Alder payback, P10 / P50 / P90	Alder: chance above 24 months	Birch payback, P10 / P50 / P90	Birch: chance above 18 months
One window	11.2 / 22.5 / 67.4	34%	15.8 / 18.9 / 23.5	65%
One quarter	14.6 / 19.9 / 31.3	34%	17.1 / 19.0 / 21.1	72%
Trailing year	16.9 / 19.9 / 25.0	13%	18.0 / 18.9 / 19.9	88%



Simulated payback of one window and of a trailing year, for Alder and Birch, against the payback ceiling of the segment.

Alder's true payback of 20.0 months sits inside its ceiling of 24, and one window still crosses it 34% of the time. A window with one new customer gives a payback of 67 months, and a window with none gives no payback at all. The median window reads 22.5 months, not 20.0: the flaw of averages at work. All of this is noise, and a year that pools 13 windows cuts the crossings to 13%.

Birch's true payback of 18.9 months sits above its ceiling of 18, and the longer the period, the more often the reading lands above it. A trailing year moves only about one month around the true value. At Alder the interval warns you not to react. At Birch it says that the gap is probably real, though small.

The same engine serves the objective. Write the net value of a window of acquisition as v × l − K. Here v is the discounted gross profit per euro of new ARR, l is the new ARR won in the window and K is the cost. For Alder, with the inputs above, the mean net value is about 470,000 euro, and P10, P50 and P90 are about −49,000, 425,000 and 1,038,000 euro. The chance of a loss in one window is 13.1%, about one window in eight. *Commit with uncertainty* explains the card in full.

Implementation note: the companion code computes these tables with `payback_distribution`, `decision_card` and `flip_analysis` in its value module.

Switching values

A *switching value* is the amount by which one input must move, with all others at their means, before a conclusion changes. It replaces the tornado chart and the fixed best, base and worst cases, and it asks the reverse question: how big a change flips the call? A small switching value marks a fragile decision and names the input to measure first. It moves one input at a time; the chance of a loss from the simulation covers inputs that move together.

Take the call "keep acquiring at the current cost" against "do nothing". It flips when v × I equals K.

Input	Mean	Switching value	Move needed	Inside the P10 to P90 of the input?
Value per euro of new ARR	3.68	1.34	−64%	No
New ARR in a window	200,972	73,234	−64%	Yes
Cost of the window	269,421	739,352	+174%	No

First, the mean. The value of a customer would have to fall by 64%, or the cost rise by 174%, before the call changes. Both moves lie outside what the owners believe, so the call on the mean is solid.

Second, a single window. New ARR must fall to 73,234 euro, about 1.2 deals. That level lies inside a normal window: one deal or none occurs in 15% of windows at Alder. The input that can flip a single window is the count of new customers. No measurement can improve it; only pooling windows over time can.

The same logic applies to the payback ceiling. For Alder to cross 24 months in a trailing year, the cost must rise by 20.2%, or the number of customers must fall by 16.8%, from 43.9 to 36.5. The 80% Poisson range of a year is 36 to 52, so the switching value lies inside it. For Birch to fall inside 18 months, the cost must fall by 4.7%, or the count must rise by 5.0%, from 598 to 628. Birch is close to its ceiling.

What changes with volume

The mechanics hold at any volume, but what you can conclude depends on the customers a window brings. *Decide what the evidence can resolve* defines the regimes.

Regime	What one window of unit economics tells you	How to read it
Low (Alder, 3.4 new customers a window)	Mostly noise: payback crosses its ceiling in 34% of windows although the true value is inside	Read a trailing year with an interval; use outside survival priors, because 44 first-renewal decisions a year leave a standard error of 5.4 points; no channel cells
Medium (Birch, 46 new customers a window)	A window reads payback to about ±20%; a quarter reads it to about ±10%	Read CAC by quarter and pooled; channel cells of 18 to 35 wins carry 17% to 23% error
High (Cedar, 1,417 new paying accounts a window in self-serve)	A window reads payback between 11.6 and 12.5 months (80% interval); count noise is small	The noise no longer hides anything; the remaining doubt is the future survival and the definition of the margin

High volume does not make the future known. Cedar's self-serve payback of 12.0 months sits on its 12-month ceiling, and noise only decides whether a window reads 11.6 or 12.5. The ceiling, the margin rule and the survival curve decide the answer, not the count.

Alder, Year 1 April: a payback that looks alarming

Alder, Year 1 April: about 3.5 new customers a window, 0.9 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.1 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

In April, the board of Alder reads the first quarter: 7 new customers, and 875,619 euro of new-customer sales and marketing cost. The cost per customer on the quarter is 125,088 euro, against 79,782 euro for the trailing year. The head of finance turns that into payback: 31.3 months on the quarter, against 20.0 months for the year and a ceiling of 24 months. The board reads a rise of eleven months and asks for a cut in programme spend.

The analyst accepts the arithmetic and asks how often a quarter with 7 customers happens when nothing has changed. The model expected 10.7 new customers in the first quarter, a figure that already allows for the season and the sales cycle. The 80% interval of a Poisson count with that mean runs from 7 to 15, and the chance of 7 or fewer is 16.5%, about one quarter in six. The simulation agrees: a single window or quarter crosses 24 months in 34% of cases, although the true payback is 20.0 months. The trailing year runs from 16.9 to 25.0 months at the 80% level and crosses 24 months in 13% of cases.

A lean quarter does not shrink the cost. Only 7.5% of Alder's new-customer cost is paid media, which a manager could stop in a week. People are 37.2% of it. Trade fairs, partner fees and content add 45.1%, and Alder commits all of them in advance. Cutting committed programmes in a quiet quarter would not change the quarter, and it would cut the pipeline of the quarters that follow.

The cash view is calmer. The first invoice covers 60% of the CAC, the second takes the expected cash past it, and the expected cash payback is 12 months. The liquidity worry behind the word "payback" does not hold.

The board decides to leave the programme budget as it is. The board paper changes: it shows the trailing-year payback with its interval and the count of 44 customers behind it, and it drops the single-quarter figure. It also states the value basis: one new Alder customer is worth 220,768 euro of expected gross profit over five years at 12%, which is 2.77 times the fully loaded CAC of 79,782 euro. At low volume, one window of unit economics is mostly noise, so you read the trailing year and attach an interval. The interval also tells you when to worry: if the trailing-year interval sat wholly above 24 months, the finding would be real.

Birch, Year 1 April: paid efficiency that is not

Birch, Year 1 April: about 46 new customers a window, 1.0 times the derived typical; medium volume: a quarter reads a 20% change down to the SQL to opportunity rate. The win rate reads a 30% change in 74 days, and the close rate reads a 20% change in about 7 months.

Birch's head of marketing shows a dashboard tile in the monthly review: CAC of 3,602 euro, next to a cost per lead of 111 euro. The tile divides paid media spend of 165,704 euro per window by all 46.0 new customers. The cost per lead divides paid spend by paid leads only, and paid sources deliver 52.0% of Birch's leads. The rest come from organic search, direct visits and referrals.

The tile breaks the population rule. The top holds the spend of the paid population, and the bottom holds the customers of everybody. About 23.9 of the 46.0 customers per window came from paid leads, if the lead-to-customer rate is the same for all sources. On the same population, CAC is 6,932 euro, which is 1.9 times the number on the tile. The tile gave the organic customers to the media budget.

The fully loaded CAC is 12,160 euro. The new-customer CAC ratio is 1.27, below the published median of 1.76. The accrual payback is 18.9 months, slightly above the 18-month ceiling for mid-market. A trailing year reads from 18.0 to 19.9 months at the 80% level, so the gap is small but probably real. The expected cash payback is 12 months, so billing in advance eases the gap on the cash side.

The chief financial officer asks for the value basis next. Finance at Birch uses 12% as its hurdle rate (the lowest return that finance accepts), so the default stands. At 12%, one new customer is worth 31,538 euro of expected gross profit over five years. At 8% the value is 34,295 euro, and at 16% it is 29,164 euro. The ladder follows: one opportunity is worth 10,092 euro, one SQL 5,046 euro, one MQL 1,262 euro and one lead 505 euro. A one-point gain in the MQL-to-SQL rate is worth 58,037 euro per window, before capacity limits. The sellers are close to capacity, so the sales owner reads that number with care.

Birch buys paid media on three platforms. If the lead-to-customer rate were equal across channels, a quarter would bring 34.6 wins from paid search, 24.7 from a consumer social platform and 18.4 from a professional one. Their relative standard errors are 17%, 20% and 23%. A gap of 20% between two channels is smaller than the noise of one quarter. There is also a lag. A month's SQL cohort closes 14% in the same month, 53% in the next, 24% two months later and 6.5% three months later. So cohort paid CAC for April's spend needs until July to settle. The team decides to read channel CAC by quarter, to pool it where it can, and to move no budget on one quarter's cell.

The lesson at medium volume is that channel CAC is readable only by quarter and pooled. Company-level numbers read in a quarter, but channel numbers need two or three.

Cedar, Year 1 April: what belongs in gross margin

Cedar, Year 1 April: about 1,400 new paying accounts a window, 1.5 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 31 days; the assisted motion, with 61 opportunities a window, needs about 8 months to read a 30% change.

Finance at Cedar asks a simple question in April. Customer support and the hosting of free workspaces are both counted in the cost of revenue. Should they be? The gross margin is 81.4%. Without the hosting of free workspaces it would be 83.9%, and without support it would be 84.6%. The growth team uses the margin to value its tests, so the answer moves every test.

The rule of this chapter keeps support in: it keeps the service usable for customers who pay, and 84% of support cost sits in the cost of revenue in the published survey. Free-user hosting is a judgement. A free workspace is part of the product that everybody uses, and it is also how Cedar acquires paying accounts, so a case exists for counting it as acquisition cost. Cedar picks the first view: the cost recurs every month whether or not the workspace ever pays, so it belongs to running the product. The rule is written down, and the margin is printed with the words "includes hosting of free workspaces".

If the hosting of free workspaces, 1,014,355 euro a year, moved to acquisition, the self-serve CAC would rise from 635 euro to 690 euro, which is 8.7% higher. The margin would rise to 83.9% and the value of a paying account by 3.1%. Payback would move from 12.0 months to 12.7. The cost does not disappear when it moves, and the one error that matters is to count it in both places. Because the margin already charges paying accounts for the free workspaces, the team does not subtract the hosting again from the value of a sign-up.

A new paying account is worth 1,683 euro. Of 41,422 sign-ups per window, 36% activate and 9.5% of the activated pay, so 3.4% of sign-ups become paying accounts. One sign-up is worth 58 euro. A test that lifts sign-ups by 1% adds 414 sign-ups, worth 23,842 euro per window. A test that lifts the paid rate of activated

sign-ups by 0.1 points adds 14.9 paying accounts, worth 25,097 euro per window. Free hosting costs 1.88 euro per sign-up, 3.3% of the value, so a test that adds sign-ups that never convert is not free.

How exact is the 58 euro? The conversion part is known to about ±2.6% from one window (one standard error). The lifetime part is not. The value of a paying account rests on the churn that Cedar has seen so far, and the tail beyond the first year rests on fewer customers. A three-year horizon cuts the value by 27.0%. A rate of 8% lifts it by 7.6%, and a rate of 20% lowers it by 12.3%. The lesson at high volume is that a sign-up can be priced precisely in its conversion part. The doubt that remains is about the lifetime, and more volume does not remove it.

The decision to carry forward

Carry customer value as a distribution with its cohort, horizon and valuation date. Price an action as the difference it causes against its alternative. Keep cash recovery and capacity as explicit constraints even when expected gross profit is positive.

03 · Model recurring growth

Which source of recurring growth must change to reach your objective? A revenue total cannot answer. New customers, renewals, expansion and contraction have different causes and different clocks. The decision needs a model that connects those mechanisms to the movement of the customer book.

A model is a set of explicit assumptions about how inputs become outcomes. Its arithmetic can be exact while its behavioural assumptions remain uncertain. Multiplying conversion rates describes a path under those assumptions. It does not establish that raising an upstream rate will preserve the quality of everyone who arrives downstream.

This chapter connects revenue streams, retention and growth patterns. Keep the forecast conditional: state the segment, motion and period for which each relationship is expected to hold. Use the model to locate a commercial question, then price its alternatives on the basis established in *Value customers and cash*.

Revenue models

Revenue models differ in how the price relates to the value the customer gets. Four forms are common.

- **Subscription pricing.** The buyer pays for the right to use the product for a term. The buyer never owns it. The term can be a month, a year or longer.
- **Seat-based pricing.** The subscription price grows with the number of named users, called seats.
- **Usage-based pricing.** The price grows with use, for example by records processed or actions taken.
- **Hybrid pricing.** A fixed platform fee covers the base, and a variable part covers seats, sites or usage.

A second choice is the **billing term**: how often the customer pays. A monthly plan bills every month. An annual contract is often *billed in advance*, which means the customer pays for the whole year at the start. Cash then arrives in the first month, and the company earns the revenue over the twelve months. A multi-year contract can bill each year in advance or once for all years. The billing term decides when cash arrives and how early a customer can leave. A company may also price the terms differently. Cedar, for example, gives a discount on its annual plan.

The three companies sit at different points. None of them uses usage-based pricing.

	Alder	Birch	Cedar
Price	Platform fee plus a licence for each site (hybrid)	One annual subscription price	Price per seat per month; assisted accounts sign annual contracts
Billing term	Annual (75% of customers) or three-year (25%), billed annually in advance	Annual, billed in advance, renews by itself	Monthly plans and annual plans by card; assisted contracts annual, in advance
One new contract in ARR	New deal 60,000 a year, so 5,000 a month	New customer 9,600 a year, so 800 a month	Six seats: monthly plan 864 a year (72 a month); annual plan 720 a year, then 648 from Year 1 March

A contract and a revenue line are not the same thing. A **billing line** is one product on one contract. The test below classifies billing lines, never customers, because one customer can hold a recurring plan and a one-off service. A *revenue line* is a product sold through one motion. It belongs to the funnel and is defined in *Choose*

markets and accounts. Keep the two names apart.

Recurring against re-occurring revenue

Money that arrives more than once is not always recurring. Three questions sort every billing line.

1. Does the customer pay again for continued access to something the company keeps delivering?
2. If the payments stop, does access stop?
3. Does the contract fix the date and the amount of each payment, as a fixed sum or a **committed minimum**, so that only a notice of cancellation stops them?

A line that passes all three is *recurring revenue*. A line that fails the second or third question is *re-occurring revenue*: payments repeat, but they may vary or stop at the customer's choice. A support fee on a device the customer already owns is re-occurring, because the device works without it. A usage fee with a committed minimum is recurring up to the minimum and re-occurring above it. Revenue paid once, such as an implementation project, is one-off.

The **recurring-revenue share** is recurring revenue divided by total revenue. It moves slowly. A fall means that services, overage or one-off fees grow faster than the subscription base. A jump in one month usually means that someone changed the classification rule, so log every rule change with its date.

> **Calculation**
>
> recurring-revenue share = recurring revenue ÷ total revenue

The definition moves the share more than any tolerance does. Take a company with 9,000,000 of recurring revenue, 500,000 of re-occurring revenue and 300,000 of one-off revenue (Illustration only). Counted strictly, the share is 91.8%. Counted loosely, with everything that repeats, it is 96.9%. The gap is 5.1 points. Write the classification rule first and fix it. Then set the tolerance.

For the tolerance, start with a band of 3 percentage points around the trailing level. When the company has 12 periods of history, replace it with two standard deviations of its own period-to-period change, with a floor of 1 point. Keep the band wide enough to absorb one normal services project. In the example, one extra one-off project of 300,000 lowers the strict share by 2.7 points, which stays inside the start band. Now suppose that the period-to-period change has a standard deviation of 0.56 points (Illustration only). Two standard deviations then give only 1.1 points, which is narrower than one project. In that case, set the band at the size of one normal project, here 2.7 points.

Evidence: Plausible but untested. No published standard fixes a band. A 2021 survey of private software companies found that about two thirds kept professional services at 10% of revenue or less (KeyBanc Capital Markets, 2021, *Private SaaS Company Survey*, n = 152). This source sets the scale of the band. It does not set the number.

ARR and MRR

Annual recurring revenue (ARR) is the yearly value of all active recurring contracts on one date. It is a stock, like a bank balance, not a flow. **Monthly recurring revenue (MRR)** is ARR divided by 12. Both are management measures. They are not accounting figures, and they do not equal revenue recognised in the books.

> **Calculation**
>
> ARR = sum of the annualised recurring value of all active contracts
>
> MRR = ARR ÷ 12

Three rules keep ARR honest.

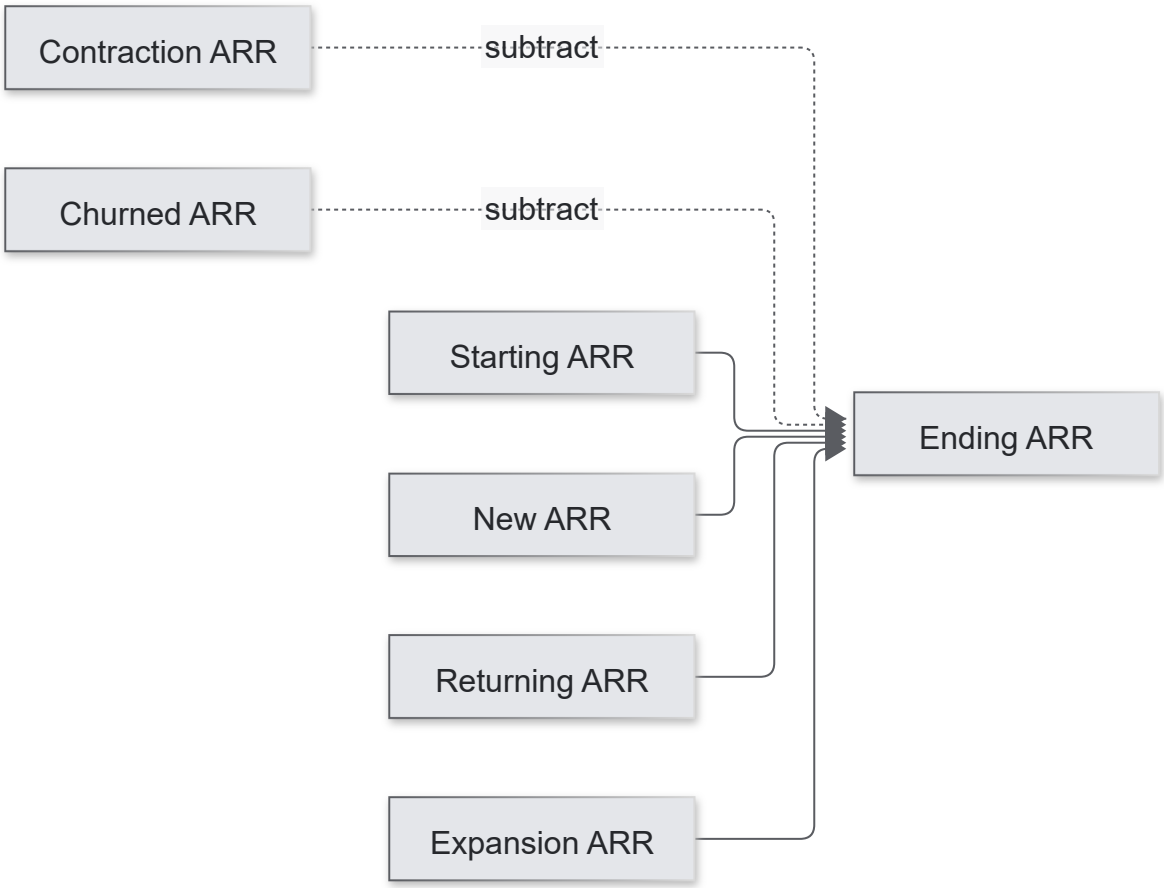
- **Count recurring lines only.** One-off services, such as implementation projects, stay out. Usage above a committed minimum stays out, because it is re-occurring.
- **Count the price the customer pays.** Use the contract price after discount, not the list price.
- **Count from the start date.** Signed contracts that have not started are tracked beside ARR, not inside it. Write the rule down and keep it.

A monthly plan needs a second note. Its ARR is its current MRR times 12. This assumes that the customer stays, and the customer has made no commitment to do so. An annual contract is a commitment for twelve months. Cedar's monthly plan with six seats counts as 864 of ARR, yet the next month it may count as nothing. ARR from monthly plans is a run rate, and the next sections show how fast it can leak.

The ARR bridge

ARR on one date differs from ARR on another date because of five movements. The **ARR bridge** carries starting ARR to ending ARR through them.

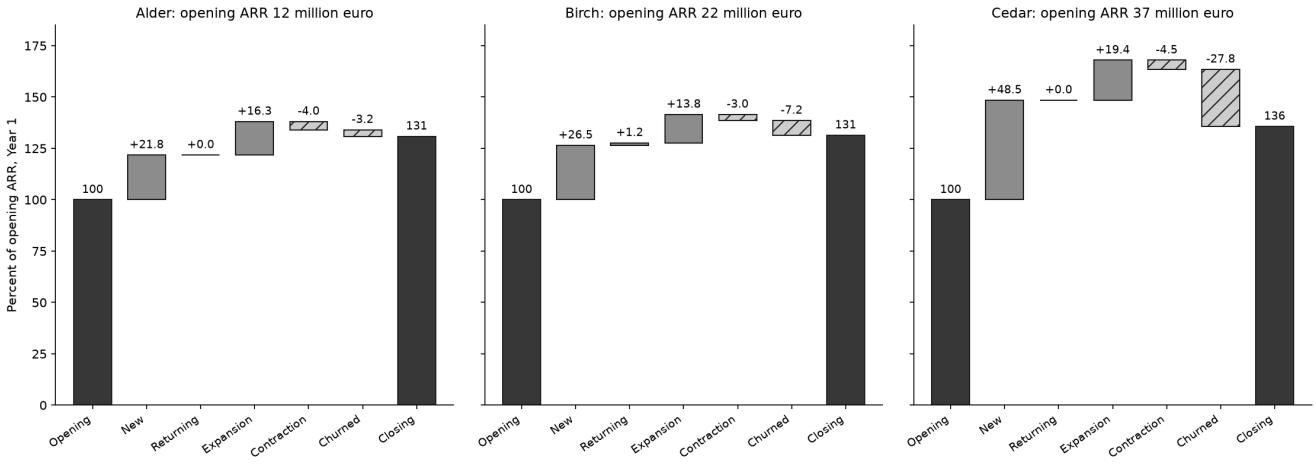
- **New ARR:** a customer that was never a customer before.
- **Returning ARR:** a former customer that comes back.
- **Expansion ARR:** more seats, usage, products or a higher price at a customer that stays.
- **Contraction ARR:** a customer that stays but pays less.
- **Churned ARR:** a customer that leaves. This includes non-renewal and non-payment.



The movements add up to **net new ARR**, the change in ARR over the period. The bridge is an identity. It must close to the euro.

> **Calculation**
>
> ending ARR = starting ARR + new + returning + expansion - contraction - churned
>
> net new ARR = new + returning + expansion - contraction - churned

Returning customers are the movement most often forgotten. A buyer who left two years ago and signs again is not new, because the company already knows the buyer and the earlier history. In the CRM, though, a returning buyer often looks like a new account. Treat returning ARR as its own movement. If you prefer to fold it into new ARR, state the rule and apply it every time. Without a rule, new ARR is overstated and the cost of winning a customer looks lower than it is.



The ARR bridge of Year 1 for the three companies, shown as a percentage of opening ARR, so that the three sizes can be compared.
Read the figure across. Alder and Birch grow by about 31% of their opening ARR, but Birch buys it with more new ARR and loses more at renewal. Cedar adds 48.5% as new ARR and loses 27.8% as churn, so a large part of what it wins only replaces what it loses.

When the bridge does not close, do not insert a balancing figure. Follow four steps.

1. Compute the gap in euro.
2. List the accounts whose opening and closing ARR differ by more than the movements explain.
3. Classify each difference: a missing movement, a change of rule, a currency effect, or a data error.
4. Fix the rule or the data, then rebuild the bridge. Log the change.

Use one fixed exchange rate for the year when contracts are in several currencies, so that currency moves do not appear as expansion or contraction. Billing is the source of the contract values. The CRM and finance records are checked against it.

Retention measures

Retention measures read the right half of the bridge. Four of them are in everyday use, and they answer four different questions.

Logo retention counts customers, not euro. It is the share of the customers present at the start that are still customers at the end. A small customer and a large customer count the same.

Gross revenue retention (GRR) counts euro and counts losses only. It is the share of the starting ARR of one group of customers that is still paid at the end. Expansion at one customer never offsets a loss at another, so GRR cannot exceed 100%.

Net revenue retention (NRR) adds expansion to GRR. It is the ending ARR of the same group of customers as a share of its starting ARR. NRR can exceed 100%, which means that the existing customers alone would grow the company.

> **Calculation**
>
> logo retention = customers at the end that were present at the start ÷ customers at the start

> GRR = (starting ARR − churned ARR − contraction ARR) ÷ starting ARR
>
> NRR = GRR + (expansion ARR ÷ starting ARR)

Do not write NRR as GRR times one plus an expansion rate. The expansion is measured against the starting ARR, and it is added. Both GRR and NRR follow only the customers that were present at the start. A customer won during the year is not in the group, even if it expands before the year ends. For this reason, expansion in the bridge is larger than the expansion inside NRR. In Alder's Year 1 book, the bridge shows 1,956,703 of expansion, and the opening group contributes 1,765,906.

The renewal rate is the fourth measure. It is the share of the contracts that came due in a period that were renewed. It is read by renewal date, not by start cohort. Alder shows why the two views differ. Only 79% of its customers come due in a year, because the three-year contracts renew every third year. Customers that are not due cannot leave. They raise logo retention (93.3%) but do not enter the renewal rate (91.6%). Both are right. They answer different questions: how did the decisions made this year turn out, and how many of last year's customers are still here?

The three companies show how far retention can differ (the book open in Year 1 January, run for twelve months).

	Logo retention	GRR	NRR	Renewal decisions a window
Alder	93.3%	92.8%	107.5%	12
Birch	90.0%	89.9%	102.1%	173
Cedar	75.0%	76.1%	92.1%	828 on annual plans

Published medians give a sense of scale. Private B2B software companies report a median GRR of about 91% (SaaS Capital, 2023, *SaaS Retention Benchmarks for Private B2B Companies*). They report a median NRR of about 101% to 102% (Benchmarkit and Pavilion, 2024, *2024 B2B SaaS Performance Metrics Benchmark Report*, 936 companies; SaaS Capital, 2023). For contracts below about 1,000 US dollars a year, the second report calls GRR "historically below 80%". Alder sits above the overall medians and close to the GRR of about 93% that SaaS Capital reports for contracts above 25,000 US dollars. Birch sits on the medians. Cedar sits below them, as its small contracts lead you to expect. The medians are percentages, so they need no currency conversion. Evidence: Standard practice for the definitions, and survey data for the medians.

Cedar's NRR of 92.1% has a plain meaning. If Cedar won no customer, its base would shrink by 7.9% in a year, about 2.9 million of ARR. New customers must fill that hole before growth begins.

Who reads these measures, and when

Each measure has a place, a rhythm and a reader. The ARR bridge is built from billing at each month end, and the owner of revenue operations (the role that keeps CRM and billing records in order) checks that it closes. The renewal calendar lists every contract with its renewal date, expected loss and named owner. Customer success reads it every week, because it shows the decisions that are still open. The hazard table (the share of customers still at risk who leave at each renewal) is rebuilt once a quarter, after the renewals of the quarter have closed. The analytics lead (who owns the measurement system and the code behind every number) reads it with the finance owner. The reason for the rhythm is simple. A rate can be read honestly only when its outcome has happened, so the bridge is a monthly record, and the survival table is a quarterly reading.

A one-screen view serves the decisions. It carries few numbers, called tiles, and each has its comparison.

Tile	Reads	Compared with
ARR bridge	Five movements and the closing check	Same month last year
Retention	Logo retention, GRR, NRR, with and without the ten largest accounts	The trailing four quarters
Renewal calendar	Contracts due in the next 90 days, owner, status	Contracts due and renewed last quarter
Survival	Share active after each renewal, with its interval	The previous quarter's curve

A tile that cannot name the decision it supports has no place on the screen. *Decide what to measure* returns to dashboards in full.

Retention as survival

One constant retention rate treats every renewal alike. Real contracts do not behave that way. A customer in its first year is a different risk from one in its fifth. This section replaces the constant rate with survival by renewal number, which is the base of every value calculation in the book.

Survival by renewal number

Annual contracts show every renewal date. We call the first renewal decision, at the end of the first contract year, renewal number 1. The **survival curve** gives the share of a group of customers that is still a customer after each renewal. The **hazard** at renewal k is the share of the customers still at risk at that renewal that leave at it.

To build the curve from billing data, follow these steps.

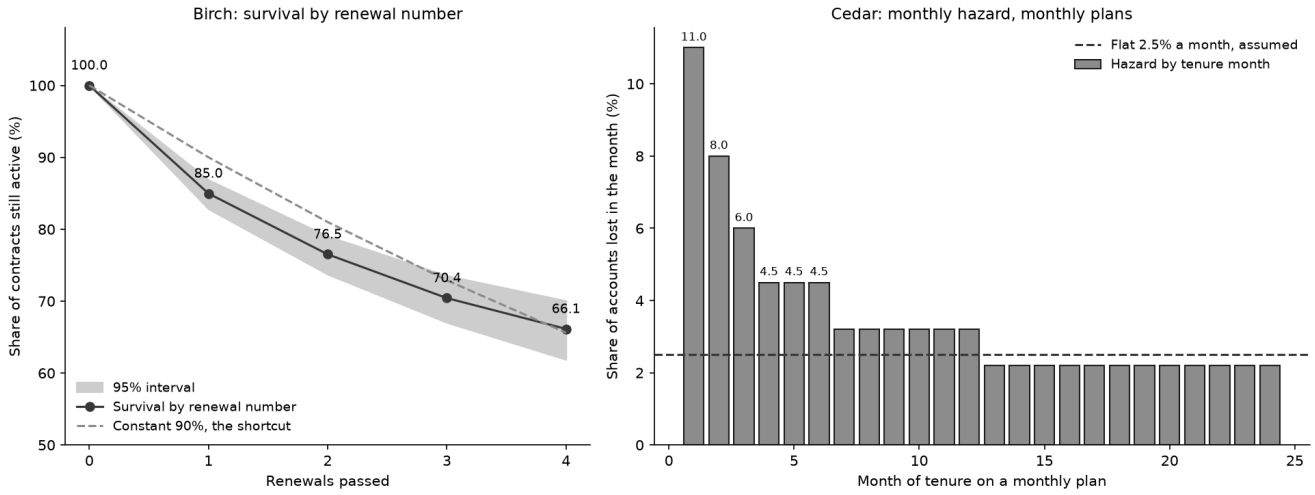
1. For each contract, record how many renewal decisions you have seen and whether the last one was a loss.
2. At each renewal number, count the contracts at risk and the contracts lost. The hazard is lost divided by at risk.
3. Multiply the shares that stay. The product is the survival after each renewal.
4. Add an interval: the range of survival values that the data cannot rule out. It comes from the number of contracts at risk, so it widens when the counts are small. Its **half-width** is the distance from the estimate to either end.

A contract that is still running has an unknown end. It is **censored**. It counts while you observe it, and it then leaves the calculation without being counted as a loss. This is the **product-limit estimate**, also called the Kaplan-Meier estimate (Kaplan and Meier, 1958). It is Standard practice.

> **Calculation**
>
> hazard(k) = contracts lost at renewal k ÷ contracts at risk at renewal k
>
> S(k) = (1 - hazard 1) × (1 - hazard 2) × … × (1 - hazard k), with S(0) = 1

Churn is the general word for a loss of customers or of ARR, so always say which form and which base you mean. Three forms need separate columns. A non-renewal is a customer that leaves. A downgrade is a customer that stays at a lower price. It is contraction, not a loss of the contract. Non-payment is a customer that leaves because a payment failed. It often has a different cause and a different fix, such as better billing recovery. Survival counts the contract as lost only when the customer leaves. The euro lost through contraction enters GRR, as shown above.

Pool all contracts first. Split the curve by segment, for example by channel or by plan, only when each segment reaches the precision that the decision needs at every renewal number it uses. As a start value, ask for a half-width of 5 points or less on each segment's survival share (Plausible but untested). Birch's example below shows why the later renewal numbers decide.



Left: survival by renewal number for Birch, with its interval, against the constant-rate shortcut. Right: the monthly hazard of Cedar's monthly plans by month of tenure, against the flat rate that finance assumed.

The hazard and its shape

The hazard has a shape, and the shape depends on the billing term.

On a **monthly plan**, the hazard is highest in the first months and then falls. Customers that never reach value leave early. Those that stay are a selected group with a lower risk, so the hazard falls as tenure grows. Cedar's monthly plans lose 11.0% of accounts in the first month, 8.0% in the second and 6.0% in the third. After a year the loss is 2.2% a month. A flat rate cannot describe this curve.

On an **annual plan**, a customer can leave only at a renewal date. If you look at it month by month, the hazard is zero for eleven months and spikes in month twelve, again in month twenty-four, and so on. This is why the book counts by renewal number for annual plans and by tenure month for monthly plans. The shapes also change what a base average means. A base that has many old customers shows a low average hazard, because the customers with a high risk left early. A new customer faces the high early hazard. The base average is the wrong number for a new customer (Linoff and Berry, *Data Mining Techniques*, chapter on survival analysis, on hazards and tenure).

Why a lifetime of 1 ÷ (1 - GRR) misleads

A common shortcut turns a retention rate into a lifetime. If retention is 90% a year, customers stay for 10 years on average.

> **Calculation**
>
> lifetime = 1 ÷ (1 - retention rate) (the shortcut)

The shortcut has four faults.

- **It mixes units.** GRR is a measure in euro. A lifetime is a measure of customers. A customer that contracts but stays lowers GRR and does not shorten its own life.
- **It assumes a constant rate.** Survival shows that the first renewals lose more customers than the later ones.
- **It reaches past the data.** Ten years of lifetime needs ten years of survival. Contracts rarely show more than a few renewals.
- **It ignores the horizon.** Every value calculation in the book stops at five years and discounts the money. A 10-year lifetime does not fit it.

Customers that leave without a trace need other tools. The Pareto/NBD model (Schmittlein, Morrison and Colombo, 1987) and the BG/NBD model (Fader, Hardie and Lee, 2005) describe buyers that may stop buying at any time, with no record of the stop. They fit repeat purchases without a contract. They do not fit a contract whose end date is on file. For contracts, use survival. A tail model for contractual retention, the shifted beta-geometric model (Fader and Hardie, 2007), can extend a curve past the data. Use it only when it beats the constant rate on a test that hides the most recent cohorts. Evidence: Standard practice.

State what lies beyond the data. A curve that runs past the last observed renewal is an assumption, not an estimate. Say so, or leave it out.

Growth streams and first principles

ARR grows from three streams. The first is **new** customers, including returning ones. The second is **expansion** at customers that stay. The third is **renewal**, and it works differently. Renewal adds no ARR in the year. It decides how much of the starting base is still there at the end. These are the **growth streams**.

Three first principles follow from the arithmetic. They agree with the pattern described by Winning by Design (Winning by Design, 2023, *Revenue Architecture*, chapter on first principles).

The growth rate falls as the base grows. Birch added 6,913,582 of net new ARR in Year 1 on a base of 22,000,000, a growth rate of 31.4%. The same euro addition on the next base of 28,913,582 gives 23.9%. Birch actually added 7,593,858 in Year 2, so growth was 26.3%. To keep 31.4%, it would have needed 9,086,200. The extra 1,492,342 is the price of a larger base. This is no failure of the funnel. The company simply has to add more euro each year to hold the same rate.

Retention and expansion take over from acquisition. A larger base means that more ARR comes up for renewal each year and more customers can expand. In Year 1, new and returning ARR made up 57% of the gross additions at Alder, 67% at Birch and 71% at Cedar. In Year 2, the share fell at Alder to 52% and at Birch to 63%. The share of expansion rose at each company. At Cedar, new and returning ARR fell from 71% to 69%.

The bucket leaks. Every euro lost has to be replaced before growth starts. This idea is known as the leaky bucket. In Year 1, losses (churn plus contraction) equalled 33% of the new and returning ARR at Alder, 37% at Birch and 66% at Cedar. At Cedar, about two euro in three of the new ARR only replaced a loss. When this

ratio is high, a gain in retention can be worth more than the same gain in acquisition. Whether it is depends on the value basis of *Value customers and cash* and the leverage method of *Find the most valuable constraint*.

The chain: customers = volume × product of rates

Customers equal the volume at the top of the chain times every step rate down to the signature. You can start the chain at any stage. Start where the flow is under your control and the counts are reliable. For Birch, that is the lead.

> **Calculation**
>
> customers = volume at the top × rate 1 × rate 2 × … × rate n

The top of the chain depends on the channel. For paid media, the chain starts at impressions. An impression is one showing of an ad, so one person can make many. The click-through rate (CTR) is the share of impressions that give a click. In budget terms, the click volume is spend divided by the cost per click, and leads follow from the click-to-lead rate. For organic and direct sources, the top is the visits and the visit-to-lead rate. Birch runs both. The paid and the unpaid parts add up to one lead volume, and from the lead on, one chain serves them both. If paid and unpaid leads convert differently, split the chain by source before you read its rates.

For Birch in Year 1 May: 2,894 leads × 40.0% × 25.0% × 80.0% × 62.5% × 32.0% gives 46.3 customers a window (Arithmetic). The chain has five step rates after the lead. In this book, an improvable rate is a step rate between two counted stages. The ratio of web sessions to ad clicks is not one of them. It is a tracking ratio, and *Decide what to measure* shows how to read it.

Close rate = opportunity rate × win rate

Sales teams often quote one close rate: customers per SQL. It hides two rates. The opportunity rate is the share of SQLs that become opportunities. The win rate is the share of opportunities that are won. Define the win rate on opportunities that have an outcome, or on a mature cohort. An open opportunity is neither a win nor a loss.

> **Calculation**
>
> close rate = opportunity rate × win rate

For Birch, 50.0% × 32.0% is 16.0%. Two different teams can show the same close rate. One qualifies tightly and wins most of what it takes on. The other qualifies loosely and loses most of it. Their pipelines differ in size and cost. The fix is in a different place: qualification for the first rate, and the proposal and negotiation for the second.

Splitting a composite rate is the first move of any diagnosis (see *Diagnose what changed*). The split can go further. Birch splits its opportunity rate into the share of SQLs that sales accepts (80.0%) and the share of accepted leads that become opportunities (62.5%). The product is again 50.0%.

Sums and products

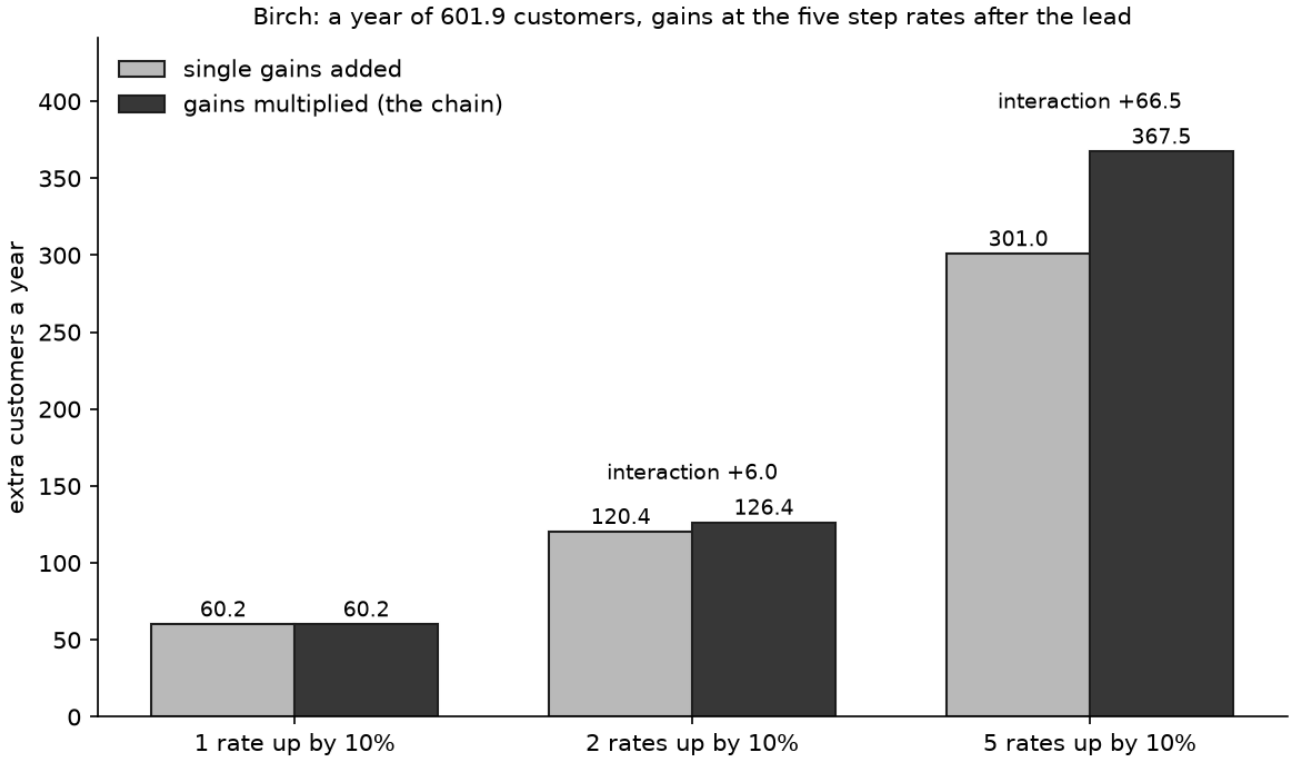
If one rate rises by 1% of itself, customers rise by 1%. This holds for every rate in the chain, because a product responds equally to a relative change in any factor. For Birch, 1% more at any of the five step rates gives about 6 more customers a year.

One percentage point is not one percent. A point on the MQL rate of 40% is 2.5% of that rate. A point on the acceptance rate of 80% is 1.25% of that rate. Express gains in relative terms before you compare them.

Several gains multiply. If two rates each rise by 10%, customers rise by 1.1 × 1.1 − 1, which is 21%, not 20%. The extra point is the interaction. With five gains of 10%, customers rise by 61%, not by 50%.

> **Calculation**
>
> combined gain = (1 + a) × (1 + b) × … − 1, where a, b, … are the relative gains at each rate

Birch wins 601.9 customers a year. Five gains of 10% add 367.5 customers. The sum of the five single gains is 301.0. The difference, 66.5 customers, is the interaction. One Birch customer is worth 31,538 euro of expected gross profit, on the value basis of *Value customers and cash*. The interaction alone is then worth about 2.1 million euro, before any capacity limit (Arithmetic).



Extra customers a year when 10% gains arrive at one, two and five of Birch's step rates: the product grows faster than the sum.

Two limits apply. The arithmetic says nothing about which gain is easy or cheap. *Find the most valuable constraint* ranks the gains by expected profit. And it assumes that each rate moves alone. In practice, more leads at the top often change the next rate, because the added leads are of a different quality. Never sum single gains, and never assume that rates are independent.

Attrition and diminishing returns

A chain with more stages has a lower cumulative rate, even when every step is strong. Five steps at 80% leave 33% (Arithmetic). Each added stage is a place where items drop out, so a stage must earn its place: it should fit the buying process and add information.

The same arithmetic runs along a sequence of contacts with one person, such as a run of nurture emails (a planned series for leads not yet ready for sales). If each send loses a share u of the people still reachable, the share left after n sends is (1 – u)^n. At 0.4% lost per send, 26 sends leave 90.1% of the list (Illustration only).

Rates have ceilings. No rate exceeds 100%, and real ceilings are lower. Birch accepts 80% of its SQLs, so that rate can gain at most 25% in relative terms. As a rule, a rate near its ceiling is harder to lift than one far below it (Plausible but untested). The gap between a rate and its ceiling is its headroom, and *Find the most valuable constraint* uses it.

Some texts model a deal as k meetings, each passed with the same probability p (Winning by Design, 2023, *Revenue Architecture*, chapter on the mathematical model). The close rate is then p to the power k. The picture shows why fewer or better meetings raise the close rate. Do not use it as a model of your data. Meetings are not alike: the first screens out weak fits and the last decides. They are not independent either, because one champion (a person at the customer who works to make the deal happen) can carry every meeting. Count the real meeting stages and read the pass rate of each (Plausible but untested).

Growth rates and patterns

Growth is a change divided by a starting value. Year-on-year growth compares ARR now with ARR a year ago. The **compound monthly growth rate** is the steady monthly rate that gives the same change over the year.

> **Calculation**
>
> year-on-year growth = ARR now ÷ ARR a year ago – 1
>
> compound monthly growth = (ARR end ÷ ARR start)^(1 ÷ months) – 1
>
> Evidence label: Arithmetic.

Do not divide a yearly rate by 12. In Year 1, Alder grows 30.9%, Birch 31.4% and Cedar 35.7%. Their compound monthly rates are 2.27%, 2.30% and 2.58%. Dividing by 12 gives 2.57%, 2.62% and 2.97%, which is 13% to 15% too high. The chapter *Model recurring growth* showed why a rate falls as the base grows. A target stated as a rate is a hidden euro target that rises every year, so state the euro target next to the rate.

Do not multiply one window by 13 either. A window is a noisy sample, and it carries the season. Alder wins 3.2 customers in a typical window at Year 1 July. Four windows in five show between 1 and 6, so 13 times one window reads anything from 13 to 78 customers, against 42 expected. Cedar shows the seasonal part. Its January runs 20% above the average month. Thirteen times its January window gives 21,341 new self-serve paying accounts for the year. The year brings 18,859, so the shortcut is 13% too high.

Growth also has stages. A **growth stage** is a phase of company life with its own main question. A common convention names the start-up, the scale-up and the mature company. Winning by Design puts the start-up below about 10 million US dollars of ARR and the scale-up above it. The book gives no sample, so this is an Assertion (Winning by Design, 2023, *Revenue Architecture*, chapter on the growth model). By size, all three example companies are scale-ups. By growth they are not: the same book gives 60% to 100% for scale-ups,

and ours grow 31% to 36%. For private B2B software companies, the median growth is 27% and the 75th percentile is 60% (Benchmarkit and Pavilion, 2024, *2024 B2B SaaS Performance Metrics Benchmark Report*, calendar year 2023). Our three sit between the median and the 75th percentile. So treat stage as context. The same book names three phases of focus: win customers at scale, grow at a sustainable cost, and keep the quality of growth. They match the three tests in *Choose markets and accounts*.

Growth over many years often follows an **S-curve**. It starts slowly, speeds up and then levels off at a **ceiling**.

> **Calculation**
>
> ARR(t) = K ÷ (1 + e^(−r × (t − t∅)))
>
> K is the ceiling, r is the steepness and t∅ is the time at which ARR is half the ceiling.
>
> The growth rate falls from the start, in proportion to 1 − ARR ÷ K.
>
> Evidence label: Arithmetic for the curve. Whether a company follows it is an Assertion.

The ceiling is the weak point. Two data points cannot tell you where the ceiling lies. Curves through the same two points with a low, a medium and a high ceiling agree for a year or two and then drift apart, as the Cedar example shows. So take the ceiling from market sizing, never from the fit. Count the accounts that you can reach. Multiply by the share of them that can realistically pay. That share is a judgement. Write it down, give a range, and test it with the leading signs of saturation: a rising cost of each new customer and a falling rate of conversion.

What changes with volume

The mechanics in this chapter hold at any volume. The bridge must close at every size, and survival by renewal number is the right form at every size. What changes is how much of a retention figure you can read.

Regime	What the volume allows	Where it holds
Low (Alder)	About 12 renewal decisions a window. A 20% change in the share of renewals lost would need about 29 years of data. Read each large renewal as a case, with a named plan, not as a rate.	Renewals of annual and multi-year contracts at about 12 a window
Medium (Birch)	The pooled survival curve is readable to within about 2 points at the first renewal. A 20% change in the loss rate needs about 1.6 years. Segment curves wait.	About 170 renewal decisions a window, with roughly 20 to 55 at each renewal number
High (Cedar)	The hazard shape by tenure month reads in one window. A 20% change in the first-month hazard needs about four windows.	Hundreds of accounts in each tenure month

Decide what the evidence can resolve explains how these numbers are derived. *Forecast conversion and timing* shows how to read rates that mature, such as these.

Alder, Year 1 February: three accounts, most of the expansion

Alder, Year 1 February: about 3.4 new customers a window, 0.9 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. A 20% change in the share of renewals lost would need about 29 years of renewals.

In January, marketing, sales and finance reported three different customer counts: 214, 203 and 190. The new chief executive asked for one answer. The ARR bridge uses the count from billing, 190, because ARR comes from the contract values that billing holds.

The head of finance builds the bridge for last year. The result looks good.

Last year	ARR (euro)
Opening ARR	9,400,000
New ARR	2,050,000
Returning ARR	0
Expansion ARR	1,230,000
Contraction ARR	390,000
Churned ARR	290,000
Closing ARR	12,000,000

Net new ARR is 2,600,000, which is growth of 27.7%. The table closes. It implies GRR of 92.8% and NRR of at most 105.9%, because some expansion can come from customers won during the year. Then the analyst lists the expansion by account. Three accounts grew by 430,000, 320,000 and 240,000. Together that is 990,000, which is 80% of all expansion and 38% of net new ARR. Without them, net new ARR would have been 1,610,000, or growth of 17.1%. NRR would have been at most 95.3%, below 100%. The ten largest accounts hold 38% of ARR, and the largest holds 11%.

The same concentration works on the way down. The renewal calendar for the coming year shows 153 renewal decisions, which covers 79% of the customers. The contract file puts expected logo retention at 93.4% and GRR at 92.8%, but the largest account is worth 1.48 times all the losses the calendar expects. Logo retention would fall by half a point if it left, and GRR by 11 points. The team cannot read these two numbers as rates. Thirteen losses a year is a count of events, and one account outweighs the rest.

The decision is small and specific, and the head of customer success makes it. Each of the ten largest accounts gets a named renewal plan with its renewal date and the key people on both sides. The monthly report shows logo retention and GRR side by side, and GRR again without the ten largest accounts. The decision log, the written record of each decision and the belief behind it, states that all ten renew at full scope. Each account's renewal month is the date by which the belief holds or fails (see *Commit with uncertainty*). A miss then becomes a case that the team can study, and not a surprise in the year-end GRR.

The lesson is about volume. With few customers, expansion and revenue retention are the story of a handful of accounts. Report logo and revenue retention apart and show concentration, so that nobody reads a rate where there is a short list.

Birch, Year 1 March: the ten-year lifetime

Birch, Year 1 March: about 48 new customers a window, 1.1 times the derived typical; medium volume: a quarter reads a 20% change down to the SQL to opportunity rate. A 20% change in the share of renewals lost needs about 1.6 years.

The board pack for the quarter says that a customer stays ten years. The figure comes from a GRR of about 90%, using the shortcut 1 ÷ (1 - 0.90). The pack uses it to argue that more spending on acquisition will pay back. The analytics lead asks for the billing data and builds the survival table.

The table holds the contracts that started in the last five years, about 1,700 in all (Illustration only: the table is built from Birch's renewal probabilities, and the counts are expected values). At the first renewal, 1,191 contracts are at risk and 179 leave. At each renewal number the result is:

Renewal	At risk	Hazard	Survival	95% interval
1	1,191	15.0%	85.0%	82.8% to 86.9%
2	663	10.0%	76.5%	73.7% to 79.1%
3	352	8.0%	70.4%	67.0% to 73.6%
4	148	6.1%	66.1%	61.9% to 70.1%

The counts fall at each step, because only older contracts have seen later renewals, and the interval widens with them. A fifth renewal has only 10 contracts at risk, so the table stops at four.

The curve does not match the ten-year claim. One contract in three is gone before the fifth year starts. Over the five-year horizon, a customer contributes 3.98 contract years, or 3.29 years when the later years are discounted at the default rate of 12%. The shortcut at 90% gives 3.39, only 2.9% more. So inside five years the shortcut is a fair quick check. The ten-year figure is the real problem. A lifetime of ten years or more needs an assumption about the years after the fourth renewal, where Birch has no data. If survival stayed at 94% a year for ever after that renewal, the lifetime would be 14.3 years, and 72% of it would lie after year five. Both the lifetime and the reasoning behind it come from a revenue measure. The board withdraws the figure, and retention becomes survival by renewal number.

The second problem comes from the bridge. Finance's bridge for the first two months has four movements, and it does not close. Its new ARR counts only first-time customers, so customers who come back have no movement. About 2.5 former customers a window return, each at about 8,000, or 20,000 of ARR a window. Over two months that is about 43,000 of ARR with no movement to carry it. In a year the missing movement is 260,000, which is 1.2% of opening ARR. In the CRM, sales had counted the same buyers as new customers, so its count of new customers was too high. Finance adds a returning movement and applies one rule: a customer with an earlier contract is returning, whatever the CRM says.

Then comes the temptation to split the curve. The analytics lead has the activation flag. Of the contracts at the first renewal, 72% were activated, with a first-renewal probability of 90%, and 28% were not, with 72%. Over the full table, the smaller group holds 333 first renewals and the gap reads clearly, with a half-width of 4.8 points. In one window, it holds 16 first renewals and the half-width is 22 points. At the third renewal, the same group holds about 100 contracts, and the half-width is 5.4 points, above the start value of 5. The gap is also an association, not a cause: customers who activate may simply fit the product better. The segment

curves need more cohorts, and the lead waits. The pooled curve is readable now, so it goes into the value calculation of *Value customers and cash*. The log entry states the expected first-renewal survival of 85.0% for the next year of renewals, with the date on which the last of those renewals closes.

Cedar, Year 1 March: the annual discount

Cedar, Year 1 March: about 1,500 new paying accounts a window, 1.5 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. The first-month hazard on monthly plans (11.0%) is known to within 2 points, and a 20% change in it needs about four windows.

Cedar sells a seat for 12 a month on a monthly plan and 10 a month on an annual plan. The annual plan has a discount of 16.7%. The growth team wants to move more customers to annual plans, because annual customers pay up front and cannot leave in the first year. Finance asks how deep the discount can go before the company loses money.

The break-even rule is simple. A monthly customer is worth the number of months it pays in the first year. An annual customer pays for twelve months at the discount price. The company is indifferent when the annual price covers the expected paid months of a monthly customer.

> **Calculation**
>
> break-even discount = 1 - (expected paid months in year one ÷ 12)

Finance used a flat monthly churn of 2.5%. A customer then pays for 10.48 months on average, and the break-even discount is 12.7%. At 16.7%, the discount is already too deep, so finance argues against going further. The 2.5% is close to what customers show after their first year, which is 2.2%. It describes the survivors.

The analytics team reads the hazard by tenure month. It is 11.0%, 8.0%, 6.0%, then 4.5% for three months, 3.2% for six months, and 2.2% after a year. At the first tenure month, 914 accounts are at risk, so the 11.0% is known to within 2.0 points. The shape is clear within a single window. The base as a whole loses 3.6% a month. The reason is its age: 55% of the monthly-plan accounts are more than a year old, and only 5.8% are in the first month. The base average is not what a new customer faces.

Along this path, 55.2% of new monthly customers are left after twelve months, and the expected number of paid months is 8.63. The break-even discount is 28.1%, more than double the flat-rate answer of 12.7%. The shape matters on top of the level. A flat rate of 4.8% would give the same twelve-month survival, but it would pay 9.27 months and a break-even of 22.8%. The flat rate understates the break-even, so it blocks a discount that pays.

The owner of self-serve pricing decides to move the annual price from 10 to 9 a seat, a discount of 25.0%. That is 3.1 points below the break-even, so a margin remains for what this view leaves out. It compares year one only. It also assumes that a customer who moves to the annual plan would have shown the monthly hazard. Customers who choose a year may be more committed, and for them the break-even is lower. Annual plans renew at 70% at the first renewal, and monthly customers that survive a year leave at 2.2% a month. The full comparison counts later years on the value basis of *Value customers and cash*. The team expects the annual share of new plans to rise from 30% to 40%. The log entry records the break-even of 28.1%, the chosen discount and an early check: the paid months of the first monthly cohort after twelve months, expected at about 8.6.

The lesson is about volume. High volume reveals a front-loaded hazard that low volume would hide. Once the data shows where a constant-rate shortcut fails, and what the error costs, there is no reason to keep it.

The decision to carry forward

Use the revenue bridge to locate the source of growth or loss. Use the behavioural model to propose an intervention. Treat the forecast as conditional on mix, survival and capacity, then observe which assumption actually changes.

04 · Choose markets and accounts

Which customers deserve a place in your growth plan? The answer must connect a customer problem, your route to that customer and the profit earned after delivery. A segment that signs quickly may be expensive to serve. An attractive enterprise account may consume scarce implementation capacity before it produces cash.

Define the account you can reach and the journey it must complete. Fit describes whether your product can deliver the required result. Intent describes evidence that the account may act. Neither is the same as the additional value your action can create.

Brian Balfour's channel-model-fit argument connects acquisition channels with how a business charges and earns (Balfour, 2017, *Get Out of the ARPU-CAC Danger Zone with Channel Model Fit*). We extend that commercial question through downstream survival and cost. Compare segments at a common cohort age. The concentration, stage definitions and go-to-market tests below keep the choice tied to an observable population.

Market and account selection

The first lever acts before any ad runs. It decides who the funnel is for.

The ideal customer profile (ICP) of *Choose the growth objective* lists traits that you can check in data. Examples are country, industry, size, process, tools and a trigger that makes a company act. Use evidence about customers who stayed and paid as well as those who signed. Compare them at a common cohort age. The customer book also reflects selection by the acquisition policy, so investigate the reachable accounts it excluded. A profile built on signed deals describes who is easy to sell to. A profile built on retained value describes who is worth selling to. The ICP describes companies. A **buyer persona** describes the people inside them, by what each one tries to achieve. We keep the two apart and tier the target accounts by fit (Winning by Design, 2023, *Revenue Architecture*, chapter on the operating model).

The **total addressable market (TAM)** is the set of companies, or the yearly revenue, that your product could serve in the market you target. Build it from the bottom: count named units in the unit that the price uses, for example plants for a product priced per plant, and not from a market report. Multiply the units by a price and you have a ceiling. *Set targets and staff capacity* showed that a ceiling is not a forecast, because the attainable share is a judgement.

The **target account list** is the named part of the TAM that you choose to work. Rank it by fit and by trigger, and give the first tier the most effort.

TAM coverage has two readings, and they answer two questions.

> **Calculation**
>
> list coverage = units on the target list ÷ units in the market outside your customers
>
> reach coverage = target accounts reached or engaged in a period ÷ target accounts on the list
>
> Evidence label: Arithmetic. Count both in one unit.

List coverage asks whether the list is the right size. Reach coverage asks whether the programme reaches the list. When list coverage is low, the list limits the volume, and widening it is the lever. High list coverage means many eligible accounts are named, not that they have been reached or engaged. Budget can still increase reach within that list. If reach is already high and measured marginal response has saturated, extra touches may reach colleagues at engaged accounts while added accounts come from deeper in the fit order. This is the quality dilution of *Find the most valuable constraint*: in a small, known market, "top of the funnel first" fails.

Check the ICP against value. Compare the ICP match rate of *Decide what to measure* with what the matching customers later became. At medium and high volume, read retention and expansion by fit, by cohort, once the cohorts are old enough (*Forecast conversion and timing*). At low volume there are too few customers for rates, so read cases: why each customer that left did leave, and what the five best customers share. Write the answer as a prior with a range.

For a programme aimed at the list, the account holdout of *Choose acquisition channels* keeps a random share of the list without programme touches. It reads engaged accounts and meetings, and only large effects.

Implementation note: company registers and commercial firmographic datasets list firms by industry, size and site. Check how old the data are and how they count sites before you size the market.

Motions and touch levels

A motion is one way of selling and serving one kind of customer. It combines a product, a route to the customer and an amount of human effort. The model needs the motion because each motion has its own chain, its own ACV, its own cycle and its own cost.

Touch level is the amount of human effort that a motion puts into each customer, across marketing, sales and customer success. The idea comes from Skok and from Winning by Design (Skok, undated, *SaaS Metrics 2.0*; Winning by Design, 2023, *Revenue Architecture*, chapter on the go-to-market model). This book uses four levels. The ACV bands are Illustration only.

Touch level	Who touches the customer	Typical ACV, euro	In this book
Self-serve	Nobody. The product sells, onboards and serves.	under 2,000	Cedar, self-serve line
Low touch	Technology first. A person helps on request.	2,000 to 8,000	none of the three
Medium touch	Specialists qualify. One seller runs a short process.	8,000 to 40,000	Birch; Cedar, assisted line
High touch	A seller with specialists, often in the field, over a long cycle.	above 40,000	Alder

Where the ACV band and the real effort disagree, the effort decides, because cost follows effort. Product-led growth, sales-led growth and product-led sales differ in where the first human enters the chain.

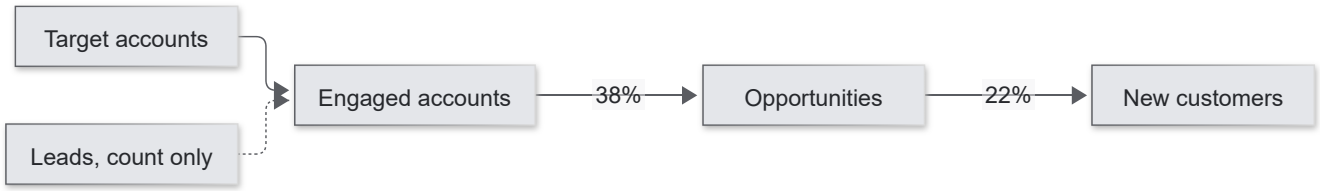
- In sales-led growth the customer meets a person before the product. A seller finds, qualifies and closes the deal.
- In product-led growth the customer uses the product before buying. What the user does inside it is the evidence for the next step.

- Product-led sales joins the two. Product use shows which accounts are ready, and a person joins after use has started, to sell a larger contract.

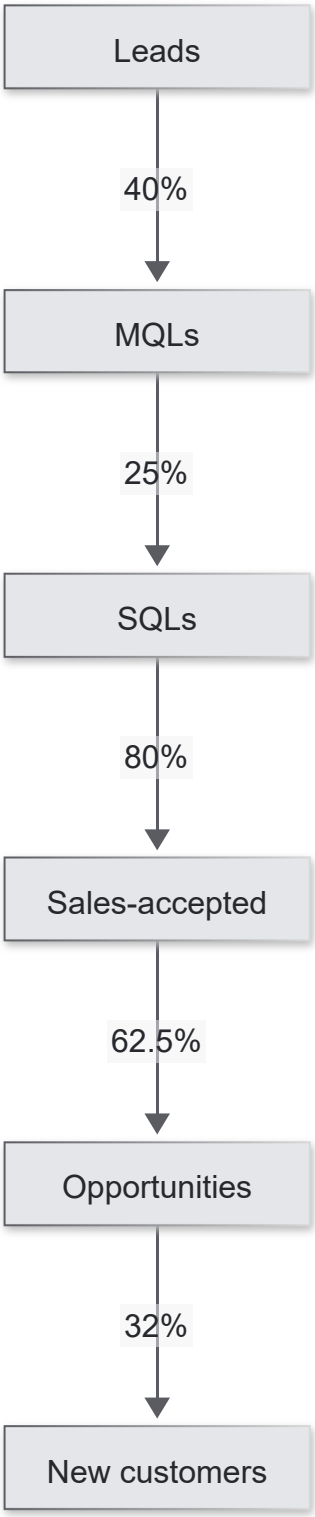
In a **sales-assisted motion**, a person joins an inbound or product-led path to qualify, demonstrate and close. Birch and Cedar's assisted line sell this way.

Product-led growth is a property of the product and the motion. It is not a campaign. It also moves the early counts into another system. In a sales-led chain the CRM holds them. In a product-led chain, product events hold them.

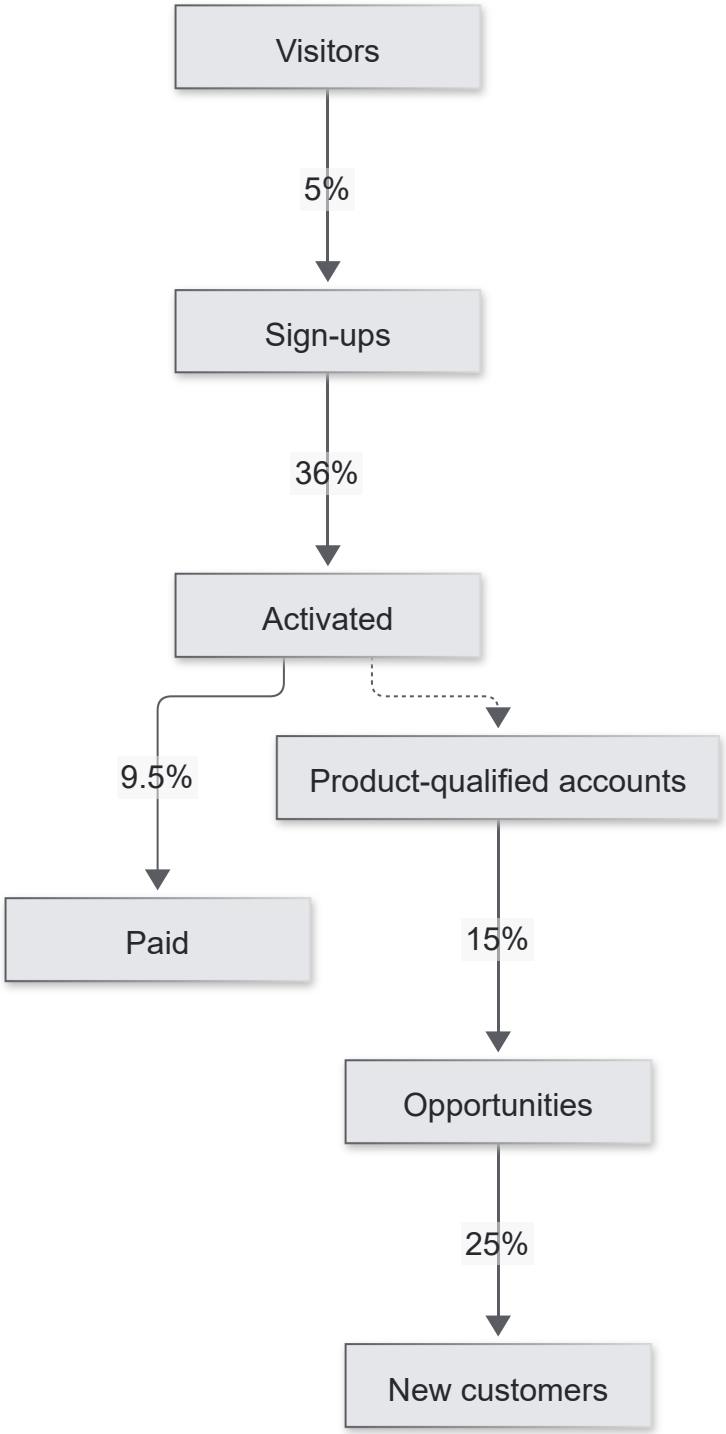
Self-serve chains need one more choice: how the free product works. A freemium plan lets users stay free for ever. A trial gives paid features for a fixed time. The two give different sign-up and payment rates. A survey of 200 software products puts a good freemium conversion to paid at 3 to 5%, and a good trial without a card at 4 to 6%. The sample is small and self-reported (ChartMogul and ProductLed, 2026, *The SaaS Conversion Report*). Cedar offers freemium plus a 14-day trial. Each sign-up starts the trial, so Cedar's trial-to-paid rate is the share of sign-ups that pay. It is 3.4%, inside the freemium span. The match is rough. The survey follows each sign-up for six months, and Cedar's 3.4% divides the payments of a window by the sign-ups of the same window. State the definition of a rate with its source, because published rates count different things.



Sales-led chain (Alder, Year 1 May).



Sales-assisted inbound chain (Birch, Year 1 May).



Self-serve chain with an assisted branch (Cedar, Year 1 May).

Cost per customer rises with human involvement. A person has a cost per hour, and each human step takes hours for each deal. Product cost is mostly fixed and spreads across all customers. Compare the fully loaded cost of winning one customer in Year 1 May. Cedar's self-serve line pays 634 euro, Birch 12,100 and Alder 79,686. The cost per customer rises about 126 times.

Contract value rises as well. The cost per euro of new annual recurring revenue goes from 0.82 at Cedar's self-serve line to 1.26 at Birch and 1.33 at Alder, about 1.6 times. A claim that human touch multiplies cost means nothing until it says cost of what, per what. For comparison, a survey of 936 private B2B SaaS companies reports a median of 1.76 euro per euro of new-customer ARR (Benchmarkit and Pavilion, 2024, *2024 B2B SaaS Performance Metrics Benchmark Report*).

Time grows with touch as well. A survey of 3.2 million opportunities, from best-performing firms only, looks at when deals close most often (Ebsta and Pavilion, 2023, *2023 B2B Sales Benchmark Report*). Small deals close most often within 31 to 60 days, medium deals within 61 to 90 days and large deals within 150 to 180 days. Our companies follow the same direction: 9 days from sign-up in Cedar's self-serve line, 34 days from SQL at Birch and 120 days from opportunity at Alder. The starts differ, so read the direction and not the gaps. A longer cycle also delays every cohort result, which is why the model carries times beside rates.

One funnel per revenue line

A revenue line is one product sold through one motion. A company with two motions runs two chains. Never blend their rates. A blended rate mixes populations with different volumes, contract values and cycles. It moves when the mix moves, even if each chain is stable. *Diagnose what changed* shows how to split mix from rate.

Assign each new customer to the motion that closed it. Write down how shared costs are allocated. Cedar shows why. Its assisted line starts from accounts that the free product already brought in. The assisted cost per customer is 4,446 euro, and that figure leaves out the cost of bringing the account into the product. The allocation is a decision. The model forces you to make it.

Account-based funnels

In B2B, a group buys. As many as 20 people can take part in a large purchase, in about nine roles (Winning by Design, 2023, *Revenue Architecture*, chapter on the operating model). This group is the buying group. Several of its members may give their contact details at different times, so one decision creates many leads.

Now follow the argument step by step. A lead funnel counts people. A decision belongs to a group. When many leads come from one account, the lead count measures activity at the account and not decisions. Leads from the same account also move together, so rates on leads are not independent observations. The honest unit is therefore the account, when three things hold. The buying group has more than one person, the target market is a known list, and the number of decisions is small. If one of the three fails, the lead can stay the unit.

An account-based funnel counts accounts through four stages. The target stage means the account is on the named list that the ideal customer profile produced. The engaged stage means the account meets a written engagement rule. The opportunity and closed-won stages work as before. Count leads beside the stages as an activity signal and not as a stage. This is account-based marketing: planned programmes for chosen accounts, not for single leads.

The account view also changes who reads the funnel. A seller reads accounts one by one. A marketer reads the share of the target list that is engaged. An analyst reads counts and, when volume allows, rates. Each reader needs a view at the right grain, and the account key joins the views.

An account can open several opportunities over time. Count the account once per stage per cycle, and define when a new cycle starts. An engagement rule needs care. It should name the roles that must act and the action that counts. Buying-group coverage, the share of expected roles with an engaged contact, is a candidate leading signal (Plausible but untested). Treat it as an early warning (a signal that moves first but is not yet shown to predict the outcome) until data grades it (see *Decide what to measure*).

Product-led companies need the same step in another form. A product-qualified account is an account whose product use reaches a defined level, for example several active users and an administrator who has invited more people. It is found in product data and not in a form. When the unit is one user, the same idea gives a product-qualified lead. The product-qualified account rate is the share of active accounts that reach that level. No reliable published benchmark exists for the step from product-qualified account to opportunity, so read it against your own history. Sales may call these accounts. Whether the call changes the outcome needs a test (see *Remove buying friction*).

Outside intent signals can enter the engagement rule as extra inputs. Show lift in a test before you pay for them (Plausible but untested). Contact-level activity is personal data. Show accounts and roles in the funnel view, not named people. This is not legal advice. Ask counsel before you link product use or web activity to named individuals.

Unit economics by channel and cohort

Unit economics are easy to compute for a company and hard for a channel, because cells get thin as you cut by channel, quarter and segment. Birch wins about 46 customers per window. Cut into three paid channels and a quarter, each cell holds 18 to 35 wins, with a relative standard error of 17% to 23%. A difference of 20% between two channels is therefore not readable in one quarter.

A second problem is time: the LTV of a channel appears only after the first renewals of its customers, twelve months after the signature. Until then you have the CAC and early signs: activation, the first invoice and the cash payback. Use the company's pooled survival curve as a prior (a starting belief that new data then update) and update it with the channel's own renewals when they arrive.

The remedy is pooling. Pool windows into quarters. Use partial pooling, which pulls a thin cell toward the pooled value in proportion to how thin it is. *Decide what the evidence can resolve* explains it. Move no budget on one cell. Compare periods with one attribution model (one fixed rule that shares the credit for a sale among earlier touches). To decide where the next euro of budget goes, use an incrementality test (a test of what the spend adds beyond what would happen anyway), as *Choose acquisition channels* explains.

A unit-economics panel

The panel below shows Alder's unit economics for the trailing year at Year 1 April. Each row has a value, an 80% interval, the count behind it, the limit and a status.

Measure	Value	80% interval	n	Limit	Status
New customers	43.9	36 to 52	44	none	count only
Fully loaded CAC	79,782	67,355 to 97,291	44	none	read the interval
New-customer CAC ratio	1.33	1.12 to 1.62	44	1.76 (median, in US dollars)	below the median
Accrual payback, months	20.0	16.9 to 25.0	44	24	inside; 13% of years above
Cash payback, months	12	by invoice date	n/a	24	inside
Value of a new customer	220,768	193,125 to 249,218	44	12%, five years	basis stated

Measure	Value	80% interval	n	Limit	Status
LTV to fully loaded CAC	2.77	2.15 to 3.43	44	none	not a rule

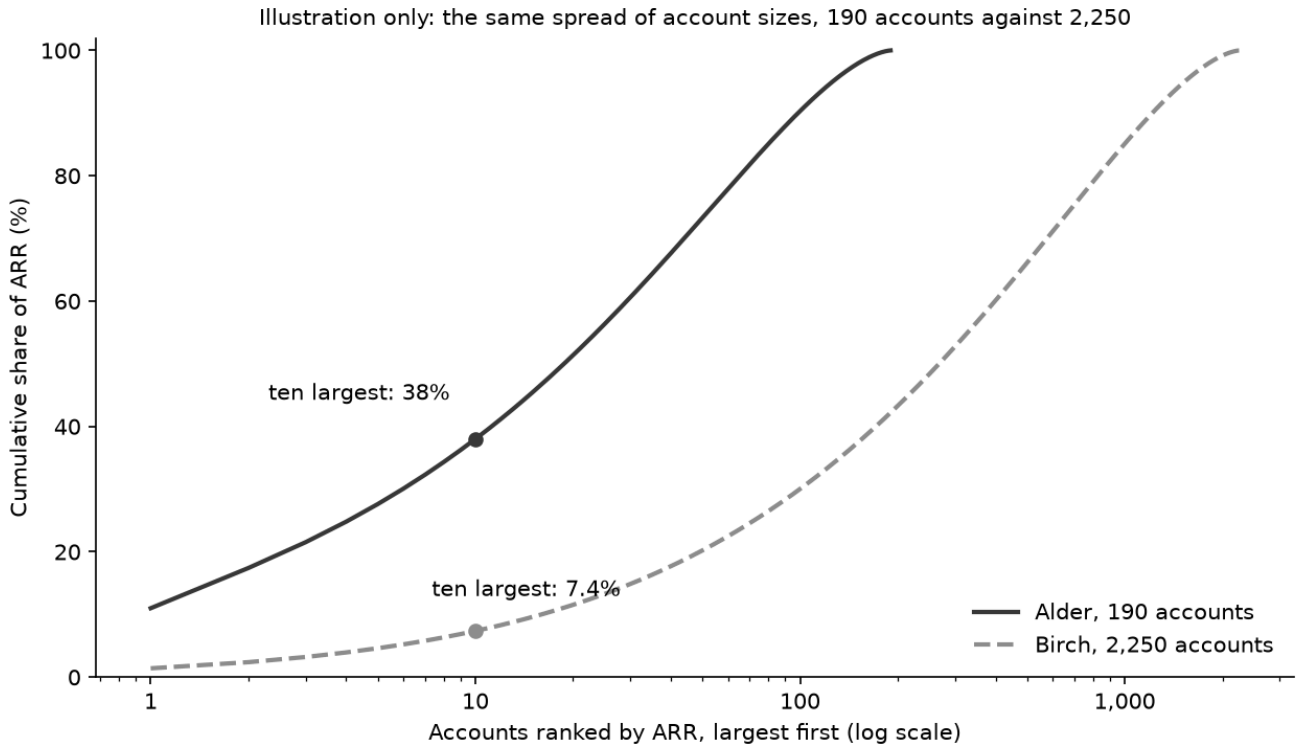
If an interval is wide, the status column says "read the interval" and no red light appears. The analytics lead owns the panel, and the finance owner reads it with the sales owner each quarter.

Concentration

Retention reads differently when a few accounts hold a large part of ARR. **ARR concentration** is the share of ARR in the ten largest accounts.

> **Calculation**
>
> ARR concentration = ARR of the ten largest accounts ÷ total ARR

The effect of one account on GRR equals its share of starting ARR. Alder's largest account holds 1,320,000, or 11.0% of the 12,000,000. If it leaves, GRR falls 11.0 points. If it halves its contract, GRR falls 5.5 points. All the losses that Alder expects in the year (churn plus contraction) come to 867,062, or 7.2% of ARR, which is less than this one account. At the same time, the account is only 0.5% of the customers, so logo retention moves by about half a point. One account changes the revenue story and barely touches the customer story.



Cumulative share of ARR by account rank. Illustration only: the account sizes of Alder follow the 11% and 38% on file plus a fitted spread, and Birch uses the same spread over 2,250 accounts.
The figure applies one spread of account sizes to two companies. With 190 accounts, the ten largest hold 38% and 19 accounts hold half of ARR. With 2,250 accounts, the same spread puts 7.4% in the ten largest, and it takes 269 accounts to reach half. The largest single account would be 1.4% of ARR. A company of

Birch's size meets a large account loss as a small event, and a company of Alder's size meets it as a turning point.

Two habits follow. Report retention on all accounts, and also without the ten largest. And report logo retention and GRR side by side, because the gap between them shows how much of the result rests on a few accounts. When counts are small, put an interval on the count of lost customers. Alder expects 13 lost customers in a year, with an 80% interval of 8 to 17. That is a range for logo retention of 91.1% to 95.8%.

The three tests of a go-to-market model

Three questions decide if a way of selling is worth feeding. Can more spend buy customers at an acceptable marginal cost? Does the spend repay in time? Do the customers keep their value? These are the **scalability, sustainability and durability tests** (the three phases in Winning by Design, 2023, *Revenue Architecture*). Read each as a forecast, with a horizon and a limit.

Test	Forecast quantity	Horizon and limit	Evidence label
Scalability	Payback of the next customer: marginal cost of the extra spend and the capacity it needs, over the contract value times gross margin	The payback ceiling of the segment: 24 months for enterprise, 18 for mid-market, 12 for small business	Standard practice for the ceilings (Bessemer Venture Partners, 2021, *Scaling to $100 Million*); Plausible but untested at the margin
Sustainability	Chance that payback of the coming 12 months stays inside the ceiling	Same ceiling; chance of a breach below 20%	Plausible but untested
Durability	Value of a new customer on the value basis, divided by its fully loaded CAC	Above 1, with the retention that would flip it	Arithmetic

The default horizon of the scalability test is 24 months, with the segment ceilings as overrides. A limit of 24 months is stricter than the survey median payback for new customers, which was 26 months (KeyBanc Capital Markets, 2021, *Private SaaS Company Survey*). A three-year horizon would pass almost any spend, so it would not separate good spend from bad. The marginal test must not be looser than the average one, because the next customer costs more than the average. Both points are Plausible but untested. For expansion, the same survey gives a median cost of 0.63 per 1.00 of expansion ARR, with 0.41 at the 25th percentile (2020 data, 98 companies). The book sets a ceiling of 0.50 and a good level of 0.40 as start values (Plausible but untested). Our companies spend 0.21 (Alder), 0.26 (Birch) and 0.09 (Cedar).

Scalability needs a response curve, which *Choose acquisition channels* builds. Birch has no measured curve yet; its budget pulses (short planned rises and falls in the spend of one channel) come in Year 2. Its media plan assumes a curve for paid media, so the test can run there as a planning reading (Illustration only). Sustainability needs the Monte Carlo draw of *Value customers and cash*, with the cost fixed and the count of customers random. At Year 1 July, the chance that the next 12 months breach the ceiling is 11% at Alder and 62% at Birch. It is 50% at Cedar's self-serve motion, because its payback of 12.0 months sits on its ceiling of 12. Durability is a ratio. At Year 1 July it is 2.78 at Alder, 2.69 at Birch and 2.66 at Cedar's self-serve motion. The chapter *Value customers and cash* read 2.77, 2.59 and 2.65 at Year 1 April. Birch's ratio moved because its fully loaded CAC fell from 12,160 to 11,705 euro in between. For Birch, every renewal rate would have to fall by about 77 points before the ratio reaches 1, so durability does not decide the question there.

A **planning view** puts these readings on one page. This is a simplified mockup for Birch at Year 1 July. The 12-month range for Birch at Year 1 July comes from the cohort-based forecast in *Set targets and staff capacity*.

Measure	Value	Range or count	Limit	Reading
ARR in 12 months	32.5 million	32.1 to 32.9 million (P10 to P90)	commit level 32.2 million	inside the plan
Seller load, trend	98%	10 sellers, 300 SQLs a window	100%	near the limit
Uncontacted SQLs	110	about 10.3 days of capacity	none yet	up from 54 in June
Payback breach, next 12 months	62%	accrual payback 18.2 months	ceiling 18 months, breach chance 20%	above the limit
Value of a customer to CAC	2.7	five years at 12%	above 1	holds

Stages defined on timestamps

A stage is a state that an item can be in. A lead is in the lead stage. An opportunity is in the opportunity stage. Each stage needs an entry criterion: one event that moves an item into the stage. The event has a timestamp. The count of a stage in a window is the number of items whose entry timestamp falls in that window.

Count by the entry timestamp and never by the current stage. A count by current stage rewrites the past. Take a lead that became an MQL in March and an SQL in April. A count by current stage removes it from March. March then shrinks every time a lead moves on, and two analysts who count on different days report different Marches. A timestamp does not change (Standard practice).

Write one entry criterion for each stage. It must be a fact that anyone can check in the data, for example "the account executive marked the lead as accepted". It must not be a judgement that changes with the reader, for example "the lead looks serious". Stages come in order, and an item enters each stage once per cycle. Some items skip a stage. A visitor who asks for a demo may become an SQL without being an MQL first. Choose one rule for skipped stages, write it down and keep it. A common rule counts the skipped stage at the moment of the next one.

Stage	Entry criterion	Unit	Timestamp comes from
Click	A visitor clicks an ad	click	the ad platform
Session	A visit to the website starts	session	web analytics
Lead	A person gives contact details, or the company records a prospect	lead	marketing automation or the CRM
MQL	The lead meets the written marketing definition	lead	marketing automation
SQL	Sales has spoken to the lead and recorded a real problem and a fit	lead	the CRM
Sales-accepted lead	An account executive accepts the SQL under the hand-off rule	lead	the CRM
Opportunity	A deal with a value and a sales stage is opened	opportunity	the CRM
Closed-won	The contract is signed	opportunity	the CRM, checked against billing

Stage	Entry criterion	Unit	Timestamp comes from
Sign-up	A person creates an account in the product	workspace	product events
Activated	The workspace reaches the activation event	workspace	product events
Paid	The first payment succeeds	workspace or account	billing
Engaged account	An account meets the engagement rule	account	the CRM and web analytics, matched
Product-qualified account	The account reaches the level of product use that signals intent	account	product events

Definitions change over time. A company tightens the MQL rule, adds a stage or moves a boundary. When that happens, a rate changes without any change in the market. Record each definition change as a dated event. Keep the old and the new definition side by side for at least one full cycle, so you can bridge the two series. Never overwrite history with the new rule unless you can recompute every past record under it. *Decide when a change needs action* shows how recorded change events keep a definition change from raising a false alarm.

Choose the unit first

Before you define any rate, choose the unit of the funnel. It is the thing that moves through the stages and gets counted. It can be a lead, a contact, an account or a workspace. The unit sets every denominator. A rate on leads and a rate on accounts answer different questions, and you cannot compare them.

Pick the unit that matches the buying decision. In a self-serve product, one person or one team decides, so the workspace is the unit. In a mid-market sale, one buyer often takes the call, so the lead works well. In a large sale, a group decides, so the account is the unit.

Count each unit once per stage. A person who fills in three forms is one lead. An account with two opportunities is one account and two opportunities. Put the unit in the name of every rate, such as "win rate per opportunity".

Never compute a rate from two populations. The numerator must be a subset of the denominator, with the same unit and the same source. Birch has 66,700 paid clicks in a window and 1,504 paid leads. Divide paid leads by paid clicks and the click-to-lead rate is 2.3%. Divide all 2,894 leads by the same clicks and it is 4.3%. The second figure is wrong, because organic leads have no paid click behind them.

Three metric families: volumes, rates and times

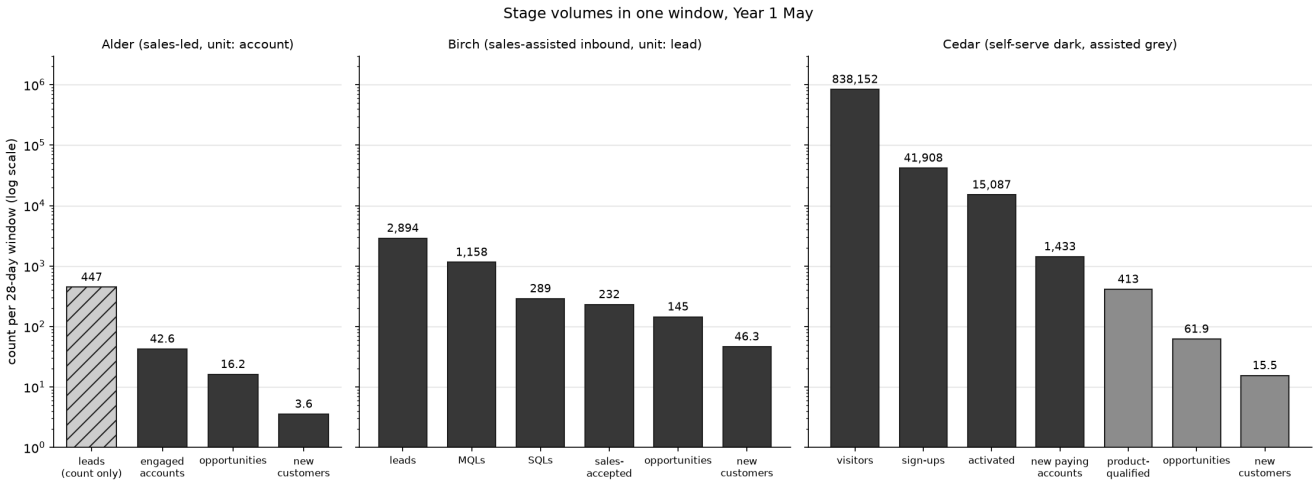
The model uses three kinds of measure. Winning by Design structures its data model the same way (Winning by Design, 2023, *Revenue Architecture*, chapter on the data model). All three are read in the 28-day window of *Choose the growth objective*.

A volume is a count of items that entered a stage in a window. Left of the knot of the bowtie, volumes are counts: leads, SQLs, opportunities. Right of the knot, the book counts euro. One step turns counts into euro: customers times average contract value (ACV). ACV is the yearly recurring value of one contract, without one-off fees. Birch wins 46 customers a window at an ACV of 9,600, which is about 444,000 euro of new annual recurring revenue. Never add or compare a count and a euro amount.

The pipeline is a different kind of volume. It is the set of opportunities that are open on one date. Pipeline volume is their count, and pipeline value is the sum of their expected contract values. Both describe a stock on a date, not a flow in a window, so do not put them in a chain of window counts.

A rate is the share of one stage that reaches the next. Many teams call it a conversion rate. A step rate compares two neighbouring stages. A cumulative rate compares a later stage with an earlier one, and it equals the product of the step rates between them. Read both. Step rates show where the chain loses items. A cumulative rate shows what the chain delivers. Print the two counts beside every rate, because a rate without counts hides whether one deal could have moved it.

Birch, Year 1 May	Count per window	Step rate	Cumulative rate from lead
Leads	2,894		100%
MQLs	1,158	40.0%	40.0%
SQLs	289	25.0%	10.0%
Sales-accepted leads	232	80.0%	8.0%
Opportunities	145	62.5%	5.0%
New customers	46	32.0%	1.6%



Stage volumes of the three companies in one window; the log scale lets stages that differ by a factor of ten thousand share one chart.

Rates also depend on how you pair the two counts. A same-period rate divides the wins of a window by the opportunities of the same window. A cohort rate follows the opportunities of one window to their outcome. The two agree only when volume is steady. *Forecast conversion and timing* explains the difference and when each one misleads. In this chapter, a rate is read on a stated cohort or on a steady flow.

A time is the wait between two timestamps. It is not the work in between. An email takes minutes to write and days to get an answer, and the time is the days. Three rules keep times honest.

- Name the start. Birch counts its sales cycle from the SQL, with a median of 34 days. Alder counts from the opportunity, with a median of 120 days. Cedar's self-serve line counts from the sign-up, with a median of 9 days. You cannot compare these medians.
- Do not add medians. Compute the total time for each item, then take the median of the totals.
- An open item is not slow. It is unfinished. *Forecast conversion and timing* shows how to read it.

The data behind the model

The model needs six kinds of record. Each carries timestamps, and the stage timestamps are what the counts come from.

Record	What it is	Stage timestamps it carries
Contact	One person	first touch
Lead	A contact, or a record of interest, in a sales context	lead created, MQL, SQL, sales-accepted
Account	One company, with an account key that links its other records	engaged, first opportunity, customer since
Opportunity	A potential sale with a value, belonging to one account	opened, closed-won or closed-lost
Subscription	A recurring contract with start, renewal and end dates	start, each renewal, end
Workspace	A space in the product where users work	sign-up, activated, first payment

Product events are time-stamped records of use. They feed the activation and product-qualified stages. A contact belongs to one account. An account holds many contacts, opportunities, subscriptions and workspaces over time. The account key is what makes these joins possible. *Decide what to measure* shows how to build it, how to keep stage events and daily snapshots, and how to rebuild any past view.

The chain is only as good as the joins. If a workspace has no account, its payment cannot reach the account funnel. If an opportunity has no lead, the lead rates miss it. State which records the chain needs, and test the joins before you read a rate.

What changes with volume

The model is the same at every volume. What you can read from it is not. The table shows how many days of flow a 20% change in each rate needs, in Year 1 May, when you compare two periods of that length. *Decide what the evidence can resolve* explains the calculation. A window is 28 days, a quarter 91 days and a year 364 days.

Line	Rate	Days of flow to read a 20% change
Alder	engaged account to opportunity	451
Alder	win rate	2,579
Birch	lead to MQL	6
Birch	SQL to sales-accepted	10
Birch	MQL to SQL	29
Birch	win rate	164
Birch	close rate	203
Cedar, self-serve	sign-up to paid	8

Line	Rate	Days of flow to read a 20% change
Cedar, assisted	product-qualified account to opportunity	154
Cedar, assisted	win rate	543

At low volume (Alder), the unit matters more than the rates, because no rate below the lead reads within a year. Choose the unit by the decision, count accounts, and read cases. At medium volume (Birch), splitting the chain pays. Several early rates read within a quarter, although the close rate needs about seven months. At high volume in the self-serve line (Cedar), every rate reads within a window, so the risk is too many readable metrics. Volume belongs to a decision and not to a company. Cedar's assisted line sits in the low-volume regime inside a high-volume company.

Alder, Year 1 May: the account is the unit

Alder, Year 1 May: about 3.6 new customers a window, 0.9 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.1 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

The new chief executive wants a data-driven funnel. The first picture that Alder draws is a lead funnel, because the marketing tools produce one. It shows 447 leads a window, 3.6% of them reaching an opportunity and 0.8% becoming customers. Those rates look like those of a mid-market firm. They hide what Alder really sells: 3.6 contracts a window, each one decided by a group, from a named list of 1,600 target accounts.

The buying group is the problem. About seven people take part in a typical purchase (Illustration only). Trade-fair scans and webinar sign-ups add colleagues who do not decide. The CRM holds about 10.5 lead records for each engaged account. A rise of 20% in leads is 89 more leads. At that ratio it could be about nine new accounts. It could also be 89 more colleagues at accounts that were engaged already. The lead count cannot tell the two apart. The January argument over three customer counts taught the same lesson: a count needs a stated unit.

Alder moves to account stages. The target stage is the named list of 1,600 accounts. An account becomes engaged on the date when two or more of its people take a defined action. Examples are a meeting with a seller at an event or an answer to a sales message. That is a rule of this example. Opportunity and closed-won follow as usual. Leads stay in the picture as a count, an activity signal and not a stage.

There is a practical limit. The CRM has no account key and no stage history yet. Both arrive in Year 1 October (see *Decide what to measure*). For now, marketing matches each engaged contact to one of the 1,600 names by company domain, and one person checks the newly engaged accounts each window by hand. At 43 accounts a window, the list is small enough to read.

Alder, Year 1 May	Per window	Step rate	Rate per lead
Leads (count only)	447		
Engaged accounts	42.6		
Opportunities	16.2	38.0% of engaged accounts	3.6%
New customers	3.6	22.0% of opportunities	0.8%

The account unit makes the funnel honest. It does not make it readable. The count of newly engaged accounts needs 263 days of flow to read a 20% change. The rate from engaged account to opportunity needs 451 days. So Alder reads accounts as cases. Which accounts are engaged? Which roles are missing? Which are stuck? A written list for the account executives replaces a rate dashboard.

Counts need care too. At 3.6 wins a window, one deal is 28% of the count, so a single signature moves any rate a long way. Alder therefore pools windows before it says that anything changed, and it reports the interval beside every count. The account list also gives a second view that the lead count never could: coverage. Of the 42.6 accounts that become engaged in a window, the sellers can see which have a decision maker engaged and which have only users. That is a case read, and it is where the next deal is won or lost.

Alder's marketer writes down the decision with the date and the two numbers to watch: newly engaged accounts and opportunities per window. Both are counts, because the rates cannot be read. At low volume with large buying groups, the account is the unit, leads are a count, and cases do the work that rates cannot.

Birch, Year 1 May: a sales-accepted stage

Birch, Year 1 May: about 46 new customers a window, 1.0 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 73 days, and the close rate reads a 20% change in about 7 months.

Marketing says the leads are good. Sales says the SQLs are thin. Between the sales development representatives and the account executives, SQLs sit for days. Until now the chain has no stage between the SQL and the opportunity. An SQL leaves its stage only when an account executive opens an opportunity. An SQL that nobody has opened looks the same as one that an executive rejected without a word. The rate from SQL to opportunity, 50%, hides three outcomes: opened, rejected and untouched.

Birch adds a sales-accepted stage after the SQL. This is Birch's own choice: *Choose the growth objective* puts acceptance before qualification. The entry criterion is a CRM action: the executive marks the SQL as accepted, or rejects it with a coded reason. Both are time-stamped. Marketing and sales owners agree the wording in writing, and the stage goes live on a stated date. Earlier months keep only the product of the two new rates. In the first window, sales accepts 232 of 289 SQLs (80.0%), and 145 of the 232 accepted leads become opportunities (62.5%). The product is still 50.0%, so history stays comparable.

Rate	Value	Published span	Days to read a 20% change
Lead to MQL	40.0%	35 to 45%	6
MQL to SQL	25.0%	13 to 51%	29
SQL to opportunity, before the split	50.0%	35 to 50%	39
SQL to sales-accepted	80.0%	none published	10
Sales-accepted to opportunity	62.5%	none published	29
Win rate	32.0%	20 to 40%	164
Close rate, SQL to won	16.0%	12 to 25%	203

The spans come from one vendor page (First Page Sage, 2025, *B2B SaaS funnel conversion benchmarks*) and other secondary sources. They repeat one vendor data set, so read them as wide ranges and not as targets.

The split pays at medium volume, in two steps. The close rate needs 203 days to read a 20% change, about seven months. Its first factor, SQL to opportunity, needs only 39 days, and its second, the win rate, needs 164. So one factor of the close rate already reads within a quarter. The new stage splits SQL to opportunity again: acceptance reads in 10 days, within one window, and acceptance to opportunity in 29. Each part reads faster than the whole, because it has more successes or a higher rate.

The new stage does not cure everything. Each executive works about 30 SQLs a window. Ten executives can work 300, and demand is 289, a load of about 96%. When sellers are this busy, a falling acceptance rate may mean load and not poor leads. The chapter *Find the most valuable constraint* returns to this limit. Birch reads acceptance by role, the whole team of executives, and never by person. *Remove buying friction* gives the full hand-off rule.

The new stage also gives each rate an owner. Marketing owns the MQL rate and reads it with sales. The sales development team owns the step from MQL to SQL. The account executives own acceptance and the rate to opportunity. A rate with a named owner gets read, and a rate without one does not. Birch adds one time metric with the stage: the hours between the SQL and the executive's answer. It is the first number to look at when leads seem to stall. Every SQL gives one wait, about 289 a window, so a large shift in the median wait should show within weeks (Plausible but untested). Count the open SQLs too, because the slowest SQLs are the ones still waiting. *Forecast conversion and timing* shows how.

Birch learns that decomposing the chain gives rates that read by quarter even when the close rate does not. The change is cheap and reversible, so the sales owner decides it by a stated date.

Cedar, Year 1 May: two motions

Cedar, Year 1 May: about 1,400 new paying accounts a window, 1.5 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 30 days; the assisted motion, with 62 opportunities a window, needs about 8 months to read a 30% change.

Cedar's growth team shows one conversion funnel for paying accounts. The sales-assist team of five account executives keeps its own sheet. Each executive can work about 15 opportunities a window, so the team can take 75 and works 62, a load of about 83%. Finance asks for one growth number. The fifteen monthly tests of January work on sign-ups, so they work on the self-serve chain. The analysts now split the company into two revenue lines, each with its own chain.

Cedar, Year 1 May	Self-serve	Assisted
Top of the chain	838,152 visitors	413 product-qualified accounts
Middle	41,908 sign-ups; 15,087 activated	61.9 opportunities
New customers per window	1,433	15.5
ACV	778	16,000
New ARR per window	1,114,487	247,747
Median cycle, with its start	9 days from sign-up	45 days from opportunity
Fully loaded cost per customer	634	4,446

The assisted line brings 1.1% of the new customers and 18.2% of the new ARR. A blended chain would hide both facts. It would also hide the difference in volume. In the self-serve line, sign-up to paid reads a 20% change in 8 days, so a test on the sign-up path can read in days. In the assisted line, the win rate needs 241 days to read a 30% change, so a test on the sales touch can read only a very large effect.

The costs need care. The assisted line looks cheaper per euro of new ARR: 0.28 against 0.82. That is misleading. The assisted cost leaves out the cost of the product-led acquisition that created the product-qualified accounts. Cedar needs a written allocation rule. The finance owner proposes it and records it. The model does not choose it.

Each line reads its own chain. In the self-serve line, 5.0% of visitors sign up, 36.0% of sign-ups activate and 9.5% of the activated pay, so 3.4% of sign-ups end as paying accounts. In the assisted line, 15.0% of product-qualified accounts become opportunities and 25.0% of those are won, so 3.75% of product-qualified accounts end as customers. The two cumulative rates look alike, but they count different units and cannot be compared. One counts workspaces that sign up, and the other counts accounts that the product has already qualified.

Cedar learns that volume belongs to a decision. The same company is a high-volume business at the sign-up page and a low-volume business at the sales desk. Each new customer now counts once, in the motion that closed it, and each experiment states which line it targets.

Alder, Year 2 August: a list that already covers the market

Alder, Year 2 August: about 3.0 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 2.9 years to read even a 30% change, and the lead count reads a 20% change in 23 days.

The marketer reviews the target list before the next budget year. Account selection is one of the two ideas that started in July (*Fund the next best action*). Two questions stand. Does the list cover the market? Do events and partners pay?

Start with the market. The six core countries hold 5,200 plants. At the last count, in Year 1 August, customers ran 1,420 of them (27.3%). That leaves 3,780 plants outside customers. The list holds 1,600 accounts. Alder's records hold no plant count per account, so the review assumes 2 plants a target account (Illustration only). On that assumption the list covers 3,200 plants. That is 84.7% of the plants outside customers, and only 580 plants sit off the list. All three numbers carry the same label. At the list price of a site licence, 21,000 euro, the whole market is worth 109.2 million euro a year, six times the ARR of 18.1 million euro. That is a ceiling, not a forecast.

Now the reach. In Year 2 August, 34.1 target accounts become engaged in a window. That is the August dip (seasonal index 0.80) on a trend of 42.6. The trend is 2.7% of the list a window; the August count is 2.1%. If no account engages twice, the trend gives 553 accounts a year, which is 34.6% of the list. Widening the list would add plants far down the fit order, the quality dilution described under market and account selection (Plausible but untested).

Two budget lines remain. Events cost 1,010,865 euro a year. A new customer is worth 220,768 euro on the value basis, so events repay if they cause 4.6 customers a year, which is 9.5% of the 48.3 new customers. On the early outcome, an engaged account is worth 18,456 euro. Events then repay if they cause 55 engaged accounts a year, which is 9.9% of the 553. The CRM cannot say that an event caused an account. It can say which engaged accounts attended one, because the account key has run since Year 1 October.

Partners source 30% of new customers, which is 14.5 a year. Their fees total 425,627 euro, so each partner customer costs 29,371 euro in fees, which is 13.3% of its value. The fee repays unless 86.7% of those customers would have come anyway. If half would have, the fee per extra customer is 58,742 euro (26.6%). The marketer believes that few would have come anyway, because partners bring plants that Alder has not reached. Nothing in the CRM proves it (Assertion).

What can Alder read? The count of engaged target accounts reads a 20% change in 263 days, and a 40% change in 66 days, which is 2.3 windows. The account holdout of *Choose acquisition channels* still runs, and its July read released the second tranche in full. Events and partners sit outside it.

The marketer is the decision owner and decides on the card, without a meeting. The card has three options: hold the list and both lines; widen the list by the 580 plants (at 2 plants a target account, Illustration only); or raise events. The list and the event plan can both change at the next booking round, so the card treats them as reversible and uses the 90% threshold. No rate can give evidence for any option. The marketer holds the list and both lines, and attaches a break-even card to each event: the engaged accounts that it must cause. The log entry states the belief that engaged target accounts keep their seasonal level. The falsifier is a fall of 40% or more below the seasonal expectation for two windows, which is 20.4 or fewer in a window like August. The confirms-by date is the end of Year 2 November, the date of the holdout read. At low volume, levers are judged on priors and early account metrics.

The decision to carry forward

Choose a reachable customer problem whose full path can earn a return. Preserve rejected and uncontacted accounts in the observation system where possible. Review the profile when mature customer outcomes challenge the value you assumed, allowing for how earlier selection shaped the evidence.

05 · Choose acquisition channels

Where should the next acquisition euro go? Compare the additional discounted gross profit caused by spending it, after incremental costs. An attributed return, an average acquisition cost and a marginal return answer different questions. The marginal return is the change caused by the next increment of spend.

Three relationships must survive the calculation. Spend changes the volume and mix of accounts. Those accounts convert and retain with their own probabilities. Sales and delivery capacity determine whether the company can realise that value. A response curve for leads alone is insufficient when additional leads wait uncontacted.

Use a common cohort age when comparing downstream outcomes. If outcomes have not matured, carry that uncertainty into the allocation. This chapter develops the causal evidence for spend changes, then connects it to channel structure and the signals used to attract demand. Its company episodes show why channel economics and signal density must be read together.

The rule: compare periods with attribution, move budget with incrementality

An attribution model shares the credit for a conversion among the touches before it (see *Decide what to measure*). A model used the same way every period lets you compare one period with the next. It does not tell you what a channel caused: the credit goes to the touches that were recorded, not to the touches that mattered.

Two published results show the size of the problem. Paid search on a brand name added almost no sales, because the buyers would have come anyway (Blake, Nosko and Tadelis, 2015, *Consumer Heterogeneity and Paid Search Effectiveness*, Econometrica). In 15 large field experiments, common observational methods often missed the experimental answer. In half of the studies they were off by a factor of three (Gordon, Zettelmeyer, Bhargava and Chapsky, 2019, *A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook*, Marketing Science). A consistent but biased rule still puts the money in the wrong place.

So the book uses one rule. Use one attribution model to compare periods. Use incrementality tests to move budget (Standard practice for the tests; the rule is the book's).

Incrementality is the part of an outcome that an action caused, against what would have happened anyway. It gives three measures.

> **Calculation**
>
> incremental conversions = conversions with the ads - conversions without the ads
>
> incremental CAC = spend ÷ incremental conversions
>
> incremental ROAS = extra gross profit from the ads ÷ spend
>
> Return on ad spend (ROAS) is credited revenue ÷ spend. It is lagging and descriptive. Do not act on it alone.
>
> Evidence label: Arithmetic.

Average incrementality asks what the whole channel adds. Holdouts and geo experiments answer it. Marginal incrementality asks what the next euro adds. Budget pulses and the response curve answer it. Budget moves on the second question. A channel can add customers on average and still add them at a loss at the margin.

An advertising effect is also small next to the normal swings of sales. Informative experiments on sales can need millions of people (Lewis and Rao, 2015, *The Unfavorable Economics of Measuring the Returns to Advertising*, Quarterly Journal of Economics). A B2B company has far fewer buyers. It must read the earliest outcome that the volume allows, as *Choose whether to test or act* taught: leads, engaged accounts or meetings, not signed contracts.

Channels, budget and campaign structure

Budget between channels

Judge a channel by the incremental value of its next euro, and use attribution only to compare periods. *Choose acquisition channels* set the rule and the designs: budget pulses, geo tests and account holdouts.

For a channel with no incrementality read yet, one number shows how much the credit may overstate before the channel loses money. The **break-even incremental share** is the share of credited conversions that must be incremental for the channel to return one euro of value for each euro spent.

> **Calculation**
>
> break-even incremental share = 1 ÷ credited value per euro
>
> Evidence label: Arithmetic.

Signal density

An **ad set** is the unit of an ad account that holds one budget, one audience and the ads that share them. The **campaign structure** is the way you split the budget into campaigns and ad sets. An ad platform learns by trial: it records who takes the action that you named and shifts delivery towards people who look like them. Each ad set learns alone, from its own events. Split a budget over many ad sets, and each one receives few events, so delivery tends to be less stable (Plausible but untested).

The number that matters is the **signal density**: the optimisation events that an ad set receives each week.

> **Calculation**
>
> events per ad set per week = conversion events that the account sends in a week ÷ ad sets that share them
>
> Evidence label: Arithmetic.

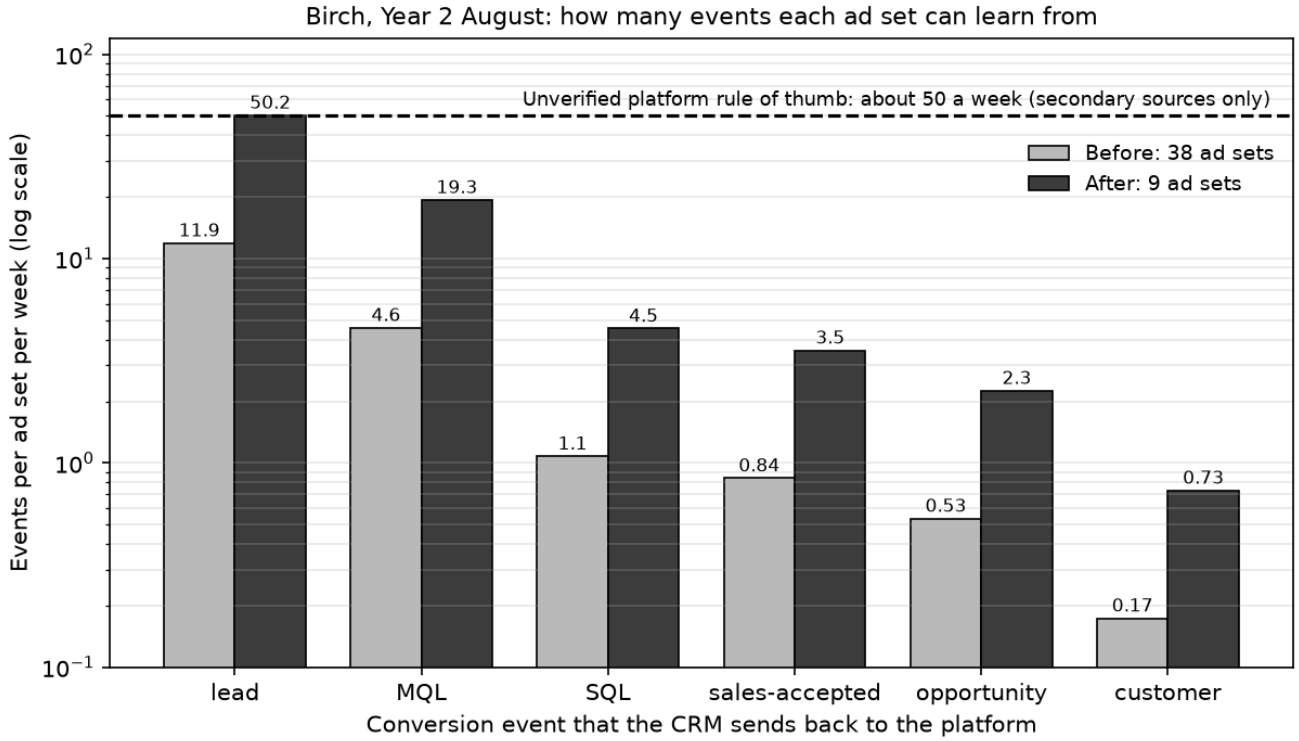
An unverified platform rule of thumb, often quoted, is about 50 events per ad set per week. We could not verify it in the platforms' own documents. The lesson does not depend on the exact figure. The count per ad set per week is a fact that you can compute, and it shows how far an ad set sits from any minimum.

Implementation note: platforms flag an ad set that receives too few events, for example with a status such as "learning limited". Read the count that your platform asks for on its help page.

Which event to send

A **conversion event** is a recorded action that you send to the platform. A lead is the first event. Later outcomes arrive from the CRM: the MQL, the sales-qualified lead (SQL), the opportunity and the customer. An **offline conversion import** sends these later events back to the platform. Each step is closer to money, rarer, and so gives each ad set fewer events.

Take Birch at Year 2 August, before and after it cuts 38 ad sets to 9. The signal-density figure shows the events that each ad set receives each week, by type. Before the cut, each ad set receives 11.9 lead events a week, which is 24% of the unverified rule of thumb, and 4.6 MQL events. After the cut, a lead event reaches 50.2 (100%), an MQL event 19.3 (39%) and an SQL event 4.5 (9%). A customer event reaches 0.73, so an ad set would need 68 weeks to collect the 50 of the unverified rule of thumb.



Signal density. Birch, Year 2 August: events per ad set per week by event type, before and after the cut from 38 to 9 ad sets. The dashed line is an unverified platform rule of thumb (about 50 a week). The counts are company facts.

The pattern holds for any minimum: the deeper the event, the larger the volume you need.

Value-based lead events

A deep event carries information about quality, and a lead event alone does not. **Value-based lead events** solve this: send the lead event with a value that equals the expected gross profit of that lead. The platform then learns from every lead, at once, and learns which leads are worth more.

> **Calculation**
>
> value of a lead = value of a customer × customers per lead
>
> value of an MQL = value of a customer × customers per MQL

> Evidence label: Arithmetic, on the value basis of *Value customers and cash*.

Birch values a customer at 31,538 euro. A lead is worth 460 euro, which is 1.5% of that. An MQL is worth 1,195 euro, which is 3.8%. The **cost per lead** is paid spend divided by paid leads. A paid lead costs Birch 86 euro of media at trend, so the average lead passes a first screen, before the cost of sales and the higher cost of the marginal lead (*Choose acquisition channels*). A flat value of 460 euro for every lead adds nothing to the lead event. Steering needs values that differ by lead type. Take them from lead cohorts that have matured (*Forecast conversion and timing*), in a few coarse bands, and refresh them each quarter (Plausible but untested start values). The bands are values, not new ad sets. *Allocate sales attention* shows how to build the score behind such values.

What consolidation costs

Fewer ad sets give each one more events, but they cost control. You can no longer steer money to a thin slice of the audience or test audiences by splitting them, because every split starves both halves. Three habits compensate.

- Test creative and pages inside the ad set, with a random split by person or visitor.
- Move budget between channels with pulses and geo tests, as *Choose acquisition channels* shows.
- Record the change as a change event (*Decide when a change needs action*). A consolidation shifts the series, and the detector must expect the shift.

A consolidation can also change the slope of a spend response, so read the curves as a description of the whole span.

Targeting and audiences

Targeting decides who sees the ad. Narrow targeting raises the share of impressions that reach a fit. It tends to raise the **cost per thousand impressions**, the price of 1,000 showings of the ad (Plausible but untested). It also shrinks the pool, and so the events that each ad set receives. Broad targeting leans on the platform to find the fit, and it works only when the signal is dense. So the best choice depends on the signal density.

Two simple rules save money at any volume. Exclude your current customers from acquisition ads, through an audience under a contract, unless the ad is about expansion. Keep audiences from overlapping, because two ad sets that reach the same people can compete in the same auctions.

An **audience** is a group of people that you target. A list of customers, visitors or leads sent to a platform is data about people, so every audience gets one **audience contract**, a written record with four fields.

- **Owner.** A role, such as the marketing owner.
- **Purpose.** What the audience is for, and what it may not be used for.
- **Consent basis.** The lawful basis for each person on it (Regulation (EU) 2016/679, Article 6).
- **Expiry.** The date on which the audience is deleted or renewed.

Three monitors keep the contract true: the match rate between your list and the platform, sudden changes in the size of the audience, and the time that a withdrawn consent takes to reach the platform. A start value for that time is 24 hours (Plausible but untested).

Lawful basis comes before any modelled value. Hashing is pseudonymisation, not anonymisation, and a model that scores people for the platform can count as profiling (*Choose acquisition channels*). Ask counsel before you send events, lists or modelled values to an ad platform. This is not legal advice.

The proof design for targeting is the audience holdout. It keeps a random share of an audience from seeing the ads and needs matched personal data and enough conversions. At low volume, judge audiences by the contract and the monitors instead.

The demand baseline

Some of the demand that your ads capture existed before the ads ran. Paid search on a brand name shows this case. Buyers who search for the brand come anyway (Blake, Nosko and Tadelis, 2015; see *Choose acquisition channels*). The demand that exists before your campaigns act is the **demand baseline**: a baseline, not a conversion claim.

One reading of it is the **branded search share**.

> **Calculation**
>
> branded search share = searches for your brand ÷ searches for all brands in the category
>
> Evidence label: Arithmetic. The denominator is an estimate; say which source gives it.

Implementation note: the webmaster report of a search engine lists the queries that led to your site, so you can count your branded queries. The count of all brands comes from keyword-volume estimates of a data provider. Check how each source defines a branded query.

Read the series as a weekly count with its season, as *Decide when a change needs action* teaches. A fall in branded searches while paid spend holds means that paid media depends on a shrinking base. A cut in prospecting may show up as a later fall in branded searches (Plausible but untested). Small companies have few branded searches, so read them by quarter and with their Poisson range.

Answer-engine visibility is how often AI answer tools name or cite your company when asked questions in your category. Measure it with a fixed list of questions, each asked many times, because the answers vary from run to run. The measure is the share of answers that name you, with a binomial interval. Record the date and the tool, because a tool changes without notice. Count sessions that arrive from these tools as a source of their own. Neither measure is a conversion claim.

Incrementality designs

Four designs cover most B2B cases. They differ in what is randomised.

Design	Randomised unit	What varies	Reads on	Main weakness
Audience holdout or platform lift study	Person or ad impression	Some people are not shown the ads	Conversions of both groups	Needs personal data matched to the platform; sees one platform
Geo experiment	Region	Ads on or off in whole regions	Sign-ups or leads by region and week	Needs many regions of similar behaviour; spill-over across borders

Design	Randomised unit	What varies	Reads on	Main weakness
Account holdout	Account	Some target accounts get no programme touches	Engaged accounts, meetings	Few accounts; buyers talk to each other
Budget pulse	Fortnight	Spend up or down for two weeks	Leads by channel and week	Reads a curve inside the tested range only

Audience holdouts. A platform lift study keeps a random share of the target audience from seeing the ads. Some studies show a placebo ad to the held-out group, so that both groups are treated alike. Others use "ghost ads". The system records when an ad would have been shown to the held-out person. This avoids the cost of the placebo (Johnson, Lewis and Nubbemeyer, 2017, *Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness*, Journal of Marketing Research). The result is clean for that platform and that audience. It needs the platform to match your conversions to its users, and the last section of this chapter says what that needs in law.

Geo experiments. Ads run in some regions and not in others. Regions are assigned at random, or matched in pairs by size and then assigned at random within each pair (Vaver and Koehler, 2011, *Measuring Ad Effectiveness Using Geo Experiments*, Google). You compare the change in results in treated regions with the change in control regions. Measuring change, not level, removes the large and lasting differences between regions.

The reading uses randomisation inference: compare the observed effect with the effects obtained under the assignments allowed by the design. *Choose whether to test or act* develops the method. For a geo test the important consequence is the floor: the number of possible draws sets the smallest p-value that the test can ever give.

Design of the draw	Regions	Possible assignments	Smallest one-sided p-value
One treated region of nine	9	9	0.111
Three treated regions of nine	9	84	0.012
12 treated of 24, free draw	24	2,704,156	0.0000004
12 treated of 24, drawn within 12 matched pairs	24	4,096	0.00024

The floor of the p-value for four geo designs. Arithmetic.

One treated region out of nine can never give a p-value below one in nine, however large its effect. With few regions, use another design. With 24 regions, matched pairs keep the groups balanced on size and still leave 4,096 draws, enough to resolve a p-value of 0.001.

Two rules protect a geo test. Choose regions where people rarely see ads meant for the next region, because spill-over pulls the two groups together and hides the effect. Then add a pre-period of several weeks before the switch and a cool-down after it. The pre-period shows the level before, and the cool-down shows any carry-over after.

Account holdouts. In account-based marketing the unit is the account. A random share of the target accounts gets no programme touches. You compare engaged accounts and meetings booked, per account. Accounts are few, so the read is coarse, and the design guards mainly against paying for a programme that does nothing. Colleagues at a held-out account may still meet a treated colleague at an event. That bias pulls the groups together, so a measured lift is, if anything, an understatement.

Budget pulses

Spend elasticity is the percentage change in an outcome for a 1% change in spend, locally around a stated spend level. Its **standard error** describes sampling uncertainty in the estimated elasticity. **Overdispersion** means counts vary more than a simple Poisson model predicts. The full noise treatment is in *Decide what the evidence can resolve*.

A **budget pulse** is a planned change in the spend of one channel for a short time. The size is up to 30%, up or down, for two weeks. The pulses are the book's always-on test for the spend response (the 30% size is a start value, Plausible but untested).

Why not read the curve from normal spend? A regression can find the slope of a curve only where spend varies. A steady plan varies a little. Worse, the small variation is not random. Managers raise spend in strong months and cut it in weak ones. So spend and results move together for reasons that have nothing to do with the ads. A pulse gives large variation, and its timing comes from a random draw, so it cannot follow demand.

> **Calculation**
>
> standard error of the elasticity = weekly noise ÷ (standard deviation of log spend × square root of the weeks)
>
> weekly noise = relative standard deviation of a week's count of leads
>
> Evidence label: Arithmetic. It is a best case: it assumes no other source of error.

Take the professional-network ads of Birch at Year 2 November. A week brings about 90 leads that carry a source. With the start overdispersion of 1.2, a week's count has a relative noise of 0.116. A steady plan that moves weekly spend by a random 5% has a standard error of 0.45 after 26 weeks and 0.16 after four years. The pulses of Figure 1 reach 0.19 after 26 weeks. A steady plan needs about three years to match them, and only if its variation is random.

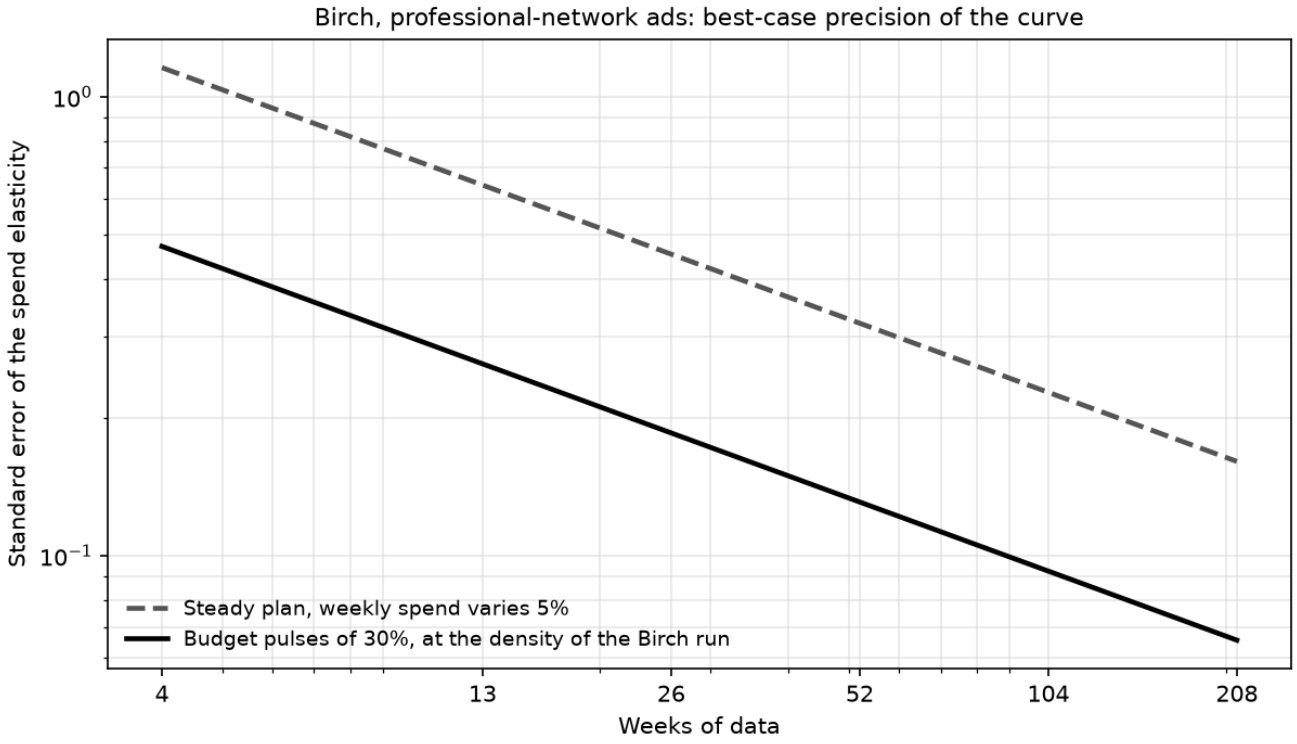


Figure 1. Birch, professional-network ads: the best-case standard error of the spend elasticity against the weeks of data. Arithmetic on the company facts.
Six rules make a pulse schedule usable.

1. **One channel at a time.** Pulse a single channel in a month. Then every change in the total has one cause.
2. **Draw the fortnight at random.** Within the month, choose the two pulsed weeks by lot. The other weeks of the month are the control weeks.
3. **Keep the schedule neutral in euro.** Rises and falls cancel over the plan, so the test costs no net budget.
4. **Read the earliest outcome.** For paid media, read leads, then translate to customers with the stage rates of the funnel.
5. **Compare inside the month.** Each month has its own level, which holds the season, the trend and any release. A change that lifts a whole month drops out of the comparison.
6. **Watch the platform.** A 30% step can reset a platform's learning. If it does, shrink the step, or vary spend by region.

A pulse also shows carry-over. **Carry-over**, also called adstock, is the effect of spend that continues in later weeks. Suppose that a week's effect decays so that half remains after two weeks (a half-life of two weeks, Illustration only). The weekly carry-over share is then 0.71. A week of spend keeps 50% of its effect after two weeks and 25% after four. The effect summed over all weeks is 3.4 times the effect of the first week. If the weeks after a pulse still differ from normal weeks, the carry-over is long, and a two-week pulse reads only part of the effect.

The spend response curve

The **spend response curve** links the spend on a channel to its results. It bends. The channel reaches the best-fitting audience first, and each later euro reaches a worse one. **Saturation** is this bend. Marketing mix models fit saturation and carry-over together (Jin and others, 2017, *Bayesian Methods for Media Mix Modeling with Carryover and Shape Effects*, Google).

Inside a narrow range, a power curve is enough (Standard practice for response functions; Hanssens, Parsons and Schultz, 2001, *Market Response Models*). Leads equal a constant times spend to the power of the **spend elasticity**, which *Find the most valuable constraint* defined. Concave means an elasticity below 1.

> **Calculation**
>
> leads = L0 × (spend ÷ S0)^e
>
> marginal cost per lead = spend ÷ (e × leads) = average cost per lead ÷ e
>
> marginal CAC = marginal cost per lead ÷ customers per lead
>
> marginal payback in months = marginal CAC ÷ (contract value × gross margin ÷ 12)
>
> Evidence label: Arithmetic, inside the tested range.

The **marginal CAC** is the cost of the next customer, not the average one. Take Birch's professional-network ads on the plan value of 0.7 at Year 2 November. The channel spends 74,567 euro a window for 484 leads, so a lead costs 154 euro on average. The next lead costs 154 ÷ 0.7 = 220 euro. A lead becomes a customer with a probability of 0.0155, which is one in 64.6. The average customer costs 9,945 euro and the next one 14,206 euro.

A new customer pays 9,600 euro a year at a gross margin of 80.5%, which is 644 euro of gross profit a month. The next customer therefore pays back in 22.1 months. The payback ceiling for Birch is 18 months (see *Set targets and staff capacity*), which allows a marginal CAC of 11,585 euro. The next customer is over the limit.

Two cautions apply to this arithmetic. First, the customers per lead is the average of the whole stream. A marginal lead may be worse (see quality dilution in *Find the most valuable constraint*). Check it by following lead cohorts to customers, by source, once the sales cycle has passed.

Second, the curve is known only where spend was varied. Spend at 70% to 130% of today is the **tested range**. Outside it, the curve is an extrapolation, and the power law has no saturation point of its own. Define the **saturation point** as the spend at which the marginal CAC reaches the payback ceiling. Headroom is the room between today's spend and that point. Report the headroom inside the tested range. Beyond it, say that the data do not know.

Marketing mix models

A **marketing mix model (MMM)** is a regression of a total result, such as weekly sign-ups, on the spend of every channel over a long history. It also holds price, season and events. Each channel gets a saturation curve and a carry-over (Jin and others, 2017). Its strengths are that it uses no personal data and sees every channel at once.

It needs about two years of weekly data and stable channels. It needs spend that varies, and varies differently in each channel. It needs measured controls for everything else that moves sign-ups. A model with six channels, each with a slope, a bend and a carry-over, has 18 parameters to fit from about 100 weeks. When the data are thin, the priors decide the result, and they can bias it (Jin and others, 2017).

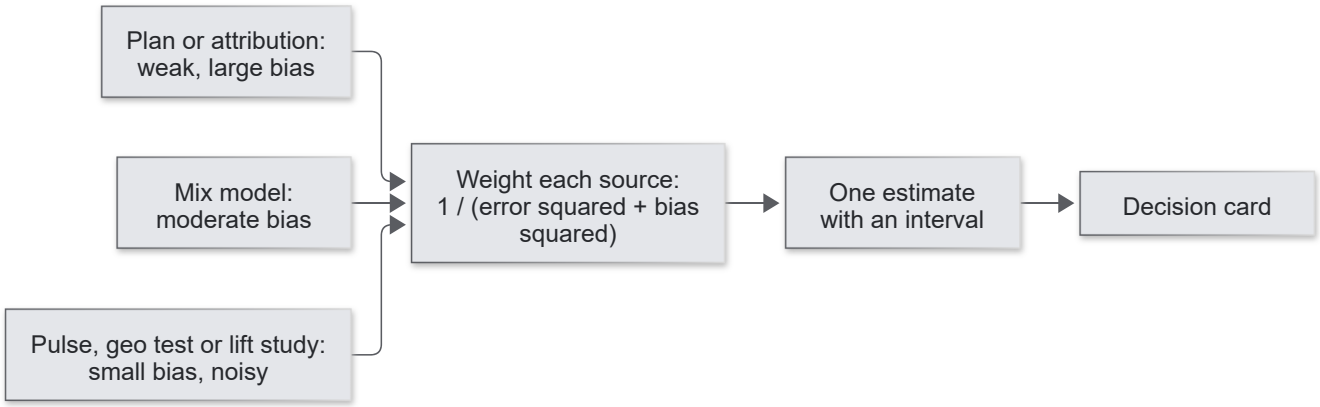
It fails in known ways. Spend that follows the season is collinear with the season. Channels that move together cannot be told apart. A new channel has no history. A model is slow to follow a change in the mix (*Marketing Measurement and Analytics*, chapter on multi-channel measurement).

Take Cedar's paid social at Year 2 May. It spends 144,581 euro a window and has 69 weeks of history. The weekly national sign-ups are 12,048, and we assume that 4% of them wander for reasons that season and trend do not explain (Illustration only). If weekly spend varies by 5%, even a model with no confounding has a standard error that is 96% of the true effect. At 10% variation it is 48%, and at 30% it is 16%. Real models do worse.

Calibration by experiments repairs this. Treat the result of a randomised test as a prior or a likelihood term of the model. Each new lift test, geo test or pulse then narrows one channel. Rotate the next test to the channel whose interval is widest. The model forecasts and allocates. The experiments anchor it.

Combining evidence

A company often holds several estimates of one quantity: a plan, a pulse read, a lift study, a mix model. Rank them by bias, but do not let the ranking decide. A small noisy holdout must not override a better model by rule. Weigh each source by its precision, with an allowance for its bias.



Evidence sources weighted into one estimate that feeds the decision card of Commit with uncertainty.

> **Calculation**
>
> weight of a source = 1 ÷ (standard error² + bias allowance²)
>
> combined estimate = sum of (weight × estimate) ÷ sum of weights
>
> combined error = square root of (1 ÷ sum of weights)
>
> Evidence label: Arithmetic. The bias allowances are Plausible but untested start values.

The **bias allowance** is the width that you add for a source because it may be wrong in a way that more data cannot cure. It is never zero. A source with a large sample but a bias allowance of 5% is worth no more than its bias. The start values are a share of the estimate. Use 5% for a randomised test and 20% for a calibrated mix model. Use 40% for a mix model that no experiment has calibrated, and 50% for an attribution report or a plan that nobody has tested. A source with a low estimate gets a small allowance, which is a weakness of this start value. Replace them with your own track record as the decision log grows (see *Commit with uncertainty*).

Take Birch's professional-network ads. The plan assumed an elasticity of 0.70, in a range of 0.50 to 0.90 whose standard deviation is 0.082. The bias allowance for an untested plan is 0.35. Its weight is 1 ÷ (0.082² + 0.35²) = 7.7. The pulses read 0.60 with a standard error of 0.25 and a bias allowance of 0.03. Their weight is 16.0.

The pulses carry 67% of the total, the plan 33%. The combined elasticity is 0.63 with a standard error of 0.21. When the plan value does not sit in the middle of its range, the mean of the range enters the combination. For a plan value of 0.80 in the same range, that mean is 0.73.

Allocation

Budget goes where the next euro earns most. For each channel, compute the **marginal gross profit per euro**: the value of a customer on the value basis, divided by the marginal CAC. Move spend from channels with a low value to channels with a high value until the values are equal. The payback ceiling and the tested range are limits on the move, and capacity in sales is another.

> **Calculation**
>
> marginal gross profit per euro = value of a customer ÷ marginal CAC

> Evidence label: Arithmetic. The value basis is the one of *Value customers and cash*.

Three habits keep the move honest.

- **Move in steps.** Move to the edge of the tested range. Run a new round of pulses at the new spend. Move again.
- **Print the gain as an upper bound.** The gain assumes four things. The leads of the new channel convert like the old ones. Channels do not interact. The curve holds. Sellers can work the extra leads. It also uses estimates that were chosen because they looked best, which overstates them. State the switching value: how much worse the new leads may convert before the move loses.
- **Check capacity.** Extra leads that nobody can work add cost and no customers. See the capacity constraint in *Allocate sales attention*.

A move of budget between channels is reversible: spend can go back within two weeks. It therefore ships at the 90% threshold of the decision card, not at 95%.

Sending data to ad platforms

Several of these designs, and the platforms' own learning, need data about people. Examples are a customer list, conversions, offline sales sent back to the click, or a modelled value per lead.

Hashing an email address does not make the data anonymous. It is pseudonymisation. **Pseudonymisation** replaces direct identifiers with codes. The data stay personal data when anyone can link them back (Regulation (EU) 2016/679, Article 4(5) and Recital 26; European Data Protection Board, 2025, *Guidelines 01/2025 on Pseudonymisation*, version for public consultation). Sending them to a platform needs a lawful basis (Article 6), and for advertising this is often consent. **Profiling** is automated processing of personal data to judge or predict aspects of a person, such as interest or behaviour. A model that scores people for the platform can fall under it.

Ask counsel before you send any data about people to an ad platform. This is not legal advice. Prefer the designs that need no such data. Budget pulses and geo experiments use spend and counts by region and week. Account holdouts hold out companies, but matching ads to the people at those companies is personal data again.

Implementation note: open-source packages exist for geo experiments (GeoLift) and for Bayesian mix models (Robyn, Meridian, PyMC-Marketing).

Tracking as a precondition of every lever

A lever that you cannot read is a lever that you cannot price. Every top-of-funnel lever depends on four things from *Decide what to measure*.

- **The tracking plan.** Fire the lead event only after the CRM confirms the lead, and count leads in the CRM.
- **Source tags.** Set UTM parameters in the ad accounts, never by hand, and store them in hidden form fields so that the lead carries them. Store the ad click identifier with the lead when you may do so, because offline conversions need it.
- **The consent capture ratio.** Read it before you compare sessions with clicks. A banner release moves it, and it can fake a rise in click-to-lead.

- **A single clock.** Set the time zone of each ad account once. Platforms cut the day at different hours, so store event times in UTC and report in one time zone.

If a data gate fails, repair the data before you read a lever (*Decide when a change needs action*). Record every release of a page, a tag or an ad set as a change event. A test that runs while tracking breaks gives a wrong answer with a tight interval.

What changes with volume

Regime	What the volume can read	The design that fits	What to decide on
Low (Alder)	Engaged accounts and meetings, only for large effects	Account holdout; early outcomes	Priors and signposts; the holdout guards against waste
Medium (Birch)	Leads by channel and week; customers only by the stage rates	Budget pulses, read on leads, over two quarters	The curve inside the tested range, weighted with the plan
High (Cedar)	Sign-ups by region and week; many regions	Geo experiments; a calibrated mix model	Measured incrementality by channel

Outside these ranges, decide on the prior, write the signposts and revisit.

Alder, Year 2 May: does the account-based programme work?

Alder, Year 2 May: about 3.9 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 2.9 years to read even a 30% change, and the lead count reads a 20% change in 24 days.

The CRM shows that most of last year's opportunities involved a programme touch. That is a description, not a cause. Accounts that look like good buyers get touched most, and good buyers also buy more. The head of sales and the marketer ask the question that budget needs: would these accounts have engaged without the programme?

They hold back 20% of the 400 programme accounts at random. That is 80 accounts with no programme touches, against 320 with them. Trade fairs and partners stay as they are for everyone. The run lasts six months. Alder reads two early outcomes: engaged accounts and meetings booked. A signed contract takes a median of 120 days from opportunity, so contracts cannot be read in the run.

The company-wide rate gives a start value for the holdout. Alder gets 42.1 engaged accounts a window from 1,600 target accounts. Over six months an account engages with a probability of 15.9%. If the programme does nothing, the holdout shows 12.7 engaged accounts and the treated group 51.0. Meetings add a second count. The baseline is 0.10 meetings per account and quarter, so the holdout expects 16 meetings and the treated group 64.

With 80 against 320 accounts, the smallest difference in the engaged share is 15.1 points, a rise of 95%, at 80% power and 5% significance. For meetings, the split reads a rate ratio of about 2. That is a coarse instrument.

The programme's break-even is much smaller. Half a year costs 260,000 euro, the paid media and the content and agency lines of Alder's budget (Illustration only). An engaged account is worth 18,456 euro. That is the customer value of 220,768 euro times the opportunity rate of 38% and the win rate of 22%. The

programme repays its cost if it adds 14.1 engaged accounts. That is 4.4 points of the treated share, a rise of 28%. The split sees a lift of that size with a power of 13%.

So the holdout cannot show that the programme pays. It can show a large effect, or the lack of any. The marketer does not drop the test for that reason. The holdout becomes a signpost, and the money is decided on priors. Written in May:

- The programme runs in two tranches (funding steps) of three months. The first tranche is funded.
- At the end of July, after three months, the holdout is read on engaged accounts and on meetings, with counts. A zero or negative difference is weak evidence. If the first tranche gives exactly its break-even lift, a difference of zero or less still turns up in 26% of runs. If the programme does nothing, it turns up in 50%. The ratio is only about two to one. The full six months raise it to about three to one (17% against 50%).
- If treated accounts show no more engaged accounts and no more meetings than held-out ones, the second tranche is halved. Otherwise it is released in full.
- "No difference" does not mean "no effect" (see *Choose whether to test or act*). The log records this reading rule.

The head of sales is the decision owner. The log entry holds the question and three options: full funding, funding in tranches with the holdout, or stop. It also holds the reading rule, which is the falsifier, and a confirms-by date in Year 2 November. At low volume, budget is decided on priors with signposts. Incrementality is read only on early outcomes, and only when it is large.

Birch, Year 2 May: where is the headroom on each paid channel?

Birch, Year 2 May: about 51 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 65 days, and the close rate reads a 20% change in about 6 months.

Birch spends 165,704 euro a window on three paid channels: professional-network ads 74,567, search ads 57,996 and social ads 33,141. *Set targets and staff capacity* gave Birch's media plan a curve and called it an assumption until the budget tests of Year 2. The attribution ranges of *Decide what to measure* overlap for every pair of channels, so credit cannot rank them. The head of marketing wants to measure the curve. The head of sales has one worry: sellers are at 107% of capacity in May.

The plan is six pulses, one a month, from Year 2 May to Year 2 October. Each channel gets one rise and one fall of 30% for two weeks. The channels follow a fixed rota, which keeps the pulses apart. The within-month comparison covers most of the hidden-trend risk that a random order would remove. Rises and falls cancel, so the six pulses move no net budget. Because the curve bends, they cost 19 leads in six months, which is 0.3 customers.

Month	Channel	Step	Weeks of the run	Extra spend, euro	Weekly leads, pulsed against other weeks
Year 2 May	Professional-network ads	up 30%	4 to 5	11,185	96.5 against 82.0
Year 2 June	Search ads	up 30%	6 to 7	8,699	178.5 against 161.0
Year 2 July	Social ads	up 30%	10 to 11	4,971	123.5 against 117.0

Month	Channel	Step	Weeks of the run	Extra spend, euro	Weekly leads, pulsed against other weeks
Year 2 August	Professional-network ads	down 30%	15 to 16	-11,185	53.5 against 65.7
Year 2 September	Search ads	down 30%	20 to 21	-8,699	114.0 against 164.5
Year 2 October	Social ads	down 30%	25 to 26	-4,971	101.5 against 127.5

Birch's pulse schedule and the weekly leads that carry a source, simulated from the company model with a fixed seed (Illustration only).

Three other changes land in the same months. The shorter form ships in May. The ad sets are consolidated in August and September (see *Improve offers and conversion*). The platforms deliver more leads per euro from October. The month effect in the read absorbs each of them, because every pulse is compared with the unpulsed weeks of its own month.

A change that alters the slope of a channel, not only its level, is not absorbed. The ad-set consolidation could do that, so the read describes the channels across the six months, not one fixed setting.

In Year 2 November the analytics lead reads the curves. Only leads that carry a source count: 74% of all leads since the consent banner of Year 1 October. A regression of weekly leads on the month and on the log of the spend multiplier gives each elasticity.

Channel	Plan (mean of its range)	Pulse read (standard error)	Weight on pulses	Combined, with 80% interval
Professional-network ads	0.70 (0.70)	0.60 (0.25)	67%	0.63 (0.37 to 0.86)
Search ads	0.80 (0.73)	0.77 (0.22)	73%	0.76 (0.50 to 0.93)
Social ads	0.60 (0.67)	0.48 (0.29)	58%	0.56 (0.28 to 0.82)

Spend elasticity of leads by channel at Year 2 November. The combination follows the bias allowances of the section on combining evidence; intervals are cut at 1, because the response is concave. Illustration only.

The intervals are wide, and they overlap. Six months of pulses cannot tell the three curves from each other or from the plan. The simulated leads follow the plan's own curves (Illustration only). So the example shows how precise a real read would be, not that a real plan was right.

As a check, we repeat the whole six months 1,000 times in simulation. The 80% interval holds the true elasticity in 79% to 80% of the runs, as it should. The run shown is not picked for its result: its seed is fixed in advance. Its three readings sit at the 35th, 48th and 32nd percentile of the 1,000 runs. So this run reads each curve a little flatter than a typical run does.

Wide curves still give a clear ranking of the next customer.

Channel	Average cost per customer	Cost of the next customer, with 80% range	Marginal payback, months	Chance inside the 18-month ceiling
Professional-network ads	9,945	15,794 (11,540 to 27,100)	24.5	10%
Search ads	4,120	5,406 (4,453 to 8,293)	8.4	98%

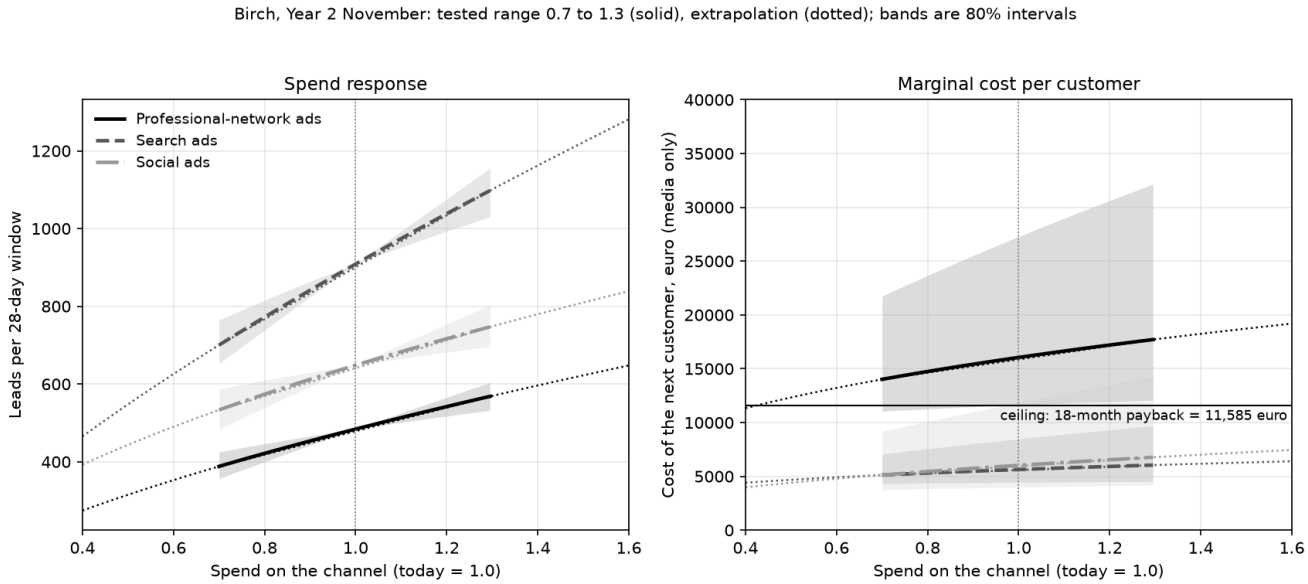
Channel	Average cost per customer	Cost of the next customer, with 80% range	Marginal payback, months	Chance inside the 18-month ceiling
Social ads	3,300	5,905 (4,009 to 11,594)	9.2	90%

Birch at Year 2 November, trend level, media cost only. The ceiling for the next customer is 11,585 euro. Illustration only.

Professional-network ads have the highest cost of the next lead with a probability of 94%. Search and social ads cannot be ranked against each other: each has the lowest cost in about half of the draws. The ranking does not depend on the curve. Even with a linear response, the average cost per customer already differs by a factor of two or more. The pulses add what the average cannot give: how fast the cost rises, and how far a move may go.

At the central estimate, the next customer from professional-network ads costs more than the ceiling even at 70% of today's spend. The chance that it sits inside the ceiling there is only 18%. So that channel has no headroom inside the tested range. Search ads stay inside the ceiling at 130% of today's spend with a chance of 96%, and social ads with 85%. Both have headroom up to the edge of the tested range, and the data do not know how far beyond it.

The spend-response figure shows the curves and the bands.



Spend response. Birch, Year 2 November: spend response and marginal cost per customer by channel. Solid lines lie in the tested range of 70% to 130% of today's spend. Illustration only.
The allocation stays inside the tested range. The best move at the central estimates cuts professional-network ads by 30%, which is 22,370 euro a window. It raises search ads by 30% (17,399 euro) and social ads by 15% (4,971 euro). Net spend does not change. At the central estimates the move loses 97 leads and gains 254, a net gain of 156 leads a window. Over the draws of the three curves the mean gain is 148 leads, with a P10 to P90 range of 72 to 220. At 64.6 leads per customer that is 2.3 more customers a window, or about 30 a year of flow.

On the value basis, a year of that flow is worth about 0.94 million euro, with a P10 to P90 of 0.46 to 1.40 million. This is an upper bound. The shift loses only if the new leads convert at less than 38% of the rate of the old ones. Sellers are at 114% of capacity in November, and the move adds about 13 SQLs a window.

The decision card has three options: do nothing, shift half now and the rest later, or shift everything now. Both shifts add leads in more than 99% of the draws, and customers too if the switching value holds. A shift is reversible, so the 90% threshold applies, and both shifts clear it. The binding limit is the sellers, not the evidence. The head of marketing decides asynchronously, by the end of the week, and writes the entry.

Half the move starts in Year 2 December, when the eleventh seller reaches full capacity. That half adds 77 leads a window, or 1.2 customers. A second round of pulses at the new spend decides the other half. The falsifier reads: if leads from search and social ads convert to customers at less than 38% of the rate of professional-network leads, reverse. The confirms-by date is Year 3 May. At medium volume, two quarters of pulses put a number on how fast each cost rises, where years of steady spend gave none.

Cedar, Year 2 May to June: does paid social add customers?

Cedar, Year 2 May: about 1,700 new paying accounts a window, 1.4 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 25 days; the assisted motion, with 68 opportunities a window, needs about 7 months to read a 30% change.

Cedar spends 144,581 euro a window on paid social, a quarter of its media. The report credits it with 8,032 sign-ups a window at 18 euro each. A sign-up from paid social becomes a paying account with a probability of 2.2%, and a paying account is worth 1,683 euro. A sign-up is worth 37.03 euro, so the report says that each euro returns 2.06 euro of gross profit. Finance wants to raise the budget. The growth lead doubts the credit: some of these people would have signed up anyway.

Cedar sells in 24 regions, enough for a geo test. The analysts sort the regions by size and pair neighbours, which gives 12 matched pairs. In each pair a draw picks the region that pauses paid social for four weeks in Year 2 May. The eight weeks before are the pre-period. The outcome is the weekly rate of sign-ups, in the test weeks against the pre-period. Matched pairs give 4,096 assignments, and the analysts list all of them. The smallest p-value is 0.00024.

The test here is simulated on the company model. The model assumes that paid social causes 60% of the sign-ups that the report credits to it (Illustration only). A real company does not know that input, which is why it runs the test.

In the simulated run, the paused regions fell 8.3 log points more than the others. Paid social therefore adds 7.9% of all sign-ups, with an 80% interval of 4.7% to 10.9%. The one-sided p-value is 0.0056, which is 23 times the floor of 0.00024. Figure 3 shows each region and the randomisation distribution.

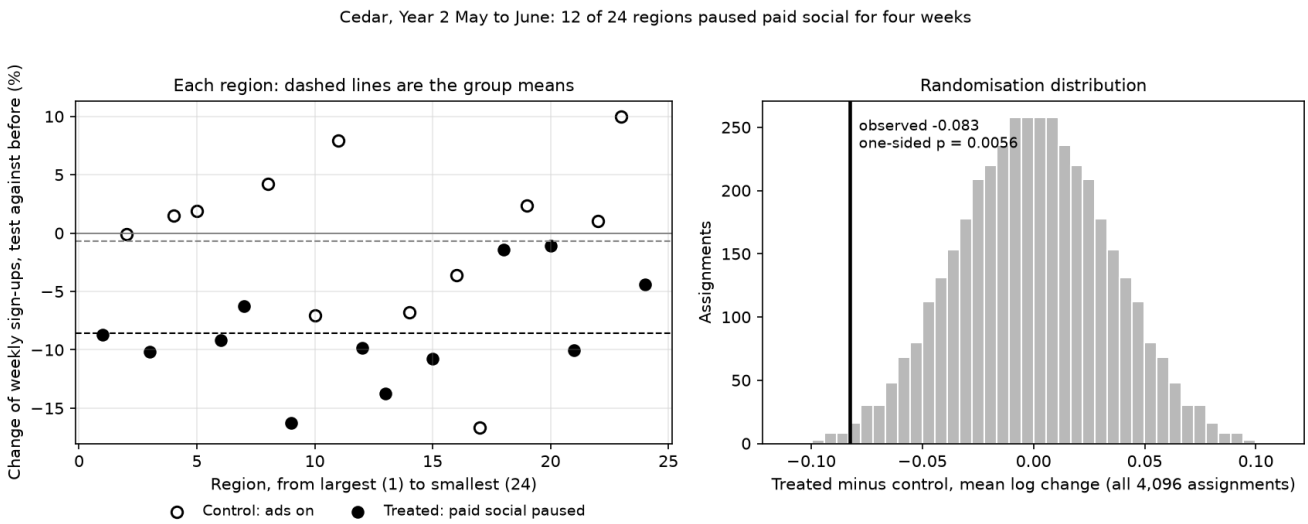


Figure 3. Cedar's geo test: the change of weekly sign-ups in each region, and the difference of the group means against all 4,096 assignments. Simulated (Illustration only).

The test means that 48% of the credited sign-ups are incremental, with an 80% interval of 28% to 65%. The seed of this run is fixed, and it is not picked for its result. Among 1,000 runs of the same test it sits at the 22nd percentile. Most runs read a larger share. Their median is 60%, the true share of the simulation, and their P10 to P90 is 38% to 81%.

The real cost of a sign-up from paid social is 37.8 euro, not 18, with a range of 27.5 to 63.4. Each euro returns 0.98 euro of gross profit, with a range of 0.58 to 1.35. The credit said 2.06. The test itself cost almost nothing. Four weeks of pause saved 83,309 euro of spend in the paused regions. The pause lost sign-ups worth about 81,666 euro.

The mix model shows why the test was needed. Its fitted estimate is 16% incremental, with a standard error of 59 points. That run sits at the 23rd percentile of 1,000 runs. The median run reads 64%, and the P10 to P90 runs from minus 15% to 137%. It cannot read the channel.

The start bias allowances are 5% for the geo test and 40% for the uncalibrated model. With them, the geo test carries 94% of the weight and the model 6%. The combined estimate is 46% incremental, with a standard error of 14 points. The mix model now takes this as the prior for paid social. It reads the other five channels from the same data.

The decision card has three options: raise the budget, hold it, or cut it. At the central estimate, a euro of paid social returns about one euro of gross profit, and the next euro returns less. A raise pays only if the next euro returns more than one euro. The chance that even the average return is at least 1 is 42%, so a raise fails the 90% bar by far. A cut pays if the next euro returns less than one euro. That chance is at least 58%, but the test read the average, not the margin, so the card cannot say how much higher it is. Neither change clears the bar.

The growth lead holds spend flat and funds a creative test, which may raise the return of every euro. The log records the belief: 0.94 euro per euro, from an incremental share of 46% with an 80% range of 28% to 64%. The break-even share is 49%. A second geo test follows the creative test. If it reads a share outside that range, the belief was wrong.

The finding depends on this one run. Had the test read the true 60%, a euro would return about 1.23 euro, and a raise would look better. A real company sees one run only. That is why one test sets a belief, not a final answer, and why the log plans a second test.

Choose acquisition channels judges channels by value per sign-up. Cedar now reads that value as incremental value, not credited value. With many regions a geo test is feasible, and a calibrated mix model becomes useful.

Birch, Year 2 August to September: ad sets with too few events

Birch, Year 2 August: about 51 new customers a window, 0.8 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 65 days, and the close rate reads a 20% change in about 6 months.

Birch runs 38 ad sets on three paid channels. It sends MQL events from the CRM so that the platforms learn from qualified leads. The head of marketing counts the events. A week brings 4.6 MQL events and 1.1 SQL events to each ad set. That is 9% and 2% of the unverified platform rule of thumb. If that rule of thumb holds, the platforms learn slowly from so few events.

The head of marketing is the decision owner and decides on a card, without a meeting. It has three options: keep the structure; cut to 9 ad sets and keep the MQL event; or cut to 9 and send value-based lead events. After the cut, only the lead event reaches the unverified rule of thumb (the signal-density figure). So only the third option gives the platform both many events and a quality signal.

Four conditions come with the choice. Counsel reviews the audience contract and the consent basis first, because the values go to the platforms as data about people. Creative and pages are tested inside the ad sets, split by visitor. Budget between channels still moves by pulses. A change event records the cut.

The same weeks hold a landing-page test on paid traffic. Birch has 54,069 paid sessions a window, and 3.34% become leads. Two windows, split in half, give 54,069 sessions in each arm. The test reads a change of 9.4% at 80% power and 5% two-sided, a rise from 3.34% to 3.65%. The composite of MQLs per session is 1.29%, and the same sessions read only 15.5% on it. So the primary is leads per session. The qualifying share goes under a guardrail with a margin, carried by the detector, as in *Choose whether to test or act*. The visitor split keeps the ad-set change from biasing the arms.

A gain of 9.4% would be 169 more paid leads a window and 15 more SQLs. At 460 euro a lead, that is 77,861 euro of expected gross profit a window. It holds only if the extra leads are as good as the average and sellers can work the extra SQLs. Sellers cannot yet. At trend the sellers run at 106% of capacity in August and 100% in September. The backlog of uncontacted SQLs grows from 48 at the end of August to 66 in September and 106 in October. From Year 2 October the simulated platforms deliver about 8% more leads per euro (Illustration only), which adds SQLs when sellers are full. So the value of the page waits for the eleventh seller, who reaches full capacity in Year 2 December. Judge the value-based events the same way, by the MQL and SQL shares of paid-lead cohorts.

In Year 2 December half of the media shift of *Choose acquisition channels* starts and cuts professional-network ads by 15%. The events per ad set on that channel fall with the spend, so the head of marketing reads the table again after each budget move. The estimate in the later-dated Birch budget episode in this chapter was 77 extra paid leads a window, read from six pulses. The simulated company runs on its true curves, and there the half shift adds 65 paid leads a window (Illustration only). The 77 is a logged belief and the 65 is an outcome. An estimate is not a measurement, but the gap does not grade the entry. Its falsifier is the conversion ratio: search and social leads that convert to customers at less than 38% of the rate of professional-network leads. It stays open until Year 3 May.

The log entry holds the belief that leads per euro hold and that the MQL share of paid leads does not fall by more than the margin. The falsifier is a fall of either beyond the detector's band. The confirms-by date is the end of Year 2 November. At medium volume, only lead-level events give each ad set a count near the minimum that platforms are said to need. Deeper events help through their value, not through their count.

The decision to carry forward

Move spend on evidence about incremental downstream value. Compare the next spend increment, allow for delayed outcomes and check that capacity can absorb the result. Keep attributed value as a reporting convention rather than treating it as the effect of the channel.

06 · Improve offers and conversion

Should you change the offer, the message or the conversion path? Begin with the customer behaviour that could change and the economic result it could produce. A page can increase form submissions by attracting people who will never become valuable customers. A shorter form can also remove useful qualification. The measured conversion gain does not settle the decision.

Define the eligible audience before changing the experience. Keep that audience stable when estimating the effect. Compare the complete paths from exposure through downstream outcomes, allowing for the time those outcomes need to mature. A useful early measure can guide a decision while later value remains uncertain, provided the record makes that limitation visible.

The mechanisms below concern offers, creative and websites. The experiment supports the commercial argument: what changed, for whom, against which alternative, and at what cost. *Choose whether to test or act* gives the shared design and sizing methods.

How to read a lever in this book

Every lever answers five questions.

1. **Mechanism.** How does the lever change what a buyer does?
2. **Metric.** What moves first, and where do we read it?
3. **Proof design.** Which design could show cause? *Choose whether to test or act* and *Choose acquisition channels* hold the designs.
4. **Readability.** What can your volume read in a window, a quarter or a year? *Decide what the evidence can resolve* holds the method.
5. **Value.** What is the change worth on the decision card of *Commit with uncertainty*, in euro on the one value basis?

A lever whose metric your volume cannot read is not a bad lever. You decide it on a prior and a signpost, and you say so on the card. Do not call a lever readable because a dashboard shows a number for it.

Figure 1 follows one visitor from the impression to the lead.

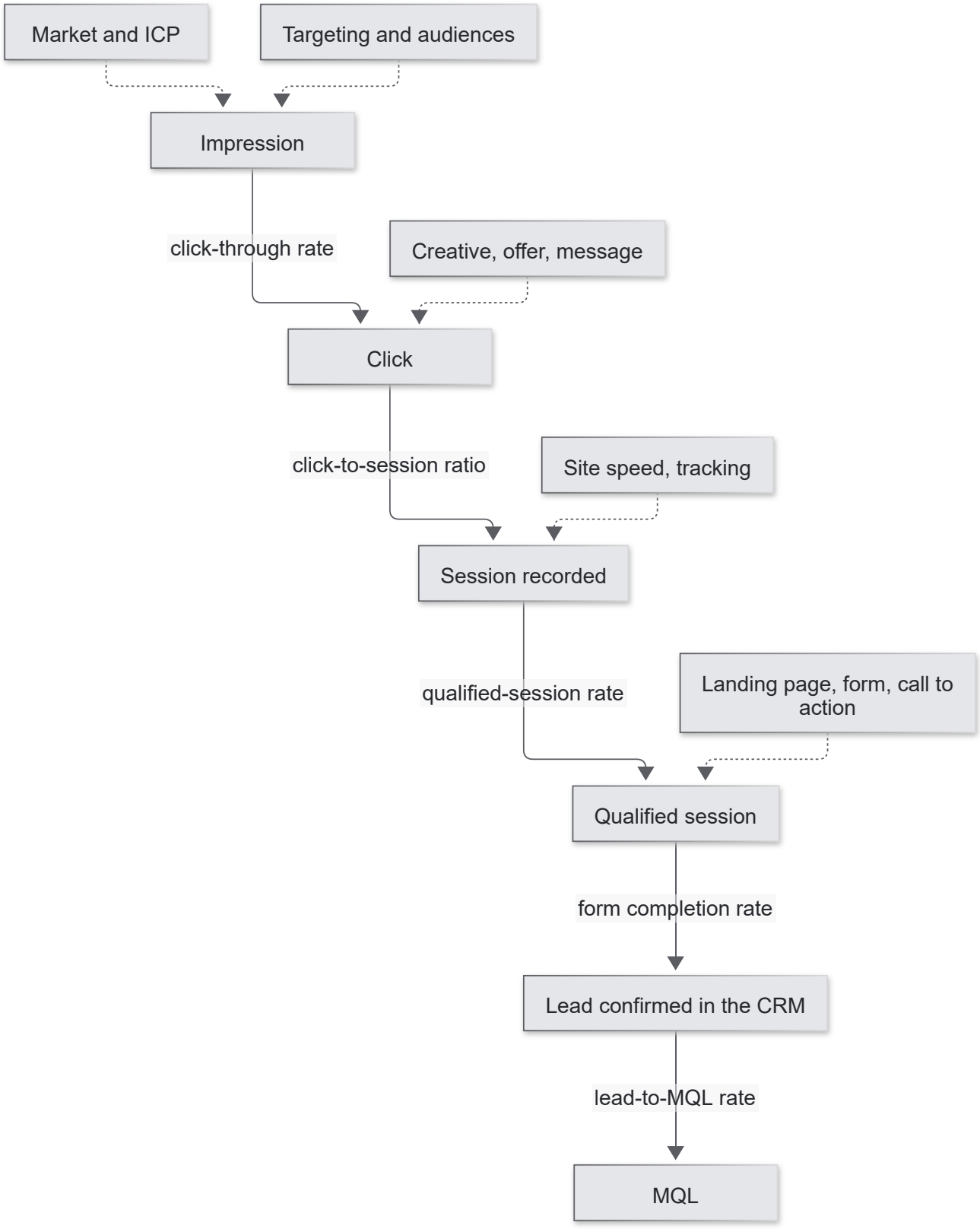


Figure 1. The path from an impression to a marketing-qualified lead (MQL). Solid arrows carry the measuring rate. Dotted arrows show where a lever acts.

Table 1 sets the levers side by side.

Table 1. The top-of-funnel levers: mechanism, first metric, proof design and the regime that can read the proof.

Lever	Mechanism	First metric	Proof design	Readable at
Market and account selection	Puts effort on accounts that fit and keep paying	ICP match rate; TAM coverage	Fit against retained value by cohort; account holdout for a programme	Low volume: cases and early account counts. Medium and high: months of cohorts
Channel budget	Moves money to where the next euro earns most	Marginal CAC; value per sign-up or lead	Budget pulses, geo test, account holdout	Pulses at medium volume; geo tests at high volume; holdout only for large effects at low volume
Campaign structure and conversion events	Gives each ad set enough events to learn, and tells it what a good lead is	Events per ad set per week; MQL share of paid leads	Before and after with a detector and a change event	Counts at any volume; cost per lead at medium and high
Targeting and audiences	Shows the ad to people who fit	Cost per qualified lead; ICP match rate	Audience holdout (personal data); contract and monitors	Medium and high
Offer, message, value proposition	Changes why a visitor acts	Click-to-lead rate; lead-to-MQL rate	Randomised page or ad test with a guardrail	Click-to-lead at medium and high; quality at high
Creative and frequency	Keeps the ad new for the person who sees it	Click-through rate by exposure; frequency	Creative test randomised by person	Clicks at high volume; sign-ups per impression at high volume; money only by quarters
Landing page, form, call to action	Removes friction between a click and a lead	Click-to-lead rate; qualified completions	A/B test by visitor with a guardrail on lead quality	Medium (large changes); high
Site speed and tracking	Lets the visitor stay long enough to be counted and to act	Click-to-session ratio; load time on real visits	Before and after against a detector; staged release	Any volume (a high-count ratio)

A channel is not a lever. It is a place where levers act, so this chapter treats the budget in a channel and the structure of its campaigns as the levers. The demand baseline is a reading, not a lever.

Offer, message, value proposition and creative

Three kinds of principle

Some marketing principles can be tested cheaply, some only at volume, and some not at all. Sort them first.

The **value proposition** is a statement of the result that the customer gets and why it is better than the alternatives. The **offer** is what you give in exchange for an action, for example a demo, a trial or a guide. Both are claims. A claim enters your plan only with a stated prediction and the metric that would refute it.

Table 2. Principles sorted by what can be tested. The metric that would refute each one is named.

Kind	Example principle	Prediction	Metric that would refute it
Testable and cheap	Message match: the page repeats the promise of the ad	Click-to-lead on paid sessions rises	Click-to-lead stays flat or falls
Testable and cheap	Specific proof: a named result with a number	Click-through rate or click-to-lead rises	No change in either
Testable and cheap	Risk reversal: a free trial, or cancellation at any time	Offer acceptance rises and lead quality holds	The MQL share of leads falls beyond a margin

Kind	Example principle	Prediction	Metric that would refute it
Testable and cheap	The must-have claim: buyers cut a long list on a few must-haves (Hauser and others, in *From Little's Law to Marketing Science*, chapter on consideration sets)	A page that leads with the must-have lifts click-to-lead	No lift
Testable only at volume	Story against feature framing	Click-through and lead quality rise	No difference in a large test
Not testable	Positioning "laws", attention-span statistics, stage models of storytelling	None stated	None

The last row is folklore: we do not plan around it. Most ideas that are testable also fail. At Microsoft about one third of experiments helped, one third did nothing and one third hurt (Thomke, 2020, *Experimentation Works*; see *Commit with uncertainty*). So price a message idea with a prior that expects little, and test the cheap ones.

Creative fatigue and frequency

Reach is the number of different people who saw an ad. **Frequency** is the average number of times each of them saw it.

> **Calculation**
>
> frequency = impressions ÷ reach
>
> Evidence label: Arithmetic. Count both over the same period and the same ad.

The **cost per click** is spend divided by the clicks that the platform bills. **Creative fatigue** is the fall in response when the same people see the same ad many times. It shows as a falling click-through rate as exposure grows. Frequency alone hides it, because frequency is an average over people. Read the click-through rate by cumulative exposure to the same ad, per person.

In a simple model, the click-through rate at the first exposure is c, and it falls by a share d with each further exposure. The model is Illustration only. A real company estimates c and d from its own exposure logs.

> **Calculation**
>
> click-through rate at exposure k = c × (1 - d)^(k - 1)
>
> marginal cost of a click at exposure k = cost per thousand impressions ÷ (1,000 × click-through rate at exposure k)
>
> Evidence label: Arithmetic, given c and d.

Take the paid social ads of Cedar at Year 2 September. The click-through rate is 0.80% at the first exposure and falls 12% with each exposure. At the tenth exposure it is 0.25%. The cost per thousand impressions is 8 euro (Illustration only), so the first click costs 1.00 euro and the click at the tenth exposure costs 3.16 euro. A person who sees the ad ten times gives an average of 0.48% over the ten exposures. A person who sees it five times gives 0.63%, which is 31% more clicks per impression. That gain holds only if fresh creative keeps the first-exposure rate (Plausible but untested).

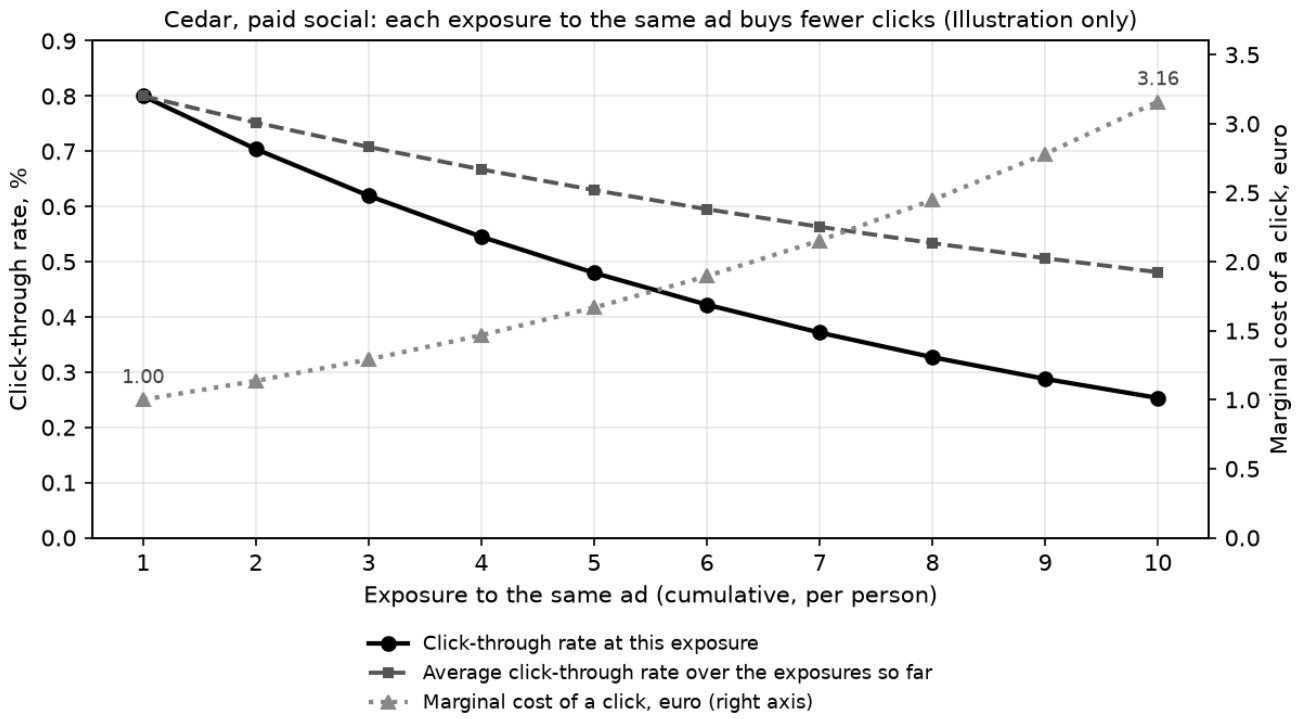


Figure 3. Cedar, paid social: the click-through rate at each exposure, the average over the exposures so far, and the marginal cost of a click. Illustration only.

Set a refresh rule from the marginal cost, not from a fixed number of exposures. Refresh the creative, or cap the frequency, when the marginal cost of a result reaches the value of that result. Use sign-ups where you can, and clicks only when you must.

Testing creative

Randomise creative by person, not by impression. One person sees the same ad many times, so impressions are not independent. The design effect of *Decide what the evidence can resolve* applies, and it grows with frequency. A count of impressions overstates what the test can read.

Read the earliest outcome that the volume allows, and set a guardrail on the next stage rate, with a margin, as the shorter-form test of *Choose whether to test or act* did. A creative that wins on clicks can lose on quality, because it may attract the wrong people.

Website conversion

A **landing page** is the page where a visitor arrives from an ad or a link. It has one job: to make one action easy. The **call to action** is that action, written as a button or a link. Two actions on one page compete for attention (Plausible but untested), so give the page one.

Conversion rate optimisation (CRO) is the practice of testing and improving the rates at which visitors act on pages and forms.

Four metrics measure the lever.

- **Click-to-lead rate.** Leads divided by clicks, from one population (*Choose markets and accounts*).

- **Qualified-session rate.** Qualified sessions divided by sessions. A qualified session is an engaged session (*Decide what to measure*) from a visitor in the target market. You know the company only when the visitor identifies it, so count this rate on identified sessions and say so.
- **Form completion rate.** Forms submitted divided by forms started. Count both as the same unit.
- **Qualified completions.** Submitted forms that later become MQLs, per session. This is the composite of *Choose whether to test or act*.

The last metric matters because a form is a filter. A shorter form raises the number of requests, and some extra requests are weak. The shorter form of Birch raised requests by 12.0% and cut the qualifying share by 3.8% (*Choose whether to test or act*). A form completion rate alone would have called that a win. Read the composite, and put the qualifying share under a guardrail.

A page test has a readable size. Compute it from the rate and the sessions that you can split, with the exact formula of *Choose whether to test or act*, not from a memorised session floor. If the smallest change that the test can read is larger than the change you can plausibly expect, the test cannot show the gain. Then decide on a prior, ship the change if it is cheap and reversible, and move the read to an earlier outcome.

Speed works through a different mechanism. A visitor who leaves before the page loads is never counted, and a tag that fires late misses the visit. So slow pages show first as a low click-to-session ratio (*Decide what to measure*). Speed is an engineering cost, and a proof by randomisation is expensive. Measure the load time on real visits, release changes in stages, and read the click-to-session ratio against a detector. Decide on a prior, with a signpost.

Pages are reversible. A change that a team can undo in a day ships at the 90% threshold of the card.

What changes with volume

The mechanics hold at every volume. What you can read does not.

Regime	What the volume reads	How to decide the lever
Low (Alder)	Counts only. The lead count reads a 20% change in 23 days. Engaged target accounts read a 40% change in 66 days. No rate below the lead reads in a year.	Judge the ICP by cases. Decide on priors, with signposts on engaged target accounts. Use the account holdout for large effects.
Medium (Birch)	A two-window split of paid traffic reads a 9.4% change in click-to-lead and a 15.5% change in qualified completions. Only the lead event reaches the unverified rule of thumb of about 50 a week.	Consolidate ad sets. Send value-based lead events. Test pages with a guardrail. Move budget with pulses.
High (Cedar)	Click-through reads a 10% change in under a day, and one window reads a 6.2% change in sign-ups per impression, both before the design effect. A quarter reads a 24% change in paying accounts from one channel.	Test creative on sign-ups with a guardrail. Judge channels by incremental value per sign-up. Test channels with geo tests.

Even at high volume, money is read slowly, so use the quick metric to choose and the slow one to check.

Cedar, Year 2 September: cheap sign-ups, credited value

Cedar, Year 2 September: about 1,900 new paying accounts a window, 1.3 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 24 days; the assisted motion, with 74 opportunities a window, needs about 7 months to read a 30% change.

Cedar spends 605,905 euro a window on paid media. Paid channels bring 24,355 sign-ups, which is 48.2% of all. They become paying accounts at 2.92%, and the other sign-ups at an implied 4.28%. The growth lead wants to know which channels earn their cost. The analysts want creative tests at scale.

Table 3. Cedar, Year 2 September: cost and value per sign-up, by channel. Value is credited: the report's sign-ups times their conversion.

Channel	Cost per sign-up, euro	Value per sign-up, euro	Value per euro	Break-even incremental share
Video ads	15	30.3	2.02	49.5%
Social ads	18	37.0	2.06	48.6%
Search ads on the second search engine	24	57.2	2.38	41.9%
Search ads	30	60.6	2.02	49.5%
Software review sites	40	84.1	2.10	47.5%
Professional-network ads	58	70.7	1.22	82.1%

The cost ranking and the value ranking nearly reverse. Video and social ads give the cheapest sign-ups and the lowest value per sign-up. Cheap volume comes from deeper in the audience, the pattern that *Find the most valuable constraint* called quality dilution.

Every value in the table is credited, so the last column carries the doubt. Professional-network ads need 82.1% of their credit to be incremental, so that channel sits closest to its limit.

Only social ads have an incrementality read. The credit says 2.06 euro for each euro. The geo test of *Choose acquisition channels* read 0.98 euro, with an 80% range of 0.58 to 1.35, and the log holds a belief of 0.94 euro. That belief rests on an incremental share of 46%, just below the break-even share of 48.6%. The test run sat at the 22nd percentile of 1,000 simulated runs, so a repeat would likely read higher. Had the test read the simulation's true share of 60% (Illustration only; Cedar cannot know it), a euro would return 1.23. The other five channels carry the label "credited, not yet tested". The next test goes where the value of information is largest (*Choose whether to test or act*). Search ads hold 40% of media, 242,362 euro a window. Part of that spend may buy searches for the company's own name, and those buyers come anyway.

Now the creative. Paid social serves 18.9 million impressions a window at 8 euro per thousand (Illustration only). A creative test on clicks reads a 10% change with 204,302 impressions in each arm, under one day of all social impressions, if impressions were independent. They are not, so the real figure is larger. Sign-ups per impression read a 6.2% change in one window, with the same caveat. Paying accounts from one channel read a 24% change in a quarter, and a 10% change needs about 17 windows. So Cedar picks a creative on sign-ups per impression, with a guardrail on the activation rate. It reads the value per sign-up of a channel once a quarter.

The growth lead holds social spend flat, as the card of *Choose acquisition channels* said, and starts the creative tests with random splits by person. The log keeps the belief of 0.94 euro per euro with its range and adds a new one: fresh creative lifts sign-ups per impression. The falsifier is a read of no lift at the size that one window can see, 6.2% before the design effect. A smaller lift stays a prior. Every channel report from now on says "credited" or "incremental" beside the value. At high volume, volume is easy to buy. Judge by value per sign-up, and say which value you mean.

Related company decisions appear in *Choose acquisition channels*, *Choose markets and accounts*. Their original dates remain explicit.

The decision to carry forward

Commit to an offer or conversion change for a defined population. Read the early outcome with downstream guardrails and a dated later read. Expand only as far as the evidence and reversibility of the decision justify.

07 · Allocate sales attention

Who should receive a sales conversation when the team cannot contact everyone? Forecasting who will buy is one input. The decision is which conversation adds the most expected gross profit for the scarce time it uses.

A prospect can buy without a call. Another can need help despite having a lower predicted purchase probability. A score trained on previous sales choices also sees a selected history. These are parts of the commercial decision, not statistical details to check after the queue has been allocated.

This chapter connects scoring, exploration and product-qualified accounts with the opportunity cost of capacity. Read its forecast scores as forecasts. Estimate the effect of a contact separately, with an appropriate comparison. Where the evidence cannot distinguish the alternatives, make the simplifying assumptions explicit and keep the decision reversible.

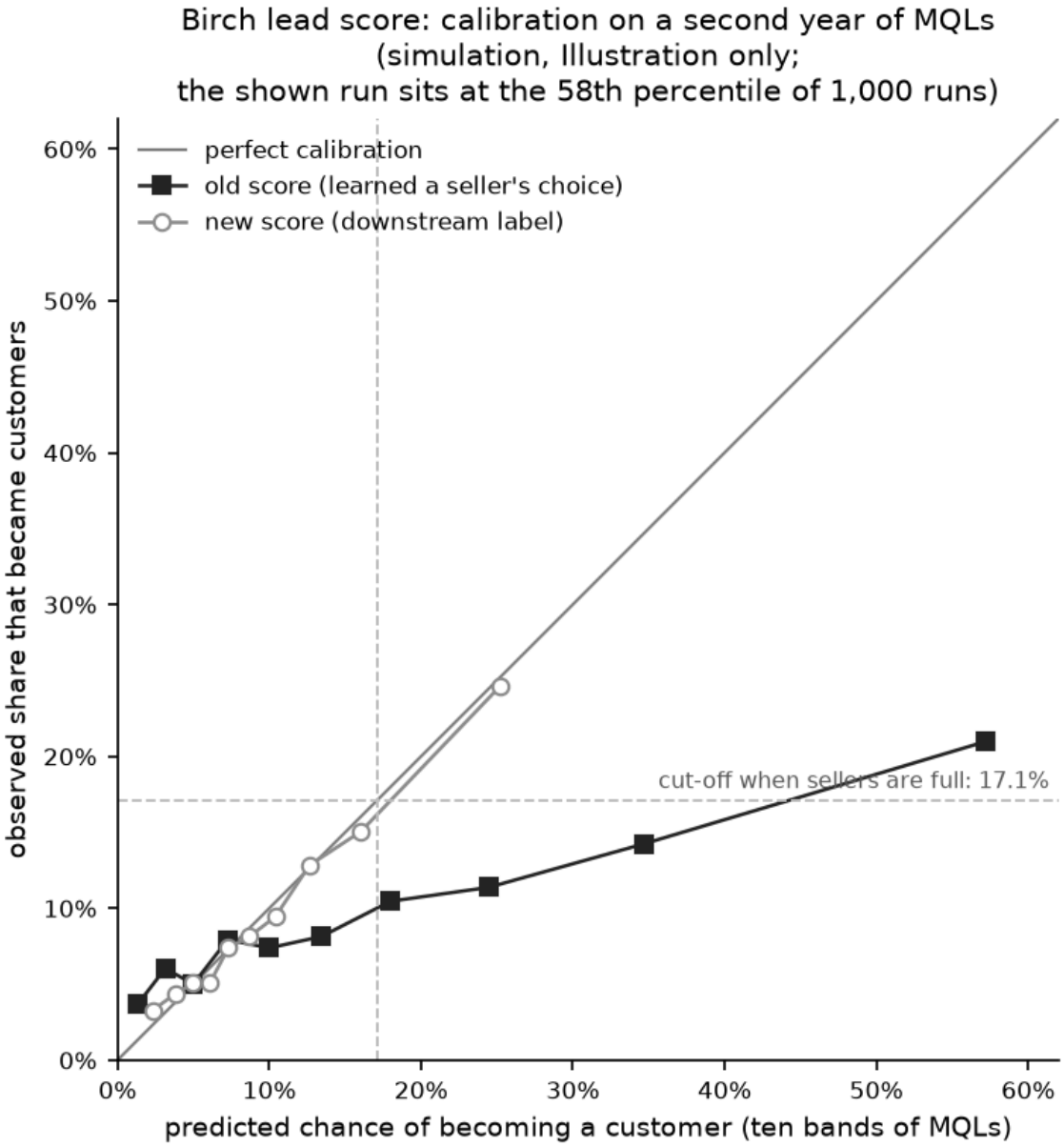
Lead and account scoring

Lead scoring gives each lead a number that stands for its chance to reach a later stage. It earns money only when something is scarce. If sellers have time for every lead, a score has little to decide. If seller time is the limit, the score decides who gets it. A score has three uses: it orders the queue, it decides who is contacted at all, and it routes a lead to the right team. Five rules keep a score honest.

Choose the label for the decision. A customer outcome is useful when predicting purchases. SQL status predicts a process choice made by sellers; use it only when that process is the prediction task. Neither label by itself establishes the additional purchases caused by contact. Outcomes under a previous targeting policy also reflect its selection. Use inputs available at the scoring date, and validate on later cohorts. A young lead is not a non-buyer because its outcome has not matured. Use a mature fixed-horizon label or an appropriate time-to-event method (see *Forecast conversion and timing*).

Keep it small, and check it on later data. Birch won 636 customers in the 12 months to Year 2 October. That supports a score with a handful of inputs, not dozens. Do not size a score by a rule such as ten customers for each input. The need depends on the number of inputs, on the share of leads that convert and on how far you shrink the fit. Train on one period, check on the next, and compare with a simple rules baseline. If the model does not beat the baseline, keep the baseline. Run the new score in shadow first. It scores the leads while the old rule still decides, and you compare the two on the later outcome. Switch only when the new score wins. Then keep a small holdback on the old rule for a quarter, so that you can see if the gain is real.

Check the calibration. A score is calibrated when leads that it rates at 20% become customers about 20% of the time. Sort the leads of a later period into ten bands by predicted chance. In each band, compare the mean prediction with the share that converted. A score that ranks well but is not calibrated misleads any rule that uses its number as a probability, such as the cut-off below.



Calibration of two scores on a second year of MQLs, in ten bands. The old score learned a seller's choice and was not recalibrated. The new score learned the customer label and was recalibrated. Illustration only.
The figure shows a simulation of Birch (Illustration only; seed 1701). Both scores see the true chance with noise. In the top band the old score predicts 57.2%, and 21.0% of those MQLs become customers. The new score predicts 25.2% in its top band, and 24.6% convert. The old score is too sure of its best leads because part of its high score is luck, and nobody removed that part. The shown run is typical. In 1,000 runs with other seeds, the observed share in the old score's top band is 34% to 38% of its prediction (P10 to P90). The shown run is at the 58th percentile of 1,000 runs.

Value the effect of contact in euro. The decision concerns expected incremental gross profit. Compare expected discounted gross profit with contact against no contact. Subtract the additional expenditure caused by contact. A predicted purchase probability can be high when the prospect would buy without a call. That probability alone does not price the conversation. Provost and Fawcett (*Data Science for Business*, chapter 7) connect model evaluation with the payoff structure of the decision; the payoff structure must match this action.

> **Calculation**
>
> contact gain = expected discounted gross profit under contact − expected discounted gross profit under no contact − incremental contact expenditure
>
> With scarce time, choose the feasible set of contacts with the greatest total contact gain.
>
> Evidence label: Arithmetic. The expected outcomes under the two actions require causal evidence or stated assumptions.

When contacts require equal time and have no interactions, rank their incremental gains and fill the available slots while the gain warrants it. If durations differ, account for each contact's time. Gain per unit of time is useful for divisible work; indivisible appointments can require a constrained portfolio calculation. Do not charge a salary cost twice if it is already in the value or remains fixed under both actions. Opportunity cost describes the alternative use of constrained time; include it once in the comparison.

The simpler purchase-probability threshold is a special case. It assumes no purchase without contact, a common customer value and a correctly measured incremental cost. Under those assumptions it compares purchase probability times value with contact cost. A ranking by purchase probability also assumes equal contact time and value. The illustrative quantities below are conditional forecast and capacity calculations under that simplification. They are not estimates of the causal gain from contacting a lead.

Birch shows both prices. A contact takes 346 euro of seller time (135,000 euro a year over 390 SQL slots), and a new customer is worth 31,538 euro. The cut-off on seller time is 1.1%. Almost every MQL clears it, so cost never binds. But sellers are full. Their load is 112% and 106 SQLs wait. A contact then displaces another SQL, and a contacted SQL closes at 17.1% and is worth 5,399 euro. The cut-off becomes 17.1%. Capacity binds, not cost. In practice you rank by score and fill the slots, and the cut-off is the chance of the last lead that fits. The 17.1% is the close rate of an average contacted SQL. The SQL that a contact displaces is the last one that fits, and its chance is lower. So 17.1% is a cautious start that sets the line a little high.

Keep exploration and a comparison. A score controls which leads sellers contact. If the outcome system sees only contacted leads, the next model learns from the previous model's selected sample. First check whether billing and account identity can also reveal purchases without contact. Those outcomes matter for the decision.

An **explore arm** contacts a random sample below the current threshold. It supplies information about outcomes under contact in that part of the population. Its **explore-arm rate** is the share that reaches the chosen outcome. When combining it with more frequently selected leads, account for their sampling probabilities and for the population to which the estimate applies. A contacted sample below the threshold is not automatically representative of all eligible leads.

To estimate the effect of contact, also retain a comparable no-contact group, ideally through random assignment among eligible accounts. Random exploration alone does not supply that comparison. Preserve assignment, attempted contact and actual contact separately. They answer different questions about the policy and its delivery.

Exploration consumes time that could serve other accounts. Price that trade-off using the binding resource. A brief first touch can reduce expenditure, but qualified leads may still need scarce seller slots. Choose an arm that can answer a commercially useful question at an acceptable cost. At modest volume it may reveal a large mistake while leaving small treatment differences unresolved. Record that limit before changing the targeting policy.

At low volume the unit of a score is the account, not the lead, and a model rarely pays. A written tier list of a few visible traits can be checked by a person. A score that nobody can check is a risk, not a lever.

Implementation note: the companion code sizes the arm and the tests with `funnelkit.sizing` (`n_per_arm`, `mde_relative_lift`). The score itself is a logistic regression with a recalibration step. Any statistics library fits it in a few lines.

Product-qualified leads and accounts

A product-qualified account is a self-serve account whose use of the product signals buying intent (see *Choose markets and accounts*). Define it from product events by a rule or a model. Validate it in two ways. First, compare with matched accounts that did not qualify. This shows that the rule finds accounts that behave differently. It does not show that a call helps them. Second, randomise the sales touch among the eligible accounts. This measures the call itself and not the selection, because sellers who pick accounts choose the promising ones.

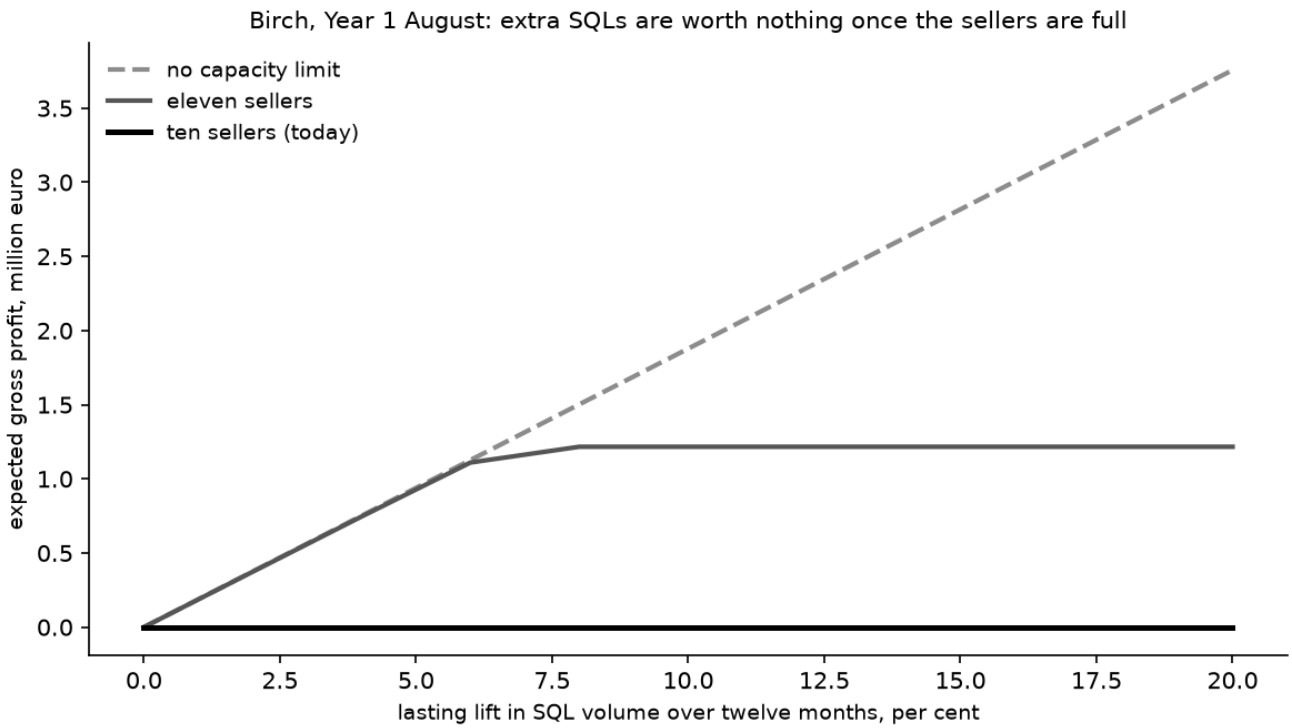
Roll people up to the account. The buying group of *Choose markets and accounts* can hold many people in several roles. Count engaged contacts by role, and read the coverage of the group. The roll-up needs the account key of *Decide what to measure*. The unit changes the count: with the account as the unit, a company with few accounts has few units.

A touch is a cost, and it competes for seller time. Value it as the gain per treated account, times the number treated, less its cost. Then check whether the sellers have room for the extra work that a good result creates. The Cedar example below shows both checks.

Capacity and its shadow price

A *capacity constraint* is a limit on how many items a team can process in a window. When the limit binds, extra volume at the top adds no customers. The sellers of Birch each work about 30 SQLs a window. Ten sellers work 300. In the company model, sellers contact new and waiting SQLs up to their capacity, and the rest wait. Waiting SQLs convert at the usual rate when someone finally contacts them. In practice they age and some are lost, so the rule understates the cost of a queue.

At the trend level of Year 1 August, Birch has 295.0 SQLs a window against a capacity of 300, a load of 98.3% and a spare 1.7%. Over the next twelve months, demand exceeds capacity in 10 months. The peak is 108.0% of capacity and the trough is 85.8%. Over the whole year, demand is 2.5% above capacity, and the queue stands at 56 SQLs on the day of the decision.



Expected gross profit of a lasting lift in SQL volume for ten sellers, eleven sellers and no limit. With ten sellers a lift adds nothing; one more seller restores it up to a lift of about 6.6%.
With ten sellers, the value is zero at every size, because every free slot already goes to the queue. With eleven sellers, it follows the line of no limit until the new slots are full, and then it stops.

The *shadow price* of a scarce resource is the value of one more unit of it. The shadow price of an SQL slot is zero while sellers have slack. It jumps to the value of an SQL when the sellers are full, because the slot then serves an SQL that would have waited. On the value ladder, an SQL at Birch is worth 5,046 euro.

> **Calculation**
>
> shadow price of a slot = value of the item that the slot would serve, if the constraint binds; otherwise 0

One more seller adds 30 slots a window, which is 390 a year. Over the next twelve months, an eleventh seller at full capacity would work 153.1 SQLs that would otherwise wait. At the value of an SQL, before any discount for delay, that is worth 772,474 euro, against a yearly cost of 135,000 euro. The seller pays for itself if it works more than 26.8 SQLs a year that would otherwise wait (Arithmetic). A new seller ramps for 3 months at half capacity. The model does not age the queue, and the cost of the seller is only the pay. So capacity is a lever like the others, with an expected profit that you can compute and rank.

The same logic applies to customer success, support and implementation hours. Relieve first the limit that binds on the most valuable items.

Birch, Year 2 October to November: a score that learned a seller's choice

Birch, Year 2 October: about 54 new customers a window, 0.8 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 62 days, and the close rate reads a 20% change in about 6 months.

The lead score of Birch was trained on SQL status, which is a seller's choice. Replacing that label aligns the forecast with customer outcomes; it does not by itself identify what contact causes. Sellers are full. Their load is 112%, 106 SQLs wait, and a waiting SQL waits 9.4 days (the backlog over 315 slots in 28 days). Each window brings 1,426 MQLs, of which 335 (23.5%) become SQLs. The other 1,091 are not passed. The analytics lead rebuilds the score on the customer label, with the 636 customers of the last 12 months. It stays small, and a later period checks it.

The sales owner uses the simplified forecast cut-off, assuming no purchase without contact and common value and contact time. On seller time it is 1.1%, and almost every MQL clears it. With full sellers it is 17.1%, so the slots set the line. The rebuilt score is in use from Year 2 December. It passes 21.15% of MQLs, about 302 SQLs a window and 33.5 fewer. Sellers accept 90% of them instead of 83%, and the win rate is 33% instead of 32.5%. A contacted SQL now closes at 18.8% instead of 17.1%, a gain of 10.1%. The owners read the accepted rate, which reads a rise of this size within a quarter. They cannot read the win rate, because a move of half a point lies far below what it can show.

The explore arm needs a price. Its design is 10% of the unpassed MQLs: 109 a window, 1,419 a year. These counts use the pass rate of Year 2 October, before the rebuilt score; with the rebuilt score the arm is 1,462 a year. If every contact took a whole SQL slot, the arm would take 34.6% of the slots of the 10.5 effective sellers (35.7% with the rebuilt score). At the price of the full cut-off, it would displace up to 7.7 million euro of expected gross profit a year. On seller time alone it would cost 491,029 euro. Birch does not pay that price, because the 34.6% holds only if every contact takes a seller slot.

The first touch of the arm is a call by sales development. A lead joins the SQL queue only when it qualifies. A first call costs about 15.10 euro. That is the 280,000 euro a year of the four sales development representatives over the 18,543 MQLs that they work in a year. It is Illustration only: it assumes that a call takes as long as an average MQL today. The arm then costs about 21,420 euro a year and adds 7.65% to the team's calls. On the time of sales development the cut-off is 0.048%, so every MQL clears it. Seller slots go only to the leads that qualify: each tenth of the arm that qualifies takes 3.46% of the slots. If the arm qualified at the MQL-to-SQL rate of 23.5%, 333 leads a year would take 8.1% of the slots. That is an upper reference, because the unpassed MQLs rank lower than the passed ones.

The arm still reads only coarsely. If it converts at the seller-time cut-off of 1.1%, it expects 15.6 customers a year, with an 80% range of 11 to 21. The sales owner reviews its size each quarter.

Speed to lead has a clock in minutes and a clock in days. The curve of Birch (Illustration only) loses 11.8 points of contact rate between 5 and 120 minutes. But a waiting SQL waits 9.4 days. The days come first.

The days then decide the next step. From Year 2 December the half media shift of *Choose acquisition channels* moves spend towards search and social ads. In the simulation's response curves it adds about 6.0 SQLs a window. That is worth 1.1 customers a window if sellers contact them (Illustration only). Birch's own estimate of extra leads in *Choose acquisition channels* was larger, and *Learn from outcomes* grades it.

In Year 2 December the sales owner rewrites the level of the hiring rule from 65 to 100 uncontacted SQLs. The new level fits eleven sellers and the rebuilt score. A level of 100 is a wait of 8.5 days at the capacity of eleven sellers. The old level of 65 was a wait of 6.1 days at ten. The backlog ends December at 89.5 and January at 95.4. At the end of Year 3 February it is 103.6. The written rule trips for the second time.

The rule works as written. The sales owner approves a twelfth seller within the week, logs the decision and holds no meeting. The wait at that backlog is 8.8 days, and 103.6 SQLs hold about 559,000 euro of expected gross profit. One more seller costs 135,000 euro a year. The hire lag is two months, so the seller starts in Year 3 April, ramps at half capacity until June and works at full capacity from July. The rule pauses until then. The backlog is 114.2 in March, 93.2 in April, 74.5 in May and 58.1 in June.

Commit with uncertainty wrote the rule. *Fund the next best action* saw it trip in Year 2 July. The second time nobody argued, because the case was written in advance. At medium volume a year of customers supports only a small score, and capacity, not cost, sets its cut-off.

Cedar, Year 2 October to November: a call to product-qualified accounts

Cedar, Year 2 October: about 1,900 new paying accounts a window, 1.3 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 23 days; the assisted motion, with 74 opportunities a window, needs about 7 months to read a 30% change.

Should sales call product-qualified accounts? The assisted-sales manager asks the question, and the growth lead owns the rule that defines eligibility. About 472.6 eligible accounts arrive each window. Without a touch, 15% of them become an opportunity. A look at accounts that sellers called would not answer the question, because sellers call the promising ones. So the analytics lead randomises the touch among the eligible accounts: half get the call, half do not.

The test reads little in two windows. Each arm holds 472.6 accounts, and the smallest rise that it reads is 47.2%. That means 15% rising to 22.1%, which is 7.1 points. The assisted motion is a low-volume business inside a high-volume company. Its win rate needs 211 days to read a 30% change.

Price the touch to see if the small effects matter. An extra opportunity is worth the win rate of 25% times 57,645 euro, which is 14,411 euro. Suppose that a touch cost as much as one whole opportunity slot, 641 euro (a seller costs 125,000 euro and has 195 slots a year). That is the cautious end, because a short call costs less. The touch then breaks even at a rise of 4.4 points, or 29.7%. Two windows cannot read that. It needs 1,130 accounts in each arm, which is 4.8 windows, about 4.4 months.

Capacity is the second check. The six assisted sellers have 90 opportunity slots a window, and their load is 82.7%. That leaves 15.6 free slots. At the smallest rise that the test reads, the touch would create 16.7 extra opportunities a window. At the break-even rise it would create 10.5. A big success would exceed the room that the sellers have. So the assisted-sales manager caps the touch at the free slots.

The design, then, is a staged test. It runs for two windows to catch a large effect or harm. It continues to about five windows to read a rise of break-even size. It reads opportunity creation, because wins take seven months to read. It also reads the revenue of the accounts later, because a touched account may have paid less than it would have paid alone. The first read is due at the end of November, and the read of break-even size about 4.4 months after the start. The lesson is that a high-volume company can still face a low-volume decision. The touch test reads only large effects, so the assisted-sales manager decides on the prior and uses the test as a guard and a signpost.

The decision to carry forward

Allocate attention by the additional value the contact can cause under the binding time constraint. Use purchase scores only under stated simplifying assumptions. Preserve exploration for coverage and a suitable comparison for causal learning.

08 · Remove buying friction

Where does a buying journey lose value that an action could recover? A rejection can reveal poor fit, missing information or a full queue. A slow response can reflect an operating constraint rather than weak demand. The same falling conversion rate supports different actions under those different mechanisms.

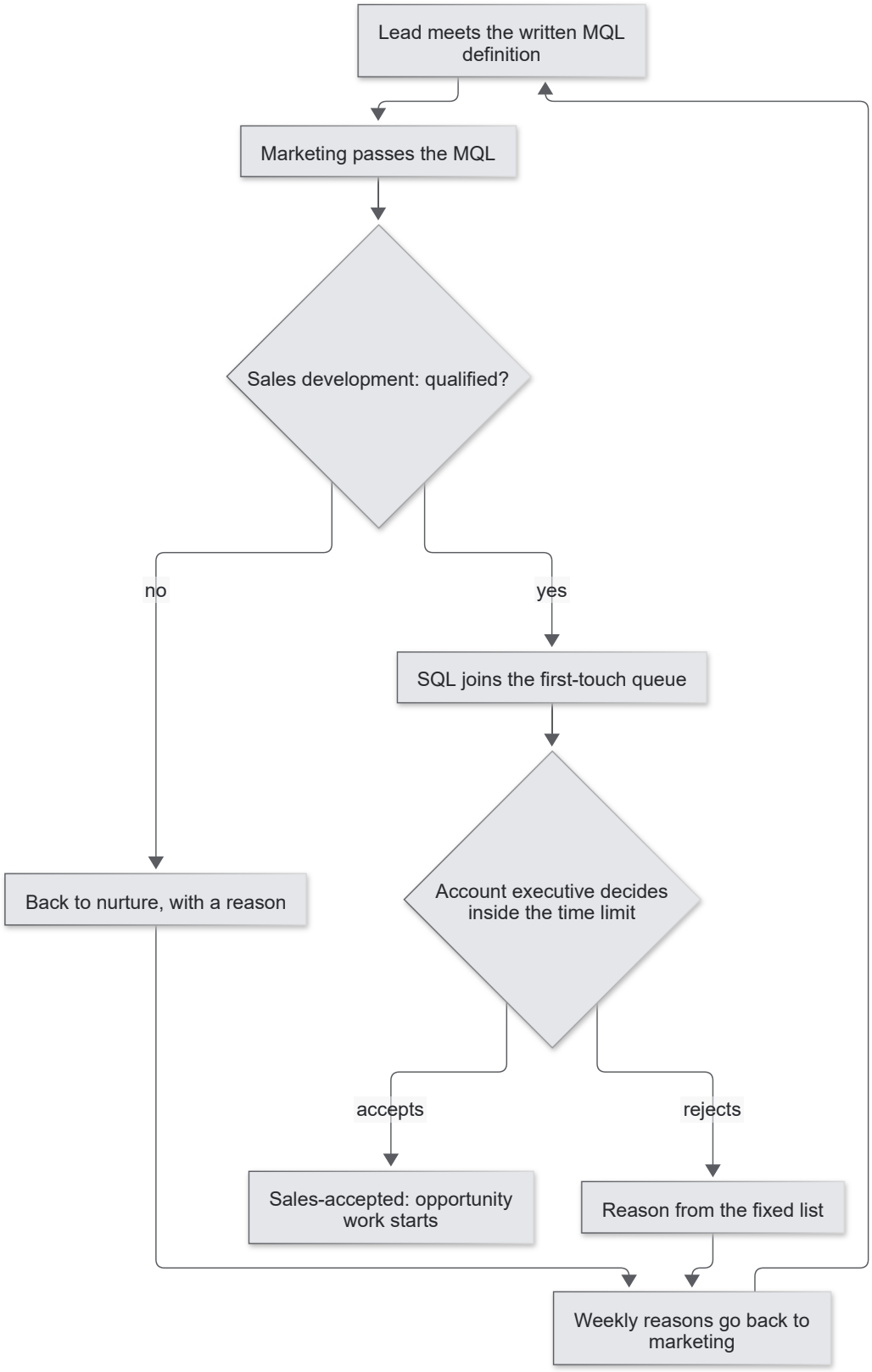
Start with a dated journey and a clear hand-off. Then inspect the rejection reasons, waiting time and downstream outcomes together. Improving process compliance is useful, but it does not by itself prove an increase in customer profit. A better recorded discovery field may improve forecasting before it improves wins.

The sections below connect qualification, nurture, response speed and discovery to decisions. Use *Allocate sales attention* for the value of assigning a contact and *Forecast conversion and timing* for the effect of unfinished deals on what you can conclude.

Qualification and the hand-off contract

Marketing finds leads. Sales development qualifies them. Account executives work them. Each step passes a record to the next owner, and each pass can fail in two ways. The record can be wrong for the receiver, or the receiver can be too busy to look at it. A **hand-off contract** is the written agreement that makes both failures visible. It has four parts.

- **The MQL definition.** What must be true of a lead before marketing passes it. *Choose markets and accounts* showed that a stage has an entry criterion and a timestamp. The contract writes the criterion in words that both owners accept.
- **The time limit.** How long the receiver has to accept or reject a record. Set the limit from the wait that your queue allows, not from a wish.
- **The rejection reasons.** A fixed short list, such as not in the profile, wrong timing, no pain found and situation unclear. A seller who rejects a record picks one. Nobody types a free text for the count.
- **The feedback.** Each week the reasons go back to the marketing owner, who may change the definition on a stated date. The change goes in the decision log.



The hand-off contract. Each pass has a rule, a time limit and a reason in words. The reasons return to the owner of the definition.

Three metrics read the contract. The SAL rate of *Choose markets and accounts* is the share of passed records that sellers accept. The **MQL rejection rate** is the share of passed MQLs that sales rejects. The third is the time that the receiver takes to answer. When sales development sits between marketing and the sellers, read

rejection at both passes. At Birch in Year 2 October, sales development passes 23.5% of MQLs as SQLs. The account executives then accept 83% of the SQLs and reject 17.0%.

Read the reasons by share, window after window. A rise in "no pain found" points at the fit of what marketing passes. A rise in "situation unclear" points at missing data about the lead. Enrichment, which adds company data from an outside source to the record, can fix it. A rise in rejections with no reason, or with no answer, points at a seller who has no time. *Choose markets and accounts* warned about this case. When sellers are full, a falling acceptance rate may measure load and not lead quality. Check the backlog before you blame the leads.

Write the MQL definition as a rule on stored fields, not as a feeling. A useful rule has three kinds of fact. Fit is a fact such as company size or industry. Intent is a fact such as a demo request or repeated visits to the pricing page. Recency says how fresh the intent is. Test the rule by reading the share of MQLs that reach the SQL stage, by lead source and by month. A source whose MQLs rarely reach the SQL stage is a source to fix or to stop.

Do not set the MQL cut-off "by agreement" between the two teams. An agreement has no price on it. *Allocate sales attention* sets out the economic rule and its assumptions. The owners write the contract, and they change it on a stated date. The analytics lead reads the rates by role and never by person. Evidence label: Standard practice. The contract follows the hand-off logic of Winning by Design (Winning by Design, 2023, *Revenue Architecture*). A before-and-after read of the accepted rate and the reasons monitors the process. It does not isolate the effect of the contract from other changes. At medium volume a large shift reads within a window and a small one within a quarter. At low volume each rejection is a case to read.

Nurture

Nurture is a planned series of messages that builds intent in leads that are not ready to talk to sales. It earns when a lead that would not have reached a seller does so later. It costs when each message loses a few readers, as the repeated-send arithmetic in *Model recurring growth* shows. A sound design has three parts.

- **One default cadence.** The best interval balances the effect of one more contact, the loss of readers and the time value of waiting. At the volume of most companies you cannot estimate that interval, so write one default and keep it.
- **A suppression arm.** A random share of the list gets no nurture for a set period. A tenth of the list for eight weeks is a start value (Plausible but untested). The gap between the nurtured and the suppressed leads is the **nurture incrementality**: what the nurture adds. A before-and-after look cannot show it, because leads mature on their own. Size the arm with the readability function, on the rate that you care about.
- **A send-health gate.** Deliverability is the share of emails that reach the inbox. Send health is the complaint rate and the bounce rate. The gate stops a campaign when either rate passes its limit. Start with the limits of your email provider and tighten them with your own history (Plausible but untested).

Value a contact as the extra chance that it moves a lead to the next stage, times the gross profit that the next stage is worth. A contact that costs more than that value should not go out. Unsubscribes are handled by rule, never by a model, and consent comes first. Counsel reviews the lawful basis of every list. This is not legal advice.

Speed to lead

Speed to lead is the time from the creation of a lead to the first human touch. A fast touch should raise the **contact rate**, the share of leads that reach a person. Many sources quote a fixed rule, such as five minutes. Treat any fixed rule as a start value. The shape of the curve differs between companies, sources and hours of the day.

Find the curve in your own data. Take the lead timestamp and the first-touch timestamp. Group leads by the minutes to the first touch. In each group compute the contact rate, with its interval. Leads that arrive in the office day differ from leads that arrive at night. So compare within the same source and the same hour band. Pool several weeks. If you can, route some leads to a second queue for a short test, so that the difference does not depend on the day. Birch's curve is Illustration only. The contact rate is 54.2% at 5 minutes, 47.1% at 60 and 42.4% at 120. It levels off near 35%.

There are two queues and two clocks. The first runs in minutes, between the lead and sales development. The second runs in days, between the SQL and the seller, as in *Set targets and staff capacity*. By Little's law the mean wait is the backlog divided by the number of SQLs that sellers contact a day. Measure both, and start with the longer one. A curve in minutes does not help while SQLs wait for days.

The sales process as data

Discovery is the part of the sale where the seller learns whether the buyer has a reason to buy. **SPICED** records five facts: Situation, Pain, Impact, Critical Event and Decision. **MEDDIC** is a qualification method with six prompts: Metrics, Economic buyer, Decision criteria, Decision process, Identify pain and Champion. Both give each fact a place in the record. A **mutual action plan** is a shared, dated plan of the steps that buyer and seller take to a signed contract and a first result. Winning by Design describes SPICED and the critical event (Winning by Design, 2023, *Revenue Architecture*).

Two facts from these methods become process metrics. A **critical event** is a dated event in the customer's business with consequences if the date is missed. A renewal date is motivating but not critical. Count a critical event only when the customer has confirmed it, in writing or in the mutual plan. **Critical-event coverage** reports three states, not two: a dated event, a confirmed "none", and unknown. A **champion** is a person at the customer who has influence and does costly acts, such as an introduction to the decider. The **champion presence rate** is the share of opportunities with such a person by the proposal date.

A mutual action plan works backwards from the customer's date. Start with the day the customer needs the result. Subtract the time for onboarding, for contract review and for the proposal. The latest signing date follows, and it is often earlier than the seller expects. A date built this way has a reason behind it, so it also tests the close date that the seller gives.

Read the two metrics with care. The evidence that they predict the close is weak. A search for independent studies found none. The published figures come from software vendors and show that deals with champions win more often. That is an association, and healthy deals attract both.

Coverage measures what the seller recorded, not what is true. A seller can type a date into a field. People fill in a field that earns them a reward. So the book labels both metrics Plausible but untested as predictors, and uses them as signposts of process quality for a manager. They are never a target for one person, and they are read by role.

You can test them. The strong test compares win rates between deals with and without a fact. To tell a 30% win rate from a 40% win rate, each group needs 356 closed deals. A cheaper test compares the error of the close date: a date taken from the customer's timeline against the seller's own date. It usually needs fewer deals, because each deal gives a number of days and not only a yes or a no. It measures forecast quality, not the win rate. Compare within a deal-size band to reduce selection. Until a test passes, decide on the prior and read the signpost.

Pipeline coverage is a related forecast check. **Pipeline coverage** is the open pipeline value divided by the new ARR still needed. At a win rate w, the coverage must be at least 1 ÷ w to reach the target on average. At Alder's win rate of 22.0% that is 4.5. At Birch's 32.5% it is 3.1. It is a signpost for the forecast of *Set targets and staff capacity*, not a lever.

The levers on one page

Table 1 gives the proof design of each lever and the volume at which it reads.

Table 1. Each lever with its metric, its proof design and where it reads.

Lever	Metric	Proof design	Where it reads
Hand-off contract	Sales-accepted rate, rejection reasons by share, time to answer	Before and after on the rates	Medium: large shifts within a window, small ones within a quarter. Low: read each case
Lead score	Calibration by band, cut-off in euro, explore-arm rate	Time-split check, explore arm	Medium: calibration reads, the gain in win rate does not. Low: a tier list
Nurture	Early-stage rates of nurtured against suppressed leads, send health	Suppression arm, send-health gate	Medium and high: early rates only
Speed to lead	Minutes to first touch, contact rate by minute band	Own-data curve, same-source comparison	Medium: pool several weeks
Touch on product-qualified accounts	Opportunity rate per eligible account	Randomise the touch	Self-serve motion: reads. Assisted motion: large effects only
Discovery quality	Critical-event coverage in three states, champion presence	Signpost, backtest on closed deals	Low: signpost only. Medium: a backtest after about a year
Price and packaging	Price realisation, discount depth, break-even close rate	Frontier, willingness-to-pay research, staged change	Low and medium: decide on the frontier
Seller capacity	Load, backlog, wait, ramped productivity	Written trigger on the backlog	Any volume with a first-touch queue

What changes with volume

- **Low volume.** No rate below the lead reads a 20% change within a year. Judge levers by signposts, cases and priors. Score accounts with a tier list. Read process quality, not win rate. Decide price on the frontier and on buyer research.
- **Medium volume.** The hand-off rates and the calibration of a score read within a quarter. The close rate reads a 20% change only in about half a year. The small gains that a better score or a better demo brings

do not read within a year. So such a gain stays a prior. A year of customers supports a small, well-checked score only.

- **High volume.** In a self-serve motion every rate reads a 20% change within a window. So touches, nurture and speed to lead can be tested. The assisted motion in the same company still behaves like a low-volume business. The barometer belongs to a decision, not to a company.
- **At every volume.** The cut-off follows the price of seller time. When sellers are full, capacity sets the cut-off, and a rule on the backlog sets the hire.

Related company decisions appear in *Allocate sales attention*, *Set prices and packages*. Their original dates remain explicit.

The decision to carry forward

Change the mechanism that explains the loss. A definition problem calls for a contract, a queue problem for a capacity response, and a persuasion problem for an intervention whose effect can be compared. Grade process improvements and commercial gains separately.

09 · Set prices and packages

Which price and package will create the most customer profit? Price changes what a customer pays, who chooses to buy, how much they use and whether they renew. Packaging also changes the cost of delivering the result. The decision must compare those complete paths.

A price list is an input to the model. The realised price, conversion, service cost and retention are outcomes. Their relationship is partly behavioural and must be estimated. A calculation can establish the break-even loss in demand without establishing how much demand will actually fall.

This chapter separates that economic frontier from the evidence about customer response. It joins new-business packaging and renewal pricing because both use the same profit basis. The customer populations and commitments differ, so the evidence and rollout decisions must remain distinct.

A price is a policy for a defined population

Define what the customer receives, the unit they pay for and when the price applies. A seat price, a site price and a platform fee expose different customer behaviour. The package can change adoption, the purchase decision and the cost of serving the account. Hold the offered result clear while evaluating its price.

Separate new customers from existing contracts. A new-business offer can change who enters the book. A renewal rise applies to customers selected by having stayed until renewal. Their baseline survival, switching costs and history differ. An effect observed in one population is not automatically transportable to the other.

Choose the comparison before inspecting the result. It can be the current offer against a revised package for eligible prospects, or the current renewal policy against a different permitted policy. Record the offer date, eligibility and assignment. Preserve the price actually offered as well as the price signed, because negotiation occurs after the offer.

Conversion and price must be valued together

For an eligible prospect, expected contribution depends on the probability of buying under the offer and the value conditional on buying. Subtract incremental costs incurred whether or not the prospect buys. Compare this with the corresponding contribution under the alternative offer.

> **Calculation**
>
> expected value of an offer = P(buy under the offer) × expected discounted gross profit given purchase − incremental offer costs
>
> offer gain = expected value of the proposed offer − expected value of the alternative offer
>
> Evidence label: Arithmetic. The behavioural probabilities and value estimates require evidence or explicit beliefs.

The conditional value includes service, survival and expansion under that offer. If the value estimate already includes a cost, do not subtract it again. When a package changes implementation effort or usage costs, recalculate those costs instead of assuming a constant margin.

Average revenue among buyers can rise because low-value buyers no longer buy. It can also fall while profit rises because the new package brings in customers who are cheaper to serve. Neither observation by itself reveals the effect on profit across all eligible accounts. Keep the denominator at the population exposed to the offer.

The economic frontier answers how far conversion or renewal could fall before the action loses value. It is a condition for a profitable decision. It does not predict where the response will land. The worked examples below show how to use that frontier when volume cannot resolve a small rate change.

Learn the response without learning the negotiation rule

Discounts are usually chosen in response to the perceived difficulty of a deal. A comparison of heavily discounted deals with full-price deals therefore mixes the discount with customer selection and seller judgement. A fitted demand curve can describe that selected history while misrepresenting the response to a changed price policy.

Where feasible, compare assigned offer policies among eligible accounts. Keep the unit of assignment at the account when multiple people buy together. Record exceptions and noncompliance. The effect of offering the policy is often the operational quantity you can implement. Conditioning the analysis only on customers who accepted discards part of that effect.

If randomisation is unsuitable, buyer research and bounded scenario analysis can inform the decision. Write which aspects they establish. An interview can reveal a rejected value proposition; it does not establish a conversion elasticity for the market. A historical comparison needs assumptions about customer mix, season and simultaneous changes. Label those assumptions instead of treating a modelled response as experimentally identified.

Use later cohorts to check the complete economics. If the new offer brings more customers who leave early, a short-term conversion gain overstates the retained contribution. If the offer changes product usage, evaluate costs at the corresponding usage level. Read equal cohort ages rather than giving the old offer more time to earn.

Connect price realisation with customer value

The signed price divided by the list price is price realisation. It describes commercial execution under the current offers and customer mix. It does not measure willingness to pay directly. Keep its interpretation separate from the conversion effect of a different policy.

A useful review asks where the contribution changes: eligibility, conversion, realised price, service cost, retention or expansion. Inspect these together, because improving one can weaken another. At low volume, a frontier and case evidence may justify a limited change while the uncertainty remains wide. At high volume, a test can estimate a response for the population and period studied. Neither setting removes the need to state the future policy and its constraints.

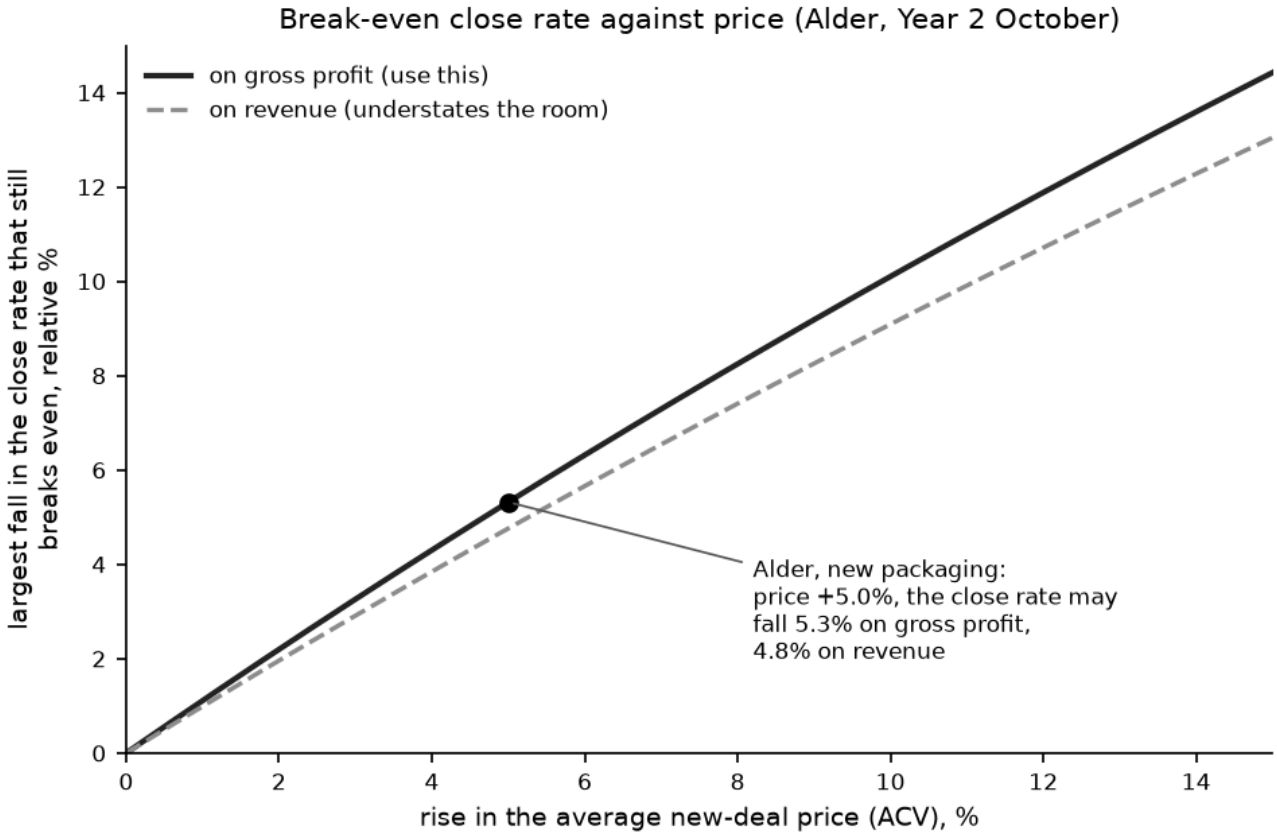
Put a dated later read on the decision record. The first read can concern acceptance or bookings. A subsequent read concerns delivery cost and retained contribution. These are observations of one commercial policy over time, so they belong to the same decision.

Pricing and packaging for new business

The price changes two things at once. It changes the gross profit of each new customer, and it changes the close rate. A higher price is worth it when the first gain outweighs the second loss. The test is the **gross-profit price frontier**: the close rate at which the new price earns as much as the old one.

> **Calculation**
>
> gross profit per customer-year = price × (1 - cost share that scales with price) - cost to serve per customer
>
> break-even close rate, as a share of the close rate before = gross profit per customer-year before ÷ gross profit per customer-year after
>
> Price break-even uses gross profit, not revenue. The revenue version is 1 ÷ (1 + price rise).
>
> Evidence label: Arithmetic. With the same retention under both prices, the five-year discounted value scales by the same ratio, so the one-year ratio holds on the value basis.

The revenue version understates the room. The cost to serve a customer does not rise with the price, so 87.5% of each extra euro of price is profit at Alder. The chart shows the gap for Alder. Do not read win rate by price point as proof. Discounts follow the weakness of a deal, and large prices go to different buyers. A comparison of win rates across price points mixes the buyers with the price.



The largest fall in the close rate that a rise in the new-deal price can absorb, on gross profit and on revenue. Arithmetic, from Alder's packaging facts.

Price is hard to randomise for new business. Buyers talk, quotes are public in some markets, and price differences between buyers may raise legal questions. Counsel reviews this. This is not legal advice. So use an evidence ladder.

1. **Price realisation.** The list price is the published price. The signed, net price is what you keep. Price realisation is the net price divided by the list price. Discount depth is one minus that share.
2. **Willingness-to-pay research. Willingness to pay** is the highest price that a buyer accepts. Interviews or surveys give a range of it. They are not demand curves, and a stated price is not a paid price.
3. **The frontier.** Write the break-even close rate. Compare it with what the volume can read. If it cannot read the break-even fall, decide on the prior.
4. **A staged change with a signpost.** Change new deals only. Keep the old price for existing customers. Write the fall in the close rate that would make you reverse. The change is reversible for new deals, so the bar of 90% applies. The section *A price rise* below gives the corresponding logic for renewals.

Pricing and packaging means the price levels and what each plan contains. The packaging also sets the limits between plans, such as the number of users or sites that each plan allows. The same frontier applies. Count the cost to serve of each package, because a package that adds work without price lowers the gross profit per customer.

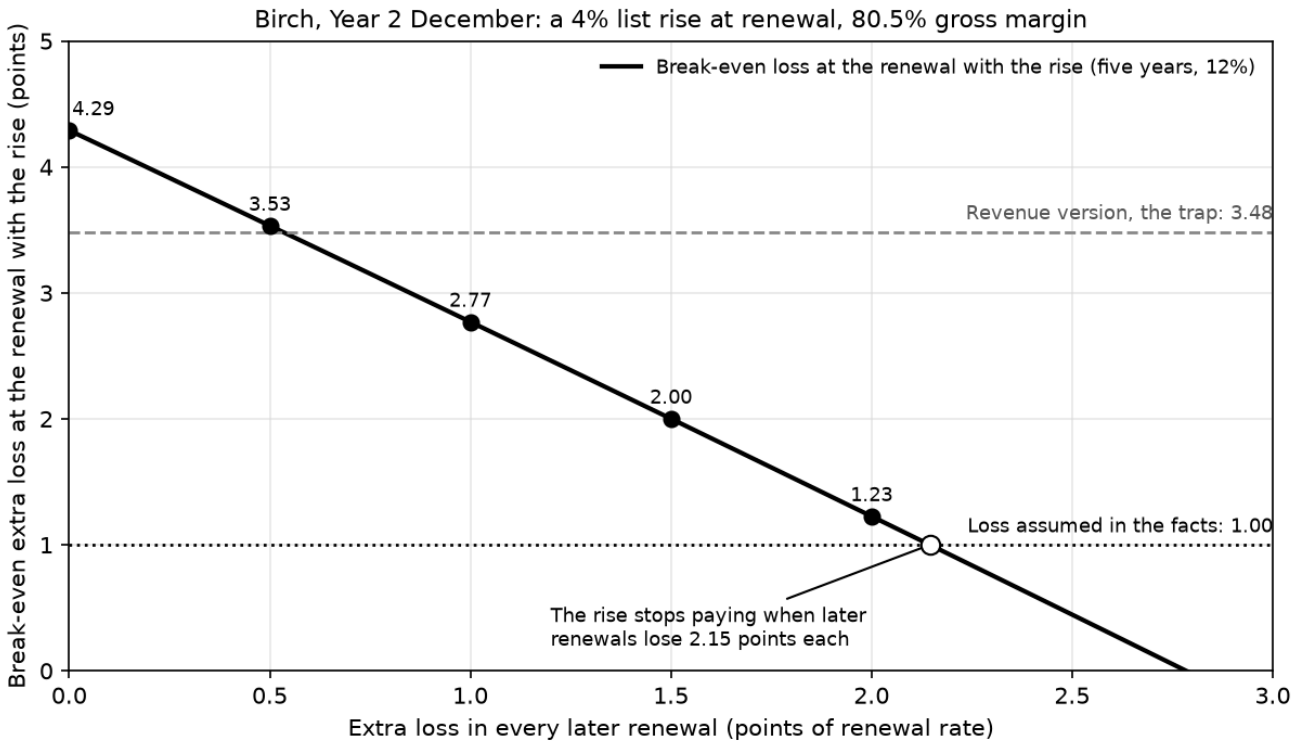
A price rise

A price rise is expansion that costs almost nothing to win. It also puts the whole renewal at risk. Take a rise of p, as a fraction of the old price. A customer that stays pays p more, and costs the same to serve. So a kept customer earns the gross margin m plus p, for each euro of the old price. A customer that leaves takes away the whole margin m. The rise breaks even when the gain on the customers who stay equals the loss of the customers who leave.

> **Calculation**
>
> break-even extra loss of renewing customers = $p \div (m + p) = p \div (1 + p - c)$
>
> Here m is the gross margin, $c = 1 - m$ is the cost share, and the loss is a share of the customers that would have renewed. In points of renewal rate: multiply by the renewal rate.
>
> Evidence label: Arithmetic.

Use gross profit, not revenue. The revenue version $p \div (1 + p)$ forgets that a lost customer takes the margin and not the price. It sets the limit too low, so it blocks rises that pay. At Birch the rise is 4% and the gross margin is 80.5%. The break-even loss is 4.74% of the renewing customers. At the average renewal rate of 90.6% that is 4.29 points of renewal rate. The revenue version gives 3.48 points.

A second question is the horizon. A customer lost today also loses all its later years. But the customers who stay pay the higher price in those years too. Both effects scale with survival. So the break-even is the same at a horizon of one year and of five years, as long as later renewal rates do not change. The break-even falls only if the rise also raises the later churn.



Birch, Year 2 December. The break-even loss at the renewal with the rise, by the extra loss in every later renewal. The dotted line is the assumed loss of 1.0 point at the renewal (Illustration only). The grey dashed line is the revenue version.
The same arithmetic gives a small table for other rises.

Table 2. Birch: break-even loss of renewing customers by size of rise, gross margin 80.5% (Arithmetic).

Price rise	2%	4%	6%	8%
Break-even loss	2.43%	4.74%	6.94%	9.04%

Price cannot be randomised across new buyers at low volume, and the win rate by price point is confounded, because discounts go to weak deals. At renewal you can do better. Hold a random share of renewing accounts out of the rise and read their renewal. This is a staged rollout with a holdback (see *Choose whether to test or act*). It is the cleanest evidence that a price move can have. A rise is hard to undo for contracts that are signed, so write the stop rule before the first quote.

Implementation note: the companion code holds the discounted value of a customer (value module) and the value of information (voi module). The break-even of a rise is a few lines on top of them.

Alder, Year 2 October: discovery quality and a new package

Alder, Year 2 October: about 4.2 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 2.8 years to read even a 30% change, and the lead count reads a 20% change in 23 days.

Two changes start this month. The account executives start to record discovery quality on every opportunity, and a new package applies to new deals only. The head of sales owns both.

The first change records the critical event in three states and the presence of a champion. At the start, 30% of opportunities have a dated event, 15% have a confirmed "none" and 55% are unknown. A champion is present in 40% (Illustration only). With 16.3 opportunities a window, that means about 4.9 with a dated event and 6.5 with a champion. Alder cannot test whether these facts help. To tell Alder's win rate of 22% from one 10 points higher, each group needs 309 closed deals. Alder opens 211.6 opportunities a year, so both groups would need 2.9 years of them. So the head of sales writes a signpost: dated events at 60% of open opportunities by Year 3 January. It measures process quality, not proof of the win. The analyst computes it each month, for the team and never for one seller. The same two facts return after the close. A champion who leaves puts the account at risk, as *Retain customers profitably* shows for Alder's largest account.

The second change is the package. The list price for a new deal rises from 66,000 to 72,000 euro (9.1%). The old list was a platform fee of 24,000 euro plus two sites at 21,000 euro. The new list is 30,000 euro plus 2.4 sites at 17,500 euro. The average new-deal price, the ACV, rises from 60,000 to 63,000 euro (5.0%). The list rises more than the signed price, so price realisation falls from 90.9% to 87.5%, and discount depth rises from 9.1% to 12.5%.

The frontier decides the question that no test can. Cost of revenue that scales with price is 12.5% of ARR. The cost to serve a customer is 5,781 euro a year, mostly people, and it does not rise with the price. Gross profit per customer-year rises from 46,719 to 49,344 euro (5.6%). The close rate can therefore fall to 94.7% of its value before and still break even: a fall of 5.3% (22.0% to 20.8%). On revenue the room looks smaller, 4.8%. Alder wins about 46.6 account-based deals a year, so the break-even loses 2.5 of them. A test of that fall would need about 19,300 opportunities in each arm. Alder opens 211.6 a year.

So the head of sales decides on the frontier and on buyer research. The changes apply to new deals only, so they are reversible, and the head of finance checks the cost lines. The head of sales does not compare win rates by price point. Discounts follow the weakness of a deal, so that comparison would only show the buyers. At low volume, process signposts lead, and the price is decided by the frontier and by what buyers say.

Birch, Year 2 December: a price rise at renewal

Birch, Year 2 December: about 62 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 56 days, and the close rate reads a 20% change in about 5 months.

The chief financial officer asks for a list price rise of 4% for renewals from Year 3 January. The head of customer success owns the renewal, so this role writes the card and decides. A first draft sets the limit with the revenue formula: the rise pays while renewals fall by less than 3.48 points. That limit is too tight. On gross profit the break-even is 4.74% of the renewing customers. Birch has 2,665 renewals a year at an average renewal rate of 90.6%, so the limit is 4.29 points of renewal rate (Arithmetic). Birch has never measured how its buyers react to a rise. So the example assumes that the rise costs the customers who get it 1.0 point of renewal (Illustration only). The break-even is 4.3 times that loss. Above a loss of 4.29 points the rise stops paying. That switching value matters more than the assumed loss.

The next check is the later years. If the rise also pushes later renewals down, the break-even falls. With each later renewal 1.0 point lower, the break-even at the renewal with the rise is 2.77 points. With 2.0 points it is 1.23. With the assumed 1.0 point at the renewal itself, the rise stops paying when each later renewal also loses 2.15 points, as the figure shows. On the value basis, 85% of renewals get the rise. Take the assumed loss of

1.0 point and later renewals unchanged. Then one year of renewals with the rise adds about 2.34 million euro of expected gross profit over five years (Illustration only). If each later renewal also lost 1.0 point, the gain would be about 1.24 million euro.

What can the held-out share tell? Birch holds out 15% of its renewing accounts at random, which is 400 renewals a year. A comparison of 400 against 2,265 reads a fall of 5.0 points in a year, and 11.2 points in a quarter. Both are larger than the break-even of 4.29 points. So the holdback can show a gross failure. It cannot tell the assumed 1.0 point from a loss just below the break-even of 4.29 points, so it cannot confirm that the rise pays. The break-even is computable, and the churn response is barely readable. The price of the holdback is a year of delay for those accounts. If they get the rise one year later, the delay costs about 120,537 euro of expected gross profit (Illustration only, given the assumed loss).

The head of customer success decides on the gross-profit break-even and the belief, not on a test result. A rise is hard to undo for signed contracts, so the 95% threshold of the decision card applies. The owner must judge a loss above 4.29 points less likely than 1 in 20, and write down why. No data supports that judgement yet. It rests on what customer success knows of the buyers. The rise goes ahead for 85% of renewals.

The log entry holds four things. The belief is a loss of 1.0 point at the renewal. The break-even is 4.29 points, and 2.15 points of extra later loss ends the case. The stop rule is a fall of 5.0 points or more between treated and held-out renewals once a year of renewals is closed. The early signposts compare the share of renewal quotes that buyers reopen, in both groups, from the first month. The link of that signpost to the renewal is Plausible but untested. *Learn from outcomes* grades the belief when the year closes. The lesson at medium volume is that the break-even can be computed today. The response cannot be read for a year, so the break-even, the written belief and the holdback make the decision.

Related company decisions appear in *Remove buying friction*, *Allocate sales attention*. Their original dates remain explicit.

The decision to carry forward

Separate a calculated break-even from an estimated price response. Carry conversion, realised price, service cost and later retention through the same comparison. State the customer population to which the new price applies before committing.

10 · Forecast conversion and timing

How many opportunities will become customers, and when will their revenue arrive? The answer changes hiring, cash planning and the value of an acquisition programme. A rate based only on completed deals can make the forecast optimistic by leaving out slow outcomes that remain open.

Use the account's observed journey at the forecast date. Do not add later stages or outcomes to information that was available then. Compare cohorts at equal age, and distinguish an outcome that has not happened yet from one that can no longer happen.

This chapter explains the time model behind a commercial forecast. It connects stage dates and competing outcomes to conversion, booking dates and work in progress. You can then decide whether to change the pipeline, allow more time or alter the capacity plan. The forecast remains conditional on the motion and customer mix that produced its estimates.

Two ways to compute a rate

Take the close rate: the share of sales-qualified leads (SQLs) that become customers. You can compute it in two ways.

The **same-period rate** counts the customers won in a period and divides by the SQLs created in the same period. Every dashboard can show it. The **cohort rate** takes the SQLs created in one period and follows that group until each member is won or lost. It divides the wins of the group by the size of the group.

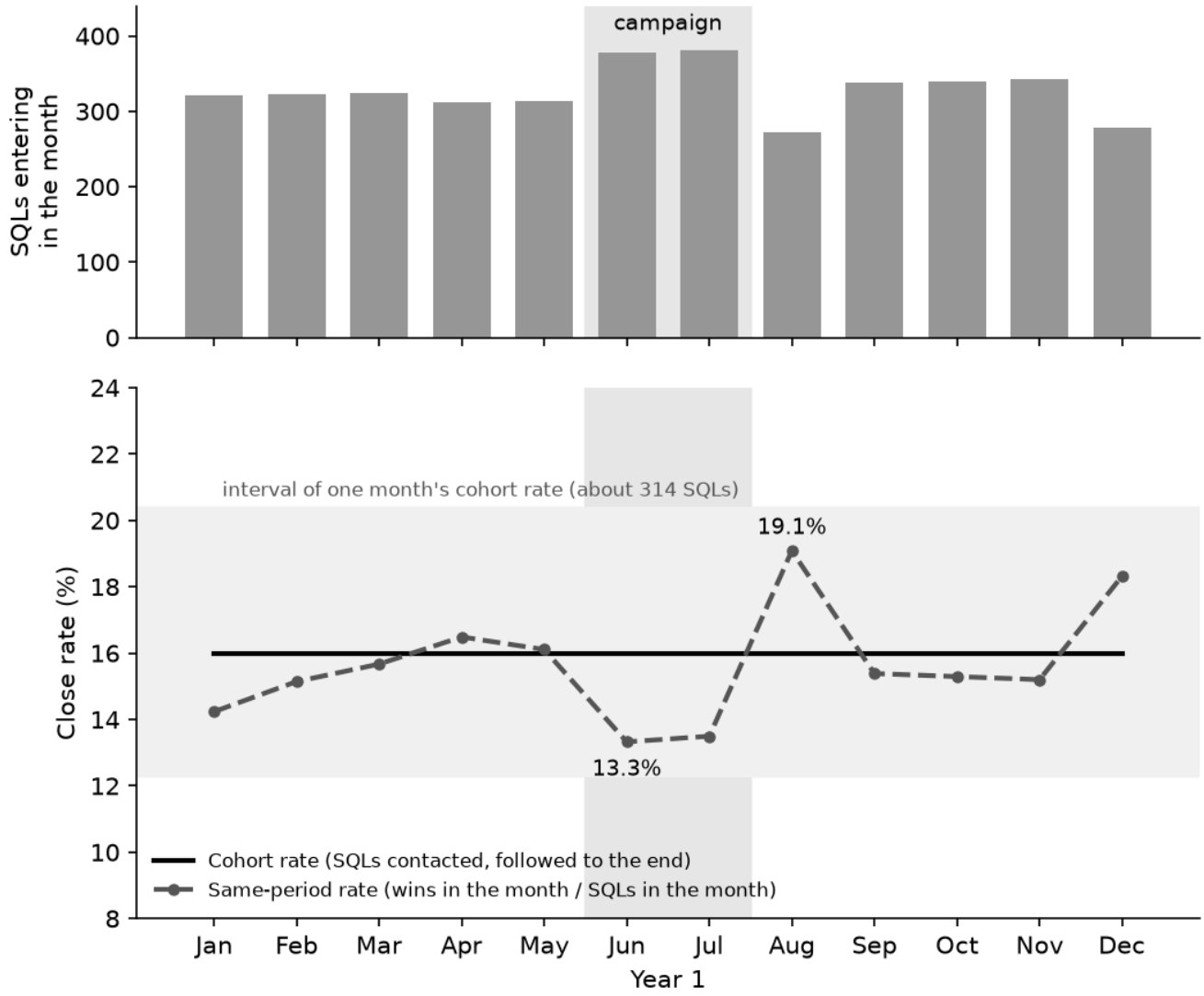
The two rates answer different questions. The same-period rate divides the wins of one group of SQLs by the size of another group. Most of this month's wins come from SQLs created in earlier months. Most of the SQLs created this month that will win do so in later months. When volume is steady, the two groups have the same size and the two rates agree. When volume changes, they do not.

At Birch, the sales cycle from SQL to win has a median of 34 days. Of all the SQLs that will win, 14% win in the month they were created, 53% in the next month, 24% in the month after that, and the last 10% later. Now let SQL volume rise 20% in one month and stay at the new level. In that first month, only 14% of the wins come from the larger cohort. The rest still come from the smaller cohorts of earlier months. Wins rise by about 3%, while SQLs rise by 20%. The reported close rate falls by 14%, and no seller has done anything differently. In the second month it is still 6% low. In the third month it is 2% low. Then it recovers.

The effect works in both directions. A 20% fall in volume lifts the reported rate by about 22% in the first month. A seasonal swing does the same every year. Even steady growth leaves a mark. The reported rate sits below the cohort rate by about the monthly growth rate times the average lag.

> **Calculation**
>
> Reported rate = cohort rate × the sum, over every lag L in months, of (share of wins that close L months after entry × volume L months ago ÷ volume now).
>
> For slow, steady growth of g per month: reported rate ≈ cohort rate × (1 − g × average lag in months).
>
> Birch has an average lag of 1.3 months. At 3% growth a month, the reported rate is 4% low. At Birch's own trend of 0.6% a month, it is 1% low.

Evidence label: Arithmetic.



Birch's inbound SQLs in Year 1: the campaign lifts SQL volume in June and July, the same-period close rate swings between 13.3% and 19.1%, and the cohort rate of the SQLs that sellers contacted stays at 16.0%.
A same-period rate still shows how volume and flow move. For the question "did selling change?", use a cohort rate. Name the method on every dashboard that shows a close rate.

A cohort rate has its own trap: it needs an age. If you count the wins in the first median cycle after entry, you see only half of the final wins. For Birch, a cohort shows 8.0% at day 34 against a final rate of 16.0%. So print the age next to every cohort rate: "close rate at day 56", or "close rate at maturity".

Cohorts, age and maturity

A **cohort** is a group of items that entered at the same time. A cohort of SQLs holds all the leads that became SQLs in one week or one month. Define a cohort only by what is true at entry: the entry date, the channel, the segment. Never define it by what happens later. A group built from the deals that closed leaves out the deals that did not, and every rate then looks better (Linoff and Berry, *Data Mining Techniques*).

The **age** of a cohort is the time since it entered. We read cohorts at fixed ages, for example day 14, 28, 56 and 90. These four ages are start values; choose ages that fit your own cycle.

The **cohort table** puts cohorts and ages together. Each row is one cohort, oldest first. Each column is one age. Each cell holds the cumulative conversion of that cohort at that age: the share of the cohort that has reached the outcome by then. We call it the cohort conversion at that age.

Birch, SQL cohorts by month of first contact, read on 30 June of Year 1. The clock starts when a seller contacts the SQL. Cell: share of the cohort won by that age. Deals are simulated from the company's parameters with a fixed seed. Illustration only.

Cohort	SQLs	Day 14	Day 28	Day 56	Day 90
January	321	0.6%	4.0%	12.5%	14.6%
February	323	0.9%	6.2%	13.0%	15.2%
March	325	1.5%	6.2%	12.0%	14.2%
April	311	1.3%	5.8%	12.5%	open
May	314	1.6%	8.0%	open	open
June	325	1.0%*	open	open	open

The lower right of the table is empty. Those cohorts have not reached those ages yet. A cell is "open" when fewer than half of the cohort is old enough to be judged. A star marks a **partial** cell: some SQLs in the cohort are still too young, so the cell uses only the older ones. Compare only complete cells.

You can read the table in three ways.

- **Along a row**, you see one cohort ageing. The curve climbs slowly, then fast around the median, then flattens. Every cohort shows this shape.
- **Down a column**, you see different cohorts at the same age. A real change in selling shows here, as a step that stays. In the day 28 column the cells read 4.0%, 6.2%, 6.2%, 5.8% and 8.0%. That looks like a trend, but each cell holds only 13 to 25 wins. The spread is the noise that such small counts make.
- **Along a diagonal**, you see all the cohorts at one calendar date. A tracking error, an outage, a price change or a holiday hits every cohort that is in flight on that day. The diagonal is a straight line only when the age step equals the cohort step, for example weekly cohorts read at week 1, 2 and 3. With other ages, mark the cells whose dates cover the event.

These readings separate three kinds of change.

Kind of change	What it follows	Where it shows	Example
Age effect	The age of the cohort	The shape of every row	Few deals close in the first two weeks
Period effect	A calendar date	A diagonal across many cohorts	An outage, a tracking break (a change in what is measured, not in what happens), a price change
Cohort effect	One group of entrants	One row, at all ages	A campaign that brings different buyers

Age, period and cohort effects mix easily, so check the change log before you explain a dent. Chapter *Diagnose what changed* shows how.

A cohort is **mature** when most of its outcomes have happened. We set the maturity age at the 90th percentile of the cycle: the age by which nine in ten of the cohort's final wins have arrived. This is a start value and a display rule. Evidence label: Plausible but untested. Each company tunes it on its own data. Birch matures at 73 days. Cedar's self-serve motion matures at 29 days and its assisted motion at 91 days. Alder matures at 243 days, about eight months. Before the maturity age, do not blank the cohort and do not wait. Print its rate with its age, and show an estimate that allows for censoring. The next section shows how.

Implementation note: the companion code holds one SQL view that builds the cohort table from stage dates and flags each cell as complete or partial.

Censoring: an open deal is not a lost deal

An open deal has no result yet. We call it **censored**: we have watched it for some time, and it has not ended. We met this idea for contracts in *Model recurring growth*, and it holds for deals in the same way: a censored deal is neither a loss nor missing data.

Two habits go wrong here. The first counts open deals as failures. On 30 June, the June cohort of Birch holds 325 SQLs and 9 wins. That is 2.8%, against a final rate of 16%. The cohort is not bad. It is young. The second habit drops every cohort that is not mature. That hides the newest 73 days, which are the days you most want to see.

The better way uses each deal only for the ages at which we have watched it. We cut age into steps, for example one week. For each step, we count the deals that were still open at the start of the step and that we watched until the step ended or the deal closed. These deals are **at risk**. The **hazard** of the step is the share of the deals at risk that win during the step. The hazard was defined for renewals in *Model recurring growth*. Here it is the same ratio, taken by week of deal age.

Birch, weekly hazard of a win, all SQLs contacted from January to June, read on 30 June of Year 1.

Week of age	Deals at risk	Wins in the week	Hazard of a win
2	1,781	20	1.1%
4	1,308	42	3.2%
6	724	32	4.4%
8	366	22	6.0%

The count at risk falls for two reasons. Deals close, and young deals have not reached that week yet. A young deal counts in the early weeks and leaves the later weeks, so it never counts as a failure.

From the hazards, you build the **cumulative conversion**: the share of the cohort won by a given age. At each step, multiply the share still open at the start of the step by the hazard of a win in that step. Add the steps up.

> **Calculation**
>
> Cumulative won at age a = sum over the steps up to a of (share still open at the start of the step × hazard of a win in the step).

> The share still open comes from the product-limit estimate in *Model recurring growth*, with any close counted as an exit.
>
> Evidence label: Standard practice.

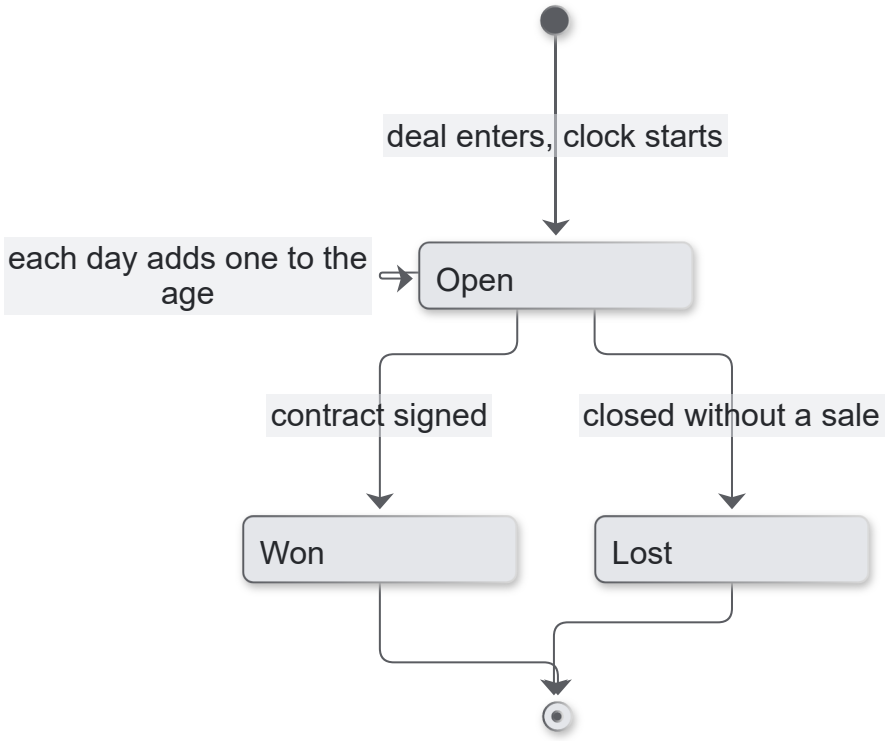
The estimate works at once, even for a young cohort, and it comes with an interval. A 95% **interval** is a range built so that, over many repeats, it holds the true value 95 times in 100. It is wide where few deals were watched, and that is the honest answer. Take the June cohort of Birch on 30 June. Its oldest SQLs are 29 days old, and 117 of them were at risk in the third week. Its estimate at day 21 is 1.8%, with an interval of 0% to 4.2%. The cohorts of January to May show 3.7% at day 21, inside that interval. The interval says how little we know yet.

You can go one step further and project the **eventual conversion**. Take the complete cohorts and measure what share of their final wins had arrived by day 28, with the day 90 count as final. For Birch, using January to March, it is 37%. The May cohort has 25 wins among 314 SQLs at day 28, which is 8.0%, with an interval of 5.5% to 11.5%. Divide by 0.37. The projected rate at maturity is 21%, with an interval of 15% to 31%. That interval is wide. Now pool April and May, the two cohorts that have not yet reached day 90. They show 6.9% at day 28, which projects to 18%, with an interval of 14% to 24%. Pooling narrows the interval because it adds wins. A single month is a hint, not a result.

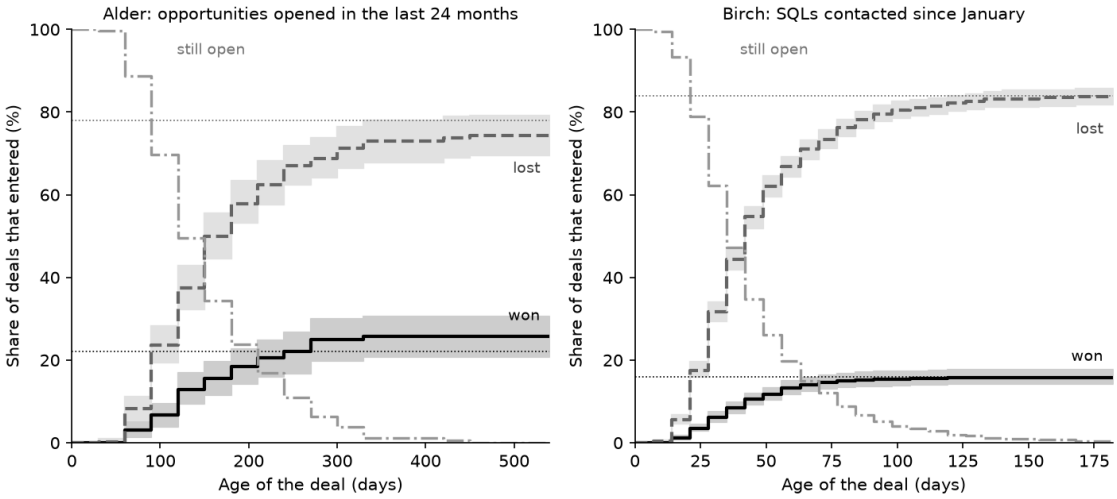
This projection assumes that a young cohort ages like the old ones. If a campaign brings slower buyers, the projection is too high. Check it by reading the cohort again at each later age. The interval above counts the noise in the new cohort only. It ignores the smaller noise in the reference share.

The pipeline as time to event

A deal is a clock with two exits. It starts when the deal enters a stage. Each day adds one to its **deal age**, the number of days since the deal opened. It ends when the deal is won or lost. The time from entry to the end is the **time to event**.



Won and lost are **competing outcomes**: a deal cannot have both. At every age, the share won, the share lost and the share still open add up to 100%. Estimate the hazard of each exit on its own. Then build each curve with the share still open at the start of each step. Do not draw the won curve as one minus a won-only curve, with lost deals treated as censored. That treats lost deals as if they could still win, and it overstates the win share.



Bands: 95% interval from resampling deals. Dotted lines: the eventual shares used to create the deals. Open deals are censored, not counted as lost. At each age, won + lost + still open = 100%.

Cumulative won, lost and still-open shares by deal age for Alder and Birch on 30 June of Year 1, with 95% bands, from deals simulated with company parameters (Illustration only).
Read the figure at one age. At day 56, Birch has won 13% of its SQLs and lost 67%, and 20% are still open. At day 180, Alder has won 18% of its opportunities and lost 58%, and 24% are still open.

The figure also shows the effect of volume. Birch has 1,919 SQLs in its curve, and the band is narrow. Alder has 388 opportunities, and the band is wide. A low-volume company reads the same chart, but it reads a wider range.

The curve gives the **cycle percentiles**. P50 is the median. P80 is the age by which 80% of the deals have closed. P90 is the age by which 90% have. Cycles are skewed: a few deals take a long time. In the example companies the 90th percentile lies at two to just over three times the median.

Take the percentiles for the outcome that the deadline is about. A close date is about a win, so use the time to a win: the won curve divided by its final level. If won and lost deals end at the same pace, the curve of all closes gives the same days. In the example companies they do, by assumption (Illustration only). In real data, lost deals can end at a different pace, so compare the two curves before you use the shortcut.

Cycle percentiles of the example companies, in days from entry to a win. The cycle shape is lognormal, an assumption that the companies share. Plausible but untested; fit your own cycle from your own log.

Motion	Cycle starts at	P50	P80	P90
Alder, account-based	Opportunity	120	191	243
Birch, inbound	SQL	34	56	73
Cedar, self-serve	Sign-up	9	19	29
Cedar, assisted	Opportunity	45	71	91

Set deadlines on P80, not on the median. By definition, half of the deals miss a date set at the median. A date set at P80 is met by four deals in five. For Birch, the median date is day 34 and the P80 date is day 56. The same holds for the loop delay of *Choose the growth objective*. The median gives the typical delay. A deadline for a decision needs P80 plus one reading window of 28 days. That is 219 days at Alder, 84 days at Birch and 47 days for Cedar's self-serve motion.

Estimate the percentiles from the curve above, with open deals censored. Never take them from the closed-won deals alone. The slow winners are still open, so they are missing from that list, and the percentiles come out too low. The Alder example below puts a number on this.

A **stalled deal** is an open deal that is older than P90. It needs a decision: act on it or close it. The published evidence supports an age limit of this kind. A sample of best-performing firms found that deals open for more than twice the average cycle had a 3% chance to close (Ebsta and Pavilion, 2023, *2023 B2B Sales Benchmark Report*). In the simulated deals of this chapter, won and lost deals share one timing, so an old deal does not become less likely to win. Real histories usually show the drop. Test your own flag: compare the win rate of deals past the line with the rate of the others.

Time in stage is the same idea inside one stage: the days a deal spends in that stage. Take the P80 of each stage, from the stage history. A stage with a long P80 is where deals wait. This needs a record of every stage change, with its date. Without that record you can measure only the whole cycle. Alder has no stage history until October of Year 1.

Forecasting bookings from deal age and stage

A forecast of bookings is a sum over the open deals. Each open deal has a chance to win within the horizon, and that chance depends on its age and stage. We call it the **conversion by age**.

> **Calculation**
>
> Expected wins = wins already booked in the period + the sum, over every open deal, of the chance that it wins within the horizon + expected wins from deals not yet opened.
>
> Evidence label: Standard practice. The assumption that open deals age as past deals did is an Assertion, so test it on past cohorts.

The work has five steps.

1. Estimate, from past deals, the chance that a deal of a given age wins within the horizon.
2. List the open deals with their ages, and add up their chances.
3. Add the wins already booked in the period.
4. Add the wins expected from deals that are not yet opened. For a horizon shorter than the cycle, this part is small.
5. Report a range, not one number. The spread of the sum gives P10, P50 and P90. The deals are separate events, so their uncertainties add up in a known way. P10 is a low case that the result should beat nine times in ten. P90 is a high case that it should beat once in ten.

If your data hold the stage history, add the stage to the condition. A deal in the proposal stage at day 60 is not like a deal in the first meeting at day 60. Condition on the stage only when each stage holds enough deals to estimate it.

A past forecast can be graded only if you can rebuild what you knew when you made it. Today's CRM shows deals as they are now, with stage, amount and close date overwritten. So store a **snapshot**, a copy of the pipeline on one date, every night. An **as-of view** then rebuilds the pipeline as it was on any past date. The chapter *Learn from outcomes* uses these views to grade forecasts.

Implementation note: the companion code includes a stage-event table and a daily snapshot table for this purpose.

The **open pipeline** is the set of deals that are still open. Group it by age band, and by stage once you have stage history. The table below shows the layout.

Alder, open pipeline by deal age in June of Year 1. Expected values in steady state, from the cycle in the table above and a win rate of 22%. This is the layout of a dashboard view.

Age band (days)	Open deals	Chance a deal wins in the next 90 days	Expected wins	Flag
0 to 60	34.0	11%	3.6	
61 to 120	24.3	15%	3.5	
121 to 191	13.5	15%	2.1	Past the median date
192 to 243	4.3	15%	0.7	Past P80
Over 243	4.7	14%	0.7	Stalled: act or close
Open deals	80.9		10.5	
Deals not yet opened			0.9	

When you sum windows into a quarter, take care. A window is 28 days. A quarter holds 3.25 windows, not 3.

To compare two forecasters of close dates, size the test. Suppose the difference between their errors has a standard deviation of 10 days, and one forecaster is better by 3 days. Then you need about 88 deals for 80% power, which means an 80% chance to see a gain of that size. Illustration only. A fixed count, such as 30 deals, is too few in one case and too many in another, so size each comparison. The chapter *Decide what the evidence can resolve* gives the sizing method.

Little's law

Little's law links the three counts of a pipeline. In a steady system, the average number of open items equals the arrival rate times the average time an item stays open (Little, 1961, Operations Research). It is a proof, so its evidence label is Arithmetic.

> **Calculation**
>
> Open items = arrivals per day × average days open.

Use it first as a data check. Alder opens about 0.58 opportunities a day. A deal stays open for 140 days on average. The law predicts about 81 open deals. If the CRM holds far more, many of those deals are dead and nobody has closed them. If it holds far fewer, check the arrival counts and the cycle.

Use it second as a capacity read. With four account executives, 81 open deals are about 20 per executive. Let the cycle grow 20% at the same arrival rate. Open deals then rise 20% to 97, which is 24 per executive. The workload grows although nobody opens more deals.

The law needs a steady system and a long enough period. It also needs the same definition of "open" on both sides. A pipeline in rapid growth or in a seasonal swing breaks the identity for a while.

Who reads these measures

The measures of this chapter have different readers and different rhythms.

Measure	What it tells	When you read it	Who uses it
Cohort conversion at age (day 14, 28, 56, 90)	How a group of entrants converts, at a stated age	Weekly or monthly, at each fixed age	Sales and marketing owners, to see a real change
Projected eventual conversion	The likely final rate of a cohort that is not yet mature	Weekly, until the cohort is mature	Growth owner, for early warnings and forecasts
Maturity	The age at which a cohort counts as final	Set once, reviewed each quarter	Analytics lead, as a display rule
Sales cycle P50, P80, P90	How long deals take	Each quarter, on a long history	Sales owner, for dates and deadlines
Time in stage	Where deals wait	Weekly, once stage history exists	Revenue operations, to find bottlenecks
Deal age	How long an open deal has been open	Daily	Account executives
Stalled deals	Open deals older than P90	Weekly	Sales owner, who decides to act or close
Open pipeline	Count of open deals by age band and stage	Daily snapshot, weekly read	Sales owner and finance owner
Bookings forecast, P10, P50 and P90	Expected wins with a range	Weekly	Sales owner and the board

What changes with volume

The mechanics hold at any volume. What changes is the size of the cohort you can read, and so the unit of the table.

Company	Cohort unit	Size of one cohort	Rate read	Smallest relative difference between two cohorts
Alder	Quarter of opportunities	53	Win rate, 22%	above 100%
Birch	Month of SQLs	314	Close rate, 16%	52% (30% with three months pooled)
Cedar, self-serve	Week of sign-ups	10,070	Activation, 36%	5%
Cedar, assisted	Quarter of opportunities	203	Win rate, 25%	49%

Smallest relative difference: 4 × the square root of (1 − p) ÷ (p × n), for two equal cohorts, 80% power and 5% significance. Arithmetic. The chapter Decide what the evidence can resolve explains the rule.

At low volume, a cohort rate says almost nothing. Alder's quarterly cohort holds about 12 wins. No quarterly difference below 100% is readable. A cohort also takes eight months to mature. So a low-volume company steers on two other things: the time-to-event model, which pools two or more years of deals, and a forecast from deal age.

At medium volume, two monthly cohorts show only a large difference, about 52%, and three pooled months show about 30%. An artefact of volume can be as large as a real change, so the cohort reading is a defence against false alarms. It is not yet a way to find small changes.

At high volume, weekly cohorts are readable, and the age can be counted in days. A period effect stands out within days. High volume does not help every motion, though. Cedar's assisted motion behaves like a low-volume business.

Alder, Year 1 June: deals that slip

Alder, Year 1 June: about 3.6 new customers a window, 0.8 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.1 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

The quarter report goes to the board. The forecast had listed deals with a close date in the quarter. Fewer deals closed than listed. Several deals "slipped" to the next quarter, and some of them won there. The head of sales explains each slip with a story about the buyer. The analyst asks a different question: when should a seller write the close date?

The CRM holds the opening date, the closing date and the outcome of every deal. It holds no stage history yet. In the last 24 months, 388 opportunities were opened. Today 77 are won, 223 are lost and 88 are still open. The analyst builds the competing-outcome curves from these deals, with the open deals censored. They are the Alder panel of the figure above. The curve gives a median of 125 days, a P80 of 205 days and a P90 of 245 days. The deals are simulated from a cycle with a median of 120 days and a P80 of 191 days, so we can check the estimate. It is 5 and 14 days high, which is inside the noise of two years of deals (see the table below).

The analyst also tries the common shortcut: percentiles of the won deals only. That gives 113, 182 and 229 days. How large is the error, and how much history does the estimate need? The analyst repeats the exercise on 300 simulated histories of each length.

Alder, estimates of P80 and P90 in days from simulated histories of different length. The company's own values are P80 191 and P90 243. Range: 5th to 95th percentile over 300 histories. Illustration only.

History	Opportunities	P80, product-limit	P90, product-limit	P80, won deals only
6 months	103	not reached in 187 of 300	not reached in 253 of 300	106
12 months	199	193 (170 to 220)	246 (210 to 292)	151
24 months	388	193 (180 to 210)	245 (225 to 270)	174
48 months	745	194 (185 to 205)	246 (230 to 260)	183

Six months of deals are too short. The curve cannot pass the age of its oldest deal, 181 days, and by that age only about three deals in four have closed. So in most histories the curve never reaches 80%. One year gives the right P80 on average, with a range of 170 to 220 days. Two years narrow it to 180 to 210 days. The won-only shortcut stays too low even at four years.

Now the analyst reads what a median date does. Half of the winners close after the median date of 120 days. The late half closes 78 days late on average. Of all winners, 34% close more than 30 days late, and 23% close more than 60 days late. The P80 date lies 71 days after the median date. Alder is not unusual: a published sample of best-performing firms reports that 37% of deals slipped in 2022 (Ebsta and Pavilion, 2023).

Little's law adds a check. About 81 deals should be open, and the CRM holds 88. A count near 81 varies by about 9 either way, so the gap is noise, and the CRM holds no large pile of dead deals. For the next 90 days, the age-based forecast expects 10.5 wins from open deals, with a P10 to P90 range of about 7 to 14. It expects 0.9 more from deals not yet opened. Of the 10.5, about 32% come from deals that are already older than the median date. A list that treats those deals as "slipped" leaves out a third of the wins.

The head of sales decides. Account executives write the P80 date as the default close date: 205 days after the opportunity opens, updated each quarter. The board sees the forecast as a range of wins, not one number. Deals older than 245 days go on a review list. Alder has no decision log yet, so the owner writes the decision and its date in the forecast note.

Birch, Year 1 June: a fall that was not a fall

Birch, Year 1 June: about 48 new customers a window, 1.0 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 71 days, and the close rate reads a 20% change in about 6 months.

A campaign has run since the start of June. It lifts SQL volume by 21%, from 314 SQLs in May to 379 in June. When the dashboard updates in early July, it shows the close rate at 13.3% for June against 16.1% for May. That is a fall of 17%. The head of sales says the new leads are poor. The head of marketing says the sellers are slow.

The analytics lead starts with the counts. Birch won about 50 customers in May and about 50 in June. Wins did not fall. The SQL count in the denominator rose.

Next, the analytics lead asks what unchanged selling would show. With the cycle shares of the mechanics section, a 21% step in volume lowers the same-period rate to about 13.6%. That is a fall of 15% with no change in selling. The dashboard fall of 17% is almost all arithmetic of the clock.

The cohort table comes next. The April and May cohorts show 5.8% and 8.0% at day 28, inside the noise of the earlier months. Pooled over January to May, the day 28 rate is 6.0%, with an interval of 5.0% to 7.3%. The June cohort is young. At day 14, only 191 of its 325 SQLs are old enough, and they hold two wins. The censoring-aware estimate uses every June SQL for the days it was watched. At day 28 it gives 11%, with an interval of 4% to 20%, because only 38 SQLs were at risk in the fourth week. The interval holds the earlier 6.0%. So June shows no sign of a fall, and it is too young to rule one out. The first complete reading arrives at the end of July, when the whole June cohort reaches day 28.

The analytics lead also sets the artefact against the noise. One month of Birch SQLs holds about 314 contacted SQLs. A cohort rate from that many SQLs, at a true rate of 16%, has an interval of 12.3% to 20.4%. Even a true cohort reading of 13.3% would lie inside it. Two monthly cohorts show a difference only when it is

above 52%. Three pooled months reduce that to 30%. So the cohort table does not prove that selling is unchanged. It says that the data show no change, and that the dashboard fall is the size that the clock alone makes.

One thing in the data is real. In June, sellers contacted 325 SQLs, but 379 arrived. Some 54 SQLs, 14% of the month, waited uncontacted at the end of June. Timing alone explains a fall of 15%. The queue explains the other 2%. If selling has not changed, the contacted June SQLs close at the usual 16%. The June cohort shows this when it matures, in the second week of September. Until then, the signal to act on is a capacity signal, not a quality signal. *Set targets and staff capacity* and *Allocate sales attention* take it up.

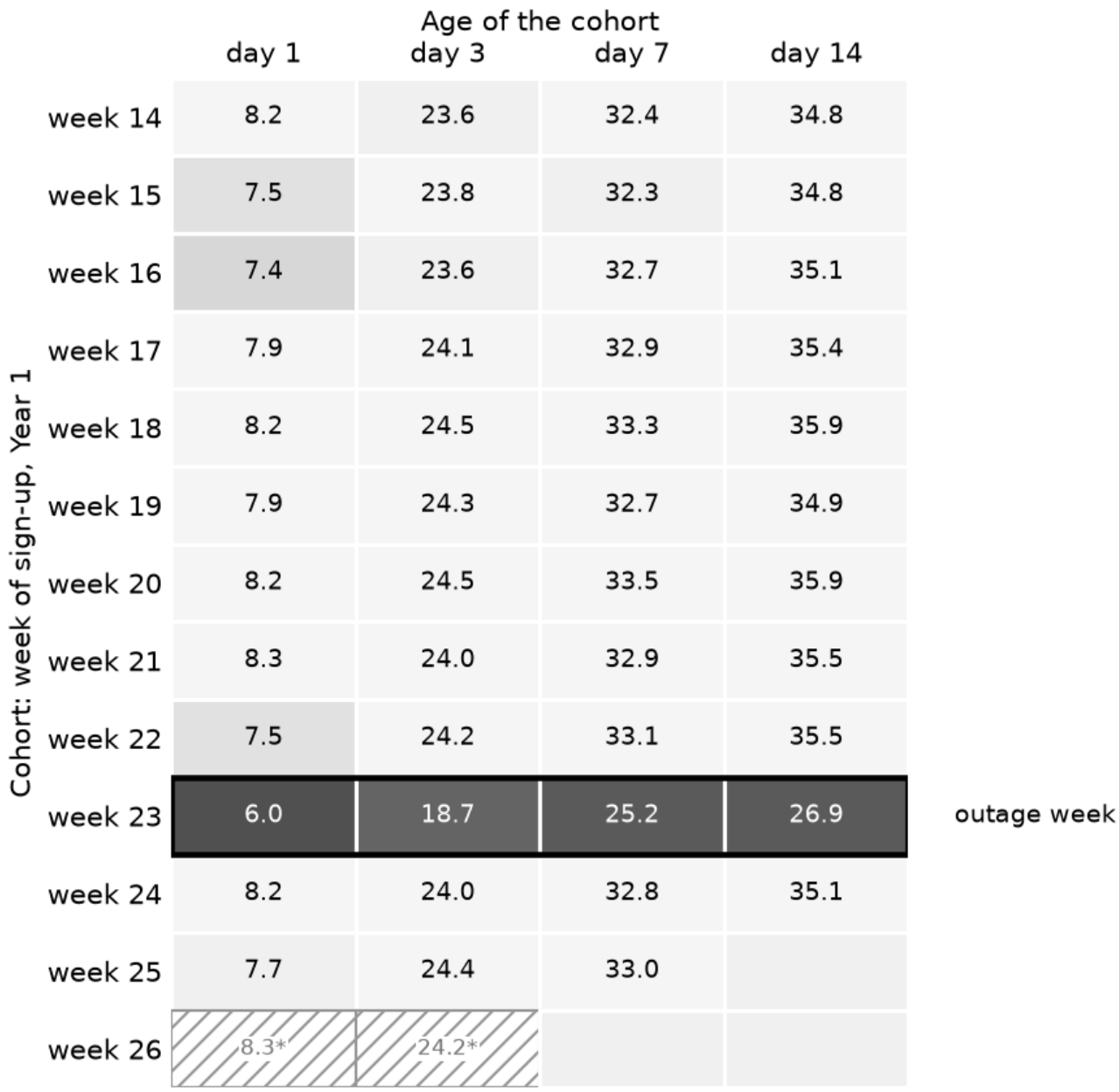
The arithmetic also predicts what follows. In July the same-period rate stays low, at 13.5%. In August the campaign has ended and summer volume is low, so the reported rate jumps to 19.1%, with no change in selling. Over Year 1 the reported rate ranges from 13.3% to 19.1%, while the cohort rate of unchanged selling stays at 16.0%.

The head of sales decides. Qualification stays as it is. The weekly dashboard now shows the cohort table and names the method of every rate. The June cohort is read at day 28, with its interval, at the end of July. The queue goes to the capacity work in the next chapters.

Cedar, Year 1 June: one bad row

Cedar, Year 1 June: about 1,300 new paying accounts a window, 1.4 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 32 days; the assisted motion, with 59 opportunities a window, needs about 8 months to read a 30% change.

Cedar's analysts keep a grid of weekly sign-up cohorts. The cell shows the share of each cohort that has activated by day 1, 3, 7 and 14. Each weekly cohort holds about 10,000 sign-ups. At that size, a weekly cohort can show a relative difference of 5%.



Cedar's weekly sign-up cohorts by age, share activated on 30 June of Year 1: the cohort of week 23 is darker than the typical cohort at every age (the activation timing is an illustration value).

On the grid, one row is dark. The cohort of week 23, the first full week of June, activates at 6.0% by day 1, against a typical 7.9%. By day 14 it reaches 26.9%, against a typical 35.3%. The fall is 22% to 25% at every age. At day 14, the gap is 17.6 standard errors. A standard error is the typical size of the noise in one number. The rows above and below look normal.

A dark row at every age is a cohort effect: something touched the people who signed up that week, and only them. A period effect would look different. If the product itself had been down in week 23, the week 22 cohort would also be dark at day 7 and day 14. About a third of the activations it makes by those ages fall in week 23. Those cells sit on the diagonal of week 23, and they are normal: 33.1% and 35.5%. So the analytics lead

looks for a cause that hit new sign-ups only. The change log shows it. The service that sends the set-up email to new sign-ups was down for most of week 23, so those sign-ups never got the link that starts the set-up. The grid shows where the dent sits, not why. The log gives the cause.

Compare this with the monthly view. The month of June as a whole shows a fall of about 5.7%, spread over the four and a third weeks of the month. A month of sign-ups can read a relative change of 2.6%, so the monthly number does show the dip. But it comes on 1 July, and it does not say which week. The weekly grid completes the day 3 cell of week 23 on 14 June. That is 17 days earlier, and it points at one week.

The outage cost real customers. If the activations were lost for good, the cohort lost about 906 activations. At the usual 9.5% of activated accounts that pay, that is about 86 paying accounts.

The growth lead decides two things. First, the change log marks week 23 as an outage week. Second, every experiment that ran in that week drops the week from its readout, or runs one week longer. About 15 experiments a month run on sign-ups, so this is a real risk to the test results. The rest of the grid stays in the baseline.

The assisted motion cannot use this reading. A weekly cohort there holds about 16 opportunities. That motion reads by quarter, in the way of Alder and Birch. High volume belongs to a motion, not to the company.

The decision to carry forward

Forecast from information available at the forecast date. Treat unfinished opportunities as unfinished, and recognise outcomes that prevent a win. Record the forecast distribution before changing the process or capacity that produced it.

11 · Set targets and staff capacity

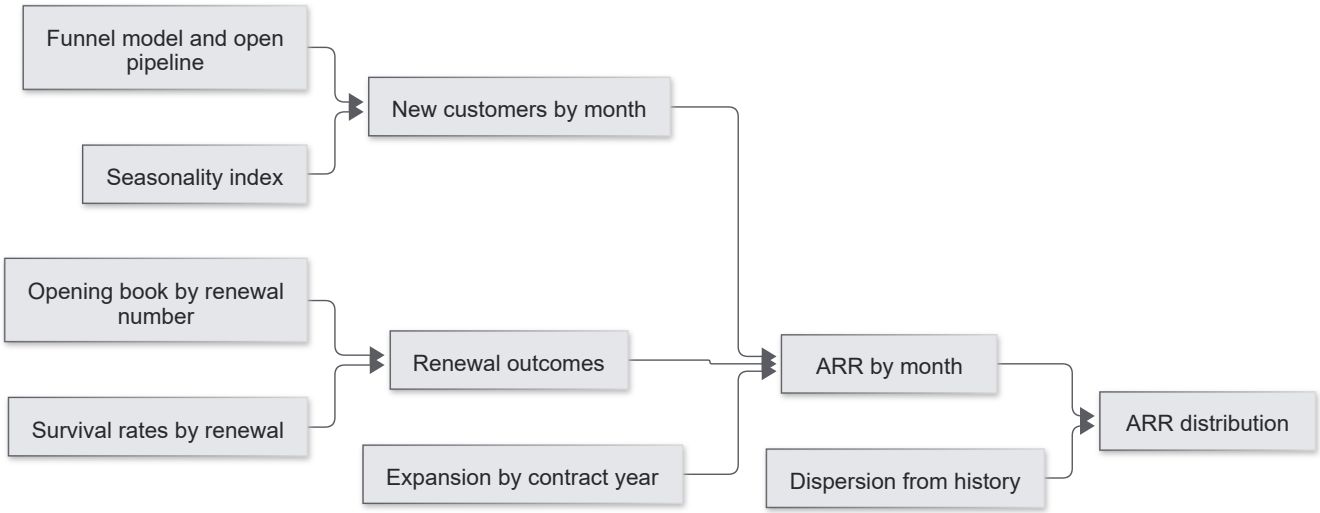
What target can the business support, and which capacity must be committed before demand arrives? A revenue ambition becomes an operating plan only when the model connects acquisition, conversion, delivery and retained revenue over time. Hiring changes that model after recruitment and ramp delays.

Treat a target as a threshold with a probability of being reached. Treat capacity as a constraint on how customer opportunities become revenue. An average number of accounts per employee does not establish the workload of your company. The service requirement and touch level determine that workload.

The forecast below converts assumptions into a distribution of outcomes. Use it to identify the commitment that must happen now, the signpost that will test it and the date when waiting becomes costly. Keep the baseline path beside the proposed plan.

A cohort-based ARR forecast

ARR in a year is the sum of two parts. The first is what the customers you have today will do: renew, shrink, grow or leave. The second is what new customers will add. A cohort-based forecast builds both parts from units, as the hybrid method of *Choose the growth objective* asks. It then compares the result with the top-down target.



The work has five steps.

1. Take the book by contract anniversary: how many customers renew in each month, at which renewal number, and with which ARR.
2. For each renewal, draw the outcome with the renewal rate for that number, as in *Model recurring growth*. Leavers take their ARR. Stayers contract or expand.
3. Add expansion for the customers that stay, by contract year.
4. Add the new customers that the pipeline model expects in each month, with the seasonal index. The open pipeline and the cycle of *Forecast conversion and timing* set the timing.
5. Repeat the draw many thousand times and read the spread of the outcomes. This is the Monte Carlo method of *Value customers and cash*.

> **Calculation**
>
> ARR at the end = ARR now − churned − contraction + expansion + new
>
> Each term is a sum over cohorts. Churned = customers due at a renewal × (1 − renewal rate for that number) × the ARR that leavers hold.
>
> Evidence label: Arithmetic for the identity. Standard practice for the cohort method.

Where the width comes from. A forecast range has several sources. The count of new customers is random. So are the size of each deal and the outcome of each renewal. The renewal rates themselves are uncertain, because you know them only from the decisions you have seen. Expansion arrives in lumps. At a company with a very large account, the renewal of that one account is a source of its own. The table switches on one source at a time. It shows the standard deviation of ARR in 12 months, as a percentage of opening ARR, at Year 1 July.

Source of noise	Alder	Birch	Cedar
Count of new customers	3.15	1.02	0.61
Deal size	0.80	0.29	0.17
Renewal outcomes	0.84	0.45	0.36
Uncertain renewal rates	0.91	0.46	0.21
Expansion	2.82	0.40	0.13
The largest account	2.00	0.00	0.00
All sources together	4.81	1.29	0.78

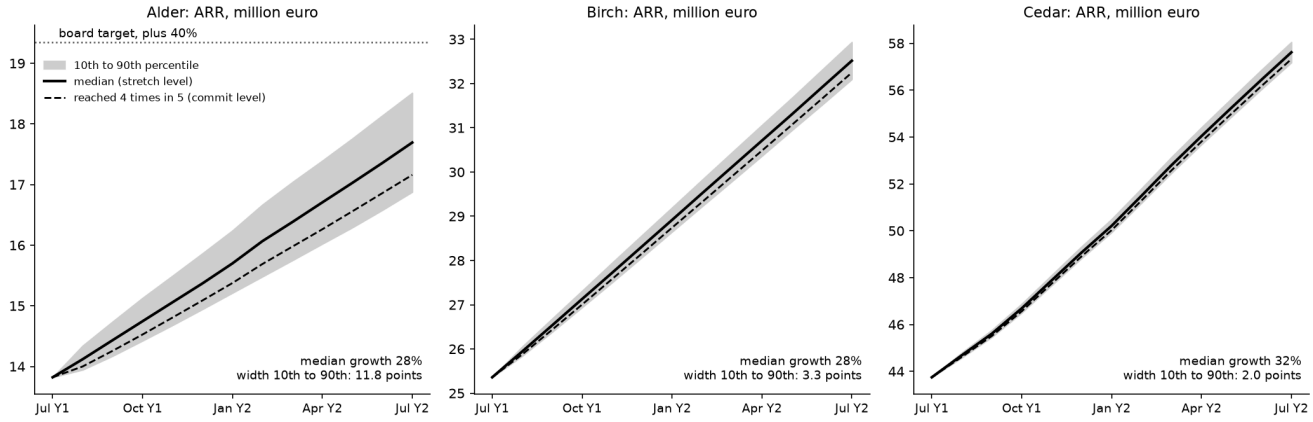
Simulated: 3,000 years for each single source, 10,000 for all sources together. The deal-size spread and the lumpiness of expansion are start values. Illustration only.

Read the table by column. At Alder, a handful of new deals, a few lumpy expansions and one very large account make up the range. At Cedar, the sources are small, and the renewal outcomes lead. The sources do not add up: variances add, so the total is smaller than the sum.

Dispersion. Real counts vary more than pure chance. We call the extra spread **dispersion** and write it as ϕ. If ϕ is 1, the variance of a count equals its mean, as in a Poisson count. If ϕ is 2, the variance is twice the mean. Estimate ϕ from at least 13 windows: the average of (actual − expected)² ÷ expected, taken over the windows. Until then, run the forecast with ϕ at 1 and at 2 and show both. The chapter *Decide what the evidence can resolve* names this effect and explains it in full. With ϕ at 2, the 12-month growth range of Alder widens from 22.1% to 33.9% to 20.9% to 35.4%. Birch widens from 26.6% to 29.9% to 26.1% to 30.4%. Evidence label for the start value: Plausible but untested.

Seasonality. The **seasonality index** of a month is its usual level divided by the average level. Cedar's January has an index of 1.20 and its July 0.85. Alder and Birch have two low months, August and December, at 0.80 for Alder and 0.85 for Birch. Estimate an index from at least two years of history. Divide each month by a centred 12-month average, average the ratios by month, and scale the indices so that they average 1. With less history, compare a month only with the same month a year earlier, never with the month before. The forecast multiplies the pipeline inflow by the index. A plan that ignores the season shows a false fall every summer and a false rise every January.

What the range leaves out. The range holds only the noise that the model knows. It does not hold a new competitor, a price change or an outage that nobody has planned. So it is a floor on the real width. Check it with a rolling backtest. Forecast each of the last eight quarters using only the data before that quarter. The 80% range should hold the outcome in 70% to 90% of the quarters. This coverage rule is a start value (Plausible but untested). Eight quarters give only a coarse check, because the coverage is itself a small, noisy count. Read a miss as a reason to look, and pool more quarters as they come. *Learn from outcomes* grades forecasts this way, which is why *Forecast conversion and timing* asked you to store snapshots.



ARR forecast from Year 1 July for the next 12 months. Grey band: 10th to 90th percentile of 10,000 simulated years. Solid line: the median. Dashed line: the level that the plan reaches four times in five. Dotted line at Alder: the board target of plus 40%.
Alder's band is wide, and its target of plus 40% sits outside it. Birch's band is narrow, and Cedar's is narrower. The 10th to 90th percentile spans 11.8 points of growth at Alder, 3.3 at Birch and 2.0 at Cedar.

Targets as probabilities

A target is one level. A forecast is a distribution. The **probability of reaching a target** is the share of simulated years that end at or above it. Take Alder at Year 1 July. The forecast reaches plus 20% with probability 95% and plus 25% with 75%. It reaches plus 30% with 33%, plus 35% with 6.5% and plus 40% with 0.6%.

Two levels of the distribution do the work. The **stretch level** is the median: the plan reaches it half the time. The **commit level** is the value reached with 80% probability, which is four times in five. It is the 20th percentile of the simulated years. Do not call it P80: in the fan, P10 and P90 are percentiles, so P80 would read as the 80th percentile. For Alder the stretch level is plus 28.0% and the commit level is plus 24.2%, which is 17.16 million euro.

Plan on the commit level and aim at the stretch level. Cash, hiring and spending rest on the commit level, because missing a commitment costs trust and forces a late cut. People aim at the stretch level, because a target that people reach half the time keeps the effort high. The value 80% is a start value (Plausible but untested). A company with a cheap, reversible plan can use a lower level. A company with a fixed debt payment may need a higher one.

When a target far exceeds the stretch level, ask what it would take. Compute the gap in euro and convert it into units: deals at the usual deal size, then opportunities and leads through the chain of rates. Compare the result with the plan. The funnel must create that pipeline a full cycle ahead, so a late target cannot be met by late effort. The Alder example below does this sum.

The alert band is the prediction interval. A **prediction interval** is the range in which a new value will fall with a stated probability. It is wider than a confidence interval, because it holds the noise of the new value as well as the error in the expected value. An alert fires when an actual leaves its band around the expected value. The expected value is the level times the seasonal index. Do not compare a window with the window before it. That doubles the variance, because both windows are noisy. Compare it with the expectation, which a baseline of many windows estimates.

> **Calculation**
>
> For a count n in a window: relative half-width = $z \times \sqrt{(\phi \times (1 + 1 \div k) \div n)}$
>
> For a rate p over n trials: relative half-width = $z \times \sqrt{(\phi \times (1 - p) \times (1 + 1 \div k) \div (p \times n))}$
>
> z is 1.96 for a 95% band. k is the number of baseline windows. The factor $(1 + 1 \div k)$ is the error of the baseline.
>
> Evidence label: Arithmetic. The level of 95% is a start value; *Decide when a change needs action* sets it from a false-alarm budget (the number of false alarms a year that you accept).

Because n depends on the metric and on the grain, each metric needs its own band for each grain. One fixed percentage cannot serve a count of 3,500 leads, a count of 48 customers and a rate on 150 opportunities. The episode *Birch, Year 1 July: a band that always rings* shows the cost.

Capacity and staffing

The plan needs people. **Capacity** is the work that a team can do in a period, for example the SQLs that a seller can work in a window. Plan from capacity, and use ratios only as a cross-check.

> **Calculation**
>
> heads = required volume ÷ (productivity per head at full ramp × attainment)
>
> Required volume comes from the funnel chain: new ARR needed ÷ deal size ÷ win rate gives opportunities. Opportunities ÷ what one seller can work gives sellers.
>
> Evidence label: Arithmetic for the identity. Plausible but untested for the start values below.

Staffing ratios

A **staffing ratio** is the number of people in one role per unit of volume or ARR. The published ratios below are start values. Replace them with your own ramped productivity from four trailing quarters.

Role and measure	Published reference	Alder	Birch	Cedar
New ARR per account executive a year	Quota 480,000 to 960,000 for mid-market and 700,000 to 2 million for enterprise; at a planned attainment near 55%, that is 264,000 to 528,000 and 385,000 to 1.1 million	665,000	599,000	654,000 (assisted motion)
Account executives per sales development representative	2.4 (Bridge Group, 2025, *SDR Models and Metrics Report*)	none	2.5	none

Role and measure	Published reference	Alder	Birch	Cedar
Accounts per customer success manager	144 under 10,000 of contract value, 49 from 10,000 to 100,000, 22 above (Gainsight, 2022, *What Is The Right CSM to Customer Ratio?*)	59	142	202 (assisted accounts)
ARR per customer success manager	Median 1.4 million, top quartile 4.2 million (Gainsight, 2022)	4.0 million	1.5 million	3.4 million (assisted accounts)

Quota ranges: Winning by Design, 2023, Revenue Architecture, chapter on the go-to-market model (Assertion, no sample). Attainment of 55% is a start value (Plausible but untested; sources differ widely). Company values at Year 1 July.

A gap of more than 30% above a published value is a reason to ask which definition or which choice explains it. Most cells sit near the references. Alder's sellers sit inside the enterprise range, and Birch's sellers are 13% above the top of the mid-market range. Birch's contract value of 9,600 is just under the line of 10,000, and its 142 accounts per manager match the low-touch reference of 144. Two cells pass the 30% gap, and each has a reason. Cedar gives its assisted accounts a thin touch: 202 accounts per manager against 49. Alder's managers carry 4.0 million of ARR, near the top quartile, because its accounts are large. Both are choices of touch level, not errors, but each company should write the choice down.

Ramp time and hiring lead time

Ramp time is the time that a new hire needs to reach full productivity. Until then the hire costs money and adds little capacity. Alder's executives ramp for six months at no output. Birch's ramp for three months at half output, and Cedar's assisted sellers ramp for four months at half output. That loses 6.0, 1.5 and 2.0 seller-months of output.

Revenue arrives later still, because a deal needs its cycle to close. A deal that the hire opens at the end of the ramp closes after the cycle of *Forecast conversion and timing*. So the time from the hire to steady wins is the ramp plus the cycle. Use the median cycle for the typical case and the cycle P80 for a plan.

Company	Ramp, months	Cycle P50, months	Cycle P80, months	Hire to a median win	Hire to a P80 win
Alder	6	3.9	6.3	9.9	12.3
Birch	3	1.1	1.9	4.1	4.9
Cedar (assisted)	4	1.5	2.3	5.5	6.3

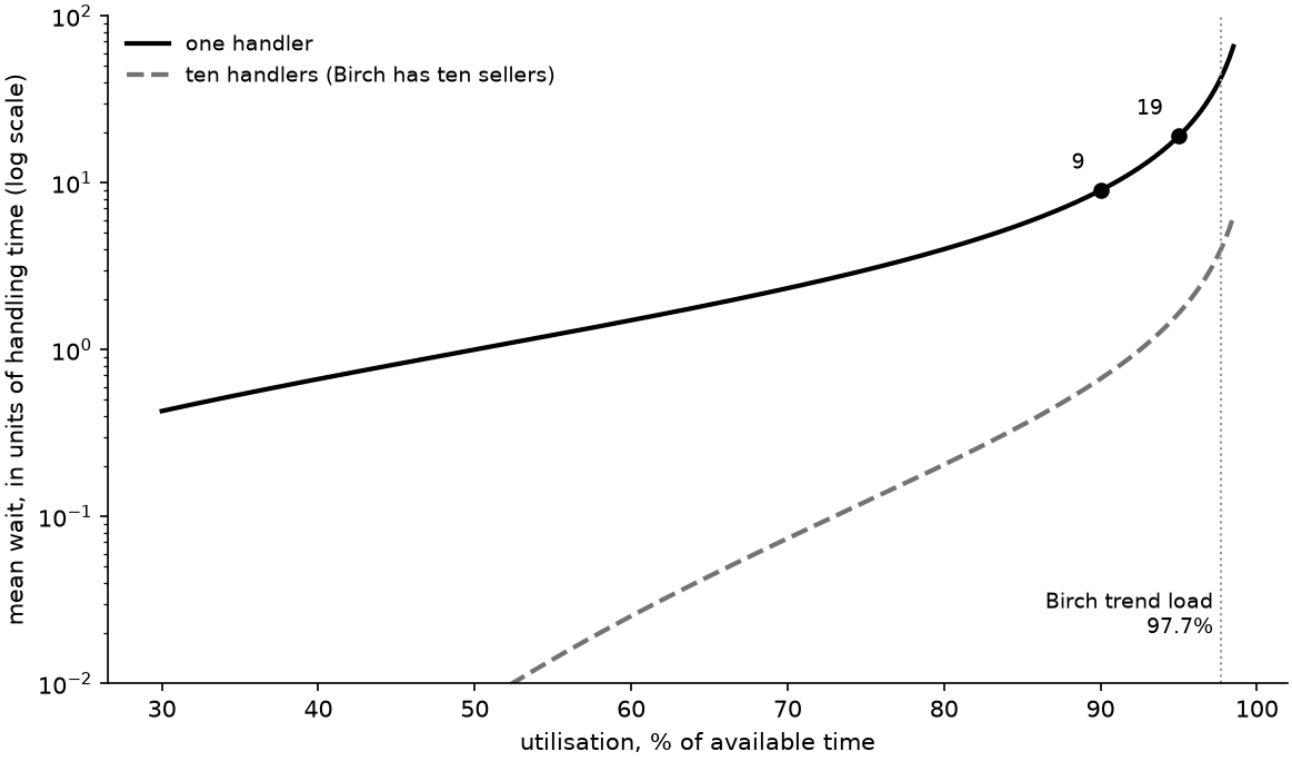
Add the time to find and hire a person, a company parameter. The lesson is that a hire decided today pays back months or a year later. A company that waits until sellers are overloaded has waited too long. Capacity is also a step function. One more seller adds one share of the team: at Birch, 30 SQLs a window, which is 10% of capacity. So hire on a written trigger, set from a leading measure such as load, and set it far enough ahead to cover the ramp. The chapter *Commit with uncertainty* calls such a trigger a signpost.

Utilisation and waiting

Utilisation is the share of available time that people spend on work. We also call it load: demand divided by capacity. Waiting does not grow in a straight line with it. For one handler with random arrivals and random handling times, the mean wait in the queue, measured in handling times, is utilisation ÷ (1 – utilisation).

Utilisation	50%	70%	80%	90%	95%	98%
Mean wait, in handling times	1	2.3	4	9	19	49

A rise from 90% to 95% makes the wait 2.1 times longer. With ten handlers, as at Birch, the wait at the same utilisation is much shorter. But it grows even faster: 2.5 times from 90% to 95%, and 3.3 times from 80% to 90%. This is Standard practice (Bhunia and others, *Advanced Optimization and Operations Research*, chapter on queueing theory). Above 100% there is no steady wait. The queue grows without limit.



Mean wait against utilisation for one handler and for ten. Arithmetic, with random arrivals and random handling times.

Use this only where a queue exists. The **first-touch queue** is the only part of selling that behaves like one: a new SQL waits until a seller makes first contact. After that, a seller works many deals at once, in short bursts. By Little's law, a Birch seller carries about 44 open SQLs and deals at once. If you model whole deals as service times, you get a queue that does not exist. For the later stages, limit the open deals per seller instead, and read the limit from your own data.

The first-touch queue is easy to measure. The backlog is the number of SQLs that wait for first contact. By Little's law of *Forecast conversion and timing*, the mean wait is the backlog divided by the SQLs contacted a day. A backlog of 110 SQLs at a capacity of 300 SQLs per 28 days is a wait of about 10.3 days.

Investor metrics in brief

Investors read three summary numbers. Each one hides something, and none is a lever.

The **growth rate** hides the size of the base and the mix of its parts. Read it beside net new ARR and the bridge of *Model recurring growth*.

The **rule of 40** says that the growth rate in percent plus a profit margin in percent should be at least 40. It treats a point of growth and a point of margin as equal, and companies define the margin in different ways: free cash flow or operating profit. Published medians are 34 for product-led and 20 for sales-led companies (Benchmarkit and Pavilion, 2024). At Birch's Year 1 growth of 31.4%, a score of 40 needs a margin of at least 8.6 points.

The **burn multiple** is net cash burned divided by net new ARR (Sacks, Craft Ventures, 2020, *The Burn Multiple*). The author calls a multiple of 2 reasonable for an early-stage company and a multiple of 5 terrible. It ignores the quality of the new ARR, and it moves with billing terms: a company that bills a year in advance burns less in the same year. At Birch's net new ARR of 6.9 million euro, a multiple of 2 means a net burn of 13.8 million euro.

The example companies give only the cost of revenue and of sales and marketing, so this chapter shows the arithmetic limits and does not score them. Investor metrics steer how money is allocated, not the work of each month. Read them each quarter as constraints, and work on the rate, the time or the cost beneath them.

Sales capacity and productivity

At Birch the limit in the middle of the funnel is the seller. *Set targets and staff capacity* showed that only first contact behaves like a queue. *Allocate sales attention* showed that a lever that adds SQLs adds nothing while sellers are full. Capacity is therefore a lever with a shadow price.

Plan it from capacity, not from a ratio. The heads that you need equal the volume to work divided by the productivity of a ramped seller. Measure productivity on ramped sellers only. A new seller starts after a hire lag and ramps at part capacity. Ratios, such as sales development staff for each seller, are cross-checks. *Set targets and staff capacity* lists them.

Four numbers describe the first-touch queue: the SQL slots that sellers have in a window, the load, the backlog and the wait. *Commit with uncertainty* wrote a hiring rule on the backlog. When the backlog at a month end passes the written level, the sales owner approves one more seller within the week. The sales owner logs the decision and holds no meeting. The rule then pauses until the new seller works at full capacity. *Fund the next best action* put the rule on a trigger sheet.

What goes into the slots matters as much as how many there are. In Year 2 January, with full sellers, 39.5 outbound SQLs that closed at 5.95% took slots from inbound SQLs that closed at 16.0%. The window won 44.0 customers instead of 48.0, as *Diagnose what changed* found.

Report productivity by role, as a pooled number for the whole team, and never by person. In a team of four sellers, a ranking by win rate is noise (see *Decide what the evidence can resolve*). A rule that puts one person's name on a result also breaks the legal rule of this book. Counsel reviews any recording of calls and any use of data about workers before you build it. This is not legal advice.

What changes with volume

The mechanics hold at any volume. The width of the forecast and the way you use it change.

Regime	What the 12-month forecast looks like	What to do
Low (Alder)	Wide: 11.8 points of growth from the 10th to the 90th percentile. A few deals, a few expansions and one large account make the width.	Commit at the commit level. Hire on written triggers, because the ramp and the cycle take a year. Bands from data are wide: a window of 3 customers reads almost nothing
Medium (Birch)	Narrow for ARR: 3.3 points. Wide for single windows: new customers carry a band of ±29% in one window and ±18% in a quarter.	Forecast ARR from cohorts. Set a band for each metric and grain. Watch load, because capacity binds
High (Cedar)	Very narrow for 12 months: 2.0 points. The doubt moves from noise to the future itself.	Short-term forecasts are tight. The ceiling and the market stay judgements. The assisted motion is low volume and needs the low-volume method

Alder, Year 1 July: plus 40% and two more sellers

Alder, Year 1 July: about 3.2 new customers a window, 0.8 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.1 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

The board meets in July. ARR is 13.82 million euro, up 15.2% since January. The board wants plus 40% over the next 12 months and two more account executives. The chief executive asks the analyst for a forecast to test both.

The analyst builds the cohort forecast from the book of about 205 customers. It lists the renewals by number, the pipeline and the cycle, and the single largest account. That account holds 1.32 million euro, which is 9.6% of ARR, and it renews in January. The plan expects 44 new customers in 12 months. The simulation gives a median ARR of 17.69 million euro in July of the next year, which is plus 28.0%. The 10th to 90th percentile runs from plus 22.1% to plus 33.9%, a width of 11.8 points. The width is the message. It comes from the count of new deals, the lumpy expansion and the one large renewal, as the table of sources showed.

Now the target. The chance of plus 40% is 0.6%. The range is a floor on the real width, so the true chance is somewhat higher, but it stays small. The commit level is plus 24.2%, or 17.16 million euro, and the stretch level is plus 28.0%. To close the gap of 1.66 million euro to the target, Alder would need about 28 more deals of 60,000 euro, so 72 new customers against 44 planned. That is 1.6 times the plan.

Next, the two sellers. Alder's four executives can work 22 opportunities a window. The plan opens 14.7 in July, which is a load of 67%. In the busiest planned month the load reaches 80%, and a single window exceeds capacity in at most 12% of cases. With two more executives, the peak load falls to 53%. So sellers do not limit Alder. Opportunities do. The two hires also come too late for this year. They ramp for six months at no output, and then a deal needs a median of 3.9 months, or 6.3 months at P80. Steady wins from a new hire start at about 9.9 months, or 12.3 months with the cycle P80. So a hire made now adds only about two months of wins inside the target year. A plan on the cycle P80 gets none. The two executives cost 370,000 euro a year. Half of that, 185,000 euro, buys no output during the ramp.

The decision is written into the forecast note, since Alder has no decision log yet. The board plans cash and hiring on the commit level of plus 24.2% and keeps plus 40% as an aim with no budget attached. It does not hire now. A written trigger replaces the request: hire one executive when the load of a quarter passes 85% in the forecast six months ahead, which is the ramp. This is a start value (Plausible but untested). The planned peak is 80%, so the trigger has not fired. The analyst rebuilds the forecast each quarter. The lesson at low volume is that the width of the interval is the message, so commit at the commit level and hire on triggers.

Birch, Year 1 July: a band that always rings

Birch, Year 1 July: about 48 new customers a window, 1.0 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 71 days, and the close rate reads a 20% change in about 6 months.

Birch's board paper carries a band of plus or minus 15% on 11 stage counts and rates. Each is compared with the window before. The band rings in nearly every quarter, and nobody trusts it. The analytics lead asks how often it would ring with no change at all. For one comparison of two windows, a count of 48 new customers breaks the band 46% of the time by chance alone. The win rate breaks it 37% of the time and the count of opportunities 19%. Leads break it almost never. Over the three comparisons of a quarter and all 11 metrics, the chance of at least one alarm is 99%.

The season adds to it. The plan itself moves by 15% or more between months four times a year. The moves are plus 23.5% in January, minus 15.0% in August, plus 23.5% in September and minus 19.0% in December. No seller did anything differently in those months.

The band is also too wide for the metrics that matter most. The table sets a 95% band from data against the fixed band. It uses the formula of the section on targets, a baseline of 13 windows and ϕ at 1.

Metric	Count in a window	Fixed band: false-alarm chance, one comparison	95% band, one window	95% band, one quarter
Leads	3,517	0.0%	±3.4%	±2.0%
SQLs	352	4.7%	±10.8%	±6.5%
Opportunities	150	19.4%	±16.6%	±9.9%
New customers	48	46.2%	±29.4%	±17.5%
MQL to SQL rate	1,407 trials	2.2%	±9.4%	±5.6%
Win rate	150 trials	37.3%	±24.2%	±14.5%

Relative half-widths. Quarter: 3.25 windows against a baseline of four quarters. Arithmetic.

A fixed band of 15% is far too wide for leads: a fall of 10% in leads would never ring, although the data show it. It is far too tight for new customers and for the win rate in one window. If the real dispersion were 2, the SQL band would widen to ±15.3%. The analytics lead proposes in writing to the growth owner: one band per metric and per grain, around the seasonal expectation. New customers are read by the quarter. The owner keeps the board paper for now and asks for a replay of the alert rules on a full year first. That rebuild, and the decision on it, comes in *Decide when a change needs action*.

The analytics lead also reads the three tests as forecasts. The chance that the next 12 months breach the payback ceiling of 18 months is 62%. For scalability, the media plan's assumed curve says that an extra 20% of paid media would add about 21 SQLs a window. Sellers are at 98% of capacity on the trend. So with the same sellers the spend gains only 1.1 customers a window, at 30,100 euro each. With one more seller, it gains 3.4 customers a window at 12,900 euro each, which is a payback of 20.0 months against the ceiling of 18. The ceiling allows a cost of 11,585 euro, so the marginal customer is 11% over it. Durability holds: a customer is worth 2.7 times its cost. The July backlog of 110 SQLs, up from 54 in June, waits about ten days. The head of marketing holds paid spend where it is until the capacity question is priced in *Allocate sales attention*. The head of sales owns the load.

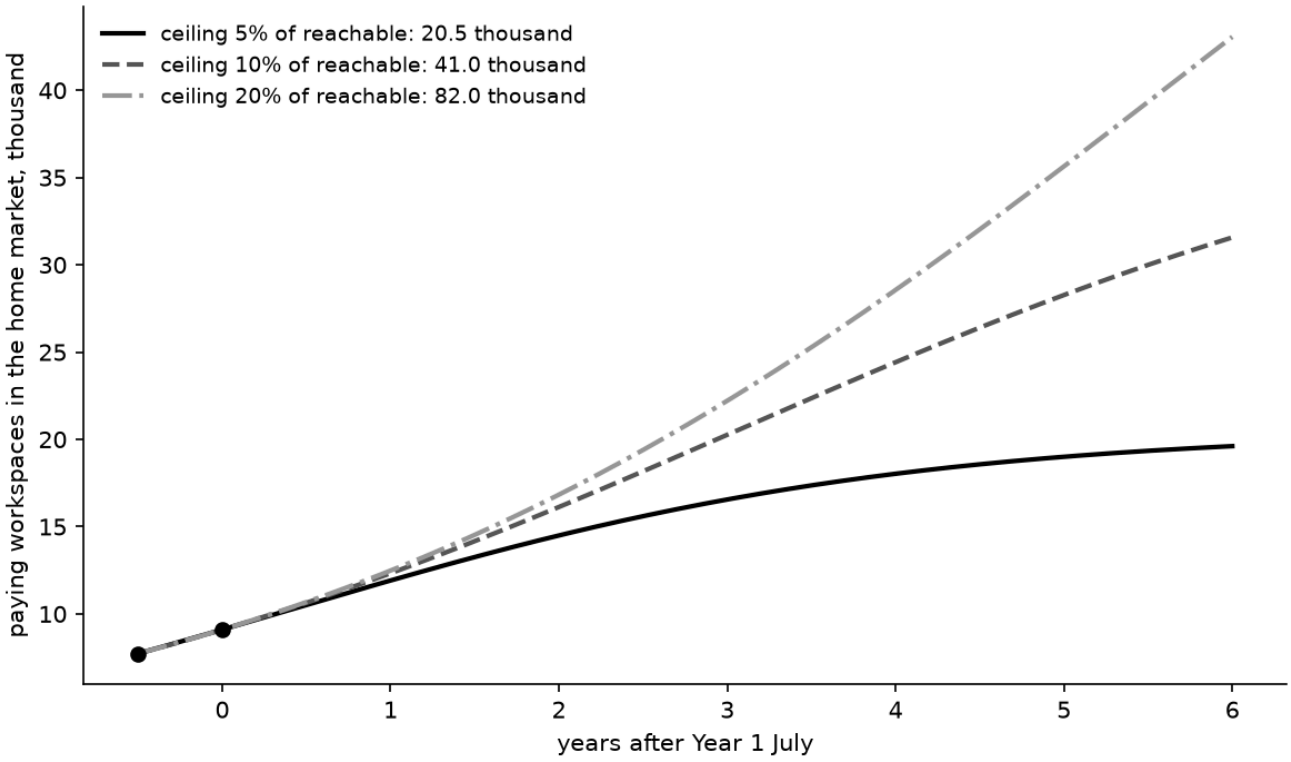
Cedar, Year 1 July: slow growth at home

Cedar, Year 1 July: about 1,200 new paying accounts a window, 1.5 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 29 days; the assisted motion, with 53 opportunities a window, needs about 8 months to read a 30% change.

The growth lead sees that seat growth in Germany is slowing. Germany brings 30% of sign-ups. The first question is the season. July has an index of 0.85 and June 0.95, so sign-ups fall 10.5% from June to July with no change in the market. The lead compares July with the same month of the year before, and not with June. Once the season is out, sign-ups still follow the plan's trend of 15% a year. The real worry is the coming years, and that is a question about the ceiling.

The second question is the ceiling. Germany has 410,000 small businesses that Cedar can reach. It has 9,100 paying workspaces, a penetration of 2.2%. Nobody knows the share of those businesses that could ever pay, so the lead tries three values. The January figure of the home market assumes that it keeps its Year 1 July share of Cedar's paying workspaces, 29.0%, which gives 7,730. That is Illustration only.

Attainable share of reachable businesses	5%	10%	20%
Ceiling, paying workspaces	20,500	41,000	82,000
Share of the ceiling reached now	44%	22%	11%
Paying workspaces after 1 year	11,916	12,329	12,483
After 2 years	14,495	16,125	16,832
After 3 years	16,556	20,264	22,214
After 6 years	19,612	31,570	43,062



One S-curve through the same two points of the home market with three ceilings. Illustration only.
The two points fit all three ceilings. After two years the projections are 14,500, 16,100 and 16,800, within 16% of each other. After six years they are 19,600 and 43,100. So the data cannot tell a ceiling from a slowdown for several years. Only a ceiling from market sizing can, and it is a judgement. A future slowdown would be a market limit only if the attainable share were near 5%. At 20% the market is far from its bend. Then the cause of any slowdown would lie elsewhere: in the funnel, the price or a competitor.

The lead also checks how much the question matters for the company. The 12-month forecast of ARR is tight: plus 30.7% to plus 32.7% from the 10th to the 90th percentile, a width of 2.0 points. A stalled home market, with sign-up growth at 10.5% instead of 15%, lowers the 12-month median by 1.2 points of opening ARR. That is inside the width of the fan. After 36 months the gap is 7.6 points against a fan of 4.1. High volume gives a tight forecast for a year, but the ceiling decides the later years.

The lead decides to set a market study as the next step. It will fix a ceiling with a range, and it will compare Germany with the countries of similar penetration among Cedar's 24 markets. Until the study is done, the plan keeps its growth of 15% and the lead reads the home market against the season each month.

The decision to carry forward

Commit capacity early enough to cover recruitment and ramp. Attach a probability to the target and a dated signpost to the commitment. Revisit the plan when its customer mix, demand or service assumptions change.

12 · Activate customers

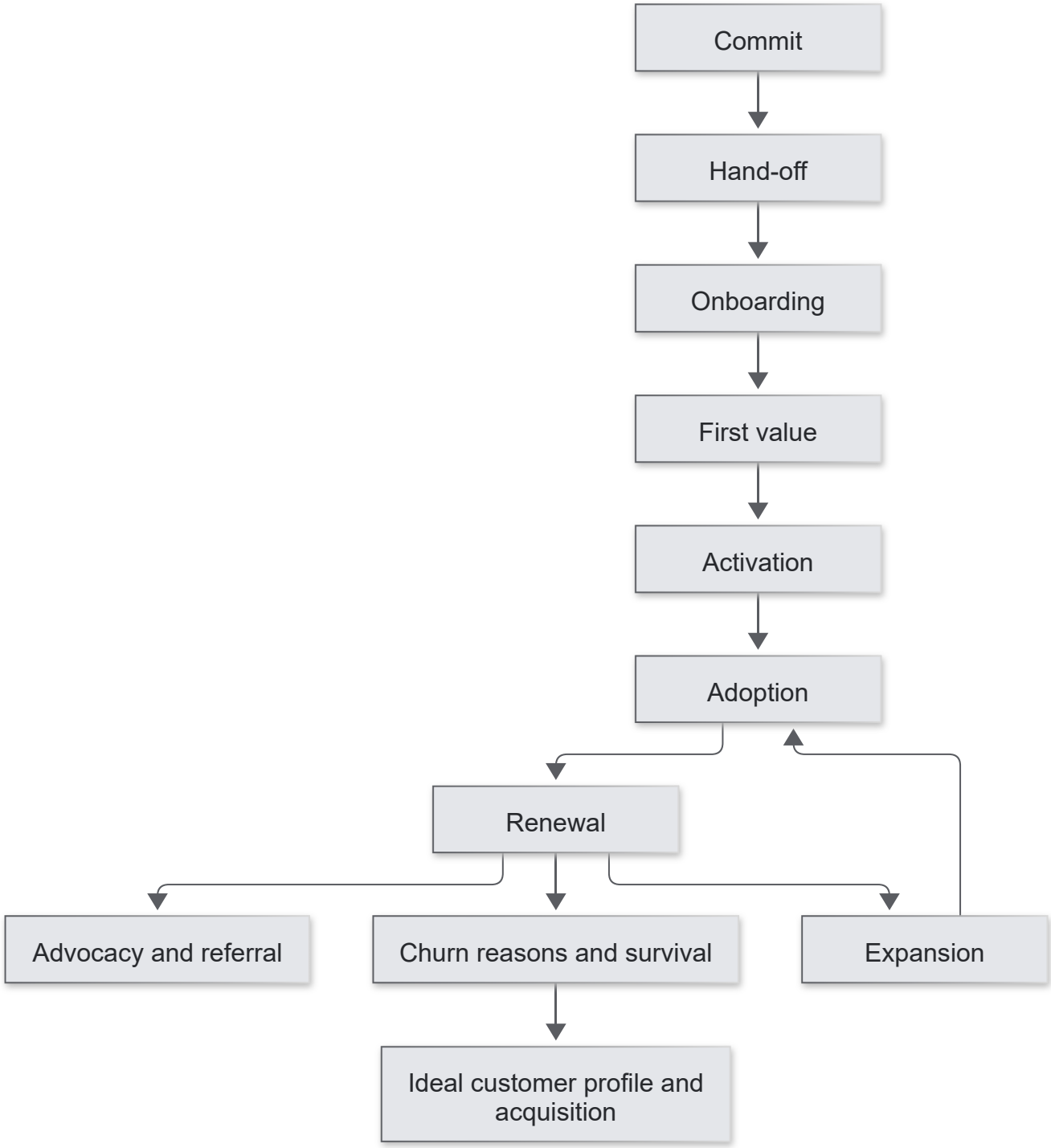
Which change will help more eligible customers obtain their first useful result? Activation is valuable because the product starts doing a job for a customer. An event is useful only if it represents that result well enough to support a decision.

Choose the time origin and the eligible population before selecting an event or a deadline. A customer who leaves before the event has a different outcome from one who remains eligible when reporting ends. Mixing those states can make a failing onboarding path look successful.

This chapter connects activation definition, time-to-event measurement and an onboarding intervention. A relationship between early activation and renewal can support a forecast. To justify changing onboarding, you need evidence about what that change causes. Keep the later renewal outcome on the decision record while using an earlier activation result as provisional evidence.

The impact journey

The right wing is a path that each customer walks. We call it the *impact journey*: hand-off, onboarding, time to first value, activation, adoption, renewal and expansion. Each step has an owner, a measure and a value in euro. The journey ends in a renewal decision, and it can end earlier if the customer leaves in the first months.



The impact journey from commit to expansion. The last arrows are the second loop: what customers do after the close changes who we win.

The hand-off is the moment when the sales owner passes a signed customer to the customer success owner. A good hand-off is a written record. It states what the buyer expects to see and by when. It names the champion and the dated event in the buyer's business that sets the pace (see *Remove buying friction* for the critical event and the champion). If the customer success owner reads only the contract, the promise made in the sale is lost.

Onboarding is the work between the signature and a working set-up. Data is connected, users are invited and the product is configured. **Time to first value** is the time from the commit until the customer first sees a result. **Activation** is the event that marks that result in our data. As *Choose the growth objective* defined it, it

is one event that the customer can see, for example the first report built from the customer's own data. For many products first value and activation are the same moment. The *time to activation* is the time from the signature, or from the sign-up, until that event.

Two measures serve this stage. The *activation rate* is the share of new customers that reach the activation event inside the window. The median time to activation says how fast they get there. The customer success owner reads both for each start month, and reads a cohort only at its own age (see *Forecast conversion and timing*). A fall in the rate for one start month is a cohort effect, for example a set-up fault in that month. Do not read it as a fall in the quality of the product.

Set the time window of activation from data, not from a rule of thumb. There is no universal value. Published figures come from vendor reports, and they measure different things. One report gives a median of about a day and a half from account creation to a first value event, in self-serve products (Userpilot, 2025, *Benchmark Report*). That is a product action, not a B2B implementation. Use start values until you have data: 14 days for self-serve, 30 days for mid-market and 60 to 90 days for enterprise (Plausible but untested). Then follow five steps.

1. Define the activation event as the first result the customer can see, and agree it at the hand-off.
2. Develop candidate definitions on earlier cohorts. Check their predictive relationship with renewal at a fixed assessment age. Validate the chosen definition on later cohorts, without using future events as inputs.
3. Measure activation and departure from the same time origin. If departure prevents activation, it is a competing outcome. Administrative censoring means observation ends while the customer is still eligible and neither outcome has occurred. Use the competing-outcome method in *Forecast conversion and timing*.
4. Choose a candidate window on development cohorts, then fix it before evaluating later cohorts. A window at the 75th to 80th percentile among eventual activators is a possible start value (Plausible but untested). It describes activators, not the activation probability of all entrants.
5. Do it for each segment, and review it every two quarters.

Cedar shows the method on its model of sign-ups. Among sign-ups that activate, the median time is 2 days, and the spread of the logarithm of the days is 0.9. The 80th percentile is 4.3 days. The 14-day trial holds 98.5% of the activations (Illustration only). The trial length is a product rule. The 4.3 days is a reading of the data, and it is the better window for the early warning.

Be careful with the link to renewal. Customers who activate fast renew more. In Birch's model the activated customers renew at 90% at the first renewal, and the others at 72% (Illustration only; see *Model recurring growth* for the read of this split). That is an association. Compare activation status measured at a fixed assessment age and renewal over a defined later period. Keep eligibility independent of later activation or renewal. Customers who fit the product better may both activate faster and stay longer. Only an experiment shows that speed causes retention, for example a faster onboarding path against the old one (see *Choose whether to test or act*).

Choose the observation before optimising the event

Set a fixed assessment age at which the activation result would inform an action. Define eligibility using information available by that age. Do not classify an account as an early success because it eventually activates after that assessment. That would use future information to make a past decision.

Build candidate activation events from a customer result that the product can record. Then compare their usefulness on development cohorts. Separate the later cohorts used for validation. Choosing both the event and the most flattering deadline on the same renewal outcomes can overstate how well activation predicts retention.

There are two questions. Does the event help predict later retained value? Does an intervention that changes the event improve that value? Customers who fit the product can both activate sooner and stay longer. An association supports the first question under its validation conditions. It does not answer the second.

Estimate activation before departure

Use a common time origin, such as eligible account creation or the contractual start of onboarding. Record the first activation event, departure and the reporting cut-off. The practical target is the probability of activating by an age before leaving that journey.

Departure that prevents activation is a competing outcome. Administrative censoring occurs when reporting ends while the customer remains eligible and neither event has occurred. It limits the observed follow-up; it is not another customer outcome. Explain any other loss to follow-up and whether it could depend on the likelihood of activation.

> **Calculation**
>
> activation probability by an age = sum over earlier intervals of P(still eligible without either outcome at the interval start) × P(activation in the interval given still eligible)
>
> Evidence label: Arithmetic for mutually exclusive first outcomes on the defined journey. Estimation from incomplete follow-up requires assumptions about observation and censoring.

This is the same competing-outcome construction used for won and lost opportunities in *Forecast conversion and timing*. A customer who has already left cannot remain in the pool that can still activate. Do not estimate the practical probability by treating those departures as ordinary independent censoring.

If customers can return, define the policy you are evaluating. A return can start a new journey, or the model can allow transitions between inactive and eligible states. Use the same definition in the denominator and the intervention. The simpler first-journey calculation answers activation before first departure; it does not answer every possible later return.

Activation-time percentiles among eventual activators describe that selected group. They do not estimate the share of all entrants who will activate. A faster median among activators can coexist with more departures. Read the probability of activation and the time distribution together.

Decide whether to change onboarding

Specify the proposed change as an action, such as a revised setup sequence, clearer invitation or additional implementation help. Compare it with the existing path among eligible accounts. Measure retained contribution after its incremental delivery costs. An intervention can increase activation while consuming enough service capacity to reduce its economic value.

An account-level comparison keeps people from the same buying organisation on a consistent path. Choose a primary outcome and keep departures visible as an outcome. If activation is the early read, record when renewal evidence will mature. A positive early result changes the belief about the intervention; it does not remove uncertainty about later profit.

When volume is too low for a useful randomised read, inspect individual journey failures and their proposed mechanisms. A setup audit can establish that a required integration fails. It cannot establish a precise uplift in renewal. Price the repair with a range and name the later evidence that could change the decision. This is still a quantitative growth decision, because the mechanism, exposure, cost and uncertainty stay connected.

Related company decisions appear in *Retain customers profitably*, *Forecast conversion and timing*. Their original dates remain explicit.

The decision to carry forward

Define a customer result, an eligible population and a time origin. Estimate activation before departure, validate the window on later cohorts and judge the onboarding intervention against an alternative. Keep later renewal evidence separate from the early signal.

13 · Retain customers profitably

Which retention action is worth taking, and for which customers? A high churn risk does not imply a high return from intervention. Some customers would stay without help. Some cannot be retained by the action available. The decision concerns the customers whose outcome can change, the value of that change and its cost.

Use health predictions to identify and understand risk. Use treatment comparisons to estimate what a save action changes. Keep those tasks distinct while evaluating both on the same discounted gross-profit basis. A discount paid to an account that would stay anyway is a cost of the policy.

The mechanisms below connect adoption, renewal and save offers. Follow accounts beyond the end of an offer to distinguish a departure prevented from one deferred. When individual treatment effects cannot be estimated credibly, a pooled or carefully defined group decision can still be useful.

Adoption and health

Adoption is the stage where the product becomes part of the customer's routine. We measure it as *adoption depth*: the share of the intended users, roles or sites that use the product in a period. A second part counts how many of the features that carry the value they use. Read it every week for each account. It is a leading measure, because it moves months before the renewal does (see *Decide what to measure* for leading and lagging measures).

A **health score** combines usage, support, sentiment and commercial signals into one number per account. It tries to predict the renewal. It is a model, and we judge it like any model. Four checks matter.

- **Later renewals.** Train it on older cohorts. Test it on later ones, in time order. A score that predicts the past only is useless.
- **Calibration.** Accounts with a predicted chance of 70% should renew about 70% of the time. Check it in score bands, as for a lead score (see *Allocate sales attention*).
- **A simple rule.** Compare the score with a plain rule, such as "no use in the last period". If the rule does as well, use the rule.
- **Lead time.** Measure the days between the warning and the renewal date. A warning that comes after the buyer decided is only a report.

Do not size the model by "10 events per variable". The evidence for that rule is weak, and a penalised fit changes what a small sample can support (van Smeden and others, 2016, *No rationale for 1 variable per 10 events criterion for binary logistic regression analysis*, BMC Medical Research Methodology 16, 163). Judge a model by its out-of-time performance. The number of losses still sets what you can learn. In the year to Year 2 December Alder loses 13.4 customers, Birch 250.7 and Cedar 8,872 (Arithmetic, from the facts). Thirteen losses a year are too few to train or test a model, so Alder uses a checklist and cases. Birch can test a score with a few inputs. Cedar can train one and recalibrate it every month.

The score tells you who is at risk. It does not tell you whom to help. A customer at high risk may leave whatever you do. The section on save plays returns to this.

Renewal

Renewal is the main event of the right wing. *Model recurring growth* gave its mechanics: the renewal calendar, the renewal rate and survival by renewal number. *Forecast conversion and timing* showed how to read rates that mature. Here we add the question of volume. How many renewals do you have, and what can they tell you?

Read survival by segment only when each segment has enough renewals. As a start value, ask for a half-width of 5 points or less (Plausible but untested; see *Model recurring growth*). The table shows what the year to Year 2 December allows. Cedar is not in it: its monthly book renews every month, so it reads cancellations, as its example shows.

Table 1. Renewals in the 12 months to Year 2 December, and the half-width of the renewal rate (95%).

	Renewals due	Renewal rate	Half-width, all renewals	Half-width, one tenth of them
Alder	177	92.4%	3.9 points	12.3 points
Birch	2,665	90.6%	1.1 points	3.5 points

Birch can read a segment of a tenth of its renewals to within 3.5 points. So its segment curves become readable when the cohorts are old enough. Alder cannot. Its pooled rate is readable to about 4 points, and any segment is noise. Segments must be known at the time of the renewal, for example the channel, the plan or the activation flag. Small segments need partial pooling (see *Decide what the evidence can resolve*).

The first renewal deserves its own attention. The hazard is highest there, as the survival tables of *Model recurring growth* show, and the customer has seen the least. A playbook for the first renewal is priced like any other lever (see *Find the most valuable constraint*). The renewal calendar sets the rhythm. Customer success reads it every week. The renewal project of an account starts when the buyer's own budget cycle starts, which can be well before the contract date.

At low volume, run each large renewal as a project. A renewal project has a named owner, a date, and a plan that is written down months before the renewal date. The plan maps the buying group at the account and names the champion and a second contact. It states the result that the customer has seen. It sets the quote and the steps to signature. The customer success owner writes the belief in the decision log before the outcome. The renewal date is the date on which the belief holds or fails (see *Commit with uncertainty*). A miss then becomes a case that teaches, not a surprise in the year-end figure.

Save plays

A **save play** is an action to keep an account that is at risk, such as an offer, a call or a workshop. Its value follows from the value basis. A play changes the chance that the account stays. If the account stays, we keep its expected gross profit. The play also has a cost.

> **Calculation**
>
> value of a play = change in the chance of staying × gross profit kept − cost of the play
>
> break-even change = cost of the play ÷ gross profit kept

> Gross profit kept is the expected discounted gross profit of the account, on the value basis of *Value customers and cash*. Evidence label: Arithmetic.

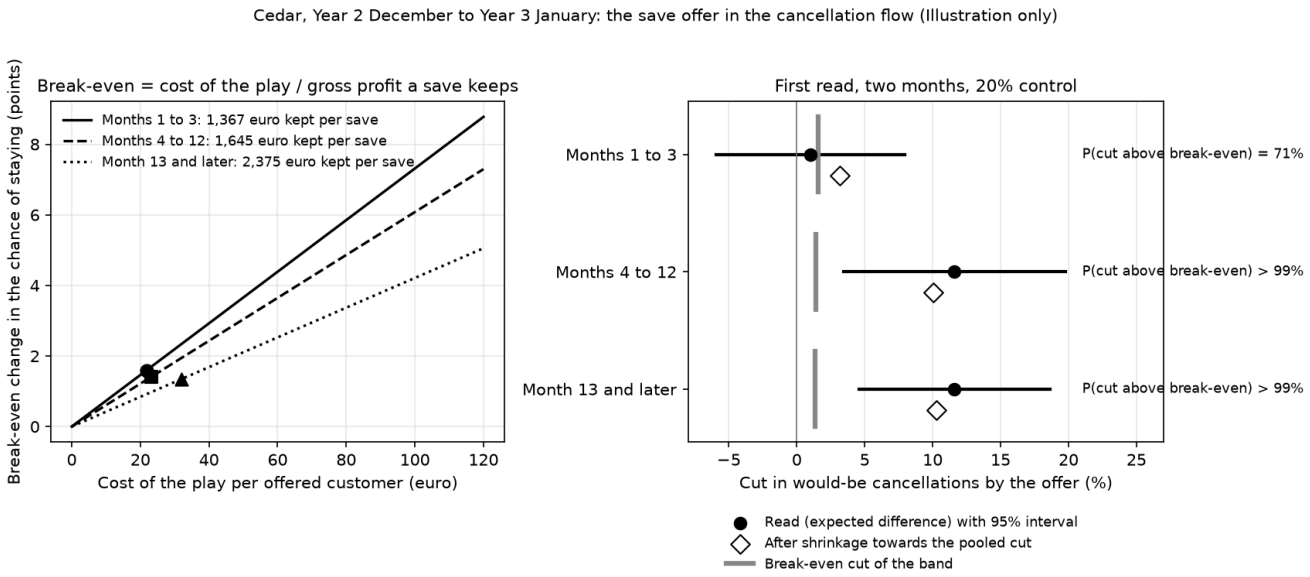
The *incremental save rate* is the change that the play causes: the share of treated at-risk accounts that stay, minus the share of control accounts that stay. The break-even change is the most useful number. It says how much the play must change behaviour before it pays. Compare it with what your volume can read. If the break-even is far below the smallest change that you can read, you decide on a belief and not on a result (see *Commit with uncertainty*).

Do not rank at-risk accounts by risk. Rank them by what the play can change. Think of three kinds of customer. Some would stay anyway, and the offer is wasted on them. Some would leave anyway, and the offer is wasted again. A few are *persuadable*: they leave without the play and stay with it. The *uplift* of a play is the difference in outcome between a treated group and a control group. It is the change that the play causes. Provost and Fawcett (*Data Science for Business*) make the same point: target by the change in the chance of staying times the value, not by the chance of leaving. Rodrigues (*Product Analytics*, the chapter on uplift modelling) shows how to estimate uplift for individuals. That needs large treated and control groups.

At most volumes you cannot learn uplift for each person. Estimate it for three or four strata instead, for example bands of tenure. Print the n and the interval of each stratum. Shrink the strata towards the pooled change, because a small stratum shows noise (see *Decide what the evidence can resolve*). Shrink only strata that are alike. A stratum that differs for a known reason is not exchangeable with the others.

To read the change, you need a control: a random share of the at-risk accounts that does not get the play. A standing control is slow and has a price. It withholds the play from customers that you could save. So time-box it. Choose its size and length by the value of information (see *Choose whether to test or act*). Stop it when more reading is worth less than it costs. Ship the play to a stratum when the chance that its change exceeds the break-even is at least 90%. A save play is easy to undo, so the lower threshold applies (see *Commit with uncertainty*). Ask the value question for each stratum apart. A stratum that fails the threshold can still be worth more reading, when it is large and its read is wide.

Two cautions apply. First, a discount can delay a departure and not prevent it. Read the staying of treated and control customers after the offer has ended, and not only during it. Second, make cancelling easy and honest. An offer must not hide the way out, and counsel reviews the flow where consumer rules apply. This is not legal advice.



Left: the break-even change in the chance of staying for the three tenure bands of Cedar's monthly plans, against the cost of the play; dots mark the cost of the offer. Right: the first read of the control, by band, with the cut after shrinkage and the break-even cut (Illustration only; the example below explains it).

The after-the-close page

One page serves the decisions of the right wing. Each tile names the decision that it supports.

Table 4. A simple layout of the page. The customer success owner reads it weekly. The finance owner and the analytics lead read it every quarter and send their questions to the owner in writing.

Tile	Reads	Compared with
Renewal calendar	Contracts due in the next 90 days, owner, stage of the renewal project	Contracts due and renewed last quarter
Survival by renewal number	Share still customers, with its interval, pooled and by readable segment	The previous quarter's curve
Time to activation	Share activated before departure inside the window, by segment, with departures and administrative censoring shown separately	The same cohort age last quarter
Adoption depth	Accounts below the depth that precedes loss	The accounts' own depth, three months ago
Health score calibration	Predicted against realised renewal, in bands	The previous calibration
Save plays	Break-even change, read change with interval, stratum, value	The break-even
Expansion and price	Whitespace, expansion rate, cost of expansion, price-rise break-even and holdback read	The ceiling of 0.50
Capacity	Renewals and plays for each manager, cost to serve	Capacity of the team

What changes with volume

The mechanics hold at any volume. What changes is how much you can read.

Regime	What the volume allows	Where it holds
Low (Alder: 177 renewals and 13 lost customers a year)	Each renewal is a case. Use renewal projects, named plans and beliefs in the log. No health model: nothing to learn from. A tenth of the renewals reads to 12.3 points, so no segment curve. Value a save plan as a case.	About 14 renewals a window
Medium (Birch: 2,665 renewals and 251 lost customers a year)	The pooled curve reads to 1.1 points, and a tenth of the renewals to 3.5 points. A health score with few inputs can be tested out of time. The price-rise break-even is computable, but a holdback of 15% reads a fall of 5.0 points only after a year.	About 205 renewals a window
High (Cedar: 8,872 lost customers a year)	Strata read inside a window or two. Uplift by stratum is readable, so plays go to customers they can change. A time-boxed control is sized by the value of information, stratum by stratum.	Hundreds of cancellations a month in each tenure band

Alder, Year 2 December: the champion leaves

Alder, Year 2 December: about 3.1 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 2.8 years to read even a 30% change, and the lead count reads a 20% change in 23 days.

The champion at Alder's largest account leaves in Year 2 December. The account holds 1,455,300 euro of ARR, which is 7.3% of Alder's ARR at Year 3 January (19.9 million euro). Its contract renews in January. The head of customer success writes a renewal project and a save plan. The plan costs 60,000 euro for executive time, customer success time and an on-site workshop.

Alder has no rate to read here. It has one account and three outcomes. The head of customer success writes a belief for each outcome, without the plan and with it (Illustration only).

Table 5. Alder, Year 2 December: the owners' belief about the largest account's renewal.

Outcome	Without the plan	With the plan
Leaves	40%	15%
Renews at half scope	40%	50%
Renews in full	20%	35%
Expected ARR kept, euro	582,120	873,180

The plan lifts the expected ARR kept by 291,060 euro. At a gross margin of 79.9% that is 232,678 euro of gross profit in the first year. After the cost of 60,000 euro the net value is 172,678 euro. We count the first year only. The kept ARR also renews in later years, so the full value on the value basis is larger, and the case only gets stronger.

The switching value is more useful than the mean. The plan pays if it delivers at least 75,055 euro of that lift, which is 25.8% of what the owners believe. A plan that works a quarter as well as hoped still breaks even. The cost is 18.9% of Alder's risk limit of 317,941 euro, which is 2% of its yearly gross profit (Plausible but untested). So the plan does not need approval from above. The head of customer success decides by a stated date and writes the entry in the log. The chief executive reads it. No meeting is held.

In Year 3 January the account renews at half scope. The logo is kept, and ARR falls by 727,650 euro. Logo retention does not move. GRR falls by 3.7 points. One account moves the revenue story and not the customer story, as *Model recurring growth* showed.

Can Alder say that the plan worked? Not from one case. The half scope has a probability of 50% with the plan and 40% without it. It is 1.25 times as likely with the plan, so it is weak evidence. The log entry of *Model recurring growth* said that all ten largest accounts renew at full scope. This renewal at half scope refutes it for one account. The process was still sound: the belief was written before the outcome, the cost was within the limit, and the renewal date was set in advance. Alder still writes the case down. A dozen such cases show patterns that one case cannot, but they are still cases and not a rate.

The contraction becomes whitespace. The sites that the account gave up are now licences that the account could take again. Across all customers, 630 plants had no licence at Year 1 August, the last count that Alder has (see *Find the most valuable constraint*). That is 11.0 million euro of ARR at the new site price and 13.2 million at the old one. The head of customer success makes the half-scope account the first target for a second champion and then for expansion. Since Year 1 February each of the ten largest accounts has a named renewal plan (see *Model recurring growth*). The plan now adds one more line to each: a second contact, named before the champion can leave. It extends the champion presence of *Set prices and packages* to the renewal. The lesson at low volume is that each renewal is a case and not a rate. Price the plan in euro, write the belief first, and keep the log.

Cedar, Year 2 December: save offers in the cancellation flow

Cedar, Year 2 December: about 1,800 new paying accounts a window, 1.3 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 23 days; the assisted motion, with 68 opportunities a window, needs about 7 months to read a 30% change.

About 827 customers on monthly plans enter the cancellation flow each month. The growth lead and the product owner of the flow test an offer: two months at half price. The offer is accepted by 30% of the customers who see it (Illustration only). A random 20% see no offer, for two months, as a control. Customers in three tenure bands differ much in value and in what the offer does, as the table shows. We call the change in the chance of staying the *cut in cancellations*: the share of would-be cancellers that the offer turns into stayers.

Table 6. Cedar, Year 2 December to Year 3 January: three tenure bands of monthly plans (Illustration only for the true cuts and for the read).

	Months 1 to 3	Months 4 to 12	Month 13 and later
Share of cancellations	34.0%	28.1%	37.9%
Gross profit a save keeps, euro	1,367	1,645	2,375
Cost of the offer per offered customer, euro	21.8	23.3	32.1

	Months 1 to 3	Months 4 to 12	Month 13 and later
Break-even cut in cancellations	1.6%	1.4%	1.4%
Cut in the read, with 95% interval	1.0% (−5.9 to 7.9)	11.6% (3.4 to 19.8)	11.6% (4.6 to 18.6)
Reliability	0.67	0.60	0.67
Cut after shrinkage	3.2%	10.0%	10.3%
Chance that the cut is above its break-even	71%	above 99%	above 99%

The first band has the highest hazard. A customer in the first three months is the one most at risk, as *Model recurring growth* showed. It is also the one the offer changes least. A ranking by risk would send the offer there first. A ranking by value of the play does the opposite.

The read in the table is the expected result under the facts, not a random draw. Over two months the control holds 112, 93 and 125 would-be cancellers in the three bands. The offered group holds 450, 372 and 502. A tenth of the customers who enter the flow stay anyway, so the control loses 90% of them (Illustration only). The two later bands read, at 2.8 and 3.2 standard errors from zero. The first band does not read, at 0.3. So we shrink the bands towards the pooled cut. The spread between the bands, estimated from the bands themselves, is 4.6 points of lost share. With only three bands this estimate is weak, so read the shrunk numbers as a guide. The first band moves from 1.0% to 3.2%.

The growth lead decides by the 90% rule for a change that is easy to undo. The two later bands (months 4 to 12, and month 13 and later) pass, with chances above 99%. The first band (tenure months 1 to 3) fails at 71%. The first band may include customers who have not reached first value. Tenure alone does not establish that mechanism; check activation status and journey evidence. The idea that the offer buys time without changing delivered value is Plausible but untested. Use the shrunk figure of 3.2% and the unpooled result as sensitivity checks rather than claiming that tenure proves nonexchangeability. On its own read the chance is 43%. Either way the band stays below the threshold, so it gets no rollout. From Year 3 February the offer goes to tenure of four months and more. The cut of the targeted offer is about 7.7%.

The value of information decides each band's control apart. In the two later bands, two more months of control are worth at most 170 and 82 euro (the expected value of perfect information). One more test costs 4,318 euro of analyst time (Illustration only). So the time box ends there. The first band is different. It is large and its read is wide. Two more months of reading, with a fifth of the band offered, are worth about 7,900 euro on the shrunk belief (the expected value of sample information, see *Choose whether to test or act*). That is more than the test costs. On the band's own read, which is wider, the value is larger still. So the growth lead keeps a small offered arm in the first band for two more months. Longer arms are worth more, but they also delay the decision, which this sum leaves out.

The staying of offered customers is read again 90 days after the discount ends, to see whether the offer saved them or only delayed them. The model of the facts lets us check the call, and the owners cannot do that in January. The true cut of the first band is 1.0%, below its break-even, and the offer would lose 2,279 euro a month there. The small arm costs a fifth of that while it runs, which is the price of the reading. Offered to all three bands it is worth 113,014 euro a month. Offered to the two later bands it is worth 115,292 euro (Illustration only). The lesson at high volume is that uplift by stratum is readable, so the offer goes to the customers the play can change. The value of information then sets the control of each stratum on its own.

The decision to carry forward

Fund the retention action for the customers it can help. Include wasted offers, service costs and the period after the offer ends. A risk score identifies a problem; incremental value decides whether the available response pays.

14 · Expand accounts and generate referrals

Should the next unit of growth come from expansion, referrals or new acquisition? An existing account can reveal unmet demand, but a larger invoice is sustainable only while the customer receives value. Referral activity creates growth only when it changes subsequent acquisition after its own costs and delays.

Model the path from an existing customer's experience to the next commercial outcome. For expansion, identify the additional unit, its adoption, service cost and survival. For referrals, identify who becomes eligible, who refers and which referred accounts become retained customers. Count the incremental contribution after rewards and delivery costs.

The customer evidence also changes who you choose to acquire. This learning return and the mechanism that generates subsequent customers are related but distinct. The sections below connect account whitespace, customer success capacity and feedback to acquisition without counting the same value twice.

Distinguish expansion, reinvestment and learning

Expansion changes revenue within the current customer book. Referral can change the arrival of future customers. Reinvestment uses contribution from one period to fund another acquisition period. Learning changes the assumptions and policies used across those paths. Separate these mechanisms before adding their values.

A referral count is an intermediate event. Estimate the additional retained contribution from the programme against referrals that would occur without it. Include reward costs and displacement from other acquisition paths. A referral programme that merely relabels an existing buyer does not create the full value of that buyer.

The reinvestment decision also needs a cash constraint. Discounted expected contribution is a valuation, not cash already available to spend. Acquisition may require payment before the customer recovers its cost. Tie reinvestment to the cash timing and capacity model rather than applying a compounding rate to gross revenue.

Use customer feedback to revise the market and account profile, as *Choose markets and accounts* explains. That feedback comes from customers selected by earlier acquisition and survival. Inspect the missing accounts and the age of the cohorts before concluding that a segment cannot succeed. A selected customer book is evidence about the policy that produced it.

Expansion and price

Expansion is growth in the recurring revenue of a customer that stays (see *Model recurring growth*). It has four sources: more seats or sites, a higher tier, another product, and a higher price. *Whitespace* is the part of an account where the product could deliver impact and does not yet, for example plants or roles that have no licence. It is the base for expansion plans. *Upsell* sells a higher tier or more of the same product. *Cross-sell* sells another product to the same customer.

Measure whitespace in the unit that the customer buys: licensed units against the units that the customer could use, such as sites, seats or modules. The *expansion rate* is the expansion ARR of a period divided by the opening ARR of the same customers (see *Model recurring growth* for its place in net revenue retention). Whitespace is the stock. The expansion rate is the flow.

Value expansion on the same basis as a new customer. A euro of ARR added to a customer that has passed its first renewal carries the survival of such customers. So it is worth more than a euro won from a new customer who may still leave (see *Find the most valuable constraint* for the worked value). Watch its cost. The *cost of expansion* is the sales and marketing cost that goes to winning expansion, fully loaded, for each 1.00 of expansion ARR. It holds account managers, the expansion share of customer success and of the sellers, and the tools and programmes. Renewal work is not in it. The book's starting ceiling is 0.50, with a starting objective at or below 0.40 (Plausible but untested). The reported 25th percentile of 0.41 provides context, not a validated operating limit for your company (KeyBanc Capital Markets, 2021, *Private SaaS Company Survey*). At Year 2 December Alder spends 0.20, Birch 0.23 and Cedar 0.07 (Arithmetic).

Advocacy and referral

Advocacy is customers who recommend you: references, reviews and referrals. The *referral rate* is the share of new customers that came from a referral. Referrals form the compounding loop of *Choose the growth objective*, where each customer brings g new customers. The referral rate is not g: g divides the referred new customers by the customers who could refer them. The three conditions of that chapter hold here: g is measured on your data, it is stable, and no referred customer is also counted in paid acquisition.

A fourth point belongs to this chapter. Referred customers come from customers that are happy, so they often stay longer for that reason alone. Do not credit the referral programme with their whole value. Compare invited and not invited customers in a random split before you judge a reward. A referral reward is a cost of acquisition, so put it in the CAC.

At low volume, references are scarce assets. Count the asks that each customer receives, so that you do not use up the goodwill of the best accounts. None of the three companies has measured g yet, so we leave it open.

Customer success capacity and cost to serve

A **customer success manager** helps customers reach impact and renew. An **account manager** owns expansion in existing accounts, usually only where expansion potential exists. Plan both from capacity, as in *Set targets and staff capacity*: accounts or ARR that one manager can carry at a given touch level, never a flat ratio. The table of published ratios is in that chapter. This chapter adds the renewal load and the cost.

Table 3. Capacity and cost to serve at Year 2 December (Arithmetic, from the facts).

	Alder	Birch
Customer success managers (full-time equivalents)	4.9	24.4
Accounts for each manager	51	about 125
ARR for each manager	4.0 million	1.47 million
Renewals due a year for each manager	36	109

	Alder	Birch
Cost to serve, share of ARR	1.65%	5.48%

Birch carried 142 accounts for each manager at Year 1 July (see *Set targets and staff capacity*) and about 125 at Year 2 December. The plan holds ARR for each manager at 1.47 million, and the average account grew, so the load in accounts fell. Cedar has 5.2 managers, for assisted accounts only. The self-serve book has no managers: its renewal work is the cancellation flow and automated messages.

The *cost to serve* is the cost of renewal and expansion work that is not in the cost of revenue (see *Value customers and cash*). The rule of that chapter holds here. Adoption work sits in the cost of revenue. Renewal and upsell work sits in the cost to serve. Count each euro in one place. A cut in customer success lifts the margin at once, and the harm shows at renewal, months later. Never manage to the margin alone.

When load passes capacity, the order of work is a choice. Give the managers' time to the plays with the highest value, not to the highest risk. That is the shadow price of *Allocate sales attention* applied to people.

The second loop

What customers do after the close is the best data on whom we should have won. This is the second loop. Churn reasons and survival by segment feed the ideal customer profile and the acquisition levers.

A **churn reason** is the coded reason that a customer left. Treat it like a loss reason. Buyers give polite reasons, coders make errors, and the labeller must be checked on a hand-checked sample before the shares are used (see *Diagnose what changed*). Print the corrected share with its n and its interval.

The survival by segment gives the value of a customer by segment. If a segment leaves earlier than the value ladder assumed, its customers are worth less. Then every lever that wins that segment must be priced again (see *Value customers and cash*). The change may touch the qualification rule, the audience of the ads and the ideal customer profile. Each is a decision with a card. Read the cohorts at equal age, and wait for them to mature before you change the profile. A change that rests on two young cohorts is noise.

Related company decisions appear in *Retain customers profitably*, *Set prices and packages*. Their original dates remain explicit.

The decision to carry forward

Price expansion and referral on incremental customer profit. Respect service capacity and the delay to retained revenue. Feed mature customer evidence back into acquisition without counting the same learning return as another stream of revenue.

15 · Find the most valuable constraint

Which constraint is worth removing first? A weak conversion rate is not automatically the best opportunity. A small feasible improvement across many valuable accounts can outweigh a large improvement in a small stage. The calculation also changes when downstream capacity prevents the extra demand from becoming revenue.

Compare interventions against the same baseline, horizon and gross-profit basis. State how the intervention changes volume, mix, rates, time or cost. Preserve dependence between those quantities: a different audience may convert and retain differently.

The marginal calculation below turns a proposed improvement into a distribution of value. Its purpose is to find where scarce work can earn most. *Allocate sales attention* applies the same opportunity-cost argument to seller time, and *Fund the next best action* applies it to the portfolio.

The question: one more percent, but where?

A *lever* is one change that a team can make to the funnel or to the customers after the sale. A new landing page (the page where a visitor arrives from an ad) is a lever. So is a qualification rule, a media budget or a plan to keep customers at their first renewal. The *leverage* of a lever is how much the value of the whole business changes when the lever works.

Choose markets and accounts showed that customers equal the top volume times the product of the step rates. A gain of 1% of itself in any one rate gives 1% more customers. Take Birch in Year 1 August, at the trend level before seasonality. It wins 47.2 new customers a window, which is 613.6 a year. One percent of that is 6.1 customers. One new customer is worth 31,538 euro of expected gross profit over five years, discounted at 12%. So one percent more customers is worth 193,522 euro for each year of flow (Arithmetic).

We call this the *marginal value* of a rate: the euro of expected discounted gross profit that a gain of one percent of itself brings. The figure is the same for each of the five step rates of Birch, from lead to MQL up to the win rate.

The same method prices measures after the sale. Birch has 689.9 first renewals due in the next twelve months, and the first-renewal rate is 85%. One percent of that rate keeps 5.9 more customers. A customer that leaves at its first renewal holds less ARR than the average, 7,906 euro in the company model. A euro of ARR that survives the first renewal is worth 3.43 euro of gross profit over the next five years. So one percent on the first-renewal rate is worth 159,159 euro. One percent on the expansion of the current book, which adds 3,182,846 euro of ARR over the next twelve months, is worth 109,260 euro.

Birch, Year 1 August	Base for a year	Extra from +1% of the rate	Marginal value, euro
Any of the five step rates of the chain	613.6 new customers	6.1 customers	193,522
First-renewal rate	689.9 first renewals due	5.9 customers kept	159,159
Expansion of the current book	3,182,846 of expansion ARR in the next 12 months	31,828 of ARR	109,260

The three figures are of the same size, so the marginal value only sets the price of a percent. The ranking comes from five other things: how much a team can plausibly gain, what the gain costs, when it pays, whether the later stages can absorb it, and whether the added units are as good as the average ones.

Headroom, the first of the five, is the gap between a rate today and a ceiling or a benchmark. A rate with more headroom can plausibly gain more.

> **Calculation**
>
> headroom = ceiling or benchmark value ÷ current value − 1

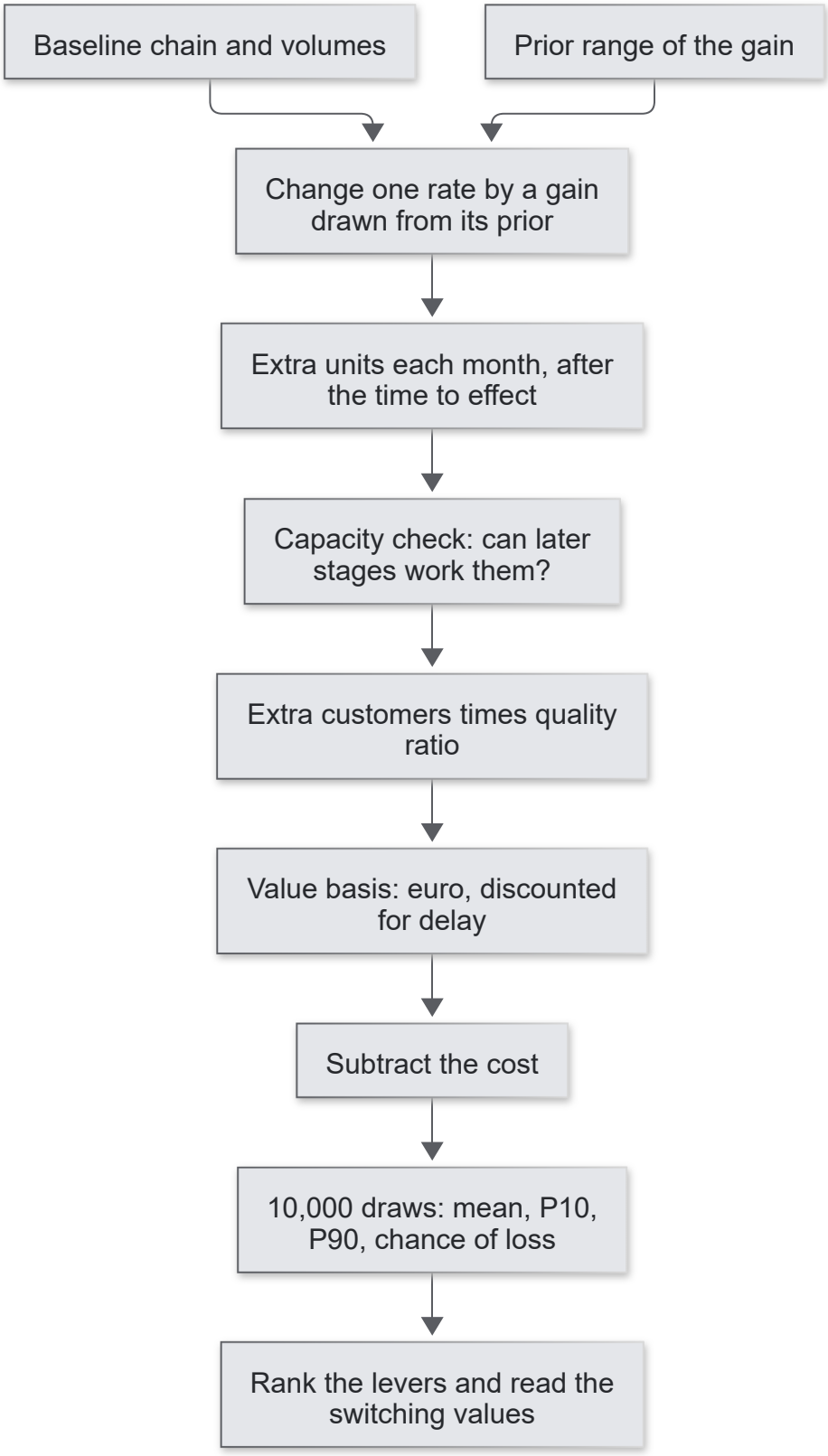
The table gives the headroom of Birch against the top of the published span for each rate in the chain. The spans come from the vendor data set that *Choose markets and accounts* discusses. Read them as wide ranges and not as targets (Plausible but untested as a measure of headroom). For the first-renewal rate we found no span that we could verify. Its ceiling here is the company's own second-renewal rate of 90%.

Rate	Birch today	Published span or ceiling	Headroom to the top
Lead to MQL	40.0%	35% to 45%	12.5%
MQL to SQL	25.0%	13% to 51%	104.0%
SQL to opportunity	50.0%	35% to 50%	0.0%
Win rate	32.0%	20% to 40%	25.0%
First renewal	85.0%	own second renewal, 90%	5.9%

A rate at the top of its span has no headroom, so a lever aimed at it has a poor prior. That is a belief: the true room is unknown until the company tries.

The leverage calculation

The exact method is to compute, for each lever, the expected profit it brings and the range of that profit. Expected profit is the gross profit of the extra customers, on the value basis, minus the cost of the lever. We use the same method for every lever: a rate before the sale, the volume at the top, a renewal or an expansion. This puts all levers in one list in euro.



The leverage calculation for one lever. A rate or a volume changes, the later stages limit what survives, the value basis converts it to euro, and a simulation gives the range.

Inputs

Each lever needs a short list of inputs. Write them on the decision card of *Commit with uncertainty*, with an owner for each.

Input	What it is	Where it comes from
Baseline chain	Volumes and rates at the trend level	The funnel model
Plausible effect	The gain in a rate as a range: P10 and P90 of the relative gain	Outside base rates (how often similar changes worked elsewhere) until the company holds about 30 recorded tests, then its own register (Plausible but untested)
Quality ratio	The value of an added unit divided by the value of an average unit	Measured on a holdout (a group kept out of the change, for comparison) where possible; a belief otherwise
Cost	One-off cost and running cost, fully loaded with people time	Finance and the stage owner
Time to effect	Months from the decision to the first extra unit, plus the cycle from the changed stage to the signature	The stage owner; the median cycle from the funnel model
Lever life	Months of extra flow that we count	A start value of 12 months (Plausible but untested)
Capacity	What the later stages can work in a window	The growth model

Two more inputs matter for the decision, but they do not enter the profit. Reversibility sets the shipping threshold on the card, the probability of a positive net value that a change needs before it ships. It is 90% for a change that can be undone and 95% for one that cannot. Readability is the time that the data need to show whether the lever worked. It sets the value of information, which is what a better answer is worth before you decide (see *Choose whether to test or act*).

The lever life is a choice. A page change may keep its gain for years or lose it in a month. The five-year horizon prices the customers, and the lever life says how long the lever keeps adding them. State the choice beside every ranking.

A worked number

Take a lever for Birch: a new landing page and form for visitors that come from paid clicks. The table follows one pass of the calculation with every input at the mean of its range.

Step	Quantity	Value
1	SQLs a window now (trend level)	295.0
2	Gain in click-to-lead on paid clicks: mean of the range (P10 2%, P90 14%)	8.0%
3	Share of all leads that come from paid clicks	52.0%
4	Gain in all leads, and so in SQLs, when later rates do not change	4.2%
5	Extra SQLs a window	12.3
6	Extra SQLs in the 12 months of lever life, with growth and season	168.4
7	Close rate from SQL to won	16.0%
8	Extra customers	26.9

Step	Quantity	Value
9	Quality ratio: mean of the range (0.75 to 1.00)	0.883
10	Extra customers of average quality	23.8
11	Value of one new customer, euro	31,538
12	Gross profit at signing, euro	750,576
13	After the delay: two months to effect, 34 days of cycle, flow spread over a year (factor 0.913), euro	685,279
14	Cost of the lever, euro	25,000
15	Expected profit before the capacity check, euro	660,279
16	Extra SQLs that ten sellers can work (the capacity check)	0
17	Expected profit after the capacity check, euro	−25,000

Read it in two parts. Steps 1 to 15 are the lever on paper, and the page looks excellent. The simulation of the same lever draws the gain, the quality ratio and the cost from their ranges. It gives a mean of 663,870 euro, a P10 of 144,011 euro and a P90 of 1,176,744 euro. The chance of a loss is 4.8%. Steps 16 and 17 are the capacity check, and they remove the whole gain. Birch's ten sellers already work all the SQLs that they can, and a queue of 56 SQLs stands at the end of August. It was 110 in July, and it shrank only because August is a slow month. Extra SQLs only lengthen the queue. If an eleventh seller at full capacity joins, the lever returns to a mean of 589,155 euro.

Price the lever on paper first, then check whether the rest of the system can take it.

After-close candidates on the same basis

Levers after the sale use the same steps. The extra units are customers kept or euro of ARR added, and not leads. The value of a kept customer is the discounted gross profit of its remaining years. There is no cycle from a changed stage to the signature, because renewals come on the contract anniversary. The check on capacity is the capacity of the customer success team and not of the sellers.

Take a first-renewal playbook at Birch. It lifts the first-renewal rate by 2% of itself on average, with a P10 of 0% and a P90 of 4%. It costs 60,000 euro and takes three months to reach the first renewal. Over the twelve months of lever life it keeps 11.7 more customers. The mean expected profit is 228,542 euro, the P10 is −63,273 euro, the P90 is 519,435 euro, and the chance of a loss is 16.2%.

Never add the values of single levers. *Choose markets and accounts* showed that gains in several rates multiply. Levers also compete for the same sellers, people and budget. Compute a combination as one scenario with one capacity check.

Why "top of the funnel first" often holds

Many teams improve the top of the funnel first. We call this rule of thumb the *leverage order heuristic*. It is not a law, but it often holds, and the calculation shows why. We do not state how often it holds, because no source measures that.

The arithmetic gives no reason for the rule, because one percent of any rate is worth the same. The five reasons lie in the other inputs.

More headroom. Rates have ceilings, and volume at the top has none until the market or the budget runs out. A win rate of 32% will not double. In the Birch table, MQL to SQL has 104% of room and SQL to opportunity has none. A rate near its ceiling is harder to lift than one far below it (Plausible but untested).

Lower cost for each gain. A gain in a conversion rate at the top costs once. A team builds a page or a rule, and every visitor that follows uses it at no extra cost. A gain in volume costs for each unit, and each unit usually costs more than the one before. The reason is that a media channel reaches the best-fitting audience first. The response of leads to spend is concave: a rise of 10% in spend gives a rise of less than 10% in leads (Standard practice).

The Birch model has such a curve. The *spend elasticity* of a channel is the percent rise in leads that a 1% rise in spend brings. Birch's channels have spend elasticities of 0.6 to 0.8. So 10% more spend buys 7.0% more paid leads (Illustration only). A real company measures the curve with budget pulses, as *Choose acquisition channels* shows. The average paid cost of one customer is 7,067 euro.

Step in spend	Extra spend per window	Extra customers per window	Marginal media cost per customer, euro	Marginal payback, months
From 100% to 110% of today	16,570	1.65	10,048	15.6
From 110% to 120%	16,570	1.61	10,304	16.0
From 120% to 130%	16,570	1.57	10,543	16.4
From 130% to 150%	33,141	3.05	10,874	16.9
From 150% to 200%	82,852	7.17	11,549	17.9

The first extra customer costs 42% more than the average one. The payback ceiling is 18 months for mid-market (Bessemer Venture Partners, 2021, *Scaling to $100 Million*). That ceiling is Standard practice for average payback; its use as a test at the margin is Plausible but untested. On this curve the marginal media cost reaches the ceiling when spend is 1.76 times today's spend. Until then, on media cost alone, a gain in volume still pays, but each step pays less. Media cost is only a floor. If the extra SQLs need another seller, add that cost; *Set targets and staff capacity* did so and found the marginal customer above the ceiling. A conversion gain such as the page costs 25,000 euro once and adds customers at no cost per visitor. This is the main reason why conversion levers at the top so often rank above buying more traffic.

Reversible and cheap. A page, a form or a targeting rule can be undone in a day. A change of price, a new sales process or a hire cannot. A reversible change meets a lower shipping threshold on the card, so it ships earlier on weaker evidence, and a wrong move costs little.

More attempts a year. A rate with a high flow shows a 20% change in days, and a rate with a low flow needs months. At Birch in Year 1 August, MQL to SQL reads a 20% change in 28 days. The team can read about 13 changes in a year, one after the other. The win rate needs 161 days, so it allows about 2. More attempts help only if the levers are worth trying, so we never use this reason alone. Reading speed counts because it raises the value of information per week.

A gain that feeds itself. A customer added at the top may bring other customers. The compounding loop of *Choose the growth objective* gives the arithmetic. If each customer brings g more, a lever that adds N customers adds N ÷ (1 − g) in all, for g below 1. With g at 0.05, the lever is worth 1.053 times its value without

the loop, which is 5.3% more. With g at 0.10 it is 1.111 times. The argument holds under three conditions. You measured g on your own data. The value of g is stable over time. And the referred customers are not already counted in paid acquisition, so that you do not count them twice. The multiplier is the same for every lever that adds new customers, so it does not change their order. It lifts them as a group against levers that add none, such as a renewal playbook. A multiplier of 1.053 cannot reverse a ranking in which two levers differ by a factor of two. If you have not measured g, leave it out.

Each of the five reasons can be false for a given company. The calculation tests them; the heuristic only suggests them.

When it fails

The heuristic fails when one of its reasons is false. Six cases matter.

Downstream capacity binds. More volume at the top helps only if the later stages can work it. If the sellers are full, extra SQLs wait. Birch in Year 1 August is the case.

Added volume is of lower quality. We call this *quality dilution*: the average value of a unit falls when you add more volume. The *quality dilution ratio* is the value of the marginal unit divided by the value of the average unit.

1. The rate that we measure comes from the units that already arrive.
2. Those units arrive through a selection. They searched for the product, they match the audience, or they met a seller at the right event.
3. More volume reaches deeper into the audience, further from the match.
4. Units from deeper in the audience convert less often.
5. So the average rate overstates the rate of the next unit, and the value of the marginal unit is below the value of the average unit.

The argument holds when the channel serves the audience in order of fit and when the audience, not the budget, limits the volume. It fails when added volume comes from a new source of the same quality, and it can reverse when a channel has improved. A conversion lever has the same problem in another form. A shorter form converts more visitors, and the extra visitors who convert may be less keen than the usual ones. So give the quality ratio of every lever an explicit range, and measure it on a holdout when you can. Cedar shows a measured case below.

The market is small and known. A company with a short list of target accounts cannot buy volume without limit. More leads then add colleagues at accounts that are already engaged, or accounts far down the list. Alder shows this below.

Unit economics are negative at the margin. A lever can add customers and still add loss, when the marginal cost exceeds the value of the customers it adds. A second test is the marginal payback against the segment's payback ceiling, which *Value customers and cash* sets as a limit and not as a goal. A lever with a positive expected profit can still breach it, and then the company may not be able to fund it.

The leak after the close is larger. When much of the base comes up for renewal, a point after the close can outweigh a point before it. In Year 1, losses at renewal equal 33% of the new and returning ARR at Alder, 37% at Birch and 66% at Cedar. Expansion also costs less to win. In Year 1 August, Alder spends 0.21 euro of fully

loaded sales and marketing cost for each euro of expansion ARR, against 1.32 euro for each euro of new-customer ARR (Arithmetic on Illustration only facts). The expansion cost ceiling of this book is 0.50.

A data gate fails. A data gate is a check on data quality and tracking that must pass before anyone reads a rate. If tracking is broken or definitions differ between systems, no rate can be trusted and no lever priced. Repair the data first. *Decide what to measure* and *Decide when a change needs action* explain the gates.

The order of work against the order of value

The ranking by expected profit says what is worth most. It does not say what to do first. The order of work can differ from the order of value for four reasons.

First, one lever may unlock another. At Birch, the landing page is worth little until the sellers have room. So capacity comes first.

Second, a lever may teach fast. A lever with a lower expected profit may come first if its result arrives sooner and settles a question that other decisions wait for. This is the value of information, and *Choose whether to test or act* prices it. A fast lever wins only when what the team learns is worth more than the difference in profit.

Third, delay costs money. *Fund the next best action* turns the weekly profit that a waiting lever loses into a cost of delay (the gross profit lost for each week that a change starts later) per unit of scarce capacity.

Fourth, two changes to one rate at the same time cannot be read apart. Sequence them, or design them together.

No score replaces the calculation. A score that multiplies impact, confidence and ease has no unit. It cannot say that a lever loses money, or that a seller queue caps its value. A ranking in euro with a range can. The owners decide from the ranking, in writing, by a stated date. A meeting is needed only if two owners conflict.

What changes with volume

The calculation is the same at every volume. What you can learn about its inputs is not.

Regime	What the ranking can use	Consequence
Low (Alder, 44 wins a year)	Priors and cases. No effect of a rate lever can be read within a year.	Rank on outside base rates and write signposts. Levers that act on a few known accounts, such as expansion, are judged by case.
Medium (Birch, 614 customers a year)	Some effects read in weeks, such as lead to MQL and MQL to SQL. The win rate reads in months.	Rank with priors, then confirm the early rates. Capacity can bind, as at Birch, so run the capacity check on every lever that adds volume.
High (Cedar, 19,295 paying accounts a year)	Every conversion lever reads within a window. The quality of added volume can be measured per channel.	Priors tighten for conversion levers. They stay wide for lifetime effects, and the margin and the horizon decide those.

Alder, Year 1 August: more leads, or more sites

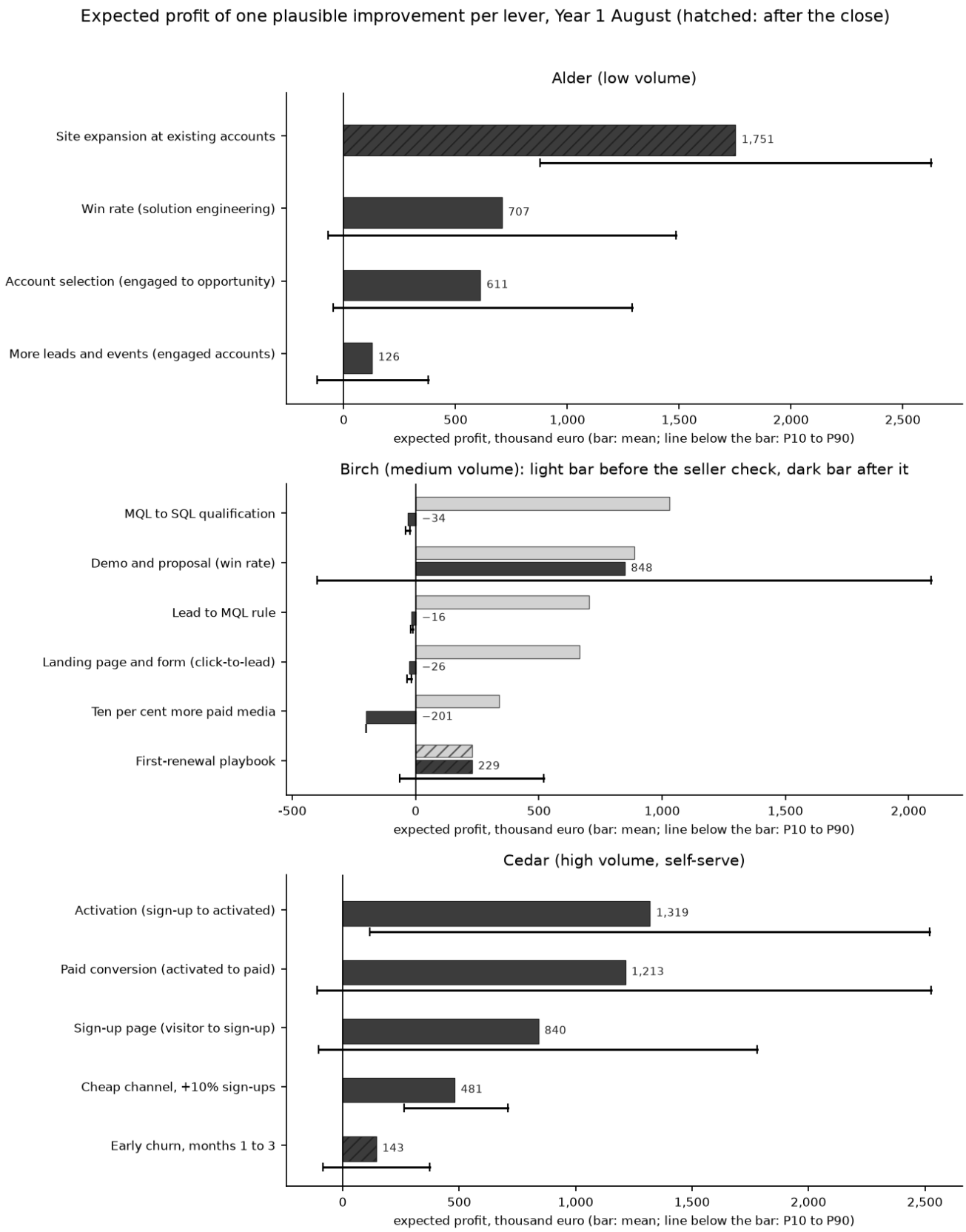
Alder, Year 1 August: about 2.7 new customers a window, 0.8 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.1 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

The marketer asks for 20% more money for events and lead generation. That is 194,397 euro on top of the 971,985 euro that the programme costs today. The analyst puts the request on the same list as three other levers.

Alder sells to a named list of 1,600 target accounts. At the trend level, before the August dip, 40.9 accounts a window become engaged, which is 532 a year, or 33.3% of the list. The CRM holds 10.5 lead records for each engaged account. So an extra 20% of leads is mostly extra colleagues at accounts that are already engaged. The lever therefore carries a gain in newly engaged accounts of only 1% to 9% (P10 and P90, Illustration only). The quality ratio is 0.50 to 0.90, as accounts deeper in the list fit less well. With these ranges the lever brings 1.6 extra customers in its life. Its expected profit is 126,135 euro, with a P10 of −119,747 euro, a P90 of 375,992 euro and a chance of a loss of 26.2%.

Now take the expansion lever. Alder's customers hold licences for 790 sites. They run 1,420 plants in all, so 630 plants have no licence yet. At the list price of 21,000 euro a site, that is 13,230,000 euro of ARR, almost as much as Alder's whole ARR of 14.1 million euro. Customers who already know the product are the best prospects for these sites (Assertion). A programme that converts 2% to 6% of them over the lever life adds 25.3 sites on average. That is 530,342 euro of ARR, or 25.3% more than the expansion that Alder's current book adds in the next 12 months. A euro of ARR added to a customer who has passed its first renewal is worth 3.82 euro of gross profit over five years. Alder spends 0.21 euro for each euro of expansion. The expected profit is 1,751,182 euro. The P10 is 877,300 euro, the P90 is 2,627,641 euro, and the chance of a loss is 0.7%.

Two other levers sit between. Better selection of accounts, from engaged account to opportunity, has an expected profit of 611,134 euro. Better work by solution engineers on the win rate has 707,225 euro. Both have P10 values below zero, because their ranges start at no gain.



Expected profit of one plausible improvement per lever for the three companies in Year 1 August, with P10 to P90 as a line below each bar and the mean, in thousand euro, on each dark bar; hatched bars act after the close. For Birch the light bars show the value before the seller check.

The first check against weak priors is the marginal value. A gain of 1% in the win rate is worth 98,193 euro, and 1% on the expansion of the current book is worth 80,161 euro. They are of the same size, so the gap is not a matter of price per percent. It comes from headroom and cost: 630 open sites against a list that is already a third engaged. The second check is the switching value. The expansion programme matches the mean of the lead lever at 1.7 converted sites. Its P10 is above the P90 of the lead lever.

The marketing budget stays as it is. The head of customer success takes the expansion programme and writes a card with a signpost: the sites converted by Year 2 February. With 44 wins a year, no rate below the lead can show the result within a year. So Alder decides on priors and logs its belief before the result. The lesson at low volume is that "top of the funnel first" fails when the market is small and known. The sites of existing customers are a larger and cheaper pool of value.

Birch, Year 1 August: conversion levers lead, until the queue says stop

Birch, Year 1 August: about 48 new customers a window, 1.0 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 72 days, and the close rate reads a 20% change in about 7 months.

The growth team lists six levers and prices each on the value basis. Four of them add SQLs: the landing page, the rule from lead to MQL, the qualification from MQL to SQL, and 10% more paid media. One is the win rate and one is the first-renewal playbook. The inputs are Illustration only. The ranges for the gain follow the headroom table, so MQL to SQL gets a range that reaches 12%.

On paper the heuristic holds. Qualification ranks first at 1,028,705 euro and the win rate second at 887,702 euro. The lead-to-MQL rule follows at 701,850 euro and the landing page at 663,870 euro. Paid media has 339,825 euro, and the first-renewal playbook comes last at 228,542 euro. Conversion changes cost once, they can be undone, and they read in days or weeks. Paid media pays less because each extra customer costs 10,048 euro at the margin against 7,067 euro on average. Its marginal payback reaches the ceiling at 1.76 times today's spend.

The capacity check changes the ranking. Demand for the 300 SQL slots exceeds them in 10 of the next 12 months, and 56 SQLs wait already. The four levers that add SQLs lose their whole gain. The landing page falls to −25,804 euro, the lead-to-MQL rule to −16,414 euro and qualification to −33,506 euro, which is about their cost. Paid media falls to −200,792 euro, because the company would pay for customers that nobody can serve. The win rate lever keeps most of its value, 848,145 euro, because it works on SQLs that sellers already handle. The playbook stays at 228,542 euro.

An eleventh seller at full capacity changes the picture again. With that seller, the landing page returns to 589,155 euro, the lead-to-MQL rule to 625,334 euro and qualification to 701,080 euro. So the heuristic holds on paper, and the calculation shows for how long: only as long as the sellers have room. At the trend level that room is 1.7% of SQL volume, which is 5 SQLs a window, and it is less in most months.

The marketing owner does not raise media spend. The sales owner brings the capacity question to a decision card. *Commit with uncertainty* shows that card in Year 2 February. Until then Birch watches the queue. The marketing owner may start the page and the form, because they are cheap, reversible and ready the day that capacity opens. Sales enablement drafts a new demo script to lift the win rate, a lever that the seller queue

does not cap. In September, *Decide what the evidence can resolve* shows what that test can read. The lesson at medium volume is that top-first holds as a ranking on paper, and that the calculation tells you where it stops.

Cedar, Year 1 August: cheap sign-ups

Cedar, Year 1 August: about 1,300 new paying accounts a window, 1.5 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 29 days; the assisted motion, with 54 opportunities a window, needs about 8 months to read a 30% change.

A marketer in the growth team finds a channel that delivers sign-ups at 12 euro each. The average paid sign-up costs 24.88 euro. The idea is to buy 10% more sign-ups there, which is 4,340 a window. The channel's own data show that 45 of its 3,750 sign-ups became paying accounts. That is 1.2%, with an 80% interval of 1.0% to 1.5% (Illustration only), against 3.42% of all sign-ups.

The quality dilution ratio is 1.2% ÷ 3.42%, which is 0.35. An average sign-up is worth 57.56 euro of expected gross profit on the value basis. A sign-up from the cheap channel is worth 20.20 euro. After its cost of 12 euro it leaves 8.20 euro. An average paid sign-up leaves 24.27 euro after its cost. So the marginal sign-up earns about a third of what an average paid sign-up earns, and it costs about half as much. The cost of a paying account from this channel is 1,000 euro, against a value of 1,683 euro, which is 1.68 times. The rate rests on 45 paying accounts, so its standard error is about 15% of its value, and the simulation gives the ratio that spread.

The expected profit of the lever is 480,701 euro, with a P10 of 264,519 euro and a P90 of 710,494 euro. No draw in 10,000 shows a loss. But Cedar's segment has a payback ceiling of 12 months, and its self-serve payback on the fully loaded CAC is already 12.0 months. The accrual payback of a paying account from the cheap channel is 19.0 months on media cost alone. The lever is profitable on the five-year basis and still breaks the limit. Cedar may test it, but it may not scale it.

Compare the four other levers. The sign-up page acts on visitors who already arrive. The other three act on the 43,398 sign-ups of a window, which Cedar already has, or on the accounts they become. A gain in activation has a range of 1% to 9% of itself and a quality ratio of 0.80 to 1.00. It costs 150,000 euro and has an expected profit of 1,319,472 euro, with a P10 of 115,898 euro and a P90 of 2,520,309 euro. Paid conversion has 1,213,191 euro. The sign-up page has 840,141 euro. A playbook that cuts early churn in the first three months of monthly plans has 143,377 euro.

Activation and paid conversion add no media cost per unit. They lift the value of sign-ups that exist. An activated sign-up is worth 160 euro, against 58 euro for a sign-up. At the marginal value, 1% on any step rate gives 192.9 paying accounts a year, worth 324,726 euro.

The ranges stay wide even at this volume. High volume measures how well the cheap channel converts, but not how much new onboarding screens will lift activation; that stays a prior until a test reads.

The growth lead decides. The cheap channel keeps a small test budget with its own tracking, and the growth team starts the activation and paid-conversion work. The log entry holds the belief for each: the gain, the quality ratio and the date of the read. The lesson at high volume is that you can measure the marginal quality of added volume, and that it is lower than the average.

The decision to carry forward

Remove the constraint with the highest feasible economic return. Recompute the ranking when the constraint moves. A larger upstream flow is valuable only if the customer system can convert, deliver and retain its contribution.

16 · Decide what to measure

Which observations would change a decision, and can you trust them? Instrumentation begins with that question. A metric needs a unit, an eligible population, a time definition and an owner who can act. A total that cannot be reconciled with the customer system cannot support a profit estimate.

Store observations at the level needed to reconstruct what was known when the action was chosen. Account identity, event history and consent conditions affect the population you observe. A gap in that history becomes uncertainty in the decision, even if a dashboard displays a precise number.

This chapter develops the shared measurement system used throughout the book. Its contracts support customer selection, experiments and later evaluation together. Keep the reporting convention for attribution stable, while using causal comparisons for decisions about what spend creates.

What to measure

Start from decisions, not from data. A company can compute thousands of metrics, and very few change what anyone does. We call a metric a **decision metric** when a named person uses it to make a decision. A **reference metric** gives context or helps a diagnosis, but it moves nobody by itself. The test is one question: if this number moves, who does what? If nobody can answer, the metric is a reference metric. It stays in the data, off the page.

About 20 decision metrics are enough to run one bowtie. The table shows a typical set. Each company picks its own, and the detail depends on its motion.

Metric	What it tells	Primary source	Read	Main user
Paid spend and clicks by channel	Cost and reach of paid demand	Ad platforms	Weekly	Marketing owner
Click-to-lead rate, one population	How well the click turns into a lead	CRM leads, platform clicks	Monthly	Marketing owner
Leads and MQLs by stage date	Volume entering the funnel	CRM stage events	Weekly	Marketing and sales owners
MQL-to-SQL rate, cohort basis	Quality at the hand-off	CRM stage events	Monthly	Marketing and sales owners
Uncontacted SQLs and seller load	Whether capacity binds	CRM	Weekly	Sales owner
Opportunity rate and win rate, cohort basis	How well selling works	CRM stage events	Quarterly	Sales owner
Sales cycle, 80th percentile	Honest deadlines	CRM stage events	Quarterly	Sales owner
New customers and new ARR	The result of the left wing	CRM, then billing	Monthly	Bowtie owner, finance
Average contract value and discount depth	Price actually realised	CRM, then billing	Monthly	Sales owner, finance

Metric	What it tells	Primary source	Read	Main user
Time to activation, 80th percentile	Speed to first value	Product events	Weekly	Customer success
Activated share by cohort	Whether new customers reach value	Product events	Weekly	Customer success, product
Renewal rate by renewal number	Survival of contracts	Billing	Monthly	Customer success
Gross and net revenue retention	Revenue kept and grown	Billing	Monthly	Bowtie owner, finance
Expansion ARR	Growth inside the book	Billing, CRM	Monthly	Account managers
ARR bridge, five movements	Where ARR came from and went	Billing	Monthly	Finance, bowtie owner
Fully loaded and cohort paid CAC	Cost to win a customer	Finance, CRM, ad platforms	Quarterly	Finance, growth owner
Payback and gross margin	Whether growth pays	Finance, billing	Quarterly	Finance
Concentration of ARR	Dependence on a few accounts	Billing	Monthly	Bowtie owner

The first column holds the *what*. The primary source says *where*. The read column says *when*, and it follows the loop delay of the stage: a weekly read of a quarterly outcome only adds noise. The last column says *who*. The bowtie owner settles conflicts between the two wings of the bowtie, as *Run the growth system* describes. The *why* is the decision each metric serves. The rest of this chapter shows *how* to compute each one safely.

Leading and lagging metrics

A **leading metric** moves first, and you can act on it this week. A **lagging metric** records the result later, and you cannot move it directly. Uncontacted SQLs lead. New customers lag. Steer with the leading metric, and check the result on the lagging one. Do not steer on a lagging metric alone: by the time it moves, its cause is weeks or months old.

A leading label is a claim: this metric predicts that outcome. The claim needs evidence. We grade it in three steps.

1. **Predictive at unit level.** On matured cohorts, the signal predicts the later outcome of each unit, beyond what channel and segment predict.
2. **Screened in a series.** The metric and the outcome move together across windows, with a lag. This is weak evidence. A year holds 13 windows. In 13 windows a correlation must reach about 0.55 to count at all, and about 0.71 to be found with 80% power (Arithmetic). Most real links are weaker, so the screen misses them. A pass does not prove a link either, because two series that both trend upwards also correlate.
3. **Causal.** An experiment or a holdout moved the metric and the outcome. The chapter *Choose whether to test or act* shows how.

Birch gives an example. In its example figures, 72% of new customers reach their first result within 30 days. Of these, 90% renew at the first renewal, against 72% of the others (Illustration only). That gap is a unit-level association. It does not prove that pushing a customer to activate will lift renewal, because customers who

activate may differ from the start. A metric without a grade is an **early warning**. A decision waits for the outcome unless the metric has been validated. We label this grading scheme Plausible but untested.

One definition per metric

Two teams can use one name for two numbers. In Year 1 January, Birch counted 1,128 MQLs a window in its marketing automation and 902 in its CRM. The gap is 226 MQLs, 20.0% of the first count, and no team owned them. An argument about whether leads are good cannot start before that gap is closed.

The cure is a **metric specification**: one written definition, with the same fields for every metric. The fields are the name, the numerator, the denominator, the population, the window, the grain, the owner, the source, the minimum sample and the limit model. The table fills them in for one metric.

Field	Close rate, cohort basis (Birch)
Numerator	SQLs of the cohort that reach closed-won within the stated age
Denominator	SQLs of the cohort that are at least that old
Population	Inbound SQLs contacted by a seller; outbound SQLs shown apart
Cohort and age	SQLs by month of first contact, read at day 28, day 56 and at maturity, day 73
Window and grain	28 days; the display grain is one month
Owner	Sales owner
Source	CRM stage events: the dates on which an SQL and a won deal entered their stage
Minimum sample	One month of SQLs, about 300; a smaller cohort prints with its interval and the note "too few to read"
Limit model	Binomial limits at the cohort size, widened by the overdispersion factor of the last 13 windows (see *Decide when a change needs action*)
Version	A version number and the date from which it holds

Populations

The **population** is the set of items that a metric counts. Two faults come from it. In the first, the top and the bottom of one rate come from different populations. In the second, one tile mixes populations that should stand apart.

The first fault gives a rate that moves with no change in behaviour. The Birch example below shows it. The second fault is the paid CAC of the chapter *Value customers and cash*. Paid spend divided by all new customers looks efficient, because organic customers sit in the denominator. Keep paid-only and all-source figures as separate tiles with separate names. The same rule applies to the **ICP match rate**, the share of leads that fit the ideal customer profile. Only enriched leads, which carry enough data to check, go in its denominator. The share of leads that are enriched is a second tile. The chapter *Improve offers and conversion* uses it.

Never average ratios

Store the two counts behind every rate. Compute the ratio at the level you show, from the sums. An average of ratios gives each group the same weight, however many items it holds.

Birch shows the error in its three paid channels in Year 1 October. Click-to-lead is 2.89% on professional-network ads, 2.70% on search ads and 1.62% on social ads. The simple average is 2.40%. The pooled rate, all paid leads over all paid clicks, is 2.25%. The average is 0.15 points too high. Social ads carry more than twice the clicks of professional-network ads, and they have the lowest rate (Arithmetic). The same rule holds when you roll weeks up to months and segments up to a total.

Cohort-correct by default

Any rate that ends in an outcome later than its entry must name its cohort and its age, as the chapter *Forecast conversion and timing* explains. So the specification of such a rate holds the entry event that defines the cohort and the ages at which it is read. A same-period rate may stay on a page for volume monitoring. It carries its name, and nobody uses it to ask whether selling changed.

Changing a definition

A definition changes through a short written request: the metric, the old and new rule, the reason, and the date. The analytics lead recomputes the last 13 windows under both rules and reports the step between them. The new rule starts at the beginning of a window, with one window computed both ways. Every chart that crosses the change carries a note, and the version number rises. Separate definitions of one field drift apart, so each definition lives in one place (Linoff and Berry, *Data Mining Techniques*).

Implementation note: keep the specifications as versioned files next to the data tests, with a review step before any change. A semantic model in a BI tool can read them. Generating views and measures from the specification is a product in itself. Do not build it before the specifications and tests exist.

Sources of truth

Each metric has one **source of truth**, also called its primary source: the system that decides its value. Other systems are secondary. We use them to cross-check and to fill gaps, never to replace the primary. Two systems that disagree produce an argument. One primary source turns the argument into a question.

System	It decides	It does not decide
Ad platforms	Spend, impressions, clicks: cost and delivery	What visitors did on your site
Web analytics	Sessions, engagement and key events of visitors who were recorded	Who became a lead or a customer
CRM	Leads, stage dates, deals and the amount at signing	Anything invoiced after the sale
Billing	Invoices, ARR and its movements, renewals	Who the buyer was before the sale
Product events	Use of the product: activation, active workspaces	Payment
Finance ledger	Every cost line	Counts of customers

The warehouse and its **semantic model** sit on top of these systems. The semantic model is one shared layer that holds the definitions, so every report uses the same formula. It computes ratios. It never originates a number, and it holds no typed-in values. A number that the model cannot compute shows as blank, never as zero.

When two sources report the same quantity, we compare them. This is **triangulation**, and it is documented: which pair, which gap, which cause. The tile shows the primary source only. The gap becomes a finding (a written record of one issue, with its evidence and one owner). The usual causes, in the order in which they occur, are:

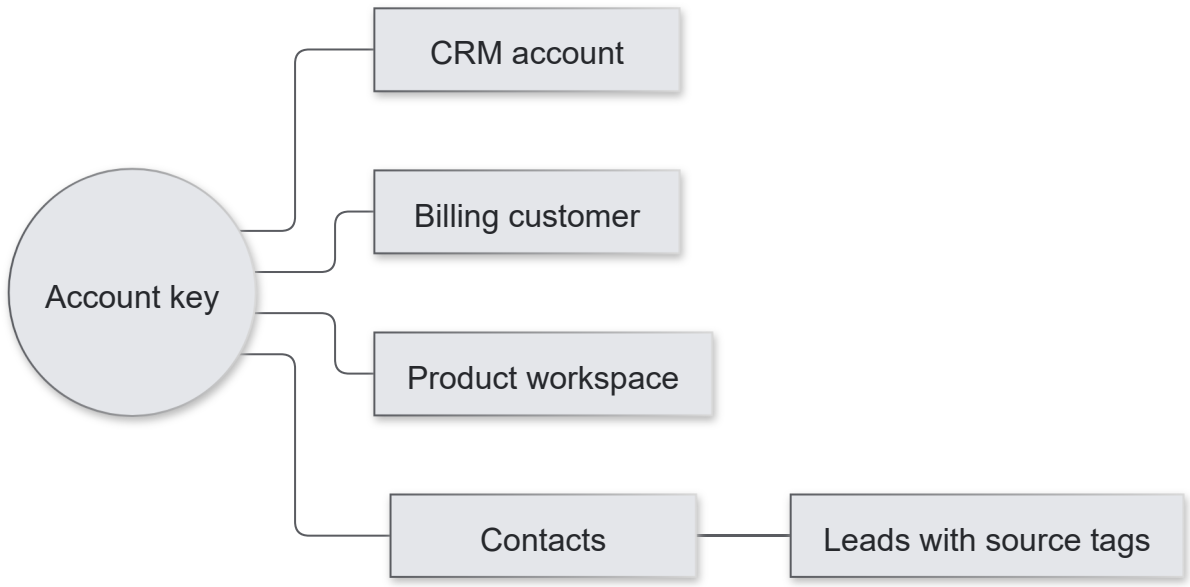
- missing or mistyped UTM tags;
- a change in consent;
- a duplicated or missing tag;
- a day boundary or time zone that differs between systems;
- a different attribution window;
- bot traffic.

A gap that suddenly closes is as suspicious as one that suddenly opens. A tag may fire twice.

Identity and history

Who is who

A B2B purchase is made by an account, and many contacts sit inside it. So the unit of record is the **account key**: one stable identifier that links every record of one customer or prospect across systems. **Identity resolution** is the work of matching records to that key. Three maps do most of it: contact to account, billing customer to CRM account, and product workspace to account.



Match deterministically first, for example on the company web domain. Send uncertain pairs to a person who knows the accounts. Count the unmatched rows and publish the count every day (Kihn and O'Hara, *Customer Data Platforms*). With the key in place, a euro of ARR can be traced back to the channel that found the

account. Contacts of one account stop counting as separate buyers, and the customer count agrees across teams. Write one definition: a paying customer is an account with an active, billed subscription.

What happened when

A CRM stage field holds only the current value. Each edit overwrites the past, so old reports cannot be reproduced. The cure is the **stage event**: one time-stamped row each time an item enters a stage. Derive it from the system's own events, not from edited fields. The table is append-only. Add a daily snapshot of the open deals. Together they let you rebuild any past view, which is the as-of view of the chapter *Forecast conversion and timing*. A late correction gets a new version with its own as-of date. Nobody edits history in silence.

Store every event time in UTC and keep one reporting time zone for the company. Convert only when you read the data. Ad platforms and web analytics tools cut the day at different places, so the same lead can fall on two different days. Compare daily numbers across tools only with an agreed mapping. Test the two clock changes of the year, because a local time can be missing or can occur twice.

Stage events record what happens to a deal. They do not record what one seller does. If you log activity by person, a works council in Austria or Germany may have a say, and results go out by role only.

Implementation note: the companion code holds the stage-event and daily-snapshot tables as SQL. For time zones, checked on 15 October 2026: in Google Ads a standard account cannot change its time zone after creation; LinkedIn reports in UTC; in Google Analytics 4 the time zone applies going forward only, and the export holds a date in the property zone and a timestamp in UTC. Azure SQL Database and Microsoft Fabric support AT TIME ZONE with Windows zone names. Fabric Warehouse cannot store a datetimeoffset column, so store UTC in a datetime2 column.

Tracking and consent

The tracking plan

The **tracking plan** is the list of events that the site and the product must send, with their names and properties. It also holds the rules for UTM tags and consent. It holds at least the page view, the clicks on key calls to action, the sign-up, the lead and the activation event. One rule matters most. Fire the lead event only after the system of record confirms the lead. Count leads from that system, not from a form event alone. Write event names in one place. Tools reserve some names for their automatic events, so send your own events under distinct names.

Implementation note: Google Analytics 4 collects form interaction events called form_start and form_submit by default, and its list of reserved names includes both. Google's pages do not say what happens when a custom event reuses a reserved name, so the safe choice is a distinct name, or switching the automatic form events off (checked 15 October 2026).

UTM parameters are tags on a link that record the source, medium and campaign of a visit. Four rules keep them true. Use one spelling per source and a closed list of media. Set the tags in the URL templates of the ad accounts, never by hand. Make sure that redirects and consent banners keep them. And store them in hidden fields of the form, so that the lead record carries them. Three signs show that the standard has slipped: a rising share of "direct" or "unassigned" traffic, fewer paid leads in the CRM than the platforms report, and source names that disagree.

What the browser hides

Web analytics sees only what the browser sends. Visitors can decline tracking. Browsers limit the cookies that follow a visitor from site to site. Ad blockers remove tags. So recorded sessions are always a subset of the visits. Do not plan the funnel around the end of third-party cookies. Some browsers already block or partition them. One large browser keeps them for users who do not change a setting. And the policies change. Treat cross-site cookie tracking as partial everywhere. Build on first-party data and on identifiers that the visitor has consented to (Kihn and O'Hara, *Customer Data Platforms*).

Implementation note: browser policy checked on 15 October 2026. Chrome keeps its current approach and does not block third-party cookies for all users (Google, 2025, *Privacy Sandbox next steps*; Google, 2025, *Update on plans for Privacy Sandbox technologies*). Safari blocks all third-party cookies and deletes script-written storage after 7 days without interaction (WebKit, undated, *Tracking Prevention*). Firefox partitions cookies by site by default (Mozilla, 2022, *Firefox rolls out Total Cookie Protection by default*). We found no official Microsoft statement on the default for Edge.

Two modes of consent handling exist in web analytics. In the first, the tag stays silent until the visitor agrees, so a visitor who says no leaves no row anywhere. In the second, the tag loads at once and sends **cookieless pings** while consent is denied: events with coarse fields and no stable identifier. The warehouse export holds only what the browser sent. In the second mode it therefore holds those pings, and you must not count them as users or sessions. The reports in the tool's own interface add modelled estimates. The export never does. So the interface and the warehouse will not match, and every report must say which of them it uses. Store the consent state with each event, so that you can split the rows.

Implementation note: in Google Analytics 4 the BigQuery export holds cookieless pings and never holds modelled data; the consent fields are ads_storage and analytics_storage in privacy_info (Google, undated, Analytics Help, *BigQuery Export* and *Behavioral modeling*; checked 15 October 2026). One point is open. Google says each session of a cookieless ping gets a different user_pseudo_id, while third-party sources say the field is null. Count the null rows in your own export.

The consent capture ratio

The **consent capture ratio** is the share of ad clicks that appear as sessions in your analytics. It falls when visitors decline, when scripts load late, when bots click, and when tags break. So it is not the consent rate, and we do not call it that.

> **Calculation**
>
> consent capture ratio = sessions recorded in web analytics ÷ clicks billed by the ad platforms
>
> The gap is 1 minus the ratio. Use one base for the gap, the share of clicks, and one band around it.

Measure the ratio for each source and each week, with an interval. Read it before you compare sessions with clicks. A fixed band for "normal" does not work, because a banner release moves the ratio by design. Record every banner release as a **change event**: a dated record of a known change that can move a metric. The chapter *Decide when a change needs action* uses change events to move the expectation instead of raising an alert. State which population every session count measures: consented, cookieless, modelled or all.

The consequence for rates is strict: never compute a rate from two populations. Click-to-lead uses platform clicks and CRM leads, which both cover everybody. Or it uses recorded leads and recorded sessions, which both cover the consenting. A fall in consent from 60% to 55% cuts recorded sessions by 8.3% and lifts CRM

leads per recorded session by 9.1%, with no change in the funnel (Arithmetic).

The engaged session

An **engaged session** is a session that shows real interaction. The common rule is a session of more than 10 seconds, or with a key event, or with at least two page views. A key event is an action that you mark as important in the tool, such as a demo request. Bounce rate is the share of sessions that are not engaged. The seconds can be changed, so write the threshold you use into the specification. Do not mark an event that almost every session has as a key event, or almost every session counts as engaged. Do not compare bounce rates across tools without checking the definitions.

Implementation note: this is the Google Analytics 4 definition (Google, undated, Analytics Help, *Engaged sessions*; checked 15 October 2026). The help page we read gives no default and no allowed range for the timer, so we print none.

Server-side collection and offline conversion import (sending a sale made outside the website back to the ad platform) can recover events that browsers lose. They do not remove the need for consent. Anything that sends data about people to an ad platform needs counsel first (see the section on personal data).

Data quality as tests

A **data-quality test** is an automatic check on a table. It returns one row for each failure, so no rows means a pass. The tests run after every load. The kinds are these:

- unique keys: one event id means one event;
- referential integrity: every event points to a known deal;
- accepted values: every stage name is on the stage list;
- null rate: not more than a stated share of won deals without an amount;
- freshness: the newest row is not older than its limit;
- volume: the last day holds between half and twice the average of the seven days before it (a start value, Plausible but untested);
- drift: no new stage names and no new UTM values without notice;
- reconciliation: two sources agree on a total.

A **data contract** supports these tests. It is a written agreement between the team that produces a table and the team that uses it. It names the fields, their meaning, their format and the allowed values. It also says how long before a change the producer gives notice. Each test belongs to one contract and one owner.

Each test has a **severity**. An error blocks. The tiles that depend on the table go grey, and no rate, alert or test result is read until the error clears. A warning shows beside the tile and goes to the owner. The chapter *Decide when a change needs action* calls this block a gate. Tiered flags of this kind are common in large experimentation platforms (Thomke, 2020, *Experimentation Works*).

Reconciliation checks that two sources agree on a total. The **reconciliation gap** is the difference as a share of the primary source.

> **Calculation**

> reconciliation gap = |primary total – other total| ÷ primary total

Alder shows it with customer counts. In Year 1 January, billing counted 190 customers, the CRM 203 and marketing 214. The CRM gap is 6.8% and the marketing gap is 12.6% (Arithmetic). Four pairs deserve a standing check:

- the CRM amount against the first invoice, for each won deal;
- platform spend against warehouse spend;
- platform clicks against recorded sessions, which is the capture ratio;
- product sign-ups against CRM and billing records.

Each pair has its own tolerance. The start values are 1% for amounts and 1 day for freshness. They are Plausible but untested, so tune them on your own history.

A test that has never failed may page nobody. So replay a month with planted faults: a duplicated event, a missing deal, a stale load, a dropped day, an invoice that is 5% off. Each fault must show up in the tests.

Speed pays. Birch spends 165,704 euro on paid media in a window, which is 5,918 euro a day. A tracking fault that a weekly read finds leaves seven days of spend, 41,426 euro, without reliable measurement. A daily test bounds that to one day. This is exposure, not loss (Arithmetic).

Five metrics describe the health of the data itself.

Metric	What it measures	Formula
Data freshness	Age of the newest row at the cut-off	Cut-off time minus latest load time
Test pass rate	Share of tests that pass in the latest load	Tests passed ÷ tests run
Reconciliation gap	Disagreement between two sources	As above
Tracking coverage	Share of real actions that analytics saw	Events in analytics ÷ matching records in the system of record
Attribution coverage	Share of leads with a recorded source	Leads with a source ÷ all leads

Tracking coverage can rise above 1 when a tag fires twice. Treat that as a fault.

Implementation note: the companion code holds these tests as one SQL view that lists the failing rows. It also holds the seeded-fault tests (replays of history with planted data faults).

Attribution as a reporting convention

An **attribution model** is a rule that shares the credit for a conversion among the touches that came before it. First-touch gives all the credit to the first touch. Last-touch gives it to the last. Position-based gives most credit to the first and the last. Data-driven fits the credit from conversion paths, and it needs enough volume and enough recorded paths. The models answer different questions, so they disagree. None of them measures what a channel caused.

Treat the credit of a channel as a range. The lower bound counts only the conversions that carry a source. The upper bound adds every conversion that carries none. The width of the range is the unattributed share, which is 1 minus the **attribution coverage**.

Calculation

lower bound = conversions credited to the channel ÷ all conversions

upper bound = (conversions credited to the channel + unattributed conversions) ÷ all conversions

Birch carries the source with the consented session, so in its set-up the attribution coverage follows the capture ratio (Illustration only). The table shows the range of each paid channel at the two ratios of the example below. In Birch's company model the true shares are 24% for professional-network ads, 45% for search ads and 32% for social ads. A real company does not know them.

Channel	Range at 62% coverage	Range at 74% coverage
Professional-network ads	14.7% to 52.7%	17.6% to 43.6%
Search ads	27.6% to 65.6%	32.9% to 58.9%
Social ads	19.7% to 57.7%	23.5% to 49.5%

Share of all paid leads, 1,630 in the window of Year 1 October. Arithmetic on the company facts.

Where ranges overlap, the data cannot rank the channels at all. Here every pair overlaps at both levels of coverage. Where they do not overlap, you may say which channel is larger in the recorded data. You still do not move budget on that.

The rule of the book is short. Use one attribution model to compare periods. Use incrementality tests to move budget. Those tests measure what a channel caused, as the chapter *Choose acquisition channels* shows. A model used consistently gives comparability from one period to the next. It does not give causation. Consistent but biased credit still misallocates money.

Dashboards for decisions

A report shows what exists. A **decision dashboard** is built backwards from decisions. Every page serves a stated decision, and every tile names its owner and the action that follows when it moves. If nobody can name the action, the tile is a reference metric and leaves the page. A dashboard that nobody acts on is a dead end (Wexler, Shaffer and Cotgreave, *The Big Book of Dashboards*).

Pages by stage and role

Build one executive page that shows the bowtie, one page for each group of stages, and one page for data health. The executive page has one row for each bowtie stage.

Birch, executive page, Year 1 October, per 28-day window. The values are computed from the company facts; the activation share is an illustration value. The status column shows the four states a tile can take, with the calibrated limits that Birch adopts in Year 1 November.

Bowtie stage	Tile	Value	95% interval	Limit status	Owner
Awareness	Paid clicks	72,317	±527	inside	Marketing
Awareness	Click-to-lead, CRM leads over platform clicks	2.25%	±0.11 points	inside	Marketing

Bowtie stage	Tile	Value	95% interval	Limit status	Owner
Education	MQLs	1,255	±69	inside	Marketing
Selection	MQL-to-SQL rate	25.0%	±2.4 points	inside	Marketing and sales
Selection	SQLs	314	±35	inside	Sales
Selection	Win rate	32.0%	±7.5 points	too few to read in a window	Sales
Commit	New customers	48	±14	inside	Sales
Onboarding and adoption	Activated within 30 days	72%	±12.7 points	too few to read in a window	Customer success
Renewal	Renewal rate, logos	90.4%	±4.1 points	inside	Customer success
Expansion	Net revenue retention, 12 months	102.3%	read yearly	not judged in a window	Account managers
Data health	Consent capture ratio	74%	±0.3 points	outside, explained by the banner release	Analytics

The selling page of Birch shows where the hand-off between marketing and sales holds.

Birch, selling page, Year 1 October, per 28-day window.

Tile	Value	95% interval	Limit status	What the owner can do
MQLs	1,255	±69	inside	Marketing: change volume or rules
MQL-to-SQL rate, cohort basis	25.0%	±2.4 points	inside	Marketing and sales: review the hand-off
SQLs	314	±35	inside	Sales: plan capacity
SQL-to-accepted rate	80.0%	±4.4 points	inside	Sales: review acceptance
Accepted-to-opportunity rate	62.5%	±6.1 points	inside	Sales: review discovery
Seller load	104.6% (314 SQLs against 300 capacity)	exact	outside	Sales: capacity decision
Uncontacted SQL backlog	84	count	no limit set yet	Sales: contact or re-qualify

The tile

Every tile prints four things. The **value** names its window and its method, cohort or same-period. The **interval** shows the noise that the volume leaves. It comes from the binomial or Poisson rule of the chapter *Decide what the evidence can resolve*. The **n** gives the counts behind the value. The **limit status** says whether the value lies inside its limits, outside them, too few to read, or blocked by a failed test. A **tile limit** is the value at which the status changes. It depends on volume and noise, not on a fixed percentage. A flat 15% is blind at the top of the funnel, where a count of tens of thousands reads far smaller changes. It is noisy at the bottom, where a rate cannot read 15% in a quarter. The chapter *Decide when a change needs action* sets the limits from an expectation model (what each metric should read, given its level and season).

Pace tiles, the detail pane and usage

A **pace tile** compares progress in the current period with the usual path or with the plan, for example ARR so far this quarter. It shows the line of the usual path with its interval, from the forecast of the chapter *Set targets and staff capacity*. A flow metric needs it, because half a quarter of progress means little without the usual shape of the quarter (Wexler, Shaffer and Cotgreave).

Each tile opens a **detail pane**. The pane shows the specification: definition, numerator, denominator, source and owner. It also shows the goal line, a link to the underlying rows, and a strip of the periods in which the target was met. The strip separates a chronic problem from a one-off.

Usage is data too. Count the views of each page and tile by role, never by person. A tile that stays inside its limits for four quarters and triggers no decision moves to the reference list. A page that nobody opens for two quarters is retired. Both numbers are start values, Plausible but untested. Counting views by role may still need counsel and the works council before the log is switched on.

The data-health page lists the tests, their start limits, their severity and their state.

Birch, data-health page. The state column shows one illustrative day, the day of the banner release (Illustration only).

Check	Start limit	Severity	State
Freshness of CRM, billing and ad data	at most 1 day old	error	pass
Unique event keys	no duplicates	error	pass
Every event points to a known deal	no orphans	error	pass
Daily event volume	0.5 to 2 times the prior 7 days	error	pass
CRM amount against first invoice, per deal	within 1%	error	pass
Consent capture ratio	weekly limits from the last 13 weeks	warning	outside, change event logged
Attribution coverage	limits from the last 13 weeks	warning	outside, change event logged

Who reads what and when

The reader decides the rhythm. Data health is read every day, by the analytics lead, before anyone else opens a page. Stage pages are read weekly by their owners, in writing, before a decision. The executive page is read monthly by the bowtie owner. The specifications and the tile list are reviewed once a quarter. The chapter *Run the growth system* gives the full rhythm.

Reader	Page	Rhythm	Decision it supports
Analytics lead	Data health	Daily	Release or hold the data; open a data fix
Marketing owner	Awareness and education tiles	Weekly	Spend, creative, landing pages

Reader	Page	Rhythm	Decision it supports
Sales owner	Selling page	Weekly	Capacity, qualification, deal reviews
Customer success owner	After-the-close page	Weekly and monthly	Onboarding and renewal actions
Finance owner	ARR bridge, CAC, payback	Monthly and quarterly	Reconciliation, budget
Bowtie owner	Executive page	Monthly	Where to look, what to stop

Reading is asynchronous. An owner reads the page, writes the decision and its reason, and the decision goes into the log. A meeting appears only when two owners conflict or when no single owner can make the commitment.

Personal data

Personal data is any information about an identified or identifiable person. Online identifiers such as cookie ids can count. The General Data Protection Regulation (GDPR) is the EU law on this. It asks for a **lawful basis** before you process personal data. **Consent** is one such basis: a free, specific, informed and clear agreement, which the person can withdraw (European Union, 2016, *Regulation (EU) 2016/679 (General Data Protection Regulation)*, Articles 4, 6 and 7). Tracking in the browser usually rests on it.

Four rules follow for a measurement system.

- Design for aggregates and accounts. Do not build any view that points at one person.
- Link web sessions to leads only for consented sessions, and tell the visitor.
- Make erasure work. A person's data must be removable from the raw layer (the first warehouse layer, each source exactly as delivered): replace the link with a token, delete the link, and rebuild the tables.
- Report results by role. Involve the works council before you log activity that could monitor staff.

Ask counsel before you build anything that links analytics to a person, sends data about people to an ad platform, or logs workers or calls. This is not legal advice.

What changes with volume

The rules of this chapter hold at every volume. What changes is what a tile can show, and where the system is most likely to fail.

Regime	What a tile can show	Typical measurement failure	Where the effort goes
Low (Alder)	Counts with ranges, and lists of named accounts	A wrong or duplicated account, history that was never recorded	Keys, stage history, a few reconciliations
Medium (Birch)	Rates at the top and middle of the funnel; the bottom only by quarter	A measurement fault as large as a real effect	Populations, capture ratio, change log, tests
High (Cedar)	Almost every rate, within days	Too many readable metrics, and false alarms from their number	Fewer tiles, evidence grades, daily tests

Each regime reads a different number of tiles with confidence, so the tile list of a company changes as its volume changes.

Alder, Year 1 October: a few well-kept records

Alder, Year 1 October: about 3.8 new customers a window, 0.8 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.1 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

Since January, three teams have given three customer counts: 214, 203 and 190 (see *Choose the growth objective*). Billing is the reference. Marketing is 24 records above it, 12.6%, and the CRM is 13 above, 6.8%. This month three records go live: a stage history in the CRM, an account key, and a map from every billing customer to a CRM account. The analyst builds the map on the company domain first. Uncertain pairs go to the head of customer success, who knows the plants. Every surplus record gets a reason: a prospect never billed, a customer under a former name, or a subsidiary of a billed parent. The unmatched rows are counted and listed every day. A customer is now an account with an active, billed subscription.

One thing cannot be rebuilt. The stage history starts today, so no past stage dates exist. Cohort tables start now, and the first cohort of opportunities reaches maturity after 243 days, in about eight months. Until then, the cycle estimates of *Forecast conversion and timing* must pool deals by their open and close dates.

What belongs on screen? Volume decides. Alder has about 3.8 wins a window, and no rate below the lead reads within a year. A page of conversion rates would show numbers that move by chance. So the page holds counts and lists of accounts.

Alder, account page, Year 1 October.

Tile	Value	Range or n	How it is read
Paying customers, billing, by account key	213	exact count	Agrees with the CRM account list except for named exceptions
ARR at the start of the month	14.8 million euro	exact	The five movements are listed each month
Largest account, share of ARR	9.4%	one account	A written case when it changes
Renewal decisions in the next three months	about 40	expected count	Each contract listed with account, date, ARR and owner
Newly engaged accounts per window	45	80% range 37 to 54	A count
Opportunities opened per window	17	80% range 12 to 23	A count
New customers, January to September	about 33	80% range 25 to 40	Every win and loss has a written reason
Open reconciliation exceptions	listed	no range	Each has an owner and a date

Three tests run at every load. The account key is unique. Every billed customer maps to exactly one CRM account. Every won deal has a first invoice, and the two amounts agree within 1% (a start value, Plausible but untested). With 213 customers, a person could check this by hand. The tests still pay, because the largest account holds 9.4% of ARR, 1,386,000 euro. A wrongly keyed record there would move the retention tile more than any rate could.

The stage history records what happens to deals, not what a seller does. No tile shows a figure for one seller, and results go out by role. In Austria and Germany, two of Alder's six countries, a works council often has a say before anyone logs activity by person. This is not legal advice.

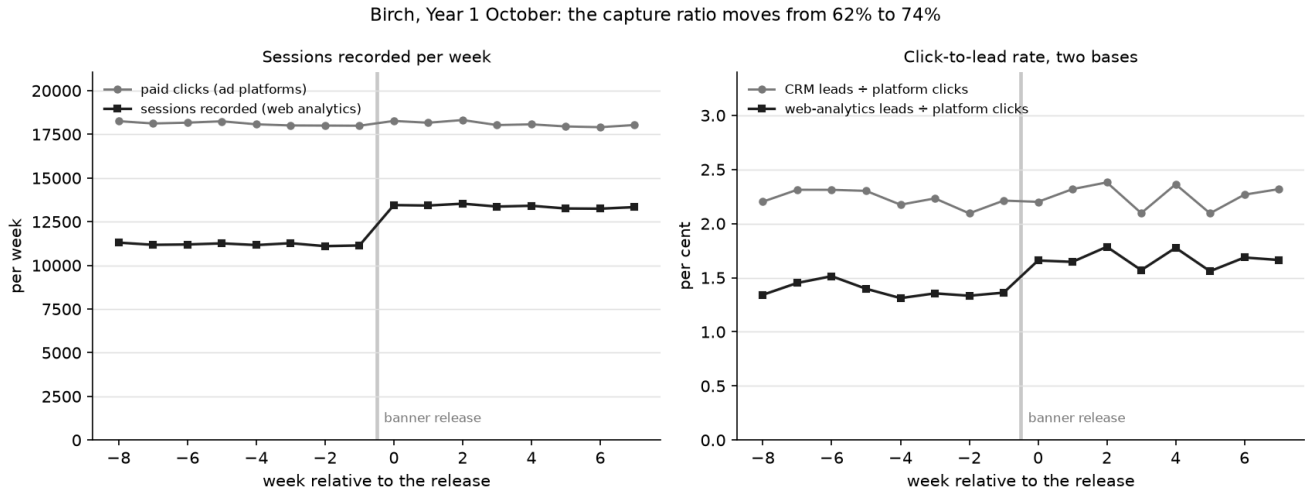
The chief executive, who owns the bowtie for now, decides. Eight tiles go into a written monthly note. No rate tile appears until volume supports one. Any change in the largest account gets a short written case. At low volume the measurement system is a few well-kept records.

Birch, Year 1 October: a rise that was a banner

Birch, Year 1 October: about 48 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 71 days, and the close rate reads a 20% change in about 6 months.

Early in October the web team releases a new consent banner. It is easier to accept than the old one, and the release is recorded as a change event with its date and the metrics it touches. Within weeks the weekly marketing note brings good news. Click-to-lead in web analytics has risen from 1.40% to 1.67%, a gain of 19.4%. The head of marketing credits the new landing pages.

The analytics lead checks the base of the rate. The dashboard counts the leads that web analytics recorded and divides them by the paid clicks that the platforms billed. The numerator exists only for visitors who consent. The denominator counts everybody. The banner moves the consent capture ratio from 62% to 74%. Recorded sessions per window rise from 44,836 to 53,514, and recorded leads rise by the same 19.4%, because 74 ÷ 62 is 1.194. No visitor behaves differently. The CRM counts every lead. Its click-to-lead is 1,630 leads over 72,317 clicks, 2.25%, before and after the release.



Birch around the banner release: recorded sessions and web-analytics click-to-lead step up, while the CRM rate stays flat (simulated weeks from the company parameters, fixed seed, Illustration only).
Birch has about 72,000 paid clicks a window, so click-to-lead reads a 20% change in about 7 days. A fault of 19.4% is large enough to read within a week. Its size cannot tell it from a real effect. Only the check of the base can.

The analytics lead proposes three changes.

1. One population. The headline tile becomes CRM paid leads over platform paid clicks, where both sides cover everybody. The second view, recorded leads over recorded sessions, stays as a diagnostic.
2. A capture-ratio tile on the data-health page, with weekly limits. The release is a change event, so the tile shows "outside, explained" and not an alarm.

3. Channel credit as a range. The source travels with the consented session, so the release also changes attribution coverage. Professional-network ads run from 14.7% to 52.7% of paid leads at 62% coverage, and search ads from 27.6% to 65.6%. After the release the ranges narrow by 12 points, and they still overlap. The data cannot rank the channels. The budget question waits for the incrementality tests of *Choose acquisition channels*.

The head of marketing decides. The claim of a landing-page gain is withdrawn. The new tile replaces the old one, and the capture-ratio tile joins the data-health page. Any click-to-lead comparison that crosses the release carries a note. At medium volume, measurement faults are as large as real effects.

Cedar, Year 1 October: nine dashboards, 25 tiles

Cedar, Year 1 October: about 1,600 new paying accounts a window, 1.4 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 28 days; the assisted motion behaves like a low-volume business.

Cedar's growth team of 12 and its two analysts keep nine dashboards that show 340 metrics. That is about 38 metrics a dashboard and 170 for each analyst. Nobody can say which of the 340 changes a decision. Adding a metric is cheap, and removing one needs an owner. The October review runs backwards from decisions. For each metric the analysts ask which decision it changes, who owns that decision, and what action follows.

Twenty-five metrics survive as tiles. Twenty are decision metrics and five are data-health tiles. The other 315 become reference metrics, which is 92.6% of the old list. They stay in the semantic model, and any of them opens from the detail pane, but none sits on a page. The 25 tiles fill three dashboards.

Dashboard	Decisions it serves	Tiles	Main readers
Growth	Spend by channel, activation, the assisted pipeline	11	Growth lead, assisted-sales manager
Revenue and retention	Retention, pricing, capacity, cost to win	9	Finance, customer success
Data health	Whether the other pages can be read	5	Analysts

The growth page holds eight self-serve tiles and three for the assisted motion. The self-serve tiles are spend by channel, the visitor-to-sign-up rate, sign-ups, activation by weekly cohort, time to activation, the paid rate, new paying accounts and value per sign-up by channel. The revenue page holds the ARR bridge, gross and net revenue retention, and survival by month of tenure. It also holds annual-plan renewal by renewal number, expansion, CAC and payback, gross margin and the assisted seller load. The health page holds freshness, test pass rate, reconciliation gap, tracking coverage and the capture ratio by country group.

Cedar's volume makes the growth tiles readable. Sign-ups per window carry a Poisson interval of about 423 around 46,642. Activated to paid, 9.5%, carries ±0.44 points. But the assisted tiles belong to a low-volume business. About 220 opportunities arrive in a quarter of 91 days, so the win rate of 25% carries ±5.7 points even then. Those tiles show "read by quarter" in their status.

Leading labels get a grade. Sign-ups are the leading label behind the 15 experiments a month, and they have no grade. In January, sign-ups rose while net new ARR stayed flat (see *Choose the growth objective*). So the sign-up tile is an early warning only. A decision on sign-ups alone waits for the outcome: paying accounts

that are still active after three months. The analysts plan a unit-level grade test on matured weekly cohorts, adjusted for channel and country.

The access log counts views by role: growth, finance, customer success and support. It never counts by person, and only role totals leave the log. The retirement rule takes its start values: four quarters for a tile and two quarters for a page (Plausible but untested).

The growth lead decides, and the dashboard list is fixed for the quarter. The risk at Cedar is not too little data. At high volume the risk is too many readable metrics.

The decision to carry forward

Collect enough history to reconstruct the decision and its alternative. Reconcile identities and quantities before judging performance. When measurement fails, repair the observation process before attributing the chart to customer behaviour.

17 · Decide what the evidence can resolve

Can the available evidence distinguish actions whose value differs enough to matter? Begin with the smallest commercial difference that would change your choice. Then work back to the observations and time required to resolve it. Counting events without defining that difference gives you a busy dashboard and no decision.

Sample size is one limit. Dependence within an account, delayed outcomes, changing customer mix and measurement error can reduce the information in a large table. A model can express uncertainty; it cannot manufacture the missing comparison.

The shared methods below connect noise, detectable effects and partial pooling. When the needed effect cannot be resolved in time, change the scope of the decision, gather another kind of evidence or make a limited commitment with explicit beliefs. Low volume changes the evidence you can obtain, not the need for economic reasoning.

Why rates are noisy

A rate is a count divided by a count. A win rate is the wins divided by the opportunities. If you run the same process twice, you get different counts, because chance decides which opportunities win. Two kinds of chance matter.

Binomial noise affects a count of successes out of a fixed number of tries. Wins out of opportunities is the standard case. **Poisson noise** affects a count of events in a period with no fixed number of tries, such as new customers in a window. In both cases the noise is large when the count is small and small when the count is large.

The **standard error** of a rate is the typical distance between the rate you measured and the true rate behind it.

> **Calculation**
>
> standard error of a rate = √(p × (1 − p) ÷ n)
>
> 95% interval = rate ± 1.96 × standard error
>
> Evidence label: Arithmetic.

The confidence interval, which the chapter *Model recurring growth* introduced, is the range of true values that the data cannot rule out. The standard error gives its width. Take Birch in Year 1 September. It has about 148 opportunities a window and a win rate of 32%. The standard error is 3.8 points. The 95% interval runs from 24.5% to 39.5%. A reported 28% in one window is therefore not a fall from 32%. It is inside the noise. Always print the interval or the count beside a rate.

For a Poisson count, the typical spread is the square root of the mean. Birch wins about 47.5 customers a window. The spread is 6.9, which is 14.5% of the mean. A window with 40 customers or with 55 customers is normal. Alder wins about 3.4 customers a window. The spread is 1.9, which is 54% of the mean. A window with five customers is not a good window. It is chance.

Two facts follow from the square roots. First, to halve the noise you need four times the events. Second, a rare event has large relative noise. A rate of 3% needs far more tries than a rate of 30% to show the same relative change.

Regression to the mean says that an extreme value tends to be followed by a less extreme one, by luck alone. We notice a segment because its rate is extreme. Part of the extremeness was luck, and the luck does not repeat. Suppose a segment of 40 SQLs shows a close rate of 25% while the pooled rate is 16%. If the true rates between segments differ by about 3 points (an assumption; Illustration only), the best forecast for the next quarter is 17.9%, not 25%. The section on small cells shows where that number comes from. It also explains why a fix applied after a bad window often seems to work: the window would have recovered anyway.

One readability function

The book uses one function to answer every question of the form "can this volume show this change?". The answer is the **readability** of a rate: how large a change must be before the volume can show it. The function has three modes and three multipliers. The modes differ in what you compare. The multipliers make the answer honest.

Mode	What you compare	Typical use
Two-arm test	Two groups that run at the same time, chosen by chance	Testing a change
Before and after	This window with the window before	A quick read of a change
Monitoring	This window with a baseline of k earlier windows	Watching for a fall

Each mode takes a baseline rate and gives one of three answers: the sample size for a change you care about, the smallest change that a given sample can show, or the days of flow needed to collect the sample. Statisticians call the smallest change that a design can see the **minimum detectable effect**, or MDE.

Two-arm test mode

In a two-arm test, units go by chance into a control arm and a test arm. Both arms run in the same weeks. Three choices set the size of the test. The first is the **significance level**: the chance that you call a difference real when it is chance. We use 5%. The second is the **statistical power**: the chance that the test finds a real difference of the size you care about. We use 80%. The third is the size of the change itself.

Under the **exact two-proportion formula**, the number of units in each arm is as follows.

> **Calculation**
>
> n per arm = $(1.96 \times \sqrt{(2 \times \bar{p} \times (1 - \bar{p})} + 0.84 \times \sqrt{(p_1 \times (1 - p_1) + p_2 \times (1 - p_2))})^2 \div (p_1 - p_2)^2$
>
> p_1 is the baseline rate, p_2 is the rate you hope for, and $\bar{p}$ is their average.
>
> 1.96 belongs to the 5% significance level (two-sided) and 0.84 to 80% power.
>
> Evidence label: Arithmetic, with the normal approximation that is Standard practice.

The formula uses two variances. The first term uses the variance when the two rates are equal, because that is the world in which a false alarm happens. The second term uses the variance of the two real rates, because that is the world in which you hope to find the change. A shortcut that uses one baseline variance for both terms is a labelled approximation. It under-sizes tests.

For a relative change r in a rate p, each arm needs about 16 × (1 − p) ÷ (p × r × r) units (the *rule of 16*; Standard practice). It comes from 2 × (1.96 + 0.84)², which is 15.7, rounded up. The chapter *Choose the growth objective* used it for a first look. Compare it with the exact formula for Birch, whose close rate is 16.0%.

Hoped change	New rate	Exact units per arm	Rule of 16	Rule is below exact by
+10%	17.6%	8,570	8,400	2.0%
+20%	19.2%	2,223	2,100	5.5%
+30%	20.8%	1,022	933	8.7%

Birch, close rate 16.0%, 80% power, 5% two-sided significance. Arithmetic.

Use the rule for a quick look and for the maps in this chapter. Use the exact formula to size a real test. The two functions in the companion code are exact inverses: the power of the test at 2,223 units per arm is 80.0%.

The weeks of a test follow from the flow, never from a guess.

> **Calculation**
>
> weeks = 2 × n per arm ÷ weekly flow of eligible units
>
> Evidence label: Arithmetic.

Birch has about 297 contacted SQLs a window at the trend level, which is 74.2 a week. We plan with the trend level because a long test runs through the seasons. Two arms of 2,223 need 4,446 SQLs. That is 59.9 weeks, or 15.0 windows, or 419 days. The outcome is a sale, so the last SQL of the test needs one more sales cycle to resolve. The 80th percentile of Birch's cycle is 56 days. The reading therefore comes after 476 days. Never state a test length as "two to four weeks". Compute it from the flow of the rate you want to read.

The same function runs in the other direction. If you split one window of Birch's SQLs in two, each arm holds 148 SQLs. That test can see a relative rise in the close rate only if it is at least 85%. A quarter of SQLs, split in two, can see a rise of 45%. Compute these limits before the test. Never compute the power after a result that disappointed you; observed power adds no information (McCullough, 2021, *Business Experiments with R*, section 4.4).

Before-and-after mode

In a before-and-after read, you compare this window with the window before. The difference of two noisy numbers is noisier than either one. Its variance is twice the variance of one window. Its standard error is √2, about 1.41, times the standard error of one window.

Take Birch's MQL to SQL rate. It is 25% on about 1,188 MQLs a window. One window has a standard error of 1.26 points. The difference between two windows has a standard error of 1.78 points. A difference must therefore exceed 3.5 points, which is 13.9% of the rate, before it separates from chance at the 5% level. Do not use the one-window spread for a two-window comparison.

Before-and-after is cheaper in time than a two-arm test. The baseline is already in hand, so you collect one period of units, not two. But it is weaker in a second way: nothing ensures that the two windows are alike. A season, a campaign or a price change moves the rate by itself. Use before-and-after only when the rate is stable, or when you have a change event (a dated record of a known change, such as a price rise) that explains the movement. The chapter *Forecast conversion and timing* shows how a same-period rate can fall without any change in selling.

Monitoring mode

A monitor compares the current period with a baseline built from k earlier periods of the same size. The baseline is also a sample, so its noise adds to the variance of the comparison.

> **Calculation**
>
> variance of the comparison = variance with a known baseline × (1 + 1 ÷ k)
>
> k = 1 gives a factor of 2 (the before-and-after case). k = 4 gives 1.25. k = 13 gives 1.077.
>
> Evidence label: Arithmetic.

The table shows the smallest fall that Birch's MQL to SQL rate can show in one window, with a one-sided test at 5% significance and 80% power. A monitor looks for falls only, which is why the test is one-sided.

Baseline	Variance factor	Smallest fall that one window can show
1 window	2.00	17.1%
4 windows	1.25	13.7%
13 windows	1.08	12.8%
Known exactly	1.00	12.3%

Birch, MQL to SQL 25.0% on 1,188 MQLs in the current window. Arithmetic.

A longer baseline helps, but only a little. After four windows, the current window carries most of the noise. To see smaller falls, you need more units in the period you watch, not a longer baseline. The chapter *Decide when a change needs action* builds the monitors that use this limit. A baseline also assumes a stable rate. A rate with a trend or a season needs a wider band.

Multipliers

Three multipliers change the sample that the formulas give.

Overdispersion is more variation than pure chance predicts. The binomial formula assumes that the true rate is the same in every window. In real funnels it is not. Campaigns start and end, a source of leads changes, and sellers take holidays. We can split the observed variance of a window rate into two parts.

> **Calculation**
>
> variance of a window rate = $p \times (1 - p) \div n + \tau^2$
>
> overdispersion factor $\phi = 1 + n \times \tau^2 \div (p \times (1 - p))$

τ is the real window-to-window wobble of the true rate. Multiply the variance and the sample size by ϕ. Multiply the width of an interval by √ϕ.

Evidence label: Arithmetic.

The first part shrinks as n grows. The second part does not. Take Birch's MQL to SQL rate with 1,188 MQLs. Suppose the true rate wobbles by 1 point from window to window (an assumption; Plausible but untested). Binomial noise alone gives a standard deviation of 1.26 points. With the wobble it is 1.61 points. The factor ϕ is 1.63. Intervals are 28% wider, and a before-and-after comparison needs 63% more units. A wobble of one point looks small. At high volume it is the largest part of the noise.

You estimate ϕ from your own history. For k windows, compute the chi-square statistic of equal rates and divide by k − 1. With 13 windows the estimate is rough: its relative spread is about 41%. Pool similar rates, or use 26 or more windows, when you can. Until you have an estimate, treat the binomial figures of this chapter as a best case. A randomised test that runs both arms at the same time cancels most of the shared wobble. Before-and-after and monitoring do not.

The design effect applies when units come in clusters. Users sit inside workspaces, contacts inside accounts, and leads inside companies. Members of one cluster resemble each other. If you assign a whole cluster to one arm, a thousand users from 200 workspaces carry less information than a thousand independent users. The **intra-cluster correlation**, ρ, measures how alike members are: 0 means they act independently, and 1 means they act as one.

Calculation

design effect = 1 + (m − 1) × ρ

m is the average number of units per cluster. Multiply the sample size by the design effect. Multiply the width of an interval by its square root.

Three contacts per account give 1.2 at ρ = 0.1 and 1.6 at ρ = 0.3.

Evidence label: Arithmetic.

Measure m and ρ in your own data. They are not constants. The factor applies when the cluster is the unit that gets the treatment. If you assign users one by one inside a workspace, a different problem appears: teammates see different variants, and the arms influence each other.

Covariate adjustment works in the other direction. If you know a baseline variable that predicts the outcome, such as past usage, you can remove part of the noise. When the correlation between covariate and outcome is r, the variance falls by the factor $1 - r^2$. A correlation of 0.5 gives 0.75, and 0.7 gives 0.51 (Deng, Xu, Kohavi and Walker, 2013, *Improving the sensitivity of online controlled experiments by utilizing pre-experiment data*).

The three multipliers combine by multiplication: sample size = exact n × ϕ × design effect × $(1 - r^2)$. Binomial figures are a floor on what you can read. Real intervals are wider because of overdispersion, clustering and drift, and narrower only when a covariate helps.

Implementation note: the companion code holds these functions in its sizing module.

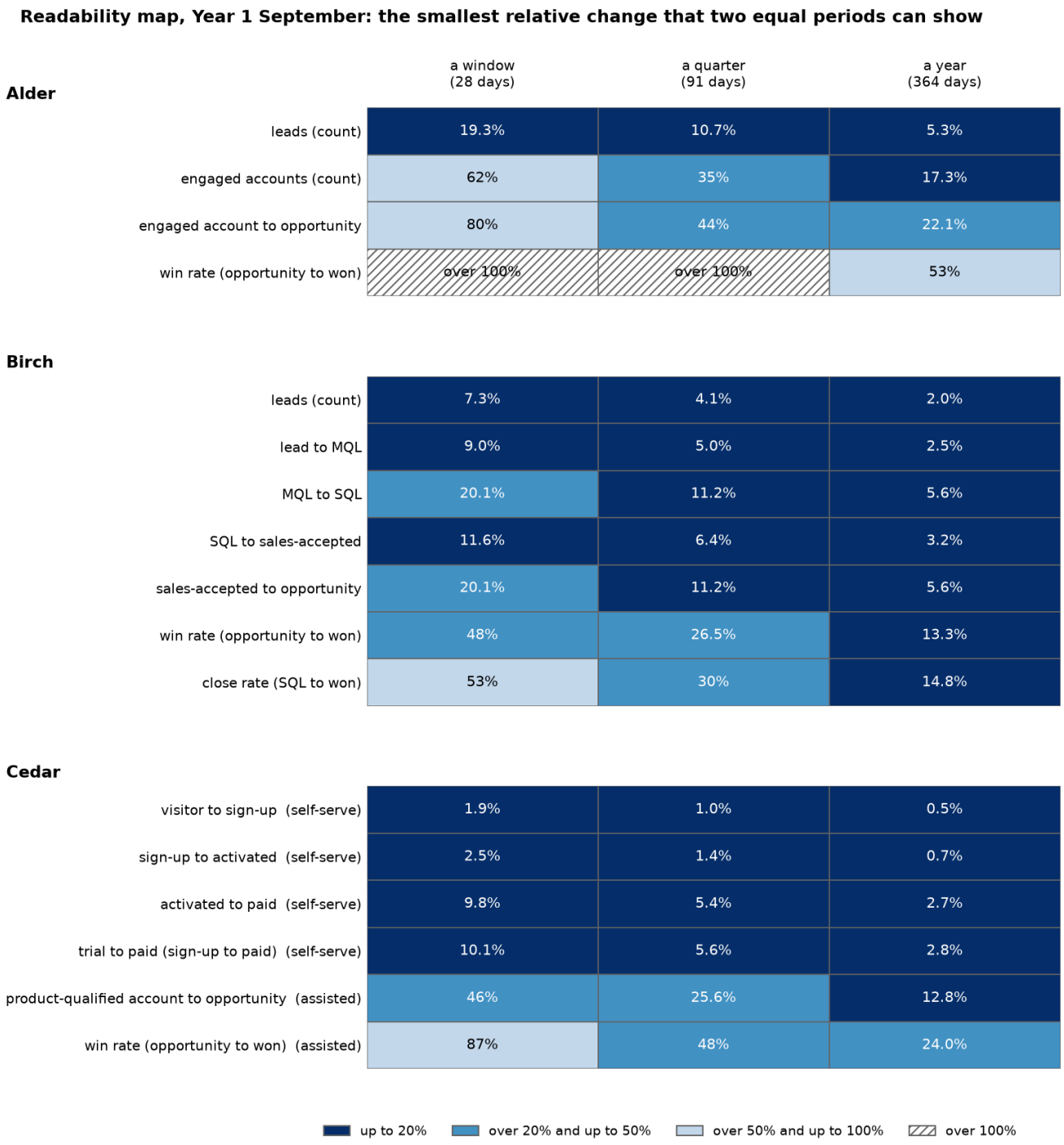
The readability map

The **readability map** is a table of the smallest change that each stage rate can show in a window, a quarter and a year. It is the full form of the barometer. Each cell comes from the same arithmetic: two equal periods, 80% power, 5% two-sided significance, and the rule of 16.

The map, the regime and the barometer all use the rule of 16 as a labelled convention, so they stay comparable. A real test, or a comparison of two named groups, uses the exact formula.

> **Calculation**
>
> smallest relative change = 4 × √((1 − p) ÷ (p × n)) for a rate p with n units in each period
>
> smallest relative change = 4 ÷ √n for a count of n events in each period
>
> Evidence label: Arithmetic, on the rule of 16 (Standard practice). For changes up to 30%, the smallest change it gives is within about 5% of the exact one, because the change goes with the square root of the sizes in the table above. Cells that show more than a 30% change are rougher.

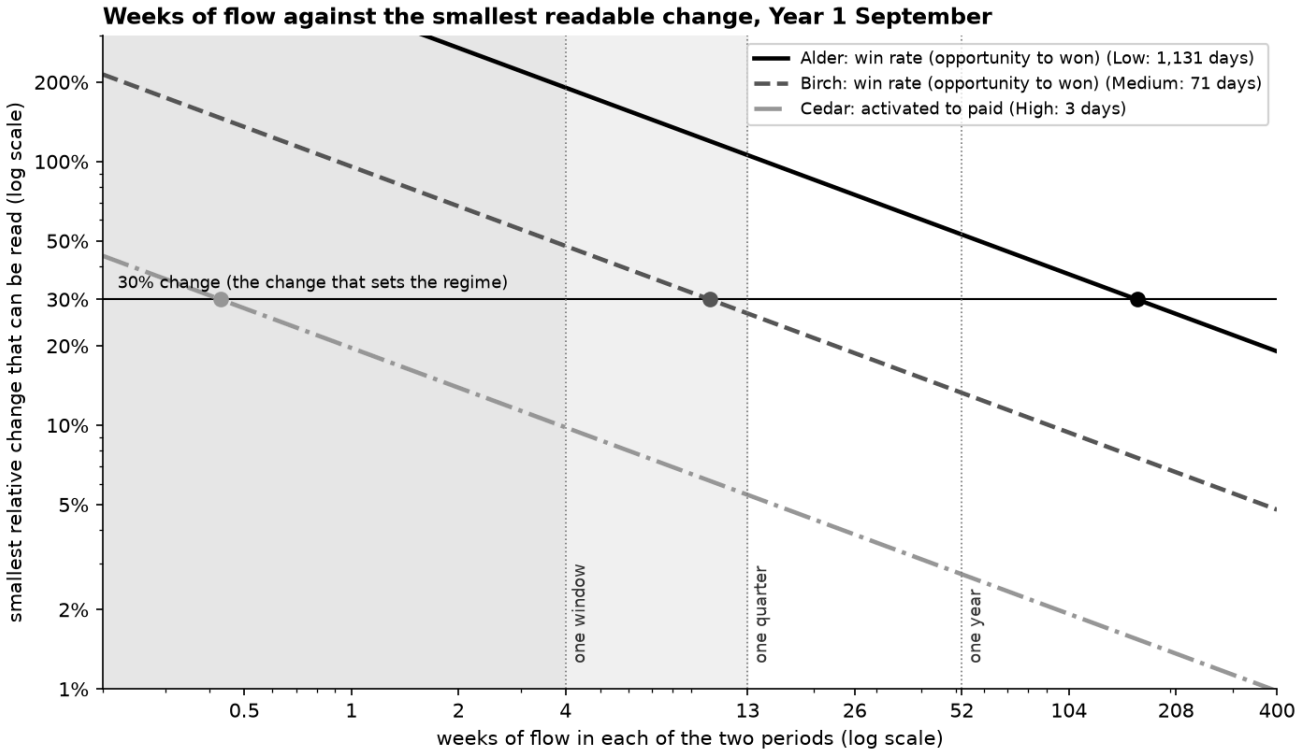
A window is 28 days. A quarter is 91 days, which is 3.25 windows. A year is 364 days, which is 13 windows. For each period, multiply the units per window by the number of windows.



The readability map for Alder, Birch and Cedar in Year 1 September, at the trend level of flow. Each cell shows the smallest relative change that two equal periods can show. Arithmetic.

Birch's top stages are readable in a window: leads change by 7.3% and the lead to MQL rate by 9.0%. Its MQL to SQL rate needs a quarter for a change under 20%: it reads 20.1% in a window and 11.2% in a quarter. The win rate reads 48% in a window, 26.5% in a quarter and 13.3% in a year. Alder reads little. Only its lead count reads a 20% change in a window. Its win rate reads only a change of 53% in a year, and no change smaller than 100% within a quarter. Cedar's self-serve rows read changes of about 10% or less within one window, while its assisted win rate reads 24% in a year.

Three cautions apply to the map. First, the map shows what the volume can read, not what the decision needs. The change that matters is the smallest one that pays for the action. The chapter *Commit with uncertainty* computes it. Second, a rate that ends in a sale resolves only after a sales cycle, so add one cycle to the time in the map. Third, the map counts independent units. At Alder, contacts cluster inside accounts, so the rates are counted on accounts. Counting leads would make the map too kind.



The smallest readable change against weeks of flow, for the last stage rate before revenue of each company. The dots mark where each curve reaches a 30% change. Arithmetic.

A company with a hundred times more units reads changes ten times smaller, not a hundred times smaller.

Volume regimes and the barometer

The map has a cell for every stage. Often we need one label for how much volume a whole company has: the **volume regime**.

The definition

The regime reads one rate. It is the last stage rate before revenue in the main motion of the company. For a sales-led motion, that rate is the win rate, from opportunity to won. For a product-led motion, it is the rate from activated to paid. We choose this stage because it has the fewest events. If it is readable, the stages above it are readable too, as long as their rates are not very small.

We count the days of flow that this rate needs to read a 30% relative change.

> **Calculation**
>
> days to read = 28 × n required ÷ n per window

> n required = 16 × (1 − p) ÷ (p × 0.30 × 0.30), the units needed in each of two equal periods
>
> High volume: 28 days or fewer. Medium volume: more than 28 and up to 91 days (one quarter, 13 weeks). Low volume: more than 91 days.
>
> Evidence label: Arithmetic for the days, on the rule of 16 as the book's convention for readability. Plausible but untested for the 30% and the two limits. Each company sets its own.

At a win rate of 25%, high volume starts at about 533 opportunities a window, which is 133 wins. Medium volume starts at about 164 opportunities, which is 41 wins. Below that the company is low. The regime follows the rate and the volume together. A low rate needs more units.

Company and motion	Units into the stage per window	Rate	Days to read a 30% change	Regime
Alder, sales-led	15.6 opportunities	22.0%	1,131 (3.1 years)	Low
Birch, inbound	148.5 opportunities	32.0%	71	Medium
Cedar, self-serve	15,806 activated accounts	9.5%	3	High
Cedar, assisted	63.9 opportunities	25.0%	234 (about 8 months)	Low

Year 1 September, trend level. Arithmetic.

Cedar, one company with two motions, shows why the rule names the main motion. The self-serve motion has most of the new customers, so Cedar is high. The assisted motion is a low-volume business inside a high-volume company, and it is read apart. As the chapter *Choose markets and accounts* showed, volume belongs to a decision, not to a company.

The choice of 30%, 28 days and 91 days is a convention. A different change, or a different comparison, moves the borders. The table shows the opportunities per window that a rate of 25% needs to read a change in one window. It uses one formula: units = 8 × m × (1 − p) ÷ (p × r × r), where r is the relative change and m is the variance factor of the comparison. The constant 8 is half of $2 \times (1.96 + 0.84)^2$, rounded up.

Relative change	Known baseline (m = 1)	13-window baseline (m = 1.08)	Previous window (m = 2)	50/50 split in one window (m = 4)
20%	600	646	1,200	2,400
25%	384	414	768	1,536
30%	267	287	533	1,067
40%	150	162	300	600

Opportunities needed in one window at a rate of 25%. Arithmetic.

The regime definition uses the previous window (m = 2). It is therefore too kind for a test with two arms in one window, which needs twice the units, and too strict for a monitor with a long baseline. Say which comparison you mean whenever you move between testing and monitoring.

Readability per decision

For each decision, ask: *can this volume read the change that matters, within the time we can wait?* The answer is one of four classes. The change reads within a window (28 days or fewer), within a quarter (91 days or fewer), within a year (364 days or fewer), or not within a year. A stage that is not readable within a year cannot guide a decision by itself. The last section of this chapter says what to do then.

The barometer

The **barometer** states the volume of a company in one short sentence. Every example in this book opens with one. Five steps build it, all from the company's own counts.

1. Pick the last stage rate before revenue in the main motion.
2. Count the units that enter that stage in a window, at the trend level, and compute the rate.
3. Compute the days to read a 30% change, and name the regime.
4. Compare the new customers per window with the derived typical number for the company's ARR and contract size. It comes from published medians (Benchmarkit and Pavilion, 2024, *2024 B2B SaaS Performance Metrics Benchmark Report*). The chapter *Choose the growth objective* shows how.
5. Name the deepest stage that reads a 20% change within a window, a quarter and a year, using the map.

The sentence has at most 30 words. It names five things: the company and the date, the new customers per window, their position against the derived typical, the regime, and the deepest stage that reads a 20% change. The period for that reading fits the regime: a window at high volume, a quarter at medium volume and a year at low volume. For a self-serve company, a new customer is a new paying account. Further detail goes into one short second sentence. Here is Birch in Year 1 September:

> Birch, Year 1 September: about 48 new customers a window, 1.0 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate.

The second sentence adds the days: "The win rate reads a 30% change in 71 days, and the close rate reads a 20% change in about 7 months."

A rate can also be compared with a published range. Birch's win rate of 32% lies in the upper half of the published span of 20% to 40%. Its MQL to SQL rate of 25% lies in the lower half of a span from 13% to 51%. These are the vendor spans that *Choose markets and accounts* discusses (First Page Sage, 2025, *B2B SaaS funnel conversion benchmarks*, and secondary sources). The published sources differ in how they define each stage, they sample companies that chose to take part, and they conflict with each other. We use a span, never a target.

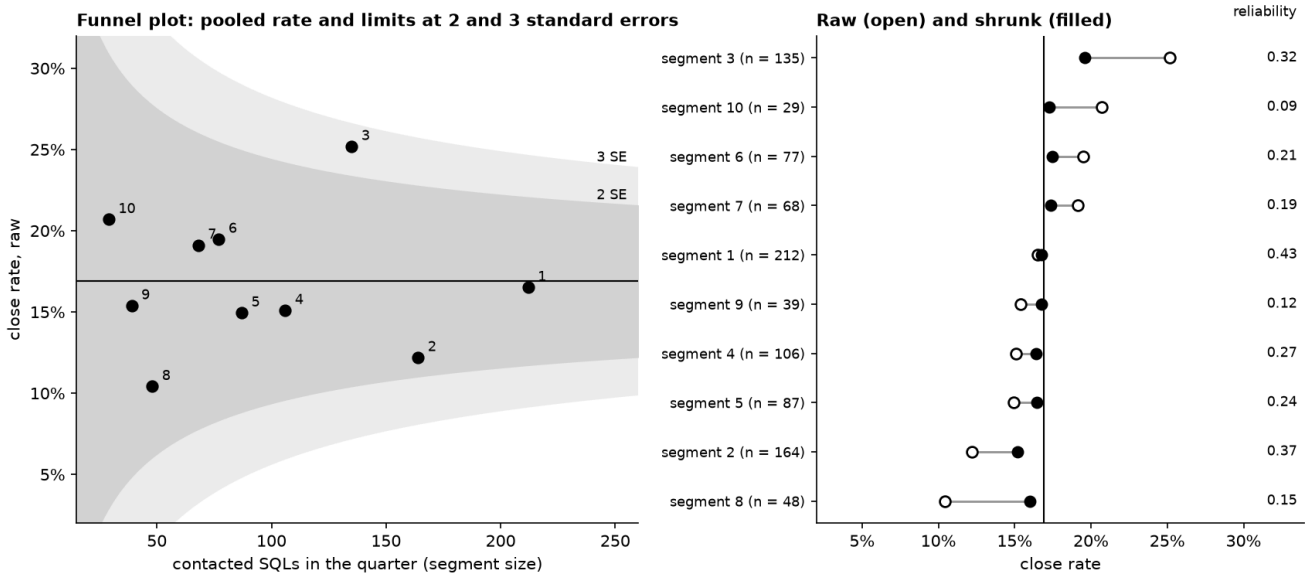
Two more cautions apply to every regime. Money metrics, such as contract value, are skewed, and they need larger samples than rates do (Kohavi and others, 2014, *Seven Rules of Thumb for Web Site Experimenters*, KDD, as cited in Thomke, 2020, *Experimentation Works*). Test them on your own data. And the days in the regime are days of flow. They do not include the sales cycle.

Small cells: partial pooling and funnel plots

Sellers, segments and channels are small cells. Each one holds only a part of the volume. If you rank them on raw rates, you rank noise. Chance alone spreads the rates of equal cells. The smallest cells spread the most, so they fill both ends of the league table.

The funnel plot

A **funnel plot** shows this directly. Plot each cell as a dot. Put its size on the horizontal axis and its raw rate on the vertical axis. Draw the pooled rate as a line. Around it, draw limits at 2 and 3 standard errors. The limits are wide for small cells and narrow for large ones, so they form a funnel. A dot inside the limits differs from the pool only by chance (Spiegelhalter, 2005, *Funnel plots for comparing institutional performance*, Statistics in Medicine, volume 24).



Ten customer segments of Birch's contacted SQLs in one quarter, with an assumed true spread of 3 points between segments. Left: funnel plot. Right: raw and shrunk close rates, with the reliability of each segment. The segment data are simulated. Illustration only.

The left panel shows ten segments from 29 to 212 SQLs. One dot, the segment at 25.2%, lies outside the 2-standard-error limits, against a pooled rate of 16.9%. Is it real? With ten dots, the chance that at least one lies outside the 2-standard-error limits is 37%, by chance alone. The better question is about the group. The **omnibus test** asks whether all cells could share one true rate. For these ten segments it gives a p-value of 0.21, so the spread is not larger than chance would make it. Use the omnibus test as the trigger to look for a difference. Use the funnel plot to see where it sits. Do not start with the largest gap.

Shrinkage

When a cell is small, its own data say little. We can borrow information from the other cells. **Partial pooling** pulls each raw rate toward the pooled rate by exactly as much as its size warrants. The pull is called **shrinkage**.

> **Calculation**
>
> shrunk rate = pooled rate + reliability × (raw rate - pooled rate)

> reliability = $\tau^2 \div (\tau^2 + \sigma^2)$
>
> τ^2 is the real spread between cells, and σ^2 is the noise within a cell, which equals $p \times (1 - p) \div n$.
>
> Evidence label: Standard practice. The method is empirical Bayes (Berger, *Statistical Decision Theory and Bayesian Analysis*, section 4.5).

The **reliability** is the share of the spread between cells that is real. A large cell has low noise and a reliability near 1, so its own rate counts. A small cell has a reliability near 0, so its rate is mostly the pool. We learn τ^2 from all the cells together. The companion code fits it with a beta-binomial model, which assumes that each cell has its own true rate, drawn from one shared spread of rates (a beta distribution).

In the segment example, the fit estimates the real spread at 2.2 points, a little below the 3 points that made the data. So the reliability runs from 0.09 for the smallest segment to 0.43 for the largest. The segment at 25.2% shrinks to 19.6%. The segment with the lowest raw rate, 10.4% on 48 SQLs, rises to 16.0%. Take a true spread of 3 points and a pooled rate of 16%. Then a cell of 29 SQLs has a reliability of 0.16, a cell of 87 has 0.37, and a cell of 212 has 0.59.

Does it help? We simulated 1,000 quarters of ten such segments. The average error of the raw rates was 3.3 points. The average error of the shrunk rates was 2.2 points. But picking the single best segment stays hard. The top raw segment was truly the best in 34% of quarters, and the top shrunk segment in 36%. So read shrinkage as a better estimate for each cell, not as a way to name a winner.

Three rules follow. Rank on the shrunk rate, never on the raw rate. Print the reliability beside every table. And shrink only cells that are alike. A segment that differs for a known reason, such as a different contract size, is not exchangeable with the others. Adjust for the reason first. Shrinkage also explains regression to the mean: the expected next value is the shrunk value. The chapter *Commit with uncertainty* uses it to correct the winner's curse of tests (the best result is overstated, because it was picked from many noisy results).

Implementation note: the companion code holds the shrinkage fit, the funnel limits (with an optional overdispersion factor) and the omnibus test in its pooling module.

When the cells are people

Sellers are cells too, and they are people. Keep to three rules. This is not legal advice.

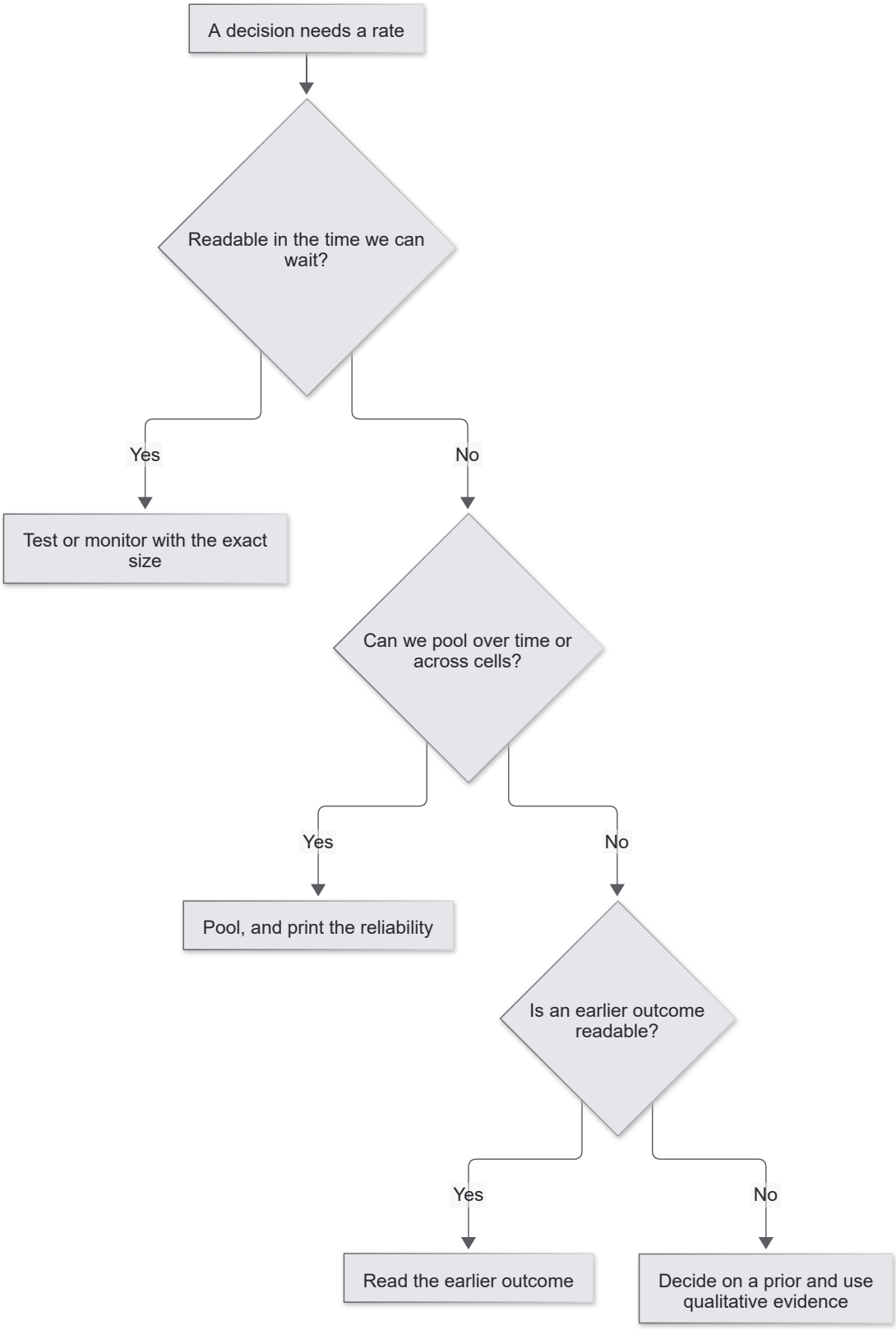
- Report results by role. Show the account executive role as a group, never one named person.
- Do not build a league table of individuals, raw or shrunk. A person who sees a rank will act on it, and the rank is mostly chance.
- Counsel must review any view of one worker before it exists. In Austria and Germany the **works council**, the elected body that represents employees, often has a say on systems that can monitor staff.

A shrunk estimate does not remove the issue. It protects people from a false ranking, but it is still output about one person.

When nothing is readable

Sometimes the stage you need is not readable in the time you can wait. This is normal at low volume. Five moves remain, in this order.

1. **Pool over time.** Use a longer window. Alder needs 40 windows of opportunities, 3.1 years, in each of two periods to read a 30% change in its win rate. A longer period risks drift and mixes old and new conditions.
2. **Pool across cells.** Use partial pooling, or merge similar segments. You learn the average and lose the differences.
3. **Read an earlier outcome.** Choose a stage that has more events and that the action can move, such as a meeting booked. You need evidence that it predicts the later result. The chapter *Decide what to measure* grades that evidence.
4. **Decide on a prior.** Use outside base rates and your own judgement, write them down, and set signposts. The chapters *Commit with uncertainty* and *Choose whether to test or act* show how.
5. **Use qualitative evidence.** Read the deals, the calls and the notes. They cannot give a rate, but they give mechanisms.



What changes with volume

The mechanics hold at every volume. The conclusions do not.

Regime	What you can read	What you do
Low (Alder)	A 20% change in the lead count in a window, and in engaged accounts in a year. No rate below the lead within a year.	Decide on priors and cases. Pool across years. Never compare individuals.
Medium (Birch)	A 20% change down to the sales-accepted to opportunity rate in a quarter. The win rate and the close rate only within a year.	Test only the stages that read. Move the read of late stages to earlier outcomes.
High (Cedar, self-serve)	A 20% change in every self-serve rate within a window, many in days.	The limit is no longer sample size. It is overdispersion, clustering and bias.

At low volume, the question is whether to decide without a reading. At medium volume, it is which stage to read. At high volume, it is whether the interval is honest. One company can hold two regimes: Cedar's assisted motion is low volume. A high-volume company with a narrow interval that ignores clustering is more wrong, and more sure, than a small company that knows it cannot tell.

Alder, Year 1 September: a ranking that chance wrote

Alder, Year 1 September: about 3.6 new customers a window, 0.8 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.1 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

The head of sales asks for a ranking of the four account executives by win rate over the last year. He wants to coach the lowest and to copy the highest. The analyst checks first whether the data can support a ranking at all.

The four sellers closed 44 wins from 198 opportunities, which is 22.2% (the split by seller is simulated; Illustration only). The interval for the role is 16.4% to 28.0%. Each seller had about 50 opportunities. At that size the standard error of one seller's win rate is 5.9 points. The gap between the highest and the lowest raw rate is 12.1 points, which looks large. But four noisy numbers always spread. The omnibus test asks if the four could share one true rate. The p-value is 0.53: four sellers with one true rate would show a spread this large or larger in about 53 years out of 100. No seller lies outside the 2-standard-error limits of the funnel plot. The largest distance is 1.1 standard errors.

The analyst does not claim that the sellers are equal: with four cells the test is weak, and a fitted spread near zero is only an absence of evidence. So the analyst asks how much a seller's rate would tell if real differences existed. Over a year, with a true spread between sellers of 3, 5 or 8 points, the reliability is 0.20, 0.42 and 0.65. Over a quarter it is 0.06, 0.15 and 0.31. Even with a large true spread, a quarter's league table is mostly noise.

How long would a fair comparison take? The analyst sizes it with the exact two-proportion formula, as for any two named groups. Two sellers who each collect a year of opportunities can show a relative difference only if it is more than 120%. The rule of 16 would give a smaller figure, because it under-sizes a two-group comparison. To tell a true 22% from a true 27%, each seller needs 1,161 opportunities, which is 23 years at 49.5 a year. To tell 15% from 27% each needs 180, which is 3.6 years. Alder's model shows about 201 opportunities a year in all. At low volume, an individual comparison is noise even over a year.

The analyst reports by role only. The four rows stay in the working file, unprinted and unranked. The report to the head of sales gives the role's win rate with its interval, the funnel plot with no dot outside the limits, and the reliability table. It says that any view of one person needs counsel first, and that in Austria and Germany a

works council often has a say. This is not legal advice.

The head of sales decides. He drops the ranking. Coaching rests on deal reviews and on the input activities of the whole role, read by role, such as meetings held and discovery questions covered. The owner writes the decision and its date in the forecast note, because Alder has no decision log yet.

Birch, Year 1 September: a test that needs fifteen windows

Birch, Year 1 September: about 48 new customers a window, 1.0 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 71 days, and the close rate reads a 20% change in about 7 months.

The head of sales has a new demo script and wants a randomised test on the close rate, hoping for a 20% lift, from 16.0% to 19.2%. The analytics lead runs the readability function before anyone writes a plan.

The exact formula gives 2,223 SQLs per arm and 4,446 in all. At 74.2 contacted SQLs a week, that is 15.0 windows of flow. With one sales cycle at its 80th percentile of 56 days, the reading comes after 476 days. The rule of 16 gives a similar picture: 4,200 SQLs, 14.1 windows. A test that ends in more than a year runs through a season, a price change and probably a new product release. It cannot give a clean answer.

A split of one window would be worse. With 148 SQLs per arm, the test sees a relative lift only if it is at least 85%. The script will not lift the close rate by 85%.

The analytics lead looks for an earlier outcome that the script could move. A demo turns a sales-accepted lead into an opportunity. That step has a rate of 62.5% on 238 leads a window. A 20% lift would take it to 75.0%, and the exact size is 215 per arm. A two-arm test needs 7.2 weeks, which is 51 days. That is readable. But the earlier outcome helps only if two things are true. The script must act on that step, and a higher step rate must lead to more wins. The analytics lead notes both as assumptions to check, and the chapter *Decide what to measure* explains how to grade such a link.

For comparison, the win rate itself needs 873 opportunities per arm, which is 11.8 windows. In a two-arm test, Birch's volume reads the middle of the funnel in weeks and the end of the funnel in more than a year.

The head of sales decides. The script goes to a two-arm test on the opportunity step, for eight weeks. The close rate is read by cohort as the deals mature, for the record, and no decision waits on it. The owner writes the expected effect in the test note before the test starts. The chapter *Choose whether to test or act* shows how to size such a choice by its value.

Cedar, Year 1 September: intervals that were too narrow

Cedar, Year 1 September: about 1,600 new paying accounts a window, 1.4 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 29 days; the assisted motion, with 67 opportunities a window, needs about 8 months to read a 30% change.

Cedar's register has just passed 30 well-recorded tests: it holds 32. The analytics lead reviews it and checks the unit of each test. A set-up flow for new workspaces was tested on user activation, which is 36% at Cedar. The flow is a workspace setting, so the test assigned whole workspaces to an arm. But the analysis counted

users and used the standard binomial interval. The intervals look too narrow, and the analytics lead asks why. The question matters more now: with 32 tests, Cedar starts to use its own register as the prior, in place of outside base rates. Wrong intervals in the register would pass into every later prior.

The cause is the design effect. A workspace holds on average 4.2 users, and the activation of two users in one workspace has a correlation of 0.06 (Illustration only). The design effect is 1 + 3.2 × 0.06 = 1.19. The true variance is 19% larger than the analysis assumed, and the true interval is 9.2% wider. A test at the 5% level with the naive interval gives a false alarm in 7.3% of runs with no real effect.

Take a 14-day test. Cedar has about 43,900 sign-ups a window at the trend level, so each arm holds about 10,980 users. Taken as independent users, the test shows a lift of 5.1% or more. With the design effect, the smallest lift is 5.5%. To read a 3% lift, each arm needs 31,207 users if users were independent, which is 39.8 days. With the design effect it needs 37,199 users per arm, which is 47.4 days. For a 5% lift in trial to paid (3.4% of sign-ups), a two-arm test needs 232 days, or 276 days with the design effect. The barometer says that Cedar reads a 10% change in this rate in 29 days. That reading compares two periods and uses the previous one as it stands. A two-arm test fills both arms from new flow, and half the change needs four times the units. So a 5% test takes most of a year, with or without clustering.

Overdispersion adds a second factor, because activation also wobbles from day to day. If that factor were 1.3 (an assumption; Plausible but untested), the combined variance factor would be 1.55 and the intervals 24% wider. The analytics lead plans to estimate the real factor from 26 weeks of A/A runs, which compare two identical arms.

More volume cannot repair this, and the problem is not only variance. If teammates share a workspace but see different variants, the arms contaminate each other, and the measured effect moves toward zero. That is bias, and bias does not shrink with n.

The growth lead decides. The test register records, for each test, the unit that was randomised and the unit that was analysed. Analysis of a workspace-level test uses workspaces as the unit, or applies the design effect. Sample sizes are multiplied by the factor before a test starts. Tests that were analysed by user are listed, and their intervals are recomputed. At high volume, the binding constraint is no longer sample size. It is whether the interval is honest.

The decision to carry forward

State the effect that would change the decision and whether it is observable in the available time. When it is not, reduce the commitment or use other evidence with explicit assumptions. Greater model complexity does not create additional independent outcomes.

18 · Choose whether to test or act

Should you buy more information before acting? An experiment is an investment that can improve a later choice. Its value depends on whether the result could change that choice, what an error would cost and how much benefit is lost while waiting.

Define the commercial action before designing the test. The eligible population, treatment, alternative and outcome must match the action you may later take. A test of a page on existing trial users cannot automatically establish the value of acquiring a new audience.

This chapter connects experimental design with the value of information. It keeps decision uncertainty in the calculation rather than requiring proof for every unknown. A precise result about an irrelevant outcome can be worth less than a coarse result that changes the funding decision.

The value of information

A test is a purchase of evidence. The price has three parts: the direct cost of people and tools, the harm to units that get the worse version, and the delay. The **value of information** is the expected gain from learning something before you decide, net of the cost and the delay of learning it. Run the test only when that number is above zero.

The **expected value of perfect information**, EVPI, is the most that any test could be worth. It is the gain from knowing the true effect before you choose. The **expected value of sample information**, EVSI, is what a test of a given size is worth on average. It is lower than EVPI, because a sample is noisy. Information never has a negative value, and a test that cannot change the choice has an EVSI of zero (*Bayesian Networks and Influence Diagrams*, chapter 11).

The third part of the price is the delay. The **cost of delay** is the gross profit that a change loses for each week that it starts later. A test holds back every gain that it says to take until its result is in. The chapter *Fund the next best action* uses the same measure to order scarce work, and treats it in full. The **expected net gain of a test** puts the three parts together.

> **Calculation**
>
> EVPI = expected value of the best choice with full knowledge − expected value of the best choice now
>
> EVSI = expected value of the best choice after the test result − expected value of the best choice now
>
> expected net gain of a test = EVSI − cost of the test − cost of delay a week × weeks until the result
>
> Run the test at the size where the net gain is highest, and only if that gain is above zero. Otherwise decide now.
>
> For a normal prior on the effect (mean μ, spread σ), a value V per unit of effect and a cost K: $EVPI = V \times \sigma \times L(z)$, with $z = |V \times \mu - K| \div (V \times \sigma)$ and $L(z) = \phi(z) - z \times (1 - \Phi(z))$, where ϕ and Φ are the density and the distribution function of the standard normal. For EVSI, replace σ by $s = \sigma^2 \div \sqrt{(\sigma^2 + se^2)}$, where se is the standard error of the test's estimate.
>
> Evidence label: Arithmetic, with the normal approximation, which is Standard practice (Berger, *Statistical Decision Theory and Bayesian Analysis*, chapter 7).

A worked case shows the parts. Cedar, in Year 2 March, considers a rebuild of its onboarding. The question is whether to test it first. Table 1 lists the inputs.

Table 1. The inputs of the worked case. Cedar, Year 2 March.

Input	Value	Where it comes from
Gross profit of 12 months of extra paying accounts, per 1% relative lift in activation	321,580 euro	Arithmetic on Cedar's facts, by the method of *Find the most valuable constraint*
Cost of the rebuild	150,000 euro	The start value that *Find the most valuable constraint* used (Illustration only)
Break-even lift in activation	0.47%	Cost ÷ value per unit of lift
Prior on the lift	mean 0%, spread 3%	Cedar's own register of 68 tests
Cost of one test	4,318 euro	10 days of analyst time at the analyst cost line over 220 working days (Plausible but untested)
Sign-ups a week	11,771	Cedar's facts, trend level
Wait for the outcome	2 weeks	Activation is read at the end of the 14-day trial

The prior says that a typical change moves activation by nothing on average, with a spread of 3%. At that prior, the lift is below the break-even of 0.47% in 56% of cases, so the rebuild has a probability of loss of 56%.

EVPI is 314,518 euro. An oracle that told Cedar the true lift would be worth that much, because Cedar would rebuild only when the lift exceeds the break-even. No test can be worth more.

Now vary the size of the test. A larger test has a smaller standard error, so its EVSI is closer to EVPI. But it runs longer. For each week until the result, Cedar loses a week of the gain in the cases where the test says to rebuild. The weeks until the result are the weeks of flow plus the 2 weeks of the trial. We price a week of delay as the expected gross profit of rebuilding after the test, spread over 52 weeks. One year of benefit is a start value (Plausible but untested). At the best size, the cost of delay is about 6,700 euro a week. Figure 1 shows the three curves.

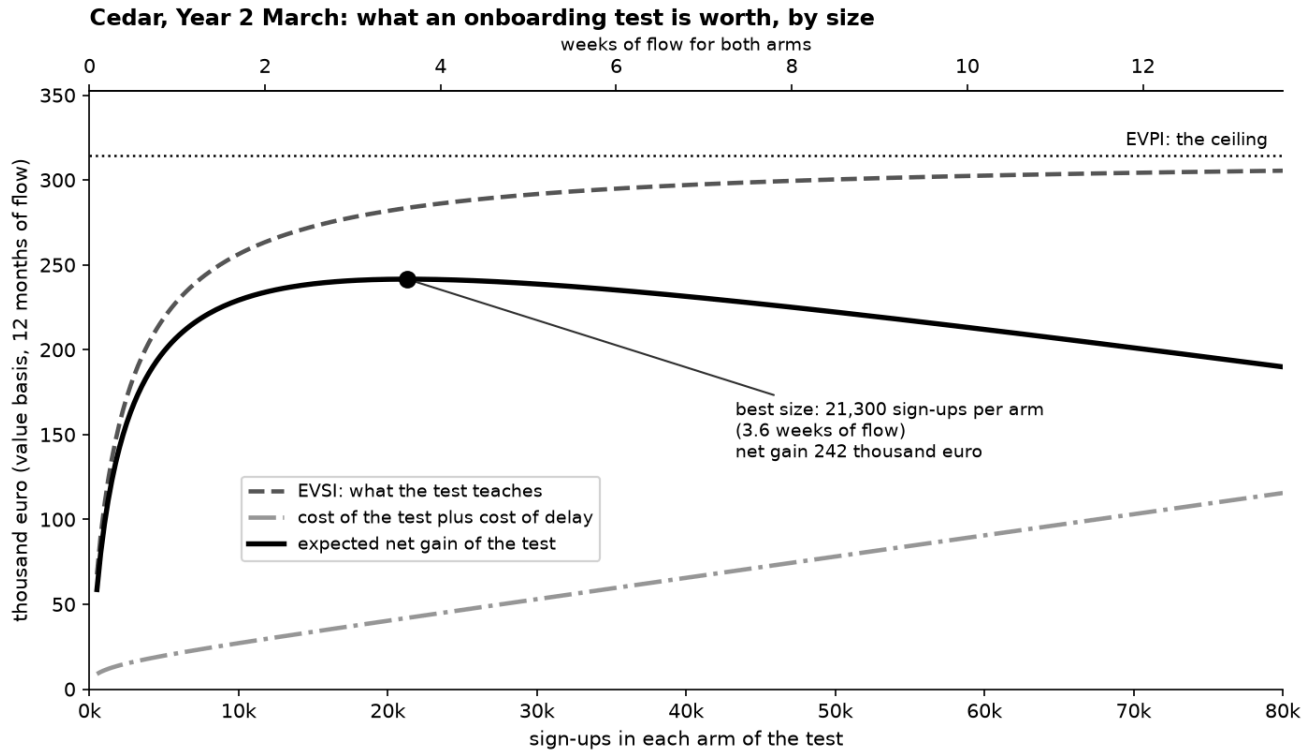
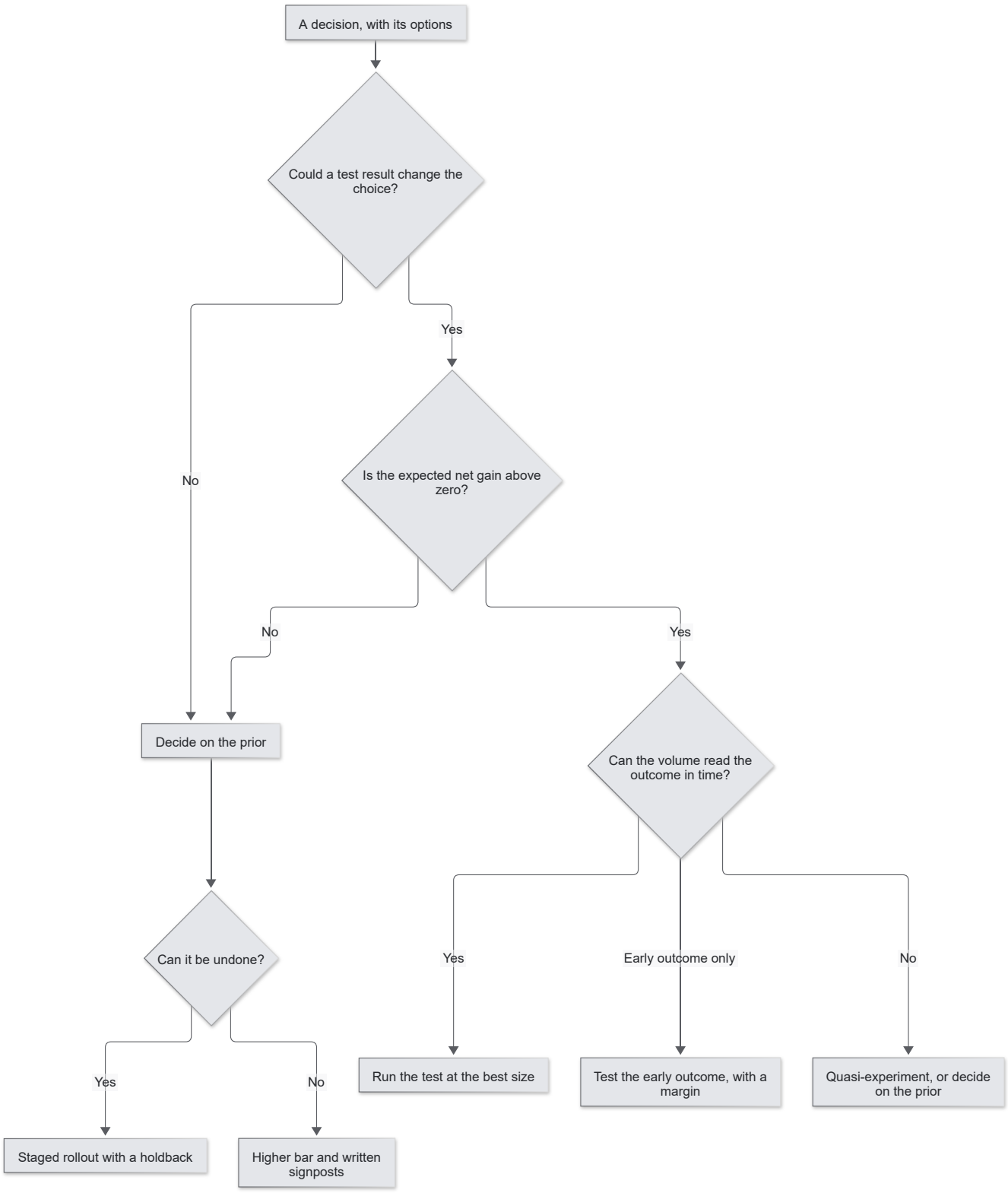


Figure 1. What a test is worth, by size. Cedar, Year 2 March. The EVSI rises and flattens. The cost of the test and the delay keeps rising. The net gain peaks at the best size. Arithmetic, with the inputs of Table 1.

The best test has 21,300 sign-ups in each arm, which is 3.6 weeks of all sign-ups. Its result is in after 5.6 weeks. Its EVSI is 283,541 euro. It costs 4,318 euro to run and 37,654 euro in delay. The expected net gain is 241,569 euro, which is 77% of EVPI. A longer test learns a little more and gains less. At 47,084 sign-ups in each arm, the test takes 8.0 weeks of flow, and its result comes after 10.0 weeks. The delay is 70,167 euro, and the net gain falls to 225,027 euro.

The answer depends on the prior, so we print its sensitivity. With a prior spread of 2%, EVPI is 188,530 euro, the best size is 29,950 per arm and the net gain is 124,248 euro. With a spread of 4%, EVPI is 441,653 euro, the best size is 16,500 per arm and the net gain is 361,985 euro. The decision to test holds in all three cases. Cedar has 68 tests in its register, so its own spread is credible.

The same arithmetic can say no. Three situations make EVSI small. The stake may be small against the price of the test. The prior may be so firm that no result would change the choice. Or the volume may be so low that no result is informative. The first example of this chapter shows the third case.



The choice between a test and a decision on the prior.

Implementation note: the companion code holds EVPI, EVSI, the test-or-decide rule and the search for the best test size in its voi module. It holds the exact sample size in its sizing module.

Deciding on the prior

When the net gain of a test is zero or negative, we decide now. Until the register holds about 30 tests, the decision card takes the prior from the outside view (see *Commit with uncertainty*). After that it uses the company's own record. This is not a guess. Outside base rates give a sober start, because many tested ideas do nothing and some do harm.

We still learn on the way, with a **staged rollout**. A staged rollout releases a change in steps. Each step keeps a **holdback**: a share of units that stays on the old version, so that we can measure the effect of the change. A **holdout** is a group kept from a treatment, for example an audience kept from an ad, so that its outcome can be compared. A holdback is a holdout inside a rollout. The holdback has a price: it withholds the change from some units. We compute that price, in expected gain, before we start.

Three rules make a staged rollout honest. First, name the earliest outcome that the volume can read, such as a meeting booked or a reply. Label it as an early warning until a grade supports it (see *Decide what to measure*). Second, write a stop rule for harm before the first step. Third, write when the next step happens, and what belief the first step will test. A staged rollout cannot prove a small gain. It can show a collapse. It keeps a record against which the belief is graded later. And it lets the signal grow as steps add units.

Design essentials

The basic design is the randomised two-version test (also called an **A/B test**). It assigns units at random to the old or new version and compares them on one primary metric. A test is valid when its arms differ only by chance and by the change. Five habits protect that.

Choose the randomisation unit. The **randomisation unit** is the thing that is assigned at random to a version. Choose the unit where the change lands and where the outcomes are correlated. A visitor is a unit. An impression is not, because impressions of one person are not independent. An account is the unit when contacts of one buying group must see one version. A workspace is the unit for a product setting that every user of the workspace shares. If units come in clusters, multiply the sample by the design effect (see *Decide what the evidence can resolve*). If units share a limited resource, such as seller time under a routing rule, the arms interfere. A **switchback design** then switches the whole market between versions over time periods, in random order.

Keep an assignment log. Assign by a stored seed and the unit identifier. Write one line for each unit: the unit, the arm, the time and the test name. Log exposure when the unit sees the version. Analyse every unit by the arm it was assigned to, whether or not it used or saw the change. This is **intention to treat**. It keeps the randomisation intact. If some control units also see the new version, the measured effect shrinks by that share.

Check the sample ratio. Compare the counts in the arms with the planned split. A **sample-ratio mismatch** is a split that differs from the plan by more than chance allows, for example 50.5 to 49.5 instead of 50 to 50. It signals broken assignment or tracking, and the test is void until the fault is fixed. Birch's demo-form test, which the examples describe, has 113,042 sessions. A fair split of that size varies by 0.149 points. A split of 50.2% gives a p-value of 0.18 and passes. A split of 50.5% gives 0.0008 and fails. We alarm at a p-value below 0.001 (a start value; Plausible but untested), which is a split beyond 50.49%.

Check balance. Compare the arms on facts known before the test: country, size, past activity. Under random assignment, differences are chance, so a large difference signals a broken assignment. Do not repair imbalance by adjusting the result. Fix the assignment and start again.

Plan the duration. The **test duration** is the weeks of flow for both arms, plus the time for the outcome to mature. The chapter *Decide what the evidence can resolve* gives the weeks as 2 × n per arm ÷ weekly flow. Run whole weeks, so that each weekday counts equally. A **novelty effect** is a gain that appears because a change is new and fades as users get used to it. A holdback read two to four weeks after shipping can show it.

Size a holdback for its real split. Take Birch's form conversion of 4.0%, and a lift of 10%. Two equal arms need 39,475 sessions each, which is 78,950 in all. A holdback of 10% needs 22,243 sessions in the holdback and 200,187 in the other arm. That is 222,430 sessions, which is 2.8 times as many.

Outcomes, primary metric and guardrails

The **primary metric** is the one metric on which the test is decided. Choose it before the test starts. One metric keeps the false-positive rate where you set it. Rates, such as the share of sessions that convert, behave better than money amounts. Money amounts are skewed, and they need far larger samples (Kohavi and others, 2014, *Seven Rules of Thumb for Web Site Experimenters*, KDD).

Some changes always raise the primary metric and may still do harm. A shorter form always raises the number of requests. The remedy is a **composite metric**: a primary that multiplies the guardrail into the result. Take the number of requests per session, and multiply it by the share of requests that qualify. The product counts a request only as far as it is good.

A **guardrail metric** is a metric that must not fall while the primary rises. A guardrail needs two things before the test starts. The first is a **non-inferiority margin**: the largest fall that you accept. Set it from break-even. The second is a sample statement: what the planned sample can show. A fall is ruled out when the one-sided 95% lower bound of the change lies above minus the margin.

> **Calculation**
>
> units per arm for a non-inferiority test, with no true difference: $n = 2 \times (1.645 + 0.842)^2 \times p \times (1 - p) \div (\text{margin} \times p)^2$
>
> smallest margin that n units per arm can rule out: $\text{margin} \times p = (1.645 + 0.842) \times \sqrt{(2 \times p \times (1 - p) \div n)}$
>
> Evidence label: Arithmetic, with the normal approximation (Standard practice). 1.645 belongs to a one-sided 5% test and 0.842 to 80% power.

A guardrail with no margin and no sample statement cannot show "no fall". Silence is not safety. "No significant fall" can mean that the sample was too small to see one. If the sample cannot rule out the margin, say so. Then choose one of three moves. Lengthen the test. Move the guardrail to an earlier step that has more events. Or watch it after launch with a detector (see *Decide when a change needs action*) and write the rollback rule in advance.

A **proxy**, also called a surrogate, is a metric that stands in for the outcome that you care about, because the outcome is slow or rare. A proxy is an early warning until a grade supports it, so a decision on a proxy alone waits for the outcome. When a test must run on a proxy, the test record names it as a proxy.

Size and looks

Size every test before launch. Use the exact formula for the smallest change that matters, and compute the weeks. If that change is smaller than what the volume can read, say so in the test record, the decision-log entry of the test. The test then answers a different question: whether the effect is large. For the small range, the decision rests on the prior and on a staged rollout.

A **look** is a planned reading of the result before the test ends. Every look is another chance of a false positive. When nothing is going on, two looks at 1.96 standard errors give an overall false-positive rate of 8.3%. Five looks give 14.2%, and ten looks give 19.3%. These are exact values for equally spaced looks, by numerical integration (Arithmetic). Peeking is priced, not forbidden.

Sequential testing reads a test at planned points and keeps the false-positive rate under control. **Alpha spending** divides the allowed 5% among the planned looks. Take two equal looks and a spending function of the O'Brien-Fleming type (Lan and DeMets, 1983). The limits are 2.77 standard errors at the first look and 1.98 at the second. An early stop then needs a large result. State, before the test, which looks may stop it and for what reason: for a clear win, for harm or for futility. A look that may stop the test only for harm spends no alpha on the win.

An unplanned extension is not a look. If a test ends unresolved and the owner wants more data, do not read the extra weeks as if the test had always been that long. Treat the result of the first part as a prior, shrunk toward the company's own record, and update it with the new data. Report the update with the label "unplanned extension", next to the planned result.

Allocation designs

Sometimes the aim is not to learn a transferable effect, but to choose the better of a few versions for a known audience. Two designs fit.

Test-and-roll tests on a small share of the audience and rolls the better version out to the rest (McCullough, 2021, *Business Experiments with R*, section 4.7). Take two versions whose true rates are independent draws around the current rate, with a spread σ. Test n units for each version, and roll the version with the higher estimate to the other N − 2n units.

> **Calculation**
>
> gain over a coin flip = (N − 2n) × sσ ÷ √π, where sσ = σ² ÷ √(σ² + p × (1 − p) ÷ n)
>
> The expected maximum of two independent normal values with spread sσ is sσ ÷ √π. Perfect knowledge would give N × σ ÷ √π.
>
> Evidence label: Arithmetic, with a normal prior and a known audience size.

The formula assumes that every unit outside the test gets the winner at once. When the outcome matures slowly, as activation does over 14 days, units that arrive during the wait miss it, so the gain is slightly overstated.

Take Cedar's first-day email, in two versions. The outcome is activation, at 36%. The audience is one window of sign-ups, 47,084. The spread of true rates is 3% of 36%, which is 1.08 points (Cedar's register spread). The best test has 3,560 sign-ups for each version, which is 15% of the audience. It gives 195 extra activations over a coin flip, against 287 under perfect knowledge. It captures 68% of the perfect gain.

A **multi-armed bandit** shifts traffic towards the better option while the test runs. Three questions decide whether it fits. Is the outcome seen within about a day? Is the aim a payoff over a known horizon, which is allocation, or an estimate, which is learning? Are there enough events for each arm at each update? If any answer is no, use another design. A bandit teaches less than a fixed test. It is not the default.

A **factorial design** varies several elements together in a planned set of combinations. Each main effect compares half of all units with the other half. So each element is read with the power of a two-arm test on the same total. Four elements take a quarter of the units that four separate tests need. The method rests on a habit of real studies: few effects are active, and main effects are larger than interactions. In a review of published factorial studies, 41% of main effects were active. Only 11% of two-factor interactions were (Li and others, 2006, as reported in McCullough, 2021, *Business Experiments with R*, chapter 8). It also has a cost. Each contrast carries a 5% false-positive risk. Four main effects give an 18.5% chance of at least one false positive, and ten contrasts give 40.1%. A combination is also noisier than one element. Its estimate adds the main effects of the elements it keeps, so with four elements its standard error doubles. So screen first, then confirm the best combination: in a two-arm test, or with a holdback after it ships. Size the confirmation on the shrunk effect of *Commit with uncertainty*, never on the raw one.

Quasi-experiments

A **quasi-experiment** estimates the effect of a change that nobody randomised. We use it when randomisation is impossible, for example for a price change for everybody, a policy, or one event. We also use it when there are too few units to randomise. A quasi-experiment compares the result with what would probably have happened. The assumption behind that comparison cannot be tested by the data alone, so the evidence is weaker. Keep the word "confirmed" for randomised tests and holdbacks (see *Decide what to measure*).

Each design has a minimum procedure.

- **Interrupted time series.** One series is read before and after a known date. Collect at least 8 weekly points before the change (a start value; Plausible but untested). Fit the level and the trend before the date, with the season. Estimate the change in level and in trend at the date, with errors that allow for autocorrelation. Add a control series that the change did not touch. Do not use it when another change happened in the same weeks, or when the date was chosen because the series looked bad. A series chosen that way would recover anyway (see *Decide what the evidence can resolve*).
- **Difference in differences.** A treated group and an untreated group are compared before and after the change. The effect is the change in the treated group minus the change in the untreated group. It needs parallel trends before the change, so plot several periods before it. It also needs no spill-over and a stable mix of units. Do not use it when the groups were chosen for how they were doing.
- **Regression discontinuity.** A rule sends units above a cut-off to one treatment and units below it to another. Examples are a lead score and a discount limit. Compare units just on each side. Check that nobody can move across the cut, that the units on both sides look alike, and that the result survives a change of window. The estimate is local. At medium volume it reads only large effects.
- **Synthetic control.** One unit is treated, and many are not. Build a weighted blend of untreated units that matches the treated unit before the change. Compare the two after the change. Judge the result by applying the same method to each untreated unit, as a placebo. Do not use it when the blend fits poorly before the change.

With few units, the usual p-value fails. **Randomisation inference** compares the observed effect with the effects from all random ways of assigning the treatment. Compute the effect for the actual assignment and again for every possible assignment. The p-value is the share of assignments with an effect as large. Birch sells in 9 countries. If one country gets the change, there are 9 possible assignments, and the smallest p-value is 1 ÷ 9, which is 0.111. No result can be significant. If 3 of the 9 are randomised there are 84 assignments, and the smallest p-value is 0.012. With 4 of 9 there are 126 assignments, and it is 0.008. The number of units sets the floor, so decide the assignment before you look.

Reading a result

Read every result as an interval against break-even and, for a guardrail, against the margin. Three outcomes are possible. The result is **superior** when the whole interval lies above break-even. It is **equivalent within the margin** when the one-sided lower bound lies above minus the margin. It is **unresolved** when the interval spans the line that decides. "Inconclusive" does not mean "no difference". It means that the data cannot tell. Never compute power after the result. Observed power adds nothing (McCullough, 2021, *Business Experiments with R*, section 4.4).

Figure 2 shows the three outcomes, and a fourth case in which the sample is too small for the margin. They come from the Birch test below.

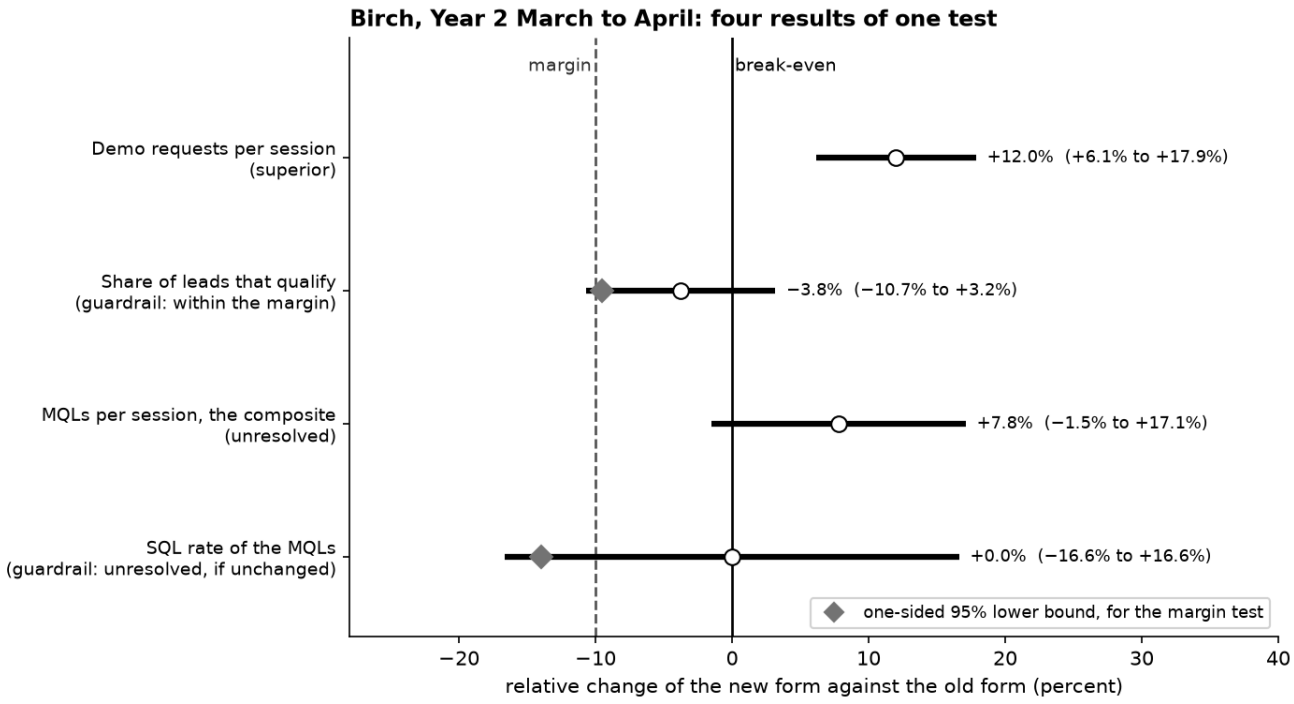


Figure 2. Four results of one test, against break-even and the margin. Birch, Year 2 March to April. The bars are 95% intervals. The diamonds are the one-sided 95% lower bounds used for margin tests. The first three rows take the effects that the example facts ship in May as the observed result. The fourth row shows what the test could read if the SQL rate did not change. Illustration only.

A report of a result has five lines. The interval and the count behind it. The assumptions of the design. The label of the evidence: randomised, holdback or quasi. What the decision card does with it: it supplies the likelihood, and the card combines it with the prior and the cost. The belief that the log recorded before the test, next to the outcome. The owner reads the report, decides by the date in the log, and writes the decision.

The short methods list

A short list of methods, by procedure, is easier to use than a long catalogue. Each entry has a data rule and a case in which not to use it.

Table 2. Thirteen methods of evidence, with their data rules.

Method	Question it answers	Data rule	Do not use when
Randomised test by visitor	Does this change move the outcome?	The exact size is met; assignment is logged	One person or account appears in many units
Randomised test by account or workspace	The same, where users share one setting	Count the clusters; apply the design effect	There are too few clusters to read the change you need
Factorial screening	Which of several elements matter?	Few active effects; main effects first	Interactions are the goal, or you cannot afford a confirmation
Staged rollout with holdback	Does harm show early?	A named early outcome and a stop rule	The change cannot be undone and no signpost is written
Test-and-roll	Which of two versions for a known audience?	The audience size and the prior spread are known	The audience is not known in advance
Bandit	Which option earns most in a short run?	Outcome within a day; enough events per update	The outcome arrives late, or you need an effect estimate
Geo experiment (ads run in some regions only)	Does a channel add outcomes?	Many regions, randomised (see *Choose acquisition channels*)	Regions are few, or they spill over
Interrupted time series	Did a level or a trend change at a known date?	A long period before, and a control series	Another change fell in the same weeks
Difference in differences	Did the treated group move more than a comparable one?	Parallel trends before the change	Groups were chosen for how they were doing
Regression discontinuity	What does a rule at a cut-off do?	Many units near the cut, which nobody can cross	People can move themselves over the cut
Synthetic control	What would one treated unit have done?	Many untreated units that match it before	The fit before the change is poor
Marketing mix model (a regression of results on the spend of every channel)	How do spend and outcomes move together?	Spend varied on purpose (see *Choose acquisition channels*)	Spend was steady, so the curve is not identified
Survival and cohort reading	When do outcomes happen, and for whom?	Dated events, with censoring	Young cohorts are compared with old ones (see *Forecast conversion and timing*)

What changes with volume

The mechanics hold at every volume. The choice of design does not. Table 3 shows, for each design, whether the volume of each company can read it. The entries come from the barometers and from the sizes in this chapter.

Table 3. Designs against volume regimes, Year 2 March, at the trend level. Two-arm sizes use the exact formula; the comparisons of periods come from the barometers.

Design	Low (Alder)	Medium (Birch)	High (Cedar)
Two-arm test on a top-of-funnel rate	Early outcome only: meetings booked read only a rise of 240% in a quarter	Feasible: the form conversion reads 8.3% in two windows	Feasible: activation reads a 3.5% change in one window
Two-arm test on the last rate before revenue	Not feasible: even a comparison of periods needs about 2.9 years to read a 30% change in the win rate	Early outcome only: a 20% rise in the close rate needs 451 days with one sales cycle	Feasible for self-serve: 52 days for a 10% rise in the paid rate. Not feasible for assisted sales: even a comparison of periods needs about 7 months for a 30% change
Factorial screening	Not feasible	Feasible on the top of the funnel	Feasible: four elements in two windows
Staged rollout with a holdback	The default	Used for slow outcomes	Used to confirm a winner after it ships
Randomisation over regions	Does not fit: the unit is the account	Large effects only: 9 countries give a smallest p-value of 0.012 at 3 of 9	Feasible: 24 regions give 2,704,156 assignments at 12 of 24
Time-series designs	Large effects only	Large effects, with a long period before	Feasible, with a control series

At low volume, test only on an early outcome, mainly to guard against harm, and decide on the prior and on cases. At medium volume, test the stages that the volume can read, and give every guardrail a margin and a sample statement. Move slow guardrails to the detector. At high volume the limit is validity, not power: the unit, the clustering, the many looks and the many contrasts.

Alder, Year 2 March: a pilot that cannot read its result

Alder, Year 2 March: about 4.0 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 2.9 years to read even a 30% change, and the lead count reads a 20% change in 24 days.

The marketer proposes a short research brief for each target account before outreach. The brief names the plant, the recent news and the likely buyers. A pilot would give briefs to 60 of the 120 accounts in the pilot group and hold 60 back, chosen at random. The early outcome is meetings booked in a quarter. The baseline is 0.10 meetings for each account in a quarter. So each arm expects 6 meetings.

The analyst asks first what the pilot can read. With 6 expected meetings in the held-back arm, the treated arm must book about 20 to separate from chance, a rise of 240% or more. The prior for the gain from research is a range from 0% to 15%, with a mean of 7.5%. It is the range that *Find the most valuable constraint* used for account selection (Illustration only). To read a rise of 7.5%, each arm would need 2,895 meetings. At 0.10 meetings for each account in a quarter, that is 28,954 account-quarters for each arm. Even with all 400 programme accounts, 200 in each arm, it would take 145 quarters, which is 36 years. No test at Alder can read the gain that matters.

Alder's meetings do not yet link to customers. So the analyst borrows the engaged-account chain as a start value (Plausible but untested). 38% of engaged accounts become opportunities, and 22% of those are won. That is 8.4% of meetings, and a new customer is worth 220,768 euro on the value basis. So a meeting is worth 18,456 euro. The programme works 400 accounts, which yield 160 meetings a year. A rise of 1% in meetings is worth 29,530 euro. The research costs 50,000 euro, which gives a break-even gain of 1.7%. At the prior mean, one year of the programme earns 171,474 euro net of its cost. The probability of loss, the chance that the true gain is below break-even, is 16%.

The first stage of a pilot costs 7,500 euro for the briefs. The held-back accounts forgo 0.45 meetings in expectation, which is 8,305 euro of value. EVPI is 14,613 euro, so no test of this choice can be worth more. EVSI for the pilot is below one euro, because no result that the pilot could show would move the belief below the break-even. The pilot as a test has no value. The analyst prints the sensitivity, because the register holds only 2 tests and the prior is a judgement. Take a prior centred on zero, in the spirit of the outside base rate. The programme would then lose 50,000 euro in expectation, EVPI would be 46,810 euro and the pilot's EVSI 337 euro. The pilot would still be worth almost nothing. So the choice rests on the prior alone, and it flips when the mean gain falls below 1.7%. That number goes on the card.

The marketer decides by the date in the log. Alder starts the first stage of a staged rollout. It briefs 60 accounts and keeps 60 back. The 8,305 euro is the price of two things. One is a stop rule: if the treated accounts book 2 or fewer of the first 12 meetings, the marketer reviews the programme. Under equal rates the chance of this is 1.9%. The other is a record. Before the quarter starts, the log states two beliefs. Treated accounts book about 6.5 meetings, with an 80% range of 3 to 10. Held-back accounts book about 6, with a range of 3 to 9. The falsifier is the stop rule. The log grades the belief in June.

Alder never reaches 30 tests, so it keeps outside base rates as priors. At low volume, a test runs on an early outcome and mainly guards against harm. The account holdout (a random share of target accounts that get no programme touches) reads engaged accounts for the whole programme (see *Choose acquisition channels*).

Birch, Year 2 March to April: a shorter form, with a margin

Birch, Year 2 March: about 51 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 68 days, and the close rate reads a 20% change in about 6 months.

Marketing wants to cut the demo request form to fewer fields. The change is cheap and can be undone in a day. The risk is quality. A shorter form brings more requests, and some may be weak. In February Birch chose better qualification over an eleventh seller (see *Commit with uncertainty*). In March, demand runs at 305 SQLs a window against capacity for 300. Extra weak leads would lengthen the queue. The hiring trigger of February, at 65 uncontacted SQLs, watches it.

The analytics lead designs the test. The unit is the visitor, assigned by a stored seed, 50 to 50. The demo page gets 56,521 sessions a window, so two windows give 56,521 sessions in each arm. The primary is a composite: MQLs per demo-page session. It is the form conversion of 4.0% times the share of leads that become MQLs, 40.0%, which is 1.6%. The form conversion alone reads an 8.3% change in two windows. The composite is harder. It reads 13.5%. A test that must read 10% on the composite needs 101,285 sessions per arm, which is 3.6 windows, or 14.3 weeks. The marketing owner wants a decision before May. The change is cheap and reversible, so Birch runs two windows and reads the composite as an interval, not as a verdict.

Two guardrails get a margin of 10%. The first is the share of leads that qualify. A fall of 10% in it would cancel a gain of 11% in requests. With 2,261 leads in each arm, the test can rule out a fall of 9.1% or more, so the margin is readable. The second is the SQL rate, 23.5% of MQLs. The test collects 904 MQLs in each arm. At that size it can rule out only a fall of 21%. To rule out 10% it needs 4,025 MQLs per arm, which is 8.9 windows. So the test cannot decide this guardrail. The plan says so. After launch the detector reads the SQL rate on all MQLs, which number 1,365 a window. Against a baseline of 13 windows, it sees a fall of 12.4% in one window and 8.8% in two. Within eight weeks a fall beyond the margin shows. The rollback rule is written first.

The plan has two looks, at the end of each window. The first look may stop the test only for a broken split or for harm, at 2.77 standard errors. A win cannot stop it, because the guardrails need the full sample. The last look then uses the full 1.96. If a win could stop it, the limits would be 2.77 and 1.98. Two looks at 1.96 would raise the false-positive rate to 8.3%. A lift of 11.4% on requests would cross the first limit and end the test before the guardrails had data. The sample-ratio check runs daily. The alarm sits at a split beyond 50.49%.

In April the test reads out. Take the effects that Birch ships in May as the observed result (Illustration only). Requests per session rise from 4.00% to 4.48%, which is +12.0%, with a 95% interval of +6.1% to +17.9%. That is superior. The share of leads that qualify falls from 40.0% to 38.5%, which is -3.8%, with an interval of -10.7% to +3.2%. The one-sided lower bound is -9.6%, inside the margin of 10%. The composite rises from 1.600% to 1.725%, which is +7.8%, with an interval of -1.5% to +17.1%. That is unresolved: the data cannot tell whether the net gain is above zero. Figure 2 shows the four results.

The marketing owner decides on the card. The interval of the composite lies mostly above zero. The downside is small. The change is reversible. The form ships in May, with the SQL rate under the detector and the rollback rule in place. The sales owner watches the queue, because the extra MQLs help only while sellers have room (see *Find the most valuable constraint*). At medium volume, a top-of-funnel test works if every guardrail has a margin and a sample statement. A guardrail that the test cannot read moves to the detector.

Cedar, Year 2 March to April: four elements at once

Cedar, Year 2 March: about 1,700 new paying accounts a window, 1.3 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 27 days; the assisted motion, with 70 opportunities a window, needs about 7 months to read a 30% change.

The growth team wants to change four onboarding elements. They are a guided checklist, a sample workspace with example data, a prompt to invite teammates and a first-day email written as a person. The value-of-information check of the worked case above says to test. Cedar uses a factorial design with 16 cells, and the outcome is activation, at 36%.

The settings belong to the workspace, so the unit is the workspace. A workspace holds 4.2 users on average, and users in one workspace correlate at 0.06. The design effect is 1.19. In two windows, Cedar gets 94,168 sign-ups in 22,421 workspaces. Each of the 16 cells holds about 1,401 workspaces, which is 5,886 users. The cells are too small to compare one by one. Each main effect uses all the users: half have the element on and half have it off, which is 47,084 users on each side. After the design effect the effective size is 39,500 per side. The smallest change that each main effect can read is 2.7%. Without the design effect it would look like 2.4%. Four separate tests of the same power would need 376,682 users, which is 8.0 windows. The factorial needs 2. High volume does not make the test cheaper in method. It makes it valid, if the unit is right.

An analysis that counted users as independent would print intervals that are too narrow. Cedar found this in Year 1 September (see *Decide what the evidence can resolve*). The analysis counts workspaces, or applies the design effect. The sample-ratio check runs on the 16 cells, with equal shares as the plan.

The register holds 68 tests, so Cedar uses its own prior: a mean of 0% and a spread of 3%. In February, the pricing-page test showed how much a raw winner overstates (see *Commit with uncertainty*). A screen across 4 main effects has an 18.5% chance of at least one false positive. So the plan screens first, then confirms.

The read comes in early May, when the last sign-ups of April finish their 14-day trial. Take the effect that Cedar ships in May as the result (Illustration only): the best combination lifts activation by about 6%. One main effect has a standard error of 0.95%. A combination of all four has twice that, 1.9%. The lift still lies 3.2

standard errors above zero.

The raw lift is still not the number to ship on. The register prior shrinks it by a factor of 0.71, to 4.3%. On that belief, the chance that the lift is above zero is 99.6%. The change is reversible, so the 90% bar applies, and the growth lead ships the combination in May, with a 10% holdback, as for the pricing page. The holdback is the confirmation, and it is sized on the shrunk 4.3%, never on the raw 6%. It needs about 10,200 users in the holdback, which takes 2.2 windows of sign-ups. With the trial, it reads after about 75 days. The log holds the shrunk belief and that date.

Once the combination ships, the team asks about a bandit for the subject line of the first-day email. The outcome that a subject line moves is a click within a day. Cedar has thousands of sign-ups a week, so the events are enough. But the click is only a proxy for activation, and the proxy has no grade. Activation arrives after 14 days. A bandit on clicks would tune a proxy. For the whole email, the activation question fits test-and-roll. The team tests two versions on 15% of one window of sign-ups, waits for activation, and rolls out the winner. At high volume, the constraint is validity, not power.

The decision to carry forward

Buy information when its expected improvement to the decision exceeds its cost and delay. Match the experiment to the action. If no feasible result would change the commitment, spend the time on a different uncertainty.

19 · Commit with uncertainty

Which commitment should you make with the evidence available now? The answer needs alternatives, a baseline and a distribution of economic consequences. A confidence measure alone cannot compare a cheap reversible action with an expensive commitment that is difficult to undo.

Write what you believe before choosing. Separate uncertainty about the effect, the exposed population, the cost and the duration. Ask which assumption would reverse the recommendation. That question tells you whether more research, a smaller commitment or immediate action is useful.

The decision card below records the calculation and its limits. The log preserves the knowledge state at the time, so a later outcome can teach you something. A documented policy threshold is a company choice, not a universal statistical law.

What a decision is

A decision is a commitment of money, time or people that you could have made in another way. It has at least two options. One option is always *do nothing*. The do-nothing option is not a pause. The world moves while you wait: a queue grows, a market closes, a contract ends. So the card measures every option against the path that the world takes if you do nothing.

Options differ in how easy it is to undo them. A decision is *reversible* when you can undo it quickly and at low cost. A new page variant is reversible: you switch it off in a day. A hire is hard to reverse. Entry into a new market is harder still. We call these *irreversible* decisions, although the word means "costly to undo" and not "impossible to undo". Reversibility sets the evidence that a decision needs.

You can often turn one large irreversible decision into a small reversible one and a later decision. This is a staged move. Stage one is cheap and carries signposts. Stage two waits for what the signposts show. A *signpost* is a dated, observable event that tells you early whether a decision is on track, with a planned response written in advance.

Every decision has one named owner. The owner is a role, such as the sales owner, and not a committee. This role is the *decision owner*. Others advise. The owner decides by a stated deadline and writes the decision in the log. *Fund the next best action* shows how owners choose which cards to work on first.

The decision card

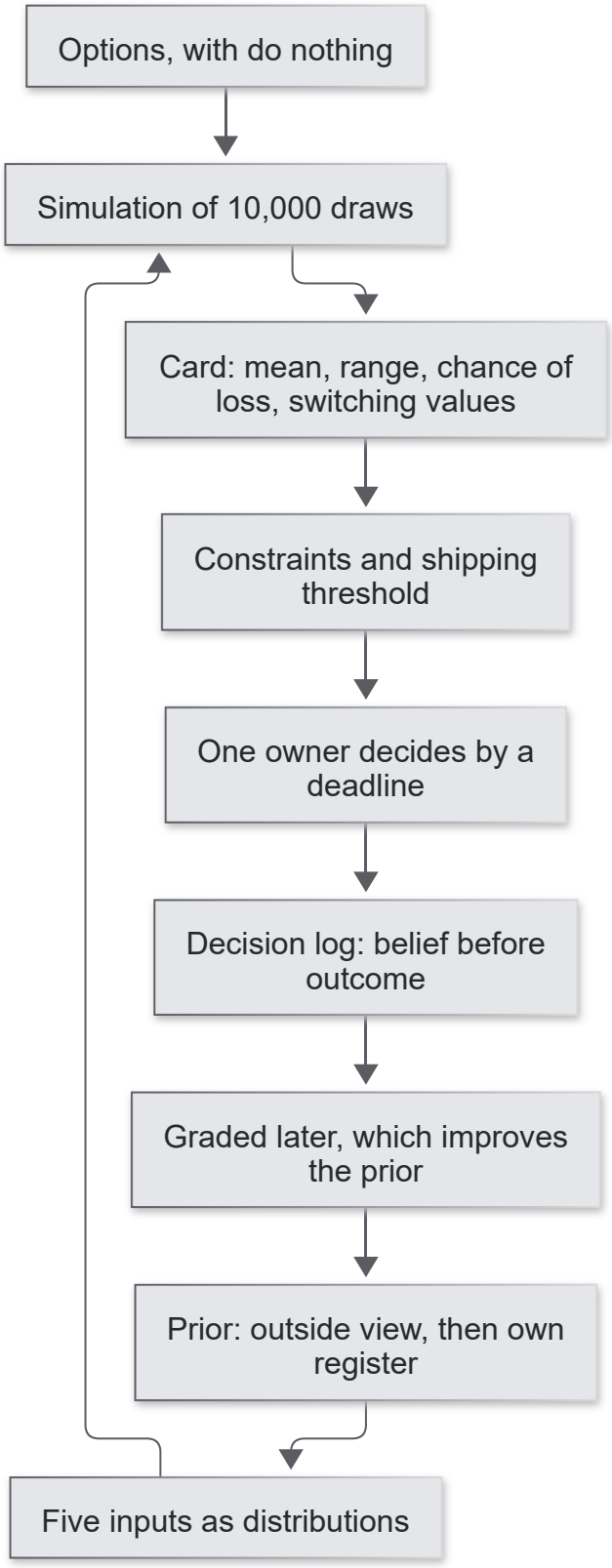
Net value as a distribution

The card prices each option as a net value, in euro, on the value basis.

> **Calculation**
>
> net value = v × l × D − K − r
>
> v: the value of one unit, in euro of discounted gross profit (for example one new customer).
>
> l: the extra units in each window, caused by the option, against doing nothing.

> D: the number of windows in which the effect counts, after the delay to the first effect.
>
> K: the cost, one-off and running, with people time included.
>
> r: the expected harm, in euro, on the other wing of the bowtie.
>
> Evidence label: Arithmetic. For the expected value approach, see Camm and others, *Business Analytics*, chapter on decision analysis.

Each of the five inputs is a belief, so each is a distribution. Where the company has data, the distribution comes from the data and carries its standard error. Where it has none, the distribution is a prior, and the card says so. A simulation of 10,000 draws turns the five distributions into one distribution of net value. All options use the same draws for the shared inputs, so most simulation noise cancels between options. The seed goes into the log, so anyone can repeat the run.



From options and inputs to the card, the threshold and the log. The log closes the loop: graded beliefs become the next prior.

Find the most valuable constraint priced one lever by its expected profit. The card keeps that calculation and adds what a decision needs: alternatives, risk, constraints, a threshold and a record.

What the card prints

Each output of the card answers a plain question.

Output	The question it answers
Expected net value (the mean)	What is the option worth on average?
P10 and P90	What range holds 80% of the outcomes?
Probability of loss	In what share of the draws is the net value below zero?
Probability of positive net value	In what share of the draws does the option pay? It is the mirror of the probability of loss.
Probability of being the best option	In what share of the draws does this option beat all the others?
Break-even effect	What is the smallest effect that pays for the option?
Switching values	By how much must one input move to change the choice?

The *break-even effect* is the effect at which net value is zero, with the other inputs at their means.

> **Calculation**
>
> break-even effect = (K + r) ÷ (value of one unit × units exposed × D)
>
> Evidence label: Arithmetic.

This is the change that matters of *Decide what the evidence can resolve*: the smallest one that pays for the action. Compare it with the smallest change that your volume can read. If the break-even effect is much smaller than what a window can read, you will never confirm that the option pays, and you decide on the prior. If it is larger, a test can settle it. The switching values of *Value customers and cash* complete the picture. The input with the smallest switching value is the one to measure first.

The mean alone misleads: two options can have the same mean and very different ranges. An option that is best in few draws can still be the safer choice. Read the mean, the range and the two probabilities together.

The card also prints the value of perfect information, the most that any measurement could add. *Choose whether to test or act* uses it to decide whether a test is worth its cost.

Table 1 shows a card in its simplest form: the card of Birch from the examples below.

Table 1. A decision card as a table: Birch, Year 2 February, from 10,000 draws (Illustration only).

Plan	Mean cost (K), euro	Mean net value, euro	P10 to P90, euro	Chance of loss	Chance of best
Hire an eleventh seller now	180,000	1,598,000	1,255,000 to 1,950,000	0.1%	33.8%
Qualification now, a seller when the backlog passes the trigger	70,928	1,715,000	577,000 to 2,875,000	2.9%	5.0%
Qualification and a seller now	211,120	1,980,000	602,000 to 3,306,000	3.5%	61.2%
Do nothing	0	0	0	0%	0.1%

Constraints

Some limits are not part of the objective but conditions that an option must meet. The card can print the chance of breaching each one.

- **Payback.** *Value customers and cash* set a payback ceiling for each segment: 24 months for Alder, 18 for Birch and 12 for Cedar. We count the effects of an option inside this ceiling. Money that comes back later is real, but cash planning does not reach it (Plausible but untested).
- **Capacity.** A stage that is full cannot work extra units. The card runs the queue month by month, as *Allocate sales attention* showed, so that extra volume at a full stage adds nothing.
- **Risk limit.** The *risk limit* is the most cash that one decision may commit before its first signpost. We set it at 2% of the yearly gross profit of the company (Plausible but untested; a company tunes it with its own log). A decision above the limit needs the approval of the next level, and it is a reason to stage. Alder's limit is about 257,000 euro, Birch's about 474,000 and Cedar's about 838,000, from the yearly gross profit of each at Year 2 February.

The change-impact review

An improvement on one wing of the bowtie can cost something on the other wing. More customers at the knot raise the load on onboarding and customer success. A tighter qualification rule changes who buys, and so who renews. Before you sign a card, ask what the option does on the other wing. If you can price the effect, put it in r, or in the quality of the added units. If you cannot price it, set a guardrail: a metric on the other wing that must not fall by more than a stated margin. *Choose whether to test or act* shows how to give a guardrail a margin. This step is the *change-impact review*: one line on the card for each option. Table 2 shows the lines of the three examples.

Table 2. The change-impact review of the three examples.

Option	What it does on the other wing	How the card handles it
Birch: qualification and an eleventh seller	More customers each year for onboarding and customer success; a different mix of buyers	The cost of revenue already holds the service cost. The mix is watched through the first-renewal rate of the new customers, which is a guardrail.
Alder: a second vertical	New customers in a segment with its own onboarding needs	The value of a customer uses the same renewal engine. The first renewals of the vertical are read as a signpost.
Cedar: a pricing-page variant	Added accounts may renew less than average	The harm enters r through a quality ratio with a mean of 0.90.

Priors

The card needs a belief about the effect before any new data arrives. We call this belief the *prior*. After you add data, the belief is the *posterior*. The question is where the prior comes from.

The first source is the *outside view*. You judge the case by what happened in similar past cases. The group of similar cases is the *reference class*. The share of cases in it that succeeded is the *base rate*. Outside base rates exist for experiments. At Microsoft about one third of experiments proved effective, one third were neutral and one third were harmful. At Google and Bing only about 10% to 20% produced a gain (Thomke, 2020, *Experimentation Works*; Standard practice).

A base rate also tells you how far to trust a significant result. A test at the one-sided 5% level with 80% power gives some false alarms among its wins. Their share depends on how many ideas truly work.

Table 3. The chance that a significant result is a false alarm, by base rate (Arithmetic, one-sided 5% level, 80% power).

Share of ideas that truly work	Share of significant wins that are false
One third	11.1%
15%	26.2%
10%	36.0%

The second source is the company's own record. When the register holds about 30 tests, the company can estimate the spread of the true effects of its own tests. It then uses that spread instead of the outside rate (Plausible but untested as a start value). The observed lifts of the tests spread by a combined amount of real differences and noise. Subtract the average noise, in squares, and what remains is the real spread. The mean of the true effects is the mean of the observed lifts. The company also tracks its *base rate of winning tests*: the share of its tests that end as significant winners.

With fewer than about 20 rows in the register, the effect rests on a prior. Then never let one euro figure stand alone. Print the prior, the range and the belief at which the choice flips. A single figure would look more exact than the evidence allows. Where no reference class exists, use your own nearest record, widen it, and label it Assertion.

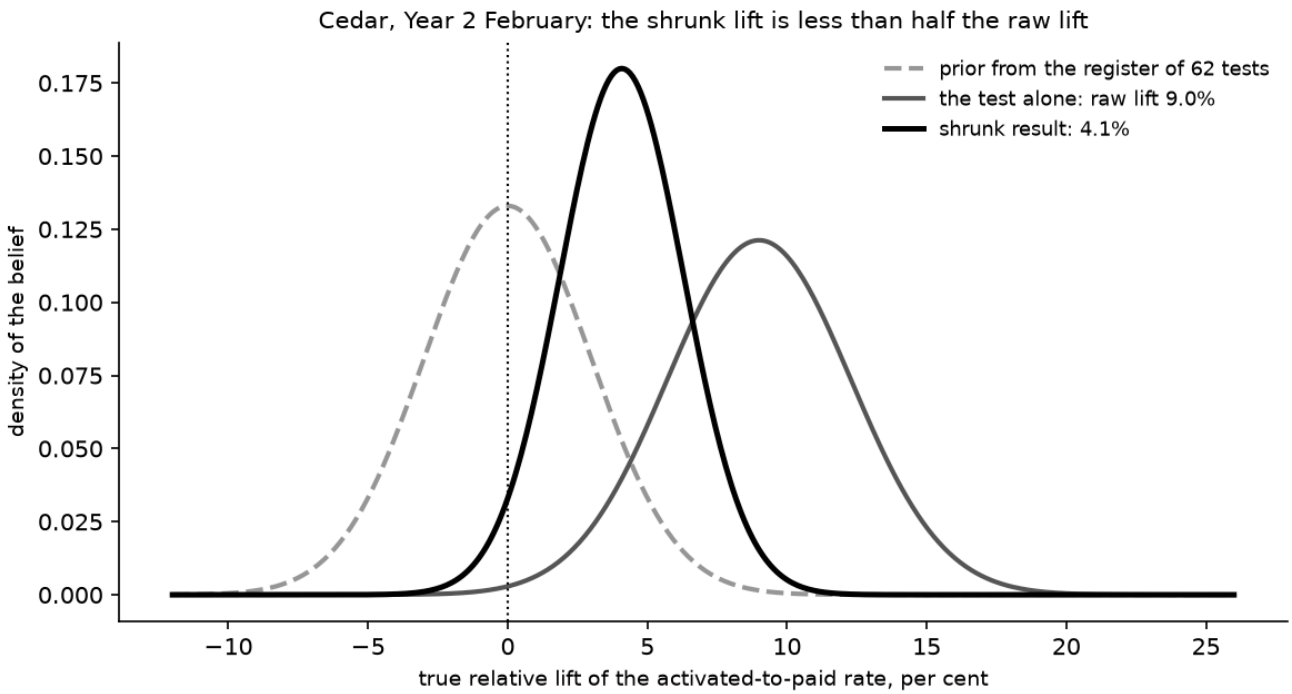
The winner's curse

The best-looking result is overstated. The reason is selection. Suppose that many tests run, each with a true effect and some noise. The test that looks best is usually one with a good true effect and good luck. You picked the luck together with the effect. The more tests you compare, and the bigger the winner, the larger the share of luck in it.

The correction is the shrinkage of *Decide what the evidence can resolve*, applied to effects. The *shrinkage factor* is the share of the observed spread that is real. It is the number that *Decide what the evidence can resolve* calls the reliability.

> **Calculation**
>
> shrinkage factor = $\tau^2 \div (\tau^2 + \sigma^2)$
>
> shrunk effect = prior mean + shrinkage factor × (observed lift − prior mean)
>
> τ is the real spread of effects in the register. σ is the standard error of the test.
>
> Evidence label: Standard practice. This is empirical Bayes (Berger, *Statistical Decision Theory and Bayesian Analysis*, section 4.5).

A noisy test has a small factor, so its result is pulled far towards the prior. A precise test keeps most of its lift. Two rules follow: plan with the shrunk effect, never with the raw one, and never reset a baseline to an observed lift. Repeated looks add to the problem. Suppose that you look at a test six times, at equal spacing, and that there is no real effect. The chance that at least one look shows a significant result is then 16%, not 5% (Arithmetic, by simulation). Plan the looks before the test. *Choose whether to test or act* shows how.



The prior from the register, the test alone and the shrunk result for the Cedar pricing-page test of the examples below. The shrunk lift sits between the prior and the raw lift, a little closer to the prior, because the noise of the test is larger than the real spread of effects.

Shipping thresholds

A card says how likely an option is to pay. A threshold says how likely is enough. We ship a change when the probability of a positive net value is at least 90% for a reversible change and at least 95% for an irreversible one. Both are start values (Plausible but untested), and a company tunes them with its own log.

The threshold is stricter than it looks, for three reasons. It reads the net value in euro, so cost and harm count. It uses the prior, so the winner's curse is already corrected. And it asks for a probability of paying, not a probability that "B beats A". A claim that B beats A with 95% probability, on a flat prior and with no cost in euro, can still ship a loss.

An option that misses its threshold has three exits. You can stage it: ship a small, reversible first step and let a signpost lift or kill the second step. You can look for more evidence, if the value of information says that a test pays (see *Choose whether to test or act*). Or you can drop it.

Apply the threshold to each part of a plan. A plan can mix a reversible part and an irreversible part. The irreversible part must meet 95% on its own, in the situation in which you would commit it. The two levels follow the cost of being wrong. If a reversible change is wrong, you lose the harm of the weeks before you notice, plus the cost of going back. If an irreversible change is wrong, you lose the whole commitment. A higher bar buys protection against the second case.

The rule has a price: it can pass up an option with a higher mean. The card shows that price to the owner.

A card near its bar needs one more line. For each input that is a belief, it prints the value at which the card would cross the bar. First ground what data can estimate. A year of counts gives the level of demand. A replay of a rule on past leads gives the share that it stops. The sample of a test gives its standard error. Give these

inputs the noise that the data show, not a typed range. Keep the rest as beliefs, and label them. Never narrow or move a belief to make a card pass. The extra line shows the owner how much the pass rests on beliefs that nobody has measured.

The decision log

The decision log is the record of decisions. Each entry is written when the decision is made, and it is not edited later. Its main job is to hold your belief before the outcome is known. After the outcome, memory changes. People recall that they expected the result, and they stop learning. Hindsight bias falls when people first list reasons why other outcomes could have happened. Only then do they judge what they would have expected (Arkes and others, 1988, reported in National Research Council, 2011, *Intelligence Analysis for Tomorrow*).

A log entry has these write-once fields.

- **Question.** What will the outcome answer?
- **Options.** The options considered, always including do nothing.
- **Belief.** The probability of an event, or the central value of an effect.
- **Interval.** The P10 and P90 of the belief.
- **Falsifier.** The result that would show that the belief is wrong (see *Diagnose what changed*).
- **Confirms-by date.** The date on which the outcome will be read.
- **Signposts.** The dated checks on the way, each with its planned response.
- **Owner.** A role.

Table 4 shows an entry. The store, not the author, stamps the date and time.

Table 4. A decision log entry: Birch, Year 2 February (Illustration only).

Field	Entry
Decision	A new qualification rule applies to all MQLs from February. A seller is hired when the backlog passes 65 SQLs at a month-end.
Question	Will the uncontacted SQL backlog stay at or below 65 at every month-end until the end of July?
Options	Hire a seller now. Qualification and a seller now. Do nothing.
Belief	Yes, with 97% probability, with demand on its trend and no other change to the funnel. The chance of crossing is 3%.
Interval	The highest month-end backlog by July: P10 0, median 8, P90 38 SQLs. With no action: about 280 at the end of July.
Falsifier	A backlog above 65 at any month-end means that qualification alone did not relieve the sellers.
Confirms-by date	The first week of August of Year 2.
Signposts	Backlog at each month-end. SQLs in each window, read each quarter.
Planned response	On a crossing, the sales owner approves the hire within five working days.

Field	Entry
Owner	The sales owner.

The log is stored as two tables that nobody edits. One holds the decisions, with the beliefs. The other holds the outcomes, written once, later, with a class: confirmed, refuted, inconclusive or void. The store stamps both. Owners appear as roles. The log holds the seed and the inputs of the card, so the card can be rerun.

Implementation note: the companion code holds the log, with views for due and graded decisions. It scores the beliefs after the outcome (see *Learn from outcomes*).

The log is not a file of minutes. It holds only decisions and forecasts that carry a belief. A forecast that a team makes anyway takes the same form: a question, a belief, an interval and a confirms-by date. The wins forecast of *Decide when a change needs action* is an example. The same form lets the owners grade forecasts and decisions together.

The log scores beliefs by class of action, and never by person. Calibration is the match between the probabilities that you state and what happens. It needs about 50 resolved entries pooled over a class (Plausible but untested). One person's score on a dozen entries is noise. *Learn from outcomes* explains the grading. At Year 2 February, Alder holds 4 entries and none is resolved. Birch holds 16 and 4 are resolved. Cedar holds 36 and 24 are resolved.

Who does what

Code computes the card. The analytics lead checks that each input has a source, an evidence label and a named owner. An agent may draft inputs and log entries from the facts, but it does not decide on money, price, people, commitments or consent. The owner reads the card, edits an input if there is a reason, and decides. The card, the decision and any edit or override go into the log, with the reason.

Decisions run asynchronously. The brief is the one-page card, and it states the deadline as a date, not a meeting. A start value is five working days (Plausible but untested). If the deadline passes without a decision, that fact is itself a finding, and the card goes to the next level with the date it stood open. The owner decides alone. A meeting is needed in two cases only. Two owners may have a real conflict that the card cannot settle. Or the commitment may be one that no single owner can make. Name such a meeting by its decision, and end it when the decision is made.

What changes with volume

The mechanics hold at any volume. What changes is where the inputs come from, and so how much the card can say.

Regime	What the card rests on	What it is good for
Low	Priors for the effects, and a handful of counts.	Naming the input that flips the choice, and setting signposts. The thresholds apply to the irreversible parts.
Medium	Some inputs from a year of data (demand, season, capacity), others still priors.	Separating the estimable inputs from the prior ones, and putting the prior ones behind a trigger.

Regime	What the card rests on	What it is good for
High	The company's own register of tests, so own base rates.	Shrinking each result by its own spread. The 90% bar can be met on a modest effect.

Alder, Year 2 February: whether to enter a second vertical

Alder, Year 2 February: about 3.5 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.0 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

The chief executive wants to enter food and beverage plants as a second vertical. The region holds 900 such plants. The core vertical is a small, known market (see *Find the most valuable constraint*). The plan adds one account executive and a small flow of engaged accounts. The decision is hard to reverse, so the bar is 95%.

The card counts 24 months, the payback ceiling of an enterprise segment, with these inputs.

Input	Belief	Source
Engaged accounts in a window	Median 2.5, log spread 0.35	The plan; the spread is Illustration only
Engaged account to opportunity	35%, a spread of 5 points	The plan; the spread is Illustration only
Win rate	20%, with P10 and P90 at 10% and 31%	Prior from the plan: 10% to 30%. The core vertical wins 22%.
Median sales cycle	130 days	The plan
Value of one new customer	165,576 euro, spread 10%	Value basis, contract value 45,000
Cost	Seller 185,000 a year; programme 3,327 for each engaged account	Alder's own cost lines

The register holds fewer than two tests, so these effects are priors, and the card prints the prior and its sensitivity instead of one euro value. Hold the other inputs at their means and vary the win rate. Over 24 months, a win rate of 10% gives −202,000 euro, 20% gives 118,000 and 30% gives 439,000. The switching value is a win rate of 16.3%. The prior puts 64% of its weight above that rate. So the vertical pays at the central belief, and by a narrow margin.

Across all draws the full commitment pays in 52.5% of cases. Its P10 is −355,000 euro and its P90 is 695,000 euro. It misses the 95% bar by a wide margin. The card also names the other flip points: a flow below 1.8 engaged accounts a window at a 20% win rate, or a value per customer 18% below the belief. The win rate cannot be read in less than 3 years, so no test settles it. The card does not count the longer value of a second market for a company with a small core market. That value is a judgement of the owners. The card prices it: if the early numbers disappoint, the worth of the second market beyond 24 months must carry the case.

The owners stage the move. Stage one runs six months. It is close to reversible, because the seller can move to core accounts if a signpost fails. It commits about 147,000 euro, which is inside the risk limit of 257,000. The full 24-month commitment would commit about 586,000 euro, which is not. The first signpost reads at six months: at least 12 engaged accounts and at least 3 opportunities. The second reads at twelve months: at least 2 new customers.

The log entry carries the beliefs. The chance of meeting the first signpost is 69%. The chance of meeting the second is 36%, with 1.3 customers expected by then, and a P10 to P90 of 0 to 3. If the first is met, the chance that the full move pays rises from 52.5% to 59%. If both are met, it rises to 88%. If the first is met and the second is not, it falls to 36%. So even two met signposts leave the 24-month card below the 95% bar. The second stage rests on the owners' judgement of the market beyond 24 months, and the log entry says so. The planned responses are written. If the first signpost is met, raise the flow of engaged accounts. If the engaged accounts arrive but the opportunities fall short, review the target list. If the engaged accounts fall short, stop and move the seller to core accounts.

The commitment crosses sales, marketing and budget, so no single owner can make it. The chief executive and the head of sales hold one meeting, named by its decision: "enter the second vertical, stage one". The meeting ends with the entry in the log. At low volume the card is mostly priors. Its value is that it names the input that flips the choice, and it gives the signposts a price.

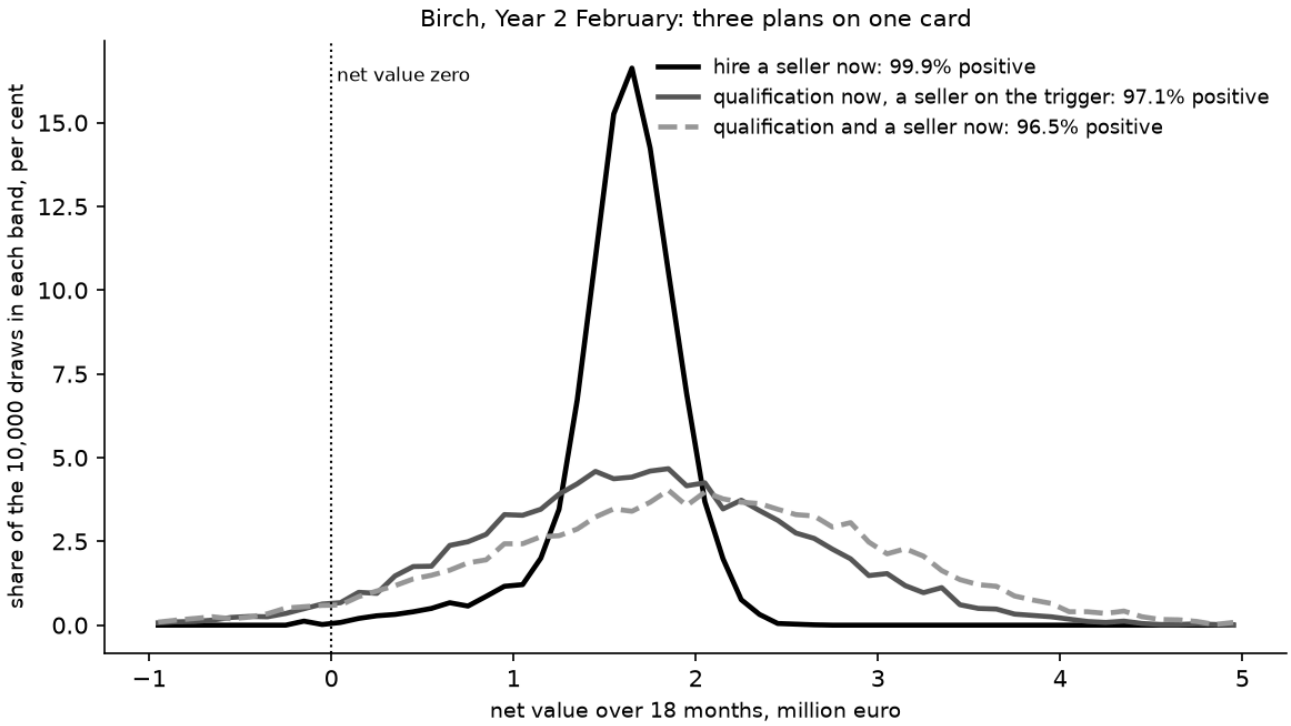
Birch, Year 2 February: a seller, better qualification, or nothing

Birch, Year 2 February: about 51 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 68 days, and the close rate reads a 20% change in about 6 months.

At the end of January, 171 SQLs waited for a seller. The load was 123%. *Diagnose what changed* found that the outbound pilot, which ended in January, explains the fall in the close rate through mix. With the pilot, sellers at full load won 44.0 customers a window. Without it, they would have won 48.0. So the gain lies in how sellers spend their time. The barometer above already counts the new rule that this card prices. The sales owner has three choices: hire an eleventh seller, tighten the qualification rule, or do nothing. The card adds two plans that combine the first two: qualification with a seller on a trigger, and both at once.

Some inputs come from a year of data. Demand follows a trend of 8% a year and a season. Ten sellers work 300 SQLs a window. A seller costs 135,000 euro a year, starts two months after approval and ramps for three months at half capacity. A year of counts, 3,949 SQLs, gives the level of demand a standard error of 1.7%, with the overdispersion factor of 1.2 from *Decide when a change needs action*. The trend is less certain, and the card gives it an error of 4 points. The owners replay the new rule on the SQLs of the last year. It would stop 6% of them, and the noise of the replay gives P10 5.5% and P90 6.5%. Customers per contacted SQL would rise 7% (P10 4%, P90 10%), to a close rate of 17.1%. The rise stays a prior, because the close rate cannot be read within a year. The rule costs 30,000 euro. It also sends the 171 waiting SQLs back to nurture, so the queue starts empty. The rise and the trend error are Illustration only. The level and the stopped share come from the data.

Table 1 gives the card for 18 months, the payback ceiling of Birch. The simulation runs the queue month by month. Hiring now has the narrowest range: it pays in 99.9% of draws. Qualification with a trigger has a higher mean and a wider range, and it pays in 97.1% of draws. The plan that does both at once has the highest mean and the best chance of being the best option, 61.2%. It pays in 96.5% of draws.



Net value over 18 months for three plans. Hiring now is the safest. The plans with qualification have more upside and a wider range.

The mean favours doing both at once. Does the rule agree? Table 5 checks the irreversible part, the seller, as a decision of its own.

Table 5. The eleventh seller as a decision of its own, with twelve months of its pay and work (Illustration only).

The hire	Chance of a positive net value	Mean net value, euro
A seller now, with no qualification	above 99.9%	1,204,000
A seller now, on top of qualification	86.6%	311,000
A seller when the backlog passes 65 SQLs	above 99.9%	1,137,000

Qualification removes the queue, so a seller hired at once on top of it has little proven work. Its chance of paying is 86.6%, below the 95% bar for a hire. A seller hired after the backlog has crossed the trigger has the work already in the queue, and the chance is above 99.9%. The trigger is the evidence that the hire needs. The plan with qualification first has a reversible first part, so its bar is 90%, and its hire clears 95% on its own.

The card falls below 90% only if demand were three times as uncertain as a year of counts shows. A trend error of 7.5 points instead of 4 also makes it fail. At a trend error of 6 points the card still pays in 93.3% of draws. The belief about the rise matters most. The card fails if that belief were 2.2 points lower, with P10 1.8% and P90 7.8%.

Hiring now also clears every bar. Against it, the qualification plan leads by about 117,000 euro on the mean, and it comes out ahead in 55% of the draws. That is close to a tie. If the extra customers of the qualification plan fall by 6.5%, hiring now leads. That move lies inside the range of the belief. So value alone does not settle the choice. The owner prefers the plan that keeps the irreversible part behind evidence. If qualification disappoints, the queue grows and the trigger hires the seller anyway.

The owner chooses qualification first, with the trigger. Against doing both at once, the owner accepts a difference in mean of about 265,000 euro, 15% of the mean of the qualification plan. The hire follows in 72% of the draws within 18 months. The trigger of 65 SQLs equals a wait of 6.1 days for first contact. It is a judgement of the owners, and the log says so. The qualification effect cannot be read for a year, so the sales owner reads the backlog and the count of SQLs instead.

Find the most valuable constraint raises one more question: what happens to SQLs that a lever adds while ten sellers are full? Take a lasting lift of 4.2% in SQLs, as in the landing-page lever, which adds 174 SQLs over twelve months. Under ten sellers the queue never clears: demand runs 5.9% over capacity in the first year of the projection and 14.3% over in the second. At month 18 the queue holds 673 SQLs, a wait of 63 days, against a median sales cycle of 34 days. With the lift it holds 847, and every SQL waits 16 days longer, so the extra SQLs are never worked, in 18 months or in 40. An eleventh seller still leaves 201 SQLs in the queue at month 18, since demand keeps growing. Birch has no data on how fast a waiting SQL goes cold. So the card shows two bounds: worked at full quality, the 174 SQLs would be worth at most 733,000 euro; gone cold, nothing. The card counts nothing. In practice Birch sends old SQLs back to nurture, as it did with the 171 SQLs of January.

The sales owner decides by the end of the week, without a meeting. The log entry of Table 4 holds the belief, the falsifier and the trigger. At medium volume some inputs come from a year of data and others stay priors. Put the prior ones behind a trigger.

Cedar, Year 2 February: a pricing-page winner

Cedar, Year 2 February: about 1,700 new paying accounts a window, 1.3 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 27 days; the assisted motion, with 70 opportunities a window, needs about 7 months to read a 30% change.

A pricing-page variant "won". The test ran its planned length of two windows at an even split, with 17,592 activated users in each arm. The analytics lead looked at it six times. At the last look, the lift in the activated-to-paid rate was 9.0%, with a standard error of 3.3 points. It would add 150 paying accounts a window. The change is reversible, so the bar is 90%.

Cedar's register holds 62 tests, more than 30, so Cedar uses its own prior. (It held 32 tests when it passed 30, as *Decide what the evidence can resolve* showed.) The true effects of its tests average 0.0% with a spread of 3.0 points (Illustration only). Take each test at the precision of this one. Then the observed lifts spread by 4.5 points, which is the square root of 3.0^2 plus 3.3^2. The shrinkage factor is 0.45. The shrunk lift is 4.1%, with a P10 to P90 of 1.2% to 6.9%. The raw result is 2.2 times larger. The register shows why. On these numbers, 11.2% of Cedar's tests end as significant winners, which is about 7 of 62. The winners' average observed lift is 7.6%, and their average true lift is 3.4%.

Two cards use the same action and two beliefs. Value per paying account is 1,683 euro. The cost to roll out is about 20,700 euro. The added accounts may renew less than average, so a quality ratio with a mean of 0.90 prices the harm on the right wing (Illustration only).

Table 6. One action, priced on two beliefs (Illustration only).

Belief about the lift	Mean net value, euro	P10 to P90, euro	Chance of positive net value
The raw result	2,770,000	1,445,000 to 4,121,000	99.7%

Belief about the lift	Mean net value, euro	P10 to P90, euro	Chance of positive net value
The shrunk result	1,239,000	342,000 to 2,150,000	96.4%

The raw card clears the bar by far. The shrunk card clears it too, by 6.4 points. The flat prior overstates the value by a factor of 2.2 and the chance of paying by 3.3 points. The break-even effect is a lift of 0.07% in the conversion rate, far below what any test can read. So the question is not the size of the lift. It is whether the lift is above zero, and the shrunk answer is 96.4%.

The length of the test matters. Read after one window, the same test would have a standard error of 4.7 points. The shrunk card would then pay in only 84.3% of draws, below the bar, so a stop at an early look would have failed. The card also fails in three other cases: a raw lift below 6.5%, a spread of the register below 1.9 points, or a standard error above 4.0 points.

The growth lead owns the decision. The change is reversible and the bar is met, so the growth lead ships the variant with a 10% holdback (a start value, Illustration only). A *holdback* keeps a share of users on the old page, so the effect can still be read later (see *Choose whether to test or act*). A holdback of 10% held for a year can read only a lift above 6.1%. An even split can read a lift above 3.6% in a year. A lift of 4.1% takes 291 days to read at an even split. So this holdback cannot confirm the belief. The log entry says so. The belief is 4.1%, with the interval 1.2% to 6.9%. The falsifier is a holdback that converts better than the variant by more than its noise. The confirms-by date is Year 3 February, when the holdback has run for a year. The grade of the effect comes from the pooled calibration of all register entries, not from this test. At high volume own base rates exist, and the shrunk effect is much smaller than the raw one.

The decision to carry forward

Choose against a dated baseline and record what could reverse the recommendation. Scale the commitment to its downside and reversibility. Preserve the original belief for later evaluation rather than replacing it with the explanation that fits the outcome.

20 · Fund the next best action

Which action receives the next unit of money, people or engineering time? Rank complete decisions on a common economic basis, then account for the resource that binds. A high expected gain can be unattractive if it consumes the capacity needed for a better opportunity.

The portfolio also determines what the business learns. Several actions may depend on the same uncertain customer behaviour or compete for the same accounts. Their risks and benefits cannot always be added independently. Testing an assumption used by several decisions can be more valuable than optimising one isolated experiment.

This chapter connects expected value, delay and constraints to funding. Make the cost of waiting visible, give each decision an owner and write the conditions for stopping or expanding the commitment. The ranking is an input to an accountable choice.

Why scoring frameworks mislead

Growth teams often score ideas with one of four scores. ICE multiplies impact, confidence and ease. RICE adds reach and divides by effort. PIE scores potential, importance and ease. PXL asks yes-or-no questions about a page. We call them *scoring frameworks*. They are standard practice, and we do not use them, for three reasons.

First, the inputs are labels and not quantities. "Impact 3" means "massive", and a score of 12 is not twice a score of 6 in euro or in anything else. Second, the score has no running cost, no capacity limit, no time to effect and no range. Third, the order depends on the scale: two teams can score the same facts on two reasonable scales and get two orders.

Take the five next changes on the growth list of Birch in Year 2 July. There is one for each stage of the chain, and the inputs come from *Find the most valuable constraint*. We leave paid media off the list while its budget pulses run, until November (see *Choose acquisition channels*). Table 1 scores the five with RICE.

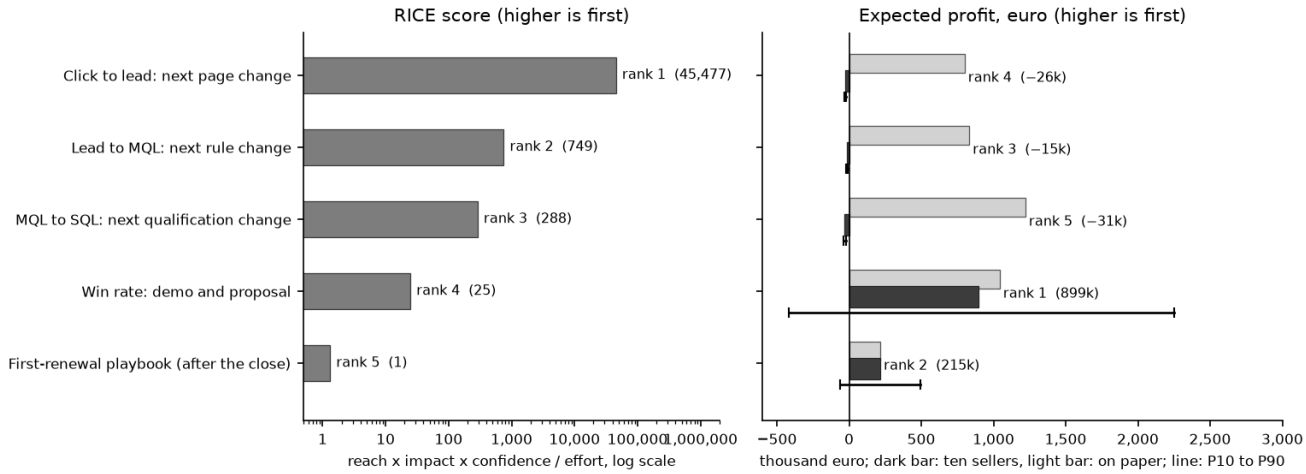
Table 1. Birch, Year 2 July: five next changes scored with RICE on the usual scale (Illustration only).

Next change	Reach, units a quarter	Impact	Confidence	Effort, person-weeks	RICE score	RICE rank	Expected-profit rank
Click to lead: next page change	246,333	3	80%	13.0	45,477	1	4
Lead to MQL: next rule change	11,690	1	50%	7.8	749.3	2	3
MQL to SQL: next qualification change	4,500	2	50%	15.6	288.5	3	5
Win rate: demo and proposal	514	2	50%	20.8	24.7	4	1

Next change	Reach, units a quarter	Impact	Confidence	Effort, person-weeks	RICE score	RICE rank	Expected-profit rank
First-renewal playbook (after the close)	164	0.5	50%	31.2	1.3	5	2

Reach is the units that enter the stage in a quarter. Impact bins the mean gain of the rate: 3 at 8% or more, 2 at 5%, 1 at 2.5% and 0.5 at 1%. Confidence is 80% when the P10 of the gain is above zero, and otherwise 50%. Effort is the one-off cost divided by 1,923 euro, the pay of one person-week of the marketing team (100,000 euro a year divided by 52).

RICE puts the page change first and the win rate fourth. The expected profit reverses the two. The rank correlation between the two orders is −0.4 (Kendall's tau, where 1 means the same order and −1 means the reverse order). Three things explain the gap.



Birch, Year 2 July. Left: RICE score of each next change, with its rank. Right: expected profit with ten sellers (dark bar, P10 to P90 as a line) and on paper before the seller check (light bar).
The first is capacity. Birch has ten sellers. Demand for their slots is 108.5% of capacity, and a queue of 79.7 SQLs waits. *Allocate sales attention* showed that an extra SQL adds no customer until a seller has room. Three of the five changes add SQLs. After the capacity check each is worth only minus its cost. No input of RICE holds capacity, so RICE cannot see this.

The second is value. RICE counts a visitor, a lead and an opportunity as one person each. A lead is worth 488 euro of expected gross profit, an SQL 5,399 euro and an opportunity 10,250 euro. Reach is largest where each unit is worth least.

The third is scale. Score the five with ICE in two reasonable ways. The first scores impact as the mean gain in points, confidence as 8 or 5, and ease as 10 minus a third of the effort weeks. The second uses five impact bands scored 2 to 10, confidence 7 or 4, and ease 9, 6 or 3 by effort band. The first ranks qualification second (150) and the lead-to-MQL rule third (140). The second ranks the rule second (216) and qualification third (192). Neither is more true.

The cure is to estimate what the score tries to estimate: expected profit, with a range, a cost, a delay and a capacity check.

The card as the ranking unit

Rank by the card of *Commit with uncertainty*. Rank first by the mean. Then read the range. Then read the probability of being the best option, which the card also prints. Here it answers a new question: how sure is the order? Put the items and "do nothing" side by side, and compare them draw by draw.

Table 2. Birch, Year 2 July: the cards of the five next changes with ten sellers, in euro (Illustration only).

Next change	Mean	P10 to P90	Probability of loss	Probability of being the best option	Mean on paper, before the seller check
Win rate: demo and proposal	898,650	−419,224 to 2,249,006	19.4%	73.8%	1,041,829
First-renewal playbook (after the close)	215,028	−61,563 to 491,689	15.8%	23.3%	215,028
Lead to MQL: next rule change	−15,477	−20,766 to −10,898	100%	0.0%	831,233
Click to lead: next page change	−25,784	−34,414 to −18,100	100%	0.0%	798,973
MQL to SQL: next qualification change	−31,012	−41,409 to −21,769	100%	0.0%	1,219,844
Do nothing	0	0	0%	2.9%	0

Read the table in three steps. First, the three changes that add SQLs lose their whole gain to the queue, so they are never the best option. Second, the win-rate card is the best option in 73.8% of draws and the playbook in 23.3%. The playbook has the lower mean and the narrower range, so it is not out. Third, doing nothing is best in 2.9% of draws, those in which both good cards lose money.

The capacity check changed the order. On paper, before the check, qualification is the best option in 38.3% of draws and the win rate in 34.2%. That is the lesson of *Find the most valuable constraint* in the language of ranks. The heuristic of the top of the funnel holds on paper. It fails when a later stage is full.

A ranking is a distribution and not a list. When two cards have close probabilities of being the best option, their order matters little. Put the one that is easier to undo first. Or let the value of information decide whether to learn before you choose (see *Choose whether to test or act*).

Only items with a card enter the ranking. An idea without a card has no rank, because it has no expected value, no cost and no range. The backlog is a list of cards and not a list of wishes.

We drew each card on its own. Cards that share a cause, such as the same sellers or the same value of a customer, must be drawn together (see *Find the most valuable constraint*).

Cost of delay

Choose whether to test or act used the cost of delay to price the weeks that a test runs. Here is the full method. A week of waiting is not free: an effect that starts a week later loses a week of its gross profit. The *cost of delay* is that loss for each week, estimated as the weekly gross profit of the effect.

> **Calculation**
>
> cost of delay per week = expected gross profit of the effect ÷ weeks of lever life
>
> cost of a delay of d weeks = cost of delay per week × d

The gross profit is the figure of the card after the capacity check and before the cost of the lever. The lever life is the 12 months of *Allocate sales attention*, counted as 52 weeks, as in *Choose whether to test or act*. When the change is taken only in some outcomes, as after a test, weight it by the chance that you take it, as *Choose whether to test or act* does.

The rule assumes that a later start does not move the end of the gain. That holds when a season, a successor change or the market ends the gain on a fixed date. If a later start only shifts the whole flow, the week is paid later, not lost, and the true cost is smaller. We take the fixed end as the start value (Plausible but untested). When an effect ends on a known date, for example a campaign, use the weeks that remain.

Take the win-rate card of Birch. Its gross profit is 939,946 euro, so each week of delay costs 18,076 euro (Arithmetic, given a lever life that is Plausible but untested). The three changes that add SQLs have a cost of delay of zero while the queue binds.

Cost of delay turns the order of work into a sum. One scarce team does items one after another. Which order loses least? Compare two neighbours, A then B against B then A. If A goes first, B waits for the whole duration of A, which costs the cost of delay of B times the duration of A. If B goes first, A waits for B. So A should go first when its cost of delay divided by its duration is larger. We call this ranking value the *cost of delay per week of work* (also called CD3).

> **Calculation**
>
> cost of delay per week of work = cost of delay per week ÷ weeks of the scarce resource that the item needs

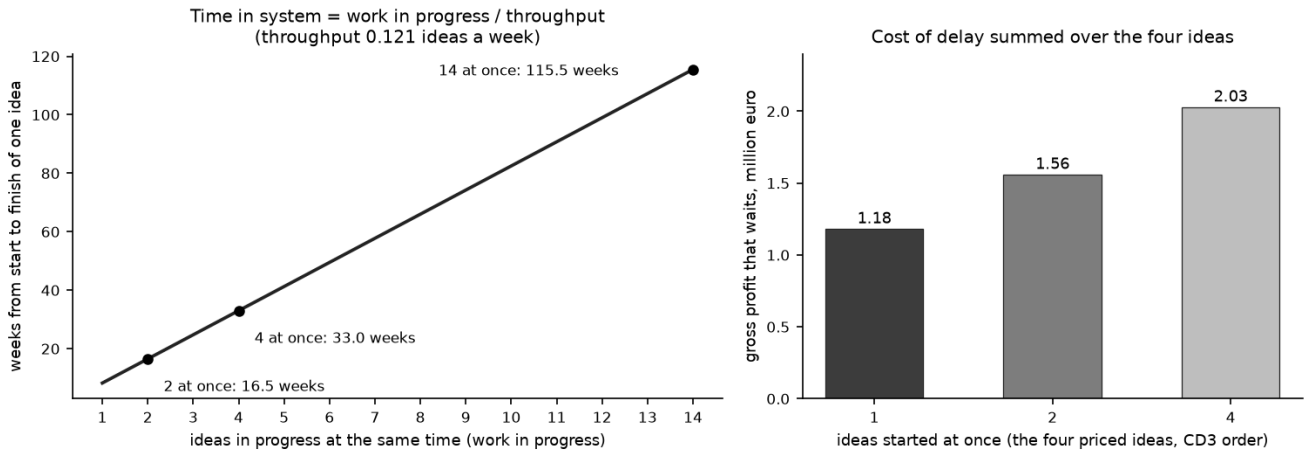
The rule is also known as weighted shortest job first (Reinertsen, 2009, *The Principles of Product Development Flow*). A short item with a modest cost of delay can beat a long item with a high one. The Alder example below shows a swap of this kind.

Work in progress is the number of items started and not finished. *Throughput* is the number of items finished in a week. *Time in system* is the weeks that one item takes from start to finish. Little's law of *Forecast conversion and timing* links the three.

> **Calculation**
>
> time in system = work in progress ÷ throughput

A team's throughput is set by its capacity. Starting more items does not finish more. Each item takes longer, and its profit starts later. *Set targets and staff capacity* showed a second effect. When load rises towards 100%, the wait grows faster than the load: for one handler, from 90% to 95% the wait is 2.1 times longer. The same holds for experiment teams (Thomke, 2020, *Experimentation Works*).

So set a limit on work in progress. Start a new item only when one finishes. Watch two numbers. One is the age of the oldest open item. The other is the weeks that an item with a high cost of delay waits before it starts. Raise an alarm when that wait passes one cycle of the team.



Alder, Year 2 July (the example below gives the numbers). Left: time in system against work in progress for a pair that finishes one idea every 8.25 weeks. Right: gross profit that waits, summed over four priced ideas, when the pair starts one, two or all four at once.

Scarce resources

When something binds, rank by value per unit of the scarce thing and not by value alone. Three resources bind most often. People bind in every company, and the unit is the person-week. Handler slots bind where a queue exists, such as the seller slots of *Allocate sales attention*, and the shadow price of that chapter prices them. Test traffic binds where a test needs many units, and the unit is the slot-week.

A *slot-week* is one week of the whole flow of eligible units, used by one test. A quarter of high traffic still holds few slot-weeks, because a test on a money outcome needs a large sample.

> **Calculation**
>
> slot-weeks of a test = units that the test needs ÷ eligible units per week
>
> EVSI per slot-week = EVSI of the test ÷ slot-weeks of the test

EVSI is the expected value of sample information of *Choose whether to test or act*. A test with a high EVSI per slot-week buys the most learning for the traffic that it uses. Choose the size of a test by the same rule. EVSI rises with the sample, but in smaller steps each time. Add traffic while the next slot-week adds more EVSI than the hurdle (defined in the next section).

Name the unit on every card. Alder counts person-weeks of one marketer and part of an analyst. Birch counts person-weeks and seller slots. Cedar counts slot-weeks. Attention also binds. An owner can read only so many results in a week, so count reading time when the list is long.

Funding without a vote

A common habit is to send the top items of a ranked list to a standing meeting for a vote. We do not do this, for three reasons. The meeting adds a wait, and the wait has a cost of delay. The list is old when the meeting reads it. And a vote replaces the number on the card with an opinion.

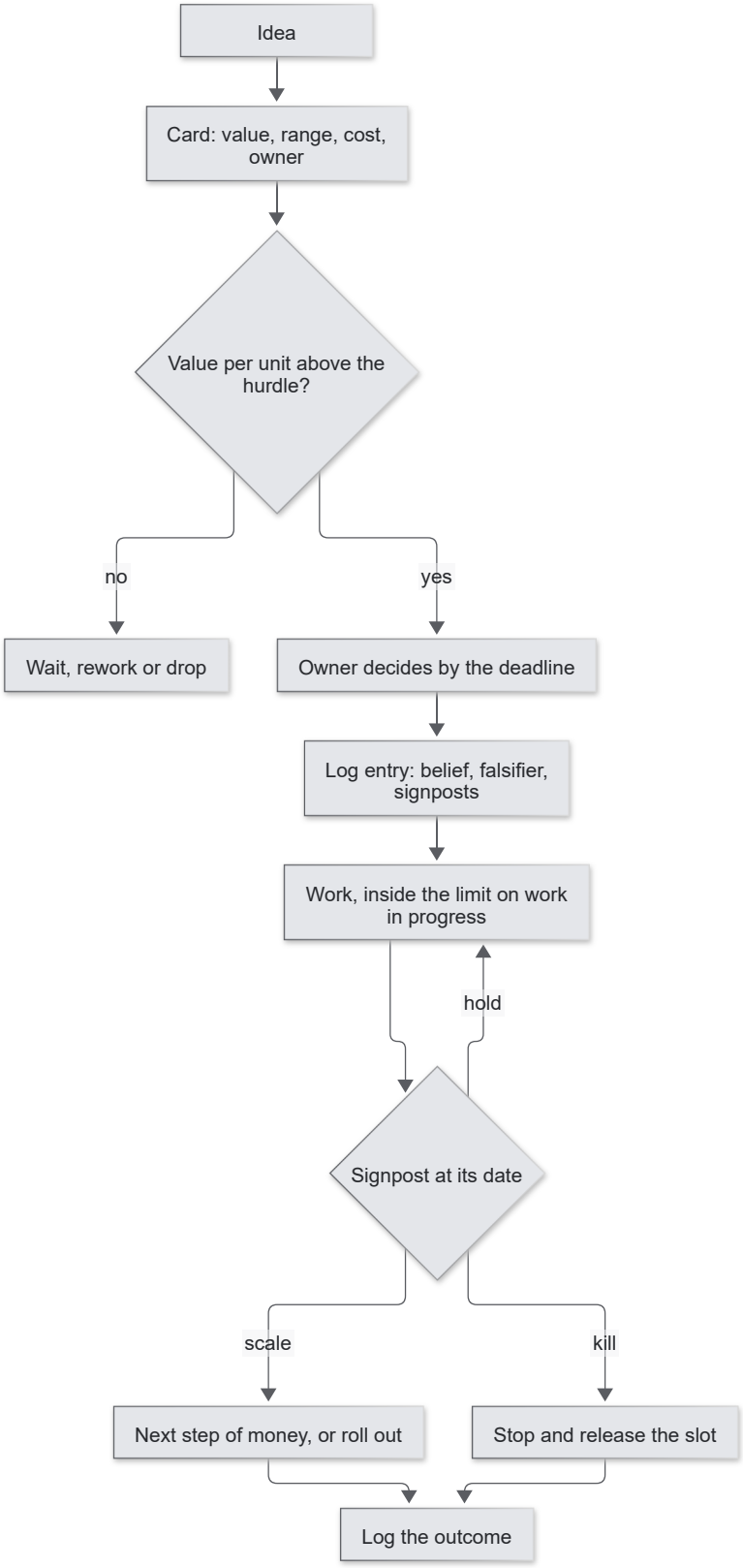
Use a *hurdle* instead. The hurdle is the minimum expected value per unit of the scarce resource that a card must clear. The named owner of a card above the hurdle funds it on the day it arrives. Set the hurdle in five steps.

1. Take last quarter's cards and sort them by expected value per unit of the scarce resource.
2. Walk down the list. Add the units that each card needs until the quarter's supply is used.
3. The value per unit of the last card that fits is the hurdle for this quarter.
4. Fund any new card above the hurdle when it arrives. Do not hold it for a ranking.
5. Each quarter, lower the hurdle if the resource sat idle. Raise it if funded items wait longer than the time in system that you accept.

A company with no history starts from a multiple of the pay of the scarce resource, for example ten times (Plausible but untested). For one slot, theory gives a reason to trust a single threshold. A well-chosen threshold earns at least half of what a chooser with hindsight earns. This is the prophet inequality (Samuel-Cahn, 1984, as summarised by Chawla and others, 2024, *Algorithmic Game Theory*, Springer). With many slots and a stable list a hurdle does far better, as the Cedar example shows (Illustration only).

The decision is asynchronous. The owner named on the card decides by a deadline. The log entry holds the belief, its interval, the falsifier, the confirms-by date and the signposts. Code computes the card, and a person decides.

A meeting has two uses. Call one when two owners compete for a resource that they share, and the hurdle cannot settle it. Call one when a commitment is larger than the authority of any one owner, for example a hire or a contract. Name the meeting by its decision, for example "who has the analyst this quarter", and end it when the decision is in the log.



From cards to funded work. The hurdle funds a card on arrival. The trigger sheet tells the owner what to do at each signpost, and the log keeps the belief and the outcome.

The hurdle funds cards one by one. It does not say whether the funded set can reach the goal. For that, compute the *probability of covering the target*. Add the net-value draws of the funded cards, and take the share of draws at or above the target. Take a target of 1,000,000 euro of net value from the quarter's funded changes. The target is a choice of the owners (Illustration only). At Birch the win-rate card alone covers it in

46.0% of draws. With the playbook added the share is 54.5%, but the playbook needs 31.2 more person-weeks and fails the hurdle. A low probability is a message. Fund more cards, lower the target or say that the target is a hope. Do not stretch the estimates.

Implementation note: the companion code computes the card, draws shared causes once for all options (`decision_card` with its shared samplers) and gives EVSI for a normal prior (`evsi_normal`).

Kill and scale triggers

A funded bet can be wrong. Decide before you fund what will tell you so, and what you will do then. Write it on a *trigger sheet*. The sheet follows plans made for deep uncertainty (Walker, Marchau and Kwakkel, in *Decision Making under Deep Uncertainty*). It has four parts.

- **Load-bearing assumptions** are the beliefs whose failure ends the case. Take them from the switching values of the card, the inputs that flip the decision.
- **Signposts** are leading measures that you can read inside the horizon. Each has a level and a date for scale, hold and kill.
- **One contingent action** belongs to each trigger: defend, correct, capitalise or reassess.
- **The strength of the link** from the signpost to the lagging profit goes into the log. Later you can check it (see *Learn from outcomes*).

Table 3. Birch, the seller card: one trigger sheet, as it stood in Year 2 July.

Part	What it holds	Birch, the seller card
Load-bearing assumption	The belief whose failure ends the case	SQL demand stays above what ten sellers can work. Demand is 108.5% of capacity.
Signpost and level	A leading measure with a level	Uncontacted SQLs at the end of a month. The written level is 65 SQLs.
Action when it trips	One contingent action	The sales owner approves one more account executive.
Hold rule	When the rule pauses	From approval until the new seller works at full capacity, the rule does not fire again.
Link to profit	How the signpost predicts profit	An SQL that waits puts at risk 5,399 euro of expected gross profit, the value of an SQL.

Staged funding pays the money in steps, called tranches. It limits the loss from a wrong bet. It also delays the right ones, and each gate has a cost of delay. Use tranches only when they pay, as the inequality below states. For a cheap, reversible item, fund in full and keep the trigger. For a large item that you cannot undo, the inequality often holds, and tranches pay. A trigger acts on a signpost, because a lagging result, such as the close rate, comes too late to trip it.

> **Calculation**
>
> use tranches when (saving from an early stop × chance the stop is correct) > (cost of delay of the gate + value lost by false stops)

When you cannot measure

At low volume, or for a lasting change, no test can read the effect. Then you decide on the prior. The steps stay; only the evidence is thinner.

1. Write the prior from outside base rates and from cases (see *Commit with uncertainty*).
2. Check the card. If its P10 is above zero, the case holds even if no rate can read the result.
3. Choose the reversible form: a staged rollout with a holdback (see *Choose whether to test or act*).
4. Read the earliest outcome that can be read, and write down what it means.
5. Set the signposts and write the belief in the log before the outcome.

Exploiting means acting on the best current belief. *Exploring* means taking an action to learn. Exploring costs the profit of the better option. So explore only when the information is worth more than that profit. When no test can read the effect, explore through a small random arm that the new rule does not touch. *Allocate sales attention* does this with a share of 10%.

Take the win-rate card of Birch. The win rate reads a 20% change in 147 days. A gain of 5% needs 2,355 days, which is 6.4 years. The card cannot wait for that read. It rests on its prior and names an earlier signpost, with the range and the date.

Portfolio risk

Each funded bet has its own risk. Together the bets have another. A *portfolio* is the set of all funded bets. Look first at the *false confirm*: a change that passes its rule although it has no effect. Suppose a rule confirms a null change with probability 2.5% (one-sided, the usual 5% two-sided rule). Then n independent null bets give at least one false confirm with this probability.

> **Calculation**
>
> chance of at least one false confirm = 1 − (1 − 0.025)^n

Table 4. Chance of at least one false confirm among n independent bets that all do nothing (Arithmetic).

Bets	1	3	5	7	10	30	60
Chance	2.5%	7.3%	11.9%	16.2%	22.4%	53.2%	78.1%

This is the worst case, where every bet does nothing. Even so, a quarter with many bets will confirm some false ones. Four controls help. Shrink each winner towards the prior (see *Commit with uncertainty*). Ask for a higher bar for changes that you cannot undo. Keep a holdback on each shipped winner. It guards against harm and a novelty effect. It rarely confirms a small lift. Record every result, losers included, in the register. Then the base rate of winners can be read (see *Learn from outcomes*).

A second risk is a shared cause. If five bets depend on one assumption, the chance that all five fail together is the chance that the assumption fails. It is not the product of five small chances. Cap the loss on the bets that share a cause.

What changes with volume

The method is the same at every volume. What limits the order of work is not.

Regime	What limits the order of work	Consequence
Low (Alder, about 3.5 customers a window)	People. Cards rest on priors with wide ranges.	Sequence by cost of delay per week of the scarce pair. Limit work in progress. Decide on priors with signposts.
Medium (Birch, about 51 customers a window)	The capacity of one stage, and the reading time of each rate.	Run the capacity check before you rank. Fund by hurdle. Some cards confirm in weeks, others never within a year.
High (Cedar, about 1,500 customers a window)	Traffic, time and attention, not data.	Rank tests by EVSI per slot-week. False confirms grow with the number of tests, so shrink and hold back.

Alder, Year 2 July: fourteen ideas, one pair of hands

Alder, Year 2 July: about 3.5 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 2.9 years to read even a 30% change, and the lead count reads a 20% change in 24 days.

The marketer keeps a list of 14 ideas. The marketer and half an analyst must do all of them. Only four of the fourteen have a card with a euro value: the four levers of *Find the most valuable constraint*, priced again for this month (Illustration only). The other ten wait until someone writes a card.

Table 5. Alder, Year 2 July: the four priced ideas in order of cost of delay per week of work (Illustration only for the weeks of the pair).

Idea	Expected gross profit, euro	Cost of delay a week, euro	Weeks of the pair	Delay cost per week of work	Chance of the highest delay cost per week of work
Site expansion at existing accounts	1,852,432	35,624	10	3,562	60.5%
Account selection (engaged to opportunity)	687,063	13,213	5	2,643	34.1%
Win rate (solution engineering)	818,649	15,743	12	1,312	4.3%
More leads and events (engaged accounts)	316,820	6,093	6	1,015	1.0%

The order of expected profit would be expansion, win rate, selection, events. The cost of delay per week of work puts account selection ahead of the win-rate work: selection needs 5 weeks of the pair, the win-rate work 12. Expansion has the highest value in 60.5% of draws and selection in 34.1%. So the first slot is clear, and the second is a close call. The call is cheap, because both items start.

The pair works on at most two ideas at once (a start value, Illustration only). Two items start now, expansion and selection. The third starts when one finishes. In the model, selection finishes in week 10, and the win-rate work starts then. Take the gross profit that waits, summed over the four items. If the pair does them one at a

time in this order, it is 1,180,555 euro. With a limit of two it is 1,559,095 euro, which is 32.1% more. If the pair starts all four at once it is 2,028,249 euro, 71.8% more. The order matters less than the limit. One at a time in the order of expected profit costs 6.8% more. The worst order costs 52.4% more.

A limit of one would cost least in this model. Alder keeps two, because work stops for reads, data and holidays, and the model does not show the cost of an idle pair. Little's law gives the wait. The pair finishes an idea every 8.25 weeks on average. With two ideas at once, each takes 16.5 weeks from start to finish. With all 14 at once, each would take 115.5 weeks, more than two years.

The marketer writes one log entry for each started item. Each entry holds the belief with its range, and a signpost with a date inside the weeks of work. When a signpost fails, the owner releases the slot at once. No rate can read these ideas, so the cards rest on priors. The lesson at low volume is that people are the bottleneck. Sequence the work by cost of delay per week of the scarce pair, and count the hurdle in person-weeks.

Birch, Year 2 July: the council retires

Birch, Year 2 July: about 51 new customers a window, 0.8 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 65 days, and the close rate reads a 20% change in about 6 months.

A fortnightly growth council has ranked the backlog with a scoring framework. In Year 2 July the owners retire both. Their last list is the one of Tables 1 and 2. Ten sellers already carry a queue of 79.7 SQLs, and the load stays above capacity in 10 of the next 12 months. A RICE order would put three changes that cannot earn at the top.

Under the new rule the owners decide on cards against a hurdle. Birch sets the hurdle at ten times the pay of a person-week, which is 19,231 euro of expected net profit per person-week (Plausible but untested). The win-rate card needs 20.8 person-weeks and has an expected net profit of 898,650 euro, which is 43,204 euro per person-week. It clears the hurdle. The playbook has 6,892 euro per person-week, so it stays on the list, and the owner of renewals may rework it into a cheaper form. The three changes that add SQLs are below zero until a seller has room. No vote is needed.

The sales owner funds the win-rate card on the day it arrives. The log entry states the belief of 898,650 euro, with a P10 of −419,224 euro and a P90 of 2,249,006 euro. The probability of loss is 19.4%. The win rate cannot show a gain of about 5% for 6.4 years. So the entry names an earlier signpost, the time that opportunities spend in their stage (see *Forecast conversion and timing*).

The old council also had a price. A fortnightly meeting makes a card wait one week on average and two weeks at most. For the win-rate card, the average wait costs 18,076 euro and the longest 36,152 euro.

In the same month, the trigger of Table 3 trips. The uncontacted SQL backlog ends the month at 79.7, against the written level of 65. The queue holds expected gross profit of about 430,400 euro (79.7 SQLs at 5,399 euro) that waits for a seller. The sales owner approves one more account executive that day and logs the decision. No meeting is held.

The February log entry gave this crossing a chance of only 3%. The shorter form, shipped in May, raised demand after that card was written; *Learn from outcomes* grades the entry. The trigger itself worked as written. The rule pauses until the new seller works at full capacity.

The weekly memo agent (an AI agent that drafts the one-page weekly summary) starts in shadow in the same month. It drafts the card tables and log entries, and nobody acts on them yet (see *Run the growth system*). The lesson at medium volume is that the order of work changes when you rank by expected value. Decisions also stop waiting for a meeting.

Cedar, Year 2 July: sixty ideas, thirteen slot-weeks

Cedar, Year 2 July: about 1,500 new paying accounts a window, 1.4 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 24 days; the assisted motion, with 59 opportunities a window, needs about 7 months to read a 30% change.

The growth team has 60 test ideas for the quarter. The quarter holds 160,321 sign-ups in 3.25 windows. That is 13.0 slot-weeks of 12,332 sign-ups each. A test sized to read a 10% change in activation needs 5,176 sign-ups, which is 0.42 slot-weeks. The quarter fits 31.0 such tests. A test of the same size on trial to paid needs 87,472 sign-ups, which is 7.09 slot-weeks. The quarter fits 1.83 of them. So the cheap outcome covers about half of the 60 ideas, and the money outcome covers fewer than two. The barometer's 24 days for trial to paid count one of two equal periods. A test splits one flow into two arms of 43,736 sign-ups, so it needs both halves.

The register holds 92 well-recorded tests. Its prior for the relative effect of one test is a mean of 0 and a standard deviation of 3%. We ship a winner at 10,000 euro (Illustration only). The value of a gain in activation rests on the quality ratio of 0.90 of the activation lever in *Find the most valuable constraint*. That ratio links activation to payment, and we have not checked the link (Plausible but untested).

Table 6. Cedar, Year 2 July: four test designs, in euro.

Test	Sign-ups needed	Slot-weeks	EVSI	EVSI per slot-week
Activation, sized for 10%	5,176	0.42	267,551	637,469
Activation, sized for 20%	1,294	0.105	159,518	1,520,276
Trial to paid, sized for 10%	87,472	7.09	300,441	42,358
Trial to paid, sized for 20%	21,868	1.77	180,098	101,565

A test on trial to paid has about the same EVSI as one on activation. It needs 16.9 times the traffic, so its EVSI per slot-week is 15.0 times lower. The size of a test is a purchase too. The test sized for 20% uses a quarter of the traffic and gives 59.6% of the EVSI. A rule that prefers the outcome nearest to money would spend the quarter on one or two tests. A rule that prices traffic spreads it over many.

Now compare ways to fill the 13.0 slot-weeks. We simulated 400 quarters of 60 seeded ideas each (Illustration only). On average a third of the ideas target trial to paid. The prior spread of each idea varies around the register value. Every test keeps the 10% size, to keep the comparison simple. A hurdle taken from the previous quarter has a mean of 375,648 euro of EVSI per slot-week. Its P10 to P90 is 246,701 to 502,620 euro. A card above it is funded on arrival. This captures 93.2% of the EVSI that a team with hindsight would capture. Funding in order of arrival captures 40.2%. Funding the trial-to-paid ideas first captures 77.2%. The choice of ideas matters more than the number of tests.

The hurdle also settles the size. For an activation idea with the register prior, the step from the 20% size to the 10% size adds 0.31 slot-weeks. It earns 343,200 euro of EVSI per extra slot-week, which is 91% of the mean hurdle. So at this hurdle the 20% size is the better buy, and the slot-weeks that it saves go to the next card. When traffic is scarce, the hurdle is high and tests get smaller.

The portfolio carries its own risk. A quarter of 30 tests gives a 53.2% chance of at least one false confirm if every idea does nothing. Under the register prior, about 6.8% of activation tests pass the 2.5% rule. Of the passes, 5.8% have a true effect below zero. The passes show a mean lift of 9.0%, while their true mean is 3.8%. So Cedar logs the shrunk effect, not the raw one. It keeps a 10% holdback on each shipped winner, as it did for the pricing page in February. The lesson at high volume is that the scarce resource is traffic, time and attention, not data. Choose tests by EVSI per slot-week on the card, and do not choose them by vote.

The decision to carry forward

Fund the best feasible combination of actions under the actual resource constraints. Reconsider shared assumptions and correlated exposure. Make stop, scale and review conditions part of the commitment itself.

21 · Decide when a change needs action

When does a change in the numbers justify attention or action? An alert uses scarce investigative capacity. A sensitive detector can create more work than the company can examine, while a slow one can leave an expensive fault in place. The operating decision is how to balance those losses.

First compare the observation with an expectation that reflects season, mix, volume and known changes. Then check whether the observation process still works. A broken event feed and a real fall in customer demand can produce the same chart.

This chapter builds detection around those decisions. Signal limits control false alarms under stated assumptions. They do not diagnose a cause or establish that an intervention will pay. An alert begins the next comparison, with an owner and a deadline.

Why "falls more than 15%" fails

Take Birch, our medium-volume company, at the trend level of Year 1 November. Suppose that nothing changes in the funnel. Each rate still varies from window to window by chance. The rule compares each window with the one before, so it adds the noise of two windows. Table 1 shows how often the rule fires on nothing. The overdispersion factor is the 1.2 start value (Plausible but untested), and the rule watches falls only. Birch's board paper puts a band of this kind on 11 stage counts and rates (see *Set targets and staff capacity*). That band rings both ways, so it fires about twice as often as a rule on falls alone.

Table 1. Birch, no real change: the fixed rule "falls more than 15% against the prior window". Simulated, 200,000 pairs of windows each (Illustration only).

Metric	Units in a window	Relative standard error of one window	Windows in which the rule fires	Fires a year
Lead to MQL	3,007 leads	2.4%	0.0%	0.0
MQL to SQL	1,203 MQLs	5.5%	1.8%	0.2
SQL count	301 SQLs	6.3%	3.5%	0.5
Win rate (opportunity to won)	150 opportunities	13.0%	19.0%	2.5
New customers (count)	48 customers	15.8%	23.5%	3.1

The same 15% is too wide for the top of the funnel and too narrow for the bottom. At the top, a 15% fall is a huge event, and the rule is blind to smaller real falls. At the bottom, a 15% move is ordinary noise. Eleven rules of this kind, one on each of Birch's 6 stage counts and 5 stage rates, fire about 8.4 times a year on nothing. People learn to ignore such a rule.

The rule has three more faults.

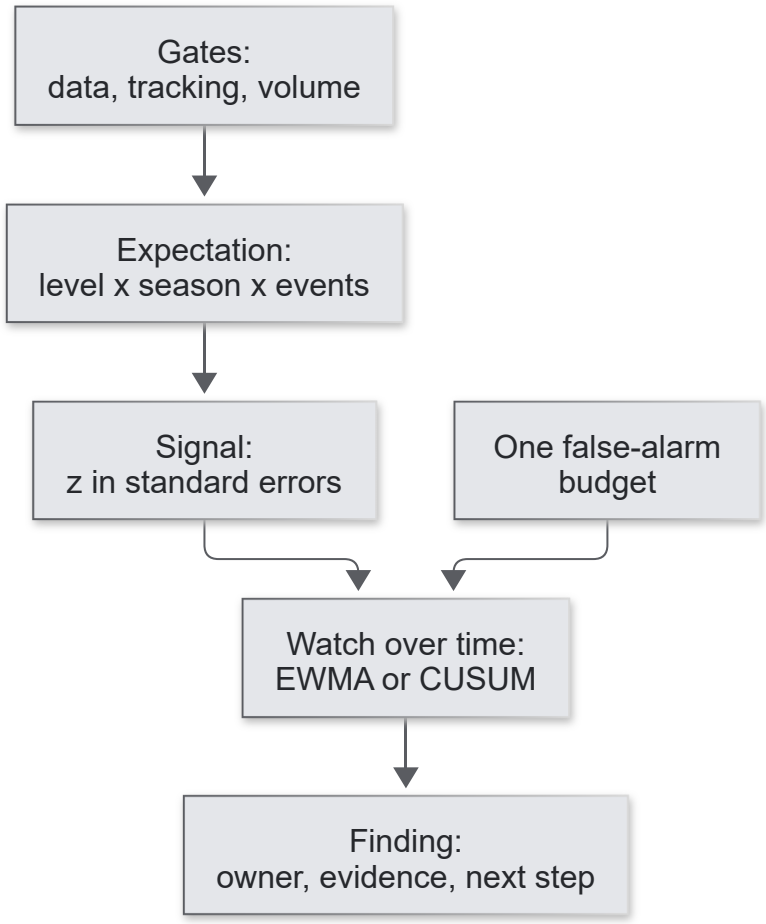
It chases luck. A window that is far above the usual tends to be followed by a lower one, by chance alone. This is regression to the mean (see *Decide what the evidence can resolve*). About 15% of windows are lucky in this sense: they sit at least one standard error above the true rate. After such a window the win-rate rule fires

in 57% of the next windows, against 19% overall. A rule that compares with the prior window punishes every good window.

It forgets. After a real fall, the new level becomes the prior window. The rule can fire only once, in the first affected window. Suppose the MQL-to-SQL rate falls 20%. The rule fires in that first window in 77% of cases and at some point in the next 13 windows in 86%. So one real fall in seven never alerts. For a 10% fall the figures are 24% and 45%. The baseline walks down with the data.

It cannot read the calendar. In December, Birch expects 257 SQLs a window against 316 in November, a fall of 18.5%. Nothing is wrong. The fixed rule fires on this fall in 68% of unchanged Decembers for SQLs, 82% for MQLs and 93% for leads. A matching rule on rises would then fire on the January recovery.

Factories met the same problem long ago. Their answer was the **control chart**: a plot of a measure over time, with limits set by the noise of the process itself. The limits come from the process and not from a tolerance that managers prefer (*Better Business Decisions from Data*, the chapter on control charts). We follow that principle.



The detection path. A failed gate stops it before any rate is read; the budget sets the limits of every watch.

Gates first

A gate is a check that must pass before you read any stage metric. Gates are fixed yes-or-no rules, not statistics. A failed gate makes every number under it meaningless. Three gates come first.

- **Data quality.** The data tests of *Decide what to measure* pass: keys are unique, the data are fresh, and the totals reconcile.
- **Tracking.** Two sources of the same event agree within their limits. Examples are sessions against ad clicks, which gives the consent capture ratio, and tracked sign-ups against accounts created in the product database.
- **Volume.** The rate is readable at this volume for the smallest fall worth acting on. This uses the monitoring mode of *Decide what the evidence can resolve*, where the baseline adds noise.

The volume gate gives different answers by company. At Birch, a win-rate monitor that reads one quarter against a baseline of four quarters sees a fall of 18% with 80% power. At Alder, the same monitor reading one year against a two-year baseline sees only 37%. In one window it sees 94%. A rate that fails the volume gate does not go on the alert list. We watch counts upstream instead, as the section on counts shows.

A tripped gate is an alert in its own right. It goes to the owner of the data, and every metric under it shows "unreadable" until the gate clears. The tracking and data gates run daily. The volume gate runs when volumes change.

The expectation model

A detector needs something to compare with. The last window is a poor choice, as the previous section showed. The **expectation model** says what the metric should be, given everything we know.

> **Calculation**
>
> Expected value = level × season × known changes.
>
> Evidence label: Arithmetic. The model is a start; keep it only if it forecasts better than the prior window in a backtest (Standard practice).

The **level** is the usual value of the metric once season and known changes are removed. We estimate it as the total of successes divided by the total of trials, where each trial is first scaled by its season and change factors.

The **season** is a regular pattern through the year. We estimate one index for each month from at least two full seasons of history. With less history, do not invent an index. Compare with a locked level and keep the band wide. Birch's calendar has a December index of 0.85 and a November index of 1.05 (Illustration only). The index alone gives a fall of 19% from November to December and a rise of 24% from December to January. Those moves are in the expectation. They are not alerts.

A **change event** is a recorded change in the business or in tracking. Examples are a launch, a price change and a consent banner release. Each event has a start, an end if it has one, the metrics it touches and the size we expect. The size multiplies the expected value while the event is active, so the event moves the expectation and raises no alert. Birch's banner redesign in Year 1 October lifted the consent capture ratio from 62% to 74%. The expected factor for measured click-to-lead is 0.74 ÷ 0.62, which is 1.19. From the release date the expectation for that metric is 19% higher. If the measured rise were larger than 19%, that surplus would be a signal. Tags, ad accounts, the CRM and the price list should write their own change events, so that nobody has to type them.

The level must not chase noise. After a lucky window, the new level should move only as far as the data are reliable. Take Birch's MQL-to-SQL rate at 25.0%. A lucky window two standard errors above it reads 27.7%. If the reliability is 0.10 (Illustration only), the expected level for the next window becomes 25.3%, not 27.7%. Reliability is the share of the spread that is real, as *Decide what the evidence can resolve* explains.

One standardised signal

The **z-score** says how far a window lies from its expectation, in standard errors.

> **Calculation**
>
> z = (observed − expected) ÷ standard error.
>
> For a rate: standard error = square root of (ϕ × expected × (1 − expected) ÷ n).
>
> For a count: standard error = square root of (ϕ × expected count).
>
> ϕ is the overdispersion factor. Evidence label: Arithmetic.

In a window, 280 of 1,203 Birch MQLs became SQLs. That is 23.3%, against an expected 25.0%. With ϕ equal to 1 the z-score is −1.38. With ϕ equal to 1.2 it is −1.26. The window is low, but chance explains it.

For detection, the overdispersion factor ϕ matters a great deal, as the stress test below shows. Estimate it from at least 13 past windows (Plausible but untested). Never use a value below 1. Until you have the history, bracket ϕ between 1 and 2, as *Set targets and staff capacity* does for forecasts. A limit needs one value, so the detector starts at 1.2, inside the bracket (Plausible but untested). Without an estimate, every false-alarm figure is too optimistic. Take a limit that is tuned for a 10% yearly budget and assumes ϕ equal to 1. If the true ϕ is 1.2, noise-only years raise an alarm in 15% of cases. At 1.5 the share is 23%, and at 2.0 it is 35%. The figures are the same for the MQL-to-SQL rate and for the win rate.

A z-score has no unit. A rate with 3,000 trials and a rate with 150 trials can then share one limit and one chart. The volume is inside the standard error.

Watching over time: EWMA and CUSUM

One window of z sees only a large jump. A small, lasting fall needs memory, which two standard tools give.

The **exponentially weighted moving average (EWMA)** is a running average of z. It gives the newest window a weight λ and the history a weight of 1 − λ. Roberts introduced it for quality control (Roberts, 1959, *Control chart tests based on geometric moving averages*, Technometrics).

> **Calculation**
>
> EWMA(t) = λ × z(t) + (1 − λ) × EWMA(t − 1), scaled to standard-normal units.
>
> A small λ gives a long memory. A large λ gives a short one.
>
> Evidence label: Standard practice. Start value λ = 0.2 (Plausible but untested).

The **cumulative sum (CUSUM)** adds up the distance below the expectation. It ignores small distances, because it takes away an allowance k each window. It never goes below zero. Page introduced it (Page, 1954, *Continuous inspection schemes*, Biometrika).

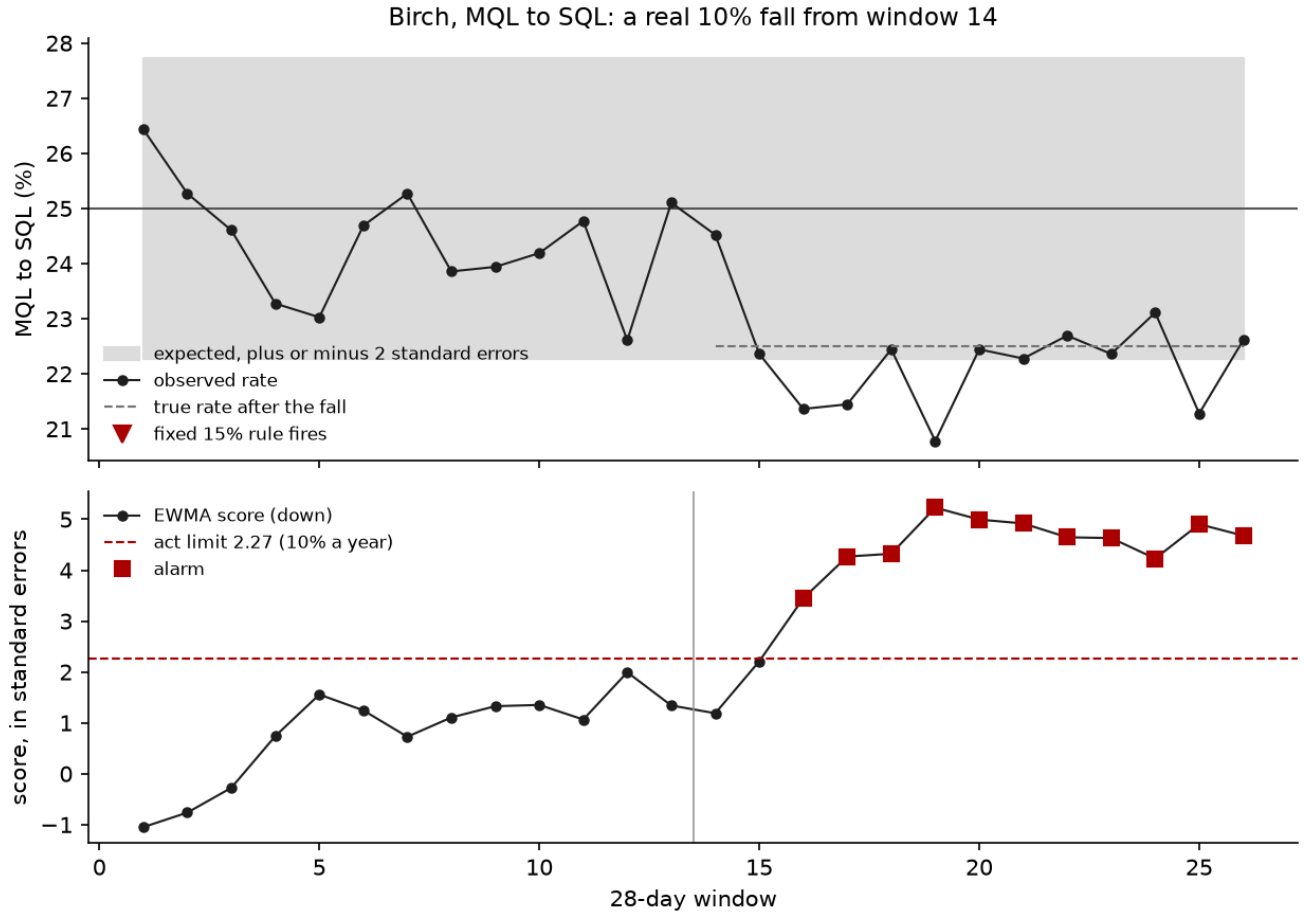
Calculation

CUSUM(t) = max(0, CUSUM(t − 1) + (−z(t)) − k), to watch for falls.

Set k to half of the smallest fall worth catching, in standard errors.

Evidence label: Standard practice. Start value k = 0.5, which suits a fall of one standard error (Plausible but untested).

Both tools turn a stream of small hints into one number that grows when a shift lasts. The chapter on control charts in *Better Business Decisions from Data* notes the same: a cumulative sum shows small changes more clearly than single readings. Figure 1 shows the EWMA on a real fall of 10% in Birch's MQL-to-SQL rate. A 10% fall is 1.8 standard errors. Many single windows after the fall still look normal. In this run the fixed rule never fires. The EWMA crosses its limit in the third window of the fall.



Birch's MQL-to-SQL rate with a real 10% fall from window 14. The upper panel shows the rate and the expectation. The lower panel shows the EWMA score and its limit. Simulated with a fixed seed (Illustration only).

Table 2 gives the median number of windows from the start of the fall to the first alarm, counting the first affected window as one. The limits are tuned to the same false-alarm budget, which the next section explains.

Table 2. Birch, median windows to the first alarm after a real fall. EWMA with λ = 0.3, CUSUM with k = 0.5, ϕ = 1.2, 10% yearly budget. Simulated (Illustration only).

Rate	Fall of 10%	Fall of 20%	Fall of 30%
MQL to SQL, EWMA	3	1	1
MQL to SQL, CUSUM	2	1	1
Win rate, EWMA	7	3	2
Win rate, CUSUM	6	3	2

The two tools perform alike. Pick one for each metric and keep it. The EWMA is easier to draw and explain. The CUSUM is natural when you want to know when a drift began. Table 2 keeps k at its start value for both rates. A 20% fall is 1.53 standard errors on the win rate and 3.66 on MQL to SQL. So the rule in the box would give k = 0.77 and 1.83. Tune k to the fall you act on once the tool is in use.

The choice of λ is a trade. On the win rate, a λ of 0.2 finds a 10% fall after 8.7 windows on average and a 20% fall after 3.5. A λ of 0.3 takes 8.9 and 3.3. So a larger λ is better for the falls that the company acts on, which at Birch are 20% and more. That is why Birch uses 0.3.

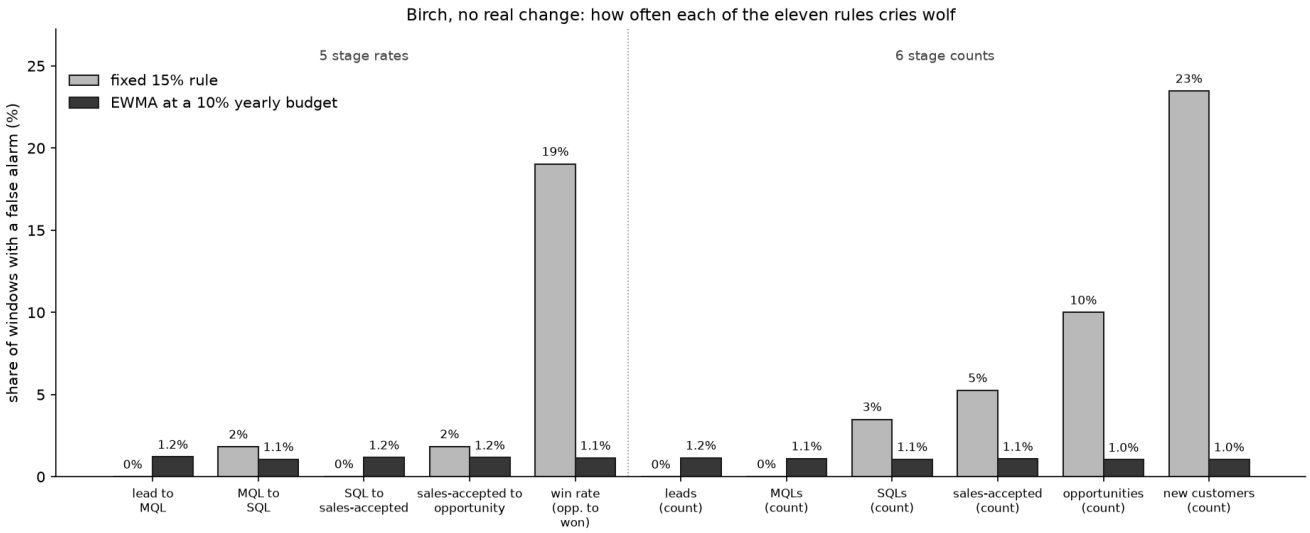
A **detection delay** is the time between a real change and its alert. It is the price of fewer false alarms. A looser limit alerts sooner but raises more false alarms. For a 20% win-rate fall, a budget of 20% a year gives a mean delay of 2.75 windows. A budget of 10% gives 3.31, and 5% gives 3.87. The medians are 3, 3 and 4 windows.

One false-alarm budget

A limit is not a number you pick by habit, such as 2 or 3 standard errors. Set it from a **false-alarm budget**: the chance that pure noise raises at least one alarm in a year. The default is 10% for each rate. It is a start value (Plausible but untested), to be tuned on the company's own data.

The method is a simulation. For one metric, take its expected rate, its volume per window and its ϕ. Simulate many years of noise only, 13 windows each. Compute the largest score in each year. Set the limit at the 90th percentile of those yearly maxima. Then, by construction, 10% of noise-only years raise an alarm. The companion code does this in one function.

Figure 2 shows the result for all eleven of Birch's metrics: the five stage rates and the six stage counts. The limits for all eleven rules lie between 2.2 and 2.3 standard errors. In noise-only windows the tuned EWMA fires in about 1% of windows. The fixed rule fires in 19% to 24% of windows on the win rate and on the new-customer count. On the opportunity count it fires in 10%.



Birch, no real change: the share of 28-day windows with a false alarm, for the fixed 15% rule and for an EWMA tuned to a 10% yearly budget. Simulated (Illustration only).
A company has many rules, and their false alarms add up. Birch states the budget for the whole list in the unit that people feel: one false alarm a quarter, so four a year (Plausible but untested). Each rule spends 10% a year, so four a year pays for 40 rules. Birch starts with eleven rules, one on each stage count and stage rate, which spend about 1.1 false alarms a year. The budget leaves room for 29 more rules, for channels and segments, before Birch must tighten anything.

Count alarms and alarm windows apart. An EWMA that crosses its limit often stays above it for a few windows. In simulated years with no real change, the eleven rules raise 1.6 alarm windows a year, more than their 1.1 alarms. One alarm makes one finding, even when it lasts several windows. Report alarms against the budget, and show the windows on the chart.

Two practices keep the budget honest.

- **Split the budget across rules.** Give each rule a share of the total. The even split is the plain start. The split may weigh a rule by the cost of a miss, so that a rate with costly delays gets a looser limit. That weighting is Plausible but untested.
- **Count the metrics, not only the noise in each.** When the list grows, false alarms grow with it. This is **multiplicity**. The even split is a little too strict when metrics move together. That is a safe error. The Cedar example below shows the scale.

When many rules watch segments and channels, the number of rules can explode. Another method then controls the **false discovery rate**: the expected share of false findings among all findings. It was introduced by Benjamini and Hochberg (Benjamini and Hochberg, 1995, *Controlling the false discovery rate*, Journal of the Royal Statistical Society, Series B). It finds more real changes when many metrics move at once, but it promises no limit on false alarms a quarter. We prefer the plain split because anyone can audit it.

Counts and rare events

A count of events in a window, such as deals won, has Poisson noise, and for a small mean the normal curve is a poor guide. Use the exact Poisson distribution for the limits.

Alder shows why. In Year 1 November it expects 3.8 wins in a window. The chance of 3 or fewer wins, when nothing has changed, is 47%. The chance of 2 or fewer is 27%. The 99% range for the count runs from 0 to 10 wins. One deal moves a window by 26%. A rate built on such counts moves even more, because decided deals vary too.

At low volume, detect on counts and on activities, not on rates. Choose counts that come earlier in the funnel and have larger means. The EWMA can then work on them. Table 3 shows Alder's choices.

Table 3. Alder, Year 1 November: counts per window and the median windows an EWMA needs to find a real fall. λ = 0.3, ϕ = 1.2, 10% yearly budget. Simulated (Illustration only).

Count	Mean per window	Relative standard error	20% fall	30% fall	40% fall
Engaged accounts	45.5	16%	4	2	2
Opportunities created	17.3	26%	6	4	3
Wins	3.8	56%	8	7	6

A 20% fall in opportunities is found in 89% of cases and a 20% fall in wins in only 50%, within 16 windows. A rate on wins stays out of the alert list. It stays on the dashboard as a trailing-year figure with its interval. The wins count stays on the list as a slow watch.

An activity count is a leading metric. *Decide what to measure* explains the grade of evidence that a leading label needs. Gates also matter more at low volume, because one wrong record is a large part of the data.

Rates that mature

A rate that ends in a sale is not final when the window ends. *Forecast conversion and timing* showed that the same-period close rate swings with volume. So never run a detector on a same-period close rate. Run it on cohorts at a fixed age.

Each age is a trade. A cohort read young has few outcomes and a noisy rate, but it arrives early. A cohort read old has many outcomes, but it arrives late. The **maturity lag** is the time from a real change to its alarm, including the time that a cohort needs to reach the reading age.

Table 4 works the trade for Birch's close rate. Each cohort is one window of contacted SQLs, about 300. The final close rate is 16.0%. The cycle has a median of 34 days. We simulate a real fall of 25% in the final rate. The days to the alarm are 28 times the median windows plus the age.

Table 4. Birch, close-rate cohorts read at different ages: the maturity lag for a real fall of 25%. EWMA with λ = 0.3, 10% yearly budget. Simulated (Illustration only).

Age read (days)	Share of final wins already in	Wins per cohort	Median windows to alarm	Days from the change to the alarm
14	7%	3.3	12	350
28	37%	17.9	5	168
56	80%	38.3	3	140
73	90%	43.1	3	157

Age read (days)	Share of final wins already in	Wins per cohort	Median windows to alarm	Days from the change to the alarm
90	95%	45.5	3	174

The best age is not the oldest. Day 56 beats the maturity age of 73 days: the reading is nearly full and it arrives 17 days sooner. Even so, a real fall of 25% in the close rate reaches an alarm after about 140 days. That is nearly five months.

So watch pairs. Watch the early steps, such as SQL to sales-accepted and sales-accepted to opportunity, which finish long before the close. Watch the mature cohort rate too. If a cause acts through an early step, the early alarm comes months before the close rate can speak.

Tracking break or real change

A tracking break changes what you measure, not what happens. It looks like a sudden fall, so it is often the first explanation to rule out. Three checks separate it from a real change.

1. **The change log.** Did a release, a tag change or a banner change happen at that time? A step that starts at one timestamp points to a release.
2. **Two sources of one event.** A real fall shows in every source. A break shows in one. If tracked sign-ups fall but the product database still creates accounts at the same rate, the break is in the tag.
3. **The shape.** A break hits every channel, country and browser at once, or one whole flow. A real change spreads unevenly.

Treat a surprising rise with the same suspicion. A figure that looks unusually good is often wrong (Thomke, 2020, *Experimentation Works*, on Twyman's law). The same book recommends checking a pipeline on runs where nothing changed, as the last mechanics section does.

Several metrics that move together are a common-cause sign. Look at the change log and tracking before any lever. Do not rely on a sign-only rule such as "three stages move in the same direction". If the stages move independently, three named stages all fall by chance alone 12.5% of the time. Either direction gives 25%, one week in four. With four stages, three or more in the same direction is 62.5%. Use a joint check on the z-scores instead, and state its false-alarm rate.

When a break is confirmed, mark the affected metrics unreadable. Fix the tag. Record a change event with its start and its end. Leave the data of those days out of tests and baselines.

Leading and lagging in alerts

Each rule has two limits from the same statistic. The **act** limit is the one tuned to the budget. The **watch** limit is looser, tuned to a budget of 50% a year (Plausible but untested). For Birch's MQL-to-SQL rate, the act limit is 2.27 and the watch limit is 1.32 standard errors. Watch asks for a look. Act asks for a decision.

A lagging metric alone, such as the win rate, churn or net revenue retention, reaches watch only. It reaches act only when a leading partner also signals, or when the owner writes down the reason. A **leading pair** is a leading metric and the lagging metric that it predicts, read together. Examples are engaged accounts and opportunities, MQL-to-SQL rate and new customers, and activation and renewal. Act on the leading side of the pair.

A lagging metric shows the effect of causes that are already old. A leading metric shows the cause while there is still time to change the result.

What happens after an alert

An alert becomes a **finding**: the record of one alert after review. It states what moved, with the metric, the direction, the size, the z-score and the score. It lists the evidence: the gates, the change events nearby and the other metrics that moved. It names the likely cause, who was told, and what should move next if the cause is right. Then *Diagnose what changed* takes over.

The mockup below shows the alert list, one row for each metric, in order of severity.

Alert list, Birch. The layout is the point. Values are simulated from Birch's rates with a fixed seed (Illustration only).

Metric	Last window	Expected	z	EWMA score	Level	Owner	Change events nearby
MQL to SQL	20.0%	25.0%	−3.6	6.4	Act	Marketing stage owner	none
Win rate (opportunity to won)	24.7%	32.0%	−1.8	1.1	none	Sales stage owner	none
Sales-accepted to opportunity	62.5%	62.5%	0.0	−0.3	none	Sales stage owner	none
Data gates	all pass				none	Analytics lead	none

Each finding goes to one owner, in writing, with a deadline. One shared cause makes one finding, not one for each affected metric. Nobody calls a meeting for an alert. The exception lane (a route that shows each owner only the items that need them) of *Run the growth system* carries the findings and gate failures. All other items run without their attention. Code computes every number in the finding. A language model may draft the text of the finding from the evidence. A person decides what to do, using the decision card of *Commit with uncertainty*.

Testing the detector with noise

A detector that was never tested may page nobody or everybody. Test it before anyone relies on it, and after every change to a rule, a limit or a parameter.

1. **Noise-only runs.** A **noise-only run** is a run of the detector on data in which nothing changes, so every alarm in it is false. Simulate 20,000 years of noise from the metric's rate, volume and ϕ. Count the years with an alarm. For Birch's win rate the share is 9.9%, with a standard error of 0.2 percentage points, against a budget of 10%. The test passes.
2. **A severity sweep.** Inject real falls of 10%, 20% and 30%. Record the delay and the share found, as in Table 2.
3. **Stress.** Make the world differ from the model. Use at least the two ends of the bracket for ϕ. If the true ϕ is 1 while the monitor assumes 1.2, 6.0% of years raise an alarm. At 1.5 the share is 16%, and at 2 it is 27%.

A change of 20% in volume, either way, moves the share only to 9.8% and 10.4%. The detector is more sensitive to ϕ than to volume, so re-estimate ϕ each month.

4. **Seeded faults.** Replay a month of history with planted data faults: a tag that fires twice, a missing campaign tag, a banner release, a late CRM load and a dropped data partition. Record whether the gates catch each fault and how late.

Learn from outcomes explains the severity sweep and the seeded faults in full and turns these tests into a standing test suite for the whole system.

Run the same detector on real history with the change log, and count the alarms. On a past year with no known change, the alarms are the false alarms. Never state a false-alarm figure without saying which ϕ it assumes.

What changes with volume

Regime	What to watch	What the detector can see	Main risk
Low volume (Alder)	Counts and activities: engaged accounts, opportunities. Wins as a slow watch. Rates only as trailing-year figures with intervals.	A 30% fall in engaged accounts in 2 windows. A 30% fall in wins in 7.	One deal looks like a trend.
Medium volume (Birch)	Stage rates with an EWMA, season and change events. The close rate by cohort at day 56.	A 20% fall in MQL to SQL in one window and in the win rate in 3. The close rate in about 140 days.	Fixed percentage rules; the calendar.
High volume (Cedar)	Every rate has power. Gates daily. A budget for the whole list.	A 10% fall in daily sign-ups on 73% of days, even at a strict limit.	The number of metrics, not the noise in each.

A rule from the growth literature points the same way. When a metric needs a month to show a 20% change in a test, consider replacing it (*Growth Data Analytics Playbook*, on measurability). That rule is about tests, but it fits detection too. At medium volume, Birch's win rate and close rate fail it, at 159 and 196 days. Every step up to the opportunity passes, in 28 days or less.

Alder, Year 1 November: one lost deal

Alder, Year 1 November: about 3.8 new customers a window, 0.8 times the derived typical; low volume: no rate below the lead reads a 20% change within a year.

Stage history went live in October. The head of sales adds a weekly sales report that shows win rates by window. This week it shows a win rate of 21.4%, down 30% from 30.8%. Last window, 4 of 13 decided deals were won. This window, 3 of 14 were won. The chief executive asks what happened.

The analyst starts by asking what nothing happening looks like. Alder expects 3.8 wins in a November window. That is 3.5 at the trend level, lifted by the November index of 1.10. The chance of 3 or fewer wins is 47%. This window is almost the median. The exact 99% range runs from 0 to 10 wins. One deal is 7.7 points of win rate when 13 deals are decided. The analyst also simulates pairs of windows with no change, with about 17 decided deals in each. In 30% of such pairs the win rate falls by 30% or more.

The win rate cannot settle the question. It needs about 3.1 years to read even a 30% change. So the analyst moves the detector to the counts of Table 3: engaged accounts (45.5 a window), opportunities created (17.3) and wins as a slow check. A 30% fall in wins takes a median of 7 windows and is found in 73% of cases

within 16. Alder has too little history to estimate ϕ, so it uses 1.2 until it has 13 windows. Each of its three rules has a 10% yearly budget.

The gates matter more here. The three tests that went live in October become the data gate. They check for a unique account key, one CRM account for each billed customer, and a first invoice for each won deal. A stage-history completeness check joins them, since the win rates of this report come from that history.

The head of sales decides. The win rate leaves the weekly report. Where someone needs it, it appears as a trailing-year figure with its interval, never by window. The three new rules start. The head of sales writes the first entry in the decision log, which starts this month: "Selling has not changed. Wins in the current market from December to May will total 16 to 29, with 80% probability." The expectation model gives 22.2 wins for those six months, and the range allows for overdispersion. The log grades the entry in June, and the chief executive reads the grade, not the weekly swing.

Birch, Year 1 November to December: the fixed rules retire

Birch, Year 1 November: about 48 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 71 days, and the close rate reads a 20% change in about 6 months.

The band of plus or minus 15% on the board paper rings in nearly every quarter. In July the analytics lead proposed one band per metric around the seasonal expectation. The growth owner asked for a replay on a full year first (see *Set targets and staff capacity*). The lead now runs that replay with a falls-only form of the old rule on each of the 11 counts and rates, in simulated years at Birch's own volumes with no real change. The eleven rules fire 8.4 times a year on nothing; Table 1 and Figure 2 show the main ones. That is more than one a quarter.

The new design has four parts. The expectation is level times season times change events. Birch has the lead counts of earlier years, so it can estimate month indices. December is 0.85 and November is 1.05. The October banner redesign is a change event, with an expected factor of 1.19 on measured click-to-lead. Every rule uses an overdispersion factor: 1.2 at the start, inside the bracket of 1 to 2, and re-estimated each month from 13 windows. Every rule uses an EWMA with a weight of 0.3, tuned to a 10% yearly budget. The budget for the whole list is one false alarm a quarter, which pays for 40 rules. The same 11 metrics stay on the list, and their rules spend about 1.1 a year. New customers keep their quarterly reading on the board paper as well.

The same simulation, with real falls injected, gives the delays of Tables 2 and 4: one window for a 20% fall in MQL to SQL, 3 for the win rate, and about 140 days for a 25% fall in the close rate read at day 56. So the close rate is watched by cohort at day 56, and the early steps carry the quick warning.

Then December arrives. The expected SQL count falls from 316 to 257 a window, which is a fall of 18.5%. The band for December is 222 to 292 SQLs, two standard errors on each side. The old rule would have fired on most unchanged Decembers. The calibrated rule expects the dip and stays quiet. In January the rise of 24% is expected too.

The growth owner decides. The fixed 15% rules retire this month, and the new list goes live. The owner writes a log entry before any outcome exists. If the funnel does not change, the first quarter will raise no false alarm with a probability of 73%, and at most one with 96%. A simulation of the eleven rules from their first window gives these figures. The log grades the entry at the end of the quarter. If the list raises more, the budget or ϕ is wrong and the owner fixes it.

Cedar, Year 1 November: a tag release breaks tracking

Cedar, Year 1 November: about 1,500 new paying accounts a window, 1.4 times the derived typical; high volume: every self-serve rate reads a 20% change within a window.

A tag release breaks the sign-up event. Tracked sign-ups fall by 38%. Cedar sees about 1,605 sign-ups a day, so the measurement loses about 610 a day. Behind its three dashboards, Cedar watches 120 monitors every day. A fall of 610 is 13.9 standard errors of a day's count, or 7.6 even with a ϕ of 4. Any limit would catch it.

The first defence is not a statistical limit. It is the tracking gate. Every day, the gate compares tracked sign-ups with the accounts created in the product database. The product database still creates accounts at the usual rate. The ratio falls. The gate fails at its first daily run, one day after the release. The gate raises one finding and sends it to the owner of the tag, with the release as the nearest change event. Every monitor that uses tracked sign-ups, as numerator or denominator, shows "unreadable". None of them raises an alert of its own. Without that step, one tag fault would produce a long list of alerts, and the growth team would look for a cause in the market.

The multiplicity arithmetic shows why the grouping matters. Treat the 120 daily checks as independent. A flat limit of two standard errors on falls then raises 2.7 false alerts a day. That is 19 a week and 248 a quarter. A flat limit of three raises 0.16 a day, or 1.1 a week. Cedar's budget is one false alert a week (Plausible but untested). Divided across 120 monitors and seven days, the chance for each monitor and day is 1 in 840. That sets the limit at 3.04 standard errors. The price is some sensitivity. A real 10% fall in sign-ups is 3.7 standard errors a day. It is seen on a given day with a chance of 73%, against 95% at the flat limit of two. A real fall still shows within days.

The analytics lead decides. The daily gate and the 3.04 limit stay. Owners see findings and failed gates only. The lead records the release as a change event with a start and an end. Experiments that ran on those days drop them from their sign-up readouts, as *Forecast conversion and timing* did for the outage week. At high volume, false alarms come from the number of metrics, not from the noise in each.

The decision to carry forward

Investigate a signal when its expected consequence warrants the attention. Test the measurement first, then compare with a stated expectation. An alert identifies a question to diagnose, not a treatment to deploy.

22 · Diagnose what changed

What changed, and which response is justified? A blended rate can fall because the customer mix changed, because behaviour within a segment changed, or because the observation process failed. Those mechanisms imply different actions. Start by separating them before selecting a treatment.

Diagnosis narrows the set of plausible explanations. It does not turn an observed association into an effect caused by an action. Write the evidence that would distinguish explanations, favour checks that could change your response and keep contradictory cases visible.

The methods below combine arithmetic decomposition, process evidence and causal checks. Customer interviews can reveal a missing mechanism while remaining an imperfect estimate of its prevalence. Quantitative and qualitative evidence contribute to the same decision, with their sampling and measurement limits stated.

Mix before rate

A blended rate is an average over groups. The close rate of all SQLs is an average over routes. It moves when the rate inside a route moves. It also moves when the share of each route moves. These are two causes with two remedies: the first points at selling, the second at where the volume comes from.

Here is a small case. Two routes close 27.9% and 12.1% of their SQLs. Both rates stay the same. The mix moves from 60% and 40% to 40% and 60%. The blended close rate falls from 21.6% to 18.4%. That is a fall of 14.6% with no change in any rate (Illustration only). A fixed rule at 15% almost fires. A team that retrains its sellers fixes nothing.

The **mix effect** is the part of a change in a blended rate that comes from changes in the shares of the groups. The **rate effect** is the part that comes from changes in the rates inside the groups. A **decomposition** splits a change in a total into parts that add up to it. This one is exact.

> **Calculation**
>
> mix effect = sum over groups of (new share − old share) × (old rate + new rate) ÷ 2.
>
> rate effect = sum over groups of (old share + new share) ÷ 2 × (new rate − old rate).
>
> mix effect + rate effect = new blended rate − old blended rate, with nothing left over.
>
> Evidence label: Arithmetic (Kitagawa, 1955, *Components of a difference between two rates*, Journal of the American Statistical Association).

Why average the two periods? Each order of valuing the two changes gives a different split. The midpoint form is the mean of the two orders. It leaves no remainder, and it does not depend on which period you call first.

Use the split as a gate on the story. While the mix effect is more than half of the total movement, tell no story about the blended rate in terms of rates, skill or the proposition. Half is a start value (Plausible but untested). The total movement is the mix effect and the rate effect, each taken without its sign.

A large mix effect and a small rate effect mean that the volume moved and the selling did not. A small mix effect and a large rate effect mean the reverse. When both are large, treat them as two findings with two owners. The rate effect then becomes its own finding inside its route, read with its own noise.

A mix effect is not an error. Sometimes it is the plan: a new route was added on purpose. Then the blended rate is the wrong place to judge it. A route with a low close rate may still pay. Judge each route on its own value, on the one value basis of *Value customers and cash*.

- **Fix the dimensions in advance.** Use a short list: route, plan, channel, size band and country. If you try every cut until one splits nicely, you hunt for a story and multiply false findings (see *Decide when a change needs action* on multiplicity).
- **Split again when needed.** A mix effect on one dimension can hide a rate effect on another. A trend that holds in every group can reverse in the total (Simpson, 1951, *The interpretation of interaction in contingency tables*, Journal of the Royal Statistical Society, Series B). Split again until the groups are alike or too small to read.
- **Print the noise.** The shares are counted, so they are exact. The rates are noisy. Give the rate effect a standard error from the counts, as *Decide what the evidence can resolve* teaches.
- **Give a new group its own rate.** A group that did not exist before has no earlier rate. Give it its later rate, so that it earns a mix effect and no rate effect. Do not give it the blended rate, which invents a rate effect.

Implementation note: the companion code computes the split from counts and returns the standard error of the rate effect.

Which kind of change is it?

The split says what moved. The next question is when and how. The kinds of change of *Forecast conversion and timing* become a diagnostic question here: where the change sits narrows the likely causes.

What you see	Kind of change	Typical causes	First check
A dip along a diagonal: one calendar date, many cohorts	Period effect	An outage, a tracking break, a price change, a holiday	The change log and the tracking gate
One low row, at every age	Cohort effect	A campaign or channel that brought different buyers; a fault at entry	The source mix of that cohort
A step down a column that stays	A real change for new cohorts	A change in process, product, price or qualification	Change events at that date
Young cohorts low at young ages, catching up later	Age effect, not a fall	Outcomes that have not arrived yet	Compare at the same age
Every group's rate holds, the blend falls	Mix	A new route, plan or channel	The split above

Two earlier cases show the point. Birch's campaign of Year 1 June made the same-period close rate fall because the extra SQLs were still young. That was an age effect. Cedar's set-up email outage in one sign-up week gave one low row. That was a cohort effect. Neither case needed a story about selling.

A calendar series alone cannot tell a period effect from a cohort effect, because both lower the blended rate within a few weeks. The cohort table can, so build it before you name the kind of change.

A step that starts on one date points to a release or a decision. A drift points to a slow cause, such as a market, a competitor or wear-out. A one-window dent points to a fault. Read the shape before you fit any model.

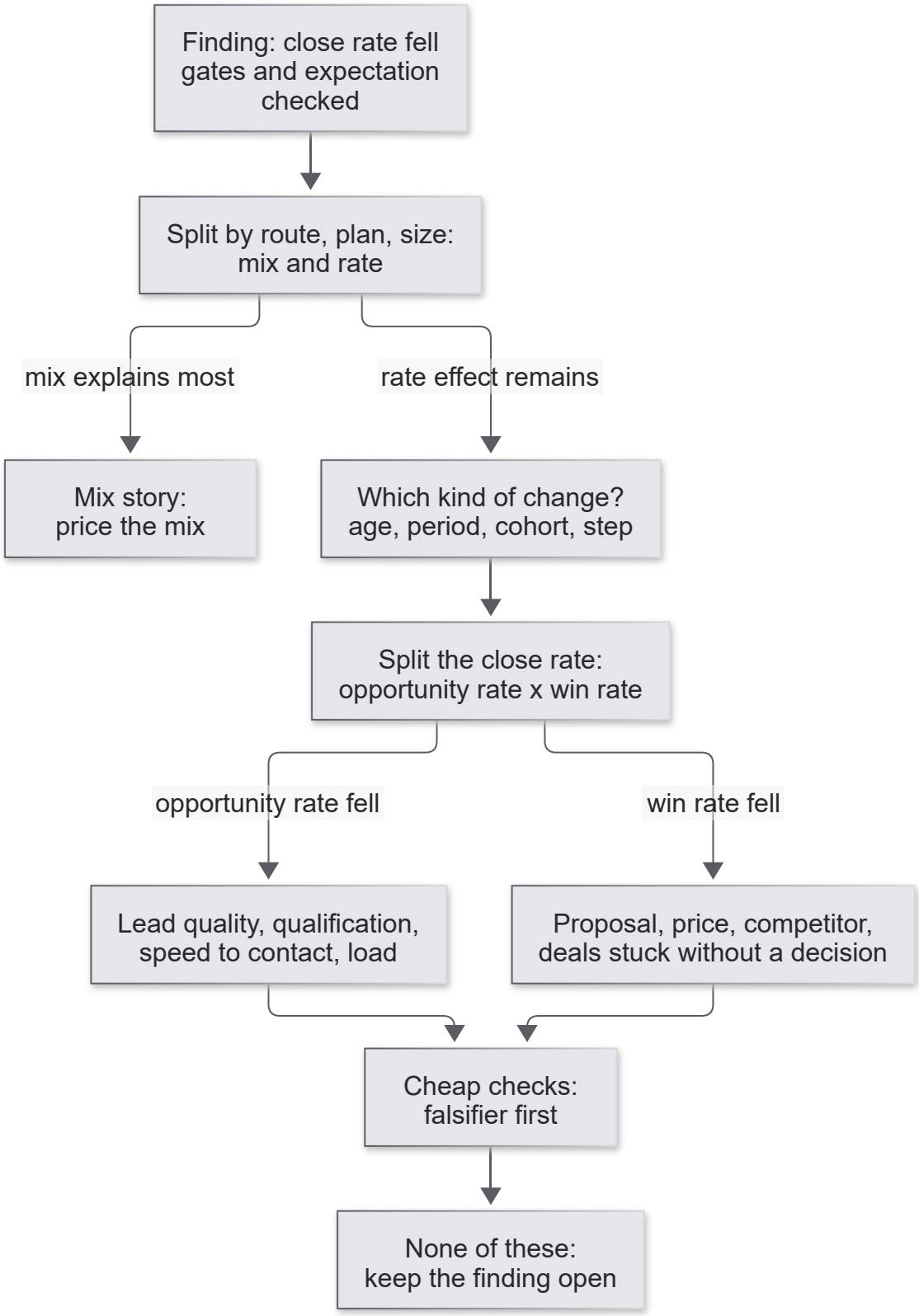
Differential diagnosis

A **differential diagnosis** is the list of causes that can produce one symptom, together with the order in which to check them. The idea is from medicine: a doctor lists the candidates and tests the cheap and likely ones first.

For a funnel, the list comes from its mechanics. A close rate is an opportunity rate times a win rate (see *Choose markets and accounts*). So a fall in the close rate has many candidates: data, chance, mix, timing, lead quality, the speed and load of sellers, the proposal or price, and events outside the company. Add every change event near the date.

Each candidate must predict something. A **separating metric** is a metric on which two candidates predict different values. Before you check a candidate, write down what it predicts on its separating metric.

For example, Birch's inbound close rate of 16.0% is an opportunity rate of 50.0% times a win rate of 32.0%. A fall of 10% in either factor lowers the product by 10%. But the two factors sit in different hands. The opportunity rate depends on the leads, on qualification and on how fast sellers reach a lead. The win rate depends on the proposal, the price and the competitor. If the opportunity rate fell and the win rate held, nobody should touch the proposal. The split costs one query on data that Birch already holds.



The differential for a fall in the close rate. The cheap split comes first. The finding stays open if no candidate survives.

The order of the checks matters, because checks cost different amounts. Order them by chance divided by cost. Take two checks. Check A finds the cause with chance 0.5 and costs 4 hours. Check B finds it with chance 0.3 and costs 1 hour. Doing A first costs 4.5 hours on average. Doing B first costs 3.8 hours

(Arithmetic; the inputs are Illustration only). The rule assumes that at most one of the causes is true, that each check settles its candidate, and that costs add.

The chances come from judgement. They are expert priors, labelled Assertion, and they serve only to order the checks. The book prints no probability for a cause until graded diagnoses exist in the decision log (see *Learn from outcomes*). Until then, the finding ranks the candidates as "likely" or "unlikely". A number would claim a precision that nobody has.

Keep an outcome called "none of these". Sometimes every check fails. Then write "unexplained", keep the finding open and widen the search to causes outside the data: a competitor, a rule change, the market. Do not pick the least bad story. When several causes act at once, the diagnosis lists each with its own check and falsifier.

Cheap causal checks

A candidate is a story until a check tests it. Before any experiment, four checks use data we already hold. Each tests a prediction that the cause must meet. None proves a cause, but each can disprove one.

First, write the falsifier. A **falsifier** is the result that would prove the diagnosis wrong. Write it before you open the data. A good falsifier has three parts: a number, a date and a reading that the volume allows. "If fewer than 25 wins appear by day 56, the diagnosis is wrong" is a falsifier. "If the effect fades" is not. Match the threshold to the size of the claim, and write down what the falsifier cannot see. Without a falsifier, every result fits the story afterwards.

A return needs the same care. A metric that falls by chance tends to return by chance (see *Decide what the evidence can resolve*). So "back to baseline" does not show that a fix worked.

The four checks follow the conditions that Hill gave for judging a cause (Hill, 1965, *The environment and disease: association or causation?*, Proceedings of the Royal Society of Medicine). A book on product analytics applies the same checks to product data (Rodrigues, *Product Analytics*, the chapter on developing heuristics in practice).

- The **dose-response check** asks whether more exposure to the cause gives a larger effect. Measure exposure as the share of units that met the cause, or compare units that met it with units that did not.
- The **placebo check** looks at units that the cause could not have touched. If they show the same change, the cause is not the explanation. The placebo group needs enough volume to read, so pool small units.
- The **timing check** asks whether the effect starts after the cause, and after the lag that the metric needs. Line the change up against the change log. A rate that matures needs cohorts at one age. A fall that starts before the cause rules the cause out.
- The **specificity check** asks whether the cause moves the metrics it can reach and leaves the others alone. If everything moved, suspect a common cause, such as tracking.

Table 1. The four cheap checks, with one example each from the three companies.

Check	Question	Example	The cause passes when
Dose-response	Does more exposure give more effect?	Cedar: Polish trial users who saw the pricing page against those who upgraded inside the product	The first group falls more

Check	Question	Example	The cause passes when
Placebo	Did units that were not exposed move too?	Cedar: the 23 other countries, pooled	They show no change that they could read
Timing	Did the effect start after the cause?	Birch: SQL cohorts before and after the outbound pilot started	Only cohorts after the start fall
Specificity	Did only the reachable metrics move?	Birch: the inbound opportunity rate and win rate beside the blended close rate	The inbound rates hold

To combine the results, do not count votes. One failed check outweighs several passed ones. A passed check only fails to disprove the cause, so it raises belief a little. A failed check removes the cause, or sends you to a changed version of it. When two checks disagree, look for a third that separates them. Stop when the cause is clear enough for the action at hand.

The checks have limits. A hidden cause that moves with the visible one can pass all four. Only an experiment breaks that link. Each check also needs its own volume: a placebo group that is too small passes everything. So ask what smallest change each check can show. That is the readability of *Decide what the evidence can resolve*.

Evidence tiers for a finding

An **evidence tier** says how strong the support for a finding is. There are three tiers. Print the tier beside the finding. A check that was not run is "not run". It is never "passed". A tier is a text label and not a probability.

Tier	What it means	What it allows
Observation	The change is real (gates and expectation pass) and fits a cause, but no check has run	More checks, and steps that cost nothing and can be undone
Checks passed	At least two different cheap checks ran, each with its falsifier written first, and none failed	Reversible actions, and the design of a test
Experiment	Randomisation, or a quasi-experiment (a comparison built around a change that nobody randomised), read as *Choose whether to test or act* teaches	Expensive or irreversible actions

The tier sets what you may propose. The decision card still has to clear its own threshold. The number of checks that make the checks-passed tier is a start value (Plausible but untested). *Choose whether to test or act* shows when a step to the experiment tier is worth its cost and its delay.

Qualitative evidence

Numbers say where a change sits. Notes often say why. Deal notes, interview notes, support tickets and coded loss reasons carry causes that no stage metric holds. Use them to build candidates and to test them.

A **loss reason** is the reason that a lost deal is coded with, taken from the notes or from an interview. Three faults are common. Buyers give polite reasons, so the stated reason is not always the real one. "No decision" is an outcome and not a reason: the buyer did not buy from anyone (see *Forecast conversion and timing* on competing outcomes). And coders make errors.

Compare two sources when you can. Code the reason from the seller's note and from the buyer's own words, in an interview or a survey. Where the two codes differ, the buyer's words usually carry more weight. The share of deals on which they differ shows how far the notes can be trusted.

Text labelling assigns codes to text, for example a loss reason to a deal note. An AI model may draft the label. A person samples the labels and measures how often they are wrong. Two rates describe a labeller. The true-positive rate is the share of true cases that it finds. The false-positive rate is the share of non-cases that it flags. Measure both on a hand-checked sample. Then correct the coded share.

> **Calculation**
>
> true share = (coded share − false-positive rate) ÷ (true-positive rate − false-positive rate).
>
> It needs a true-positive rate above the false-positive rate: the labeller must beat a coin.
>
> Evidence label: Standard practice (Rogan and Gladen, 1978, *Estimating prevalence from the results of a screening test*, American Journal of Epidemiology).

Here is one worked number. A labeller codes 30% of lost deals as "price". Its true-positive rate is 85% and its false-positive rate is 5%. The true share is (30% − 5%) ÷ (85% − 5%), which is 31.25%.

A corrected share still has sampling noise, so print its n and its interval. The two rates of the labeller carry noise too, because they come from a hand-checked sample. Ten "price" deals in 20 coded cases give a share of 50%, with an exact 80% interval of 34% to 66%. One hundred in 200 give 45% to 55%.

How many items must a person check? To show that the error rate is below 5% with 95% confidence, check 59 items and find no error (Standard practice, discovery sampling). If one error appears, the proof fails and a larger sample is needed.

Notes about workers and recordings of calls need counsel review before anyone builds a labeller on them. Works-council steps apply where they exist. Report results by role and route, never about one person. This is not legal advice. Treat coded reasons as hypotheses that can start a candidate list or support a check. A coded share alone never lifts a finding above the observation tier.

Acting on a diagnosis

A diagnosis ends in an action. Rank actions by the effect they can win back, not by the share of the change that their cause explains. The **recoverable effect** of an action is the part of the change that the action wins back.

> **Calculation**
>
> recoverable effect = size of the change × share explained by the cause × share of that part the action recovers.
>
> Then divide by cost. Count value on the one value basis of *Value customers and cash*.
>
> Evidence label: Arithmetic.

Take a fall of 100 units. Cause A explains 60% of it, but the fix recovers only half of that part. The recoverable effect is 30. Cause B explains 40%, and its fix recovers 90% of that part. The recoverable effect is 36 (Illustration only). Ranking by explained share picks A. Ranking by recoverable effect picks B.

Should you act now or check more? Compare the expected cost of acting on the wrong cause with the cost of one more check plus the value lost while you wait.

> **Calculation**
>
> Act now if (1 – chance that the cause is right) × cost of the wrong action is less than the cost of the check plus the value lost while waiting.
>
> It assumes that the check settles the question.
>
> Evidence label: Arithmetic. This is the simple form of the value of information. *Choose whether to test or act* gives the full form.

The rule needs no printed probability for the cause. The check pays only if the chance of the wrong cause is above (cost of the check + value lost while waiting) ÷ cost of the wrong action. Code computes that line. The owner judges on which side of it the chance of being wrong sits.

A reversible action is itself a check. If you can undo it cheaply and the metric answers quickly, acting may be the cheapest test. If you cannot undo it, check first. The owner of the metric decides, in writing, by a deadline. Code computes the numbers. A language model may draft the note. A person decides.

Writing it down

Every diagnosis goes into the decision log before any check is read (see *Commit with uncertainty*). The entry holds the finding, the diagnosis, the tier, the falsifier with its number, the confirms-by date, and what should move next if the cause is right. The confirms-by date is the date by which the data must show whether the diagnosis was right. The next chapter sets one for every decision.

The log records belief before the outcome, which protects against hindsight. The owner who acts writes the action and its date beside the entry. Nobody calls a meeting for this. Table 2 shows the layout, with Birch's finding of Year 2 January from the example below.

Table 2. A diagnosis entry in the decision log, written before the check is read.

Field	Entry
Finding	Read at the same age as earlier cohorts, the December and January cohorts are on course for a close rate of 14.7%, against 16.0% before
Diagnosis	Mix: the outbound pilot's SQLs close at about 5.95% and are 13.2% of contacts. The inbound rate holds at 16.0%.
Evidence tier	Observation; the split is exact, the inbound claim is open
Falsifier	Fewer than 25 inbound wins by day 56 in the January cohort
Confirms by	The end of March, when the last January SQL reaches day 56
Next, if the cause is right	The close rate by route holds at 16.0% and 5.95%; call coaching stays on hold
Owner and action	The growth owner; options go on a decision card

Learn from outcomes grades the entries later and shows which kinds of diagnosis were right. A falsifier that was written and then survived is worth more than one that was never written. A diagnosis with no falsifier cannot be graded at all.

What changes with volume

The mix split is arithmetic on counted shares, so it works at any volume. The rates inside it and the checks need volume, each in its own way.

Regime	What the split and the checks can read	How the diagnosis runs	Main risk
Low volume (Alder)	Counts only. Two groups of 5.5 expected wins each have an 80% range of 3 to 9.	Cases, deal notes and mechanisms. Falsifiers on counts and dates.	A story built on one deal
Medium volume (Birch)	The mix split is exact. A rate effect inside a route reads only large changes: one cohort at day 56 catches a fall of about a quarter, and a 20% change takes about 7 months.	Split first. A falsifier on a cohort count at a fixed age.	Retraining for a mix effect
High volume (Cedar)	A pooled placebo group reads 9.4% in a window. One small unit, such as one country, needs two windows on each side to read a fall of a quarter.	Cheap checks within days. An experiment only for costly or irreversible steps.	Many segments: choose dimensions in advance

Alder, Year 2 January: a quarter with fewer wins

Alder, Year 2 January: about 3.3 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 3.1 years to read even a 30% change, and the lead count reads a 20% change in 26 days.

The fourth quarter of Year 1 closed with 7 wins. The expectation model said 10.9. The chief executive reads a fall of about a third and asks for a cause. Alder cannot read a funnel rate at this volume. So the part-time analyst starts with the count.

Seven wins sits at the lower edge of the 80% range, which runs from 7 to 15. The 95% range runs from 5 to 18. If selling had not changed, a quarter would give 7 wins or fewer with a chance of 14.9%, about one quarter in seven. One deal is 9% of the expected wins. The count alone does not prove a fall, and it does not prove that nothing happened.

The analyst lists the candidates and orders them by chance and cost.

1. **Data.** The fourth quarter is the first with stage history, which went live in October. A look at the gates costs an afternoon.
2. **Chance.** The count above prices it: 14.9%.
3. **A thin pipeline.** Alder's median cycle is 120 days, so wins come from deals that entered months earlier. Cohorts that entered from March to September feed 82% of the expected fourth-quarter wins, which is 9.0 wins. Those months should hold 118 opportunities, with an 80% range of about 103 to 133 when we

allow for overdispersion, as *Decide when a change needs action* does. The CRM holds the creation date of each opportunity, even from before the stage history. A count costs an hour.

4. **Slow deals.** One deal in five takes more than 191 days. A deal that is open past that age is late, not lost. A list of open deals by age costs an hour.
5. **Deals lost to no decision.** Four deals ended with no decision. Even if each had kept the usual win rate of 22%, the four would have brought only 0.9 wins. That closes less than a quarter of the gap of 3.9.
6. **Mix by deal size.** A deal with one site lists at 45,000 euro and a deal with four sites at 108,000 euro. If large deals slow down or stall, the quarter's wins fall. This is a candidate, not a finding. Two groups of about 5.5 expected wins each have an 80% range of 3 to 9, so a win rate by size band cannot be read. The analyst lists the deals one by one.
7. **None of these.**

She also reads the notes of the four no-decision deals and of the open deals older than 191 days. She codes them by route and by deal size, never by seller. Four account executives cannot be compared at this volume (see *Decide what the evidence can resolve*).

The falsifier comes first, and it must match the size of the gap. Wins from these cohorts scale with their opportunities. So even 103 opportunities, the lower edge of the range, would explain only about 1 of the 3.9 missing wins. A thin pipeline explains half the gap only at about 92 opportunities, and all of it only at about 66. The analyst writes: "If more than 92 opportunities were created from March to September, a thin pipeline is not the main cause." The date is the end of January.

The head of sales owns the finding. She records the tier as observation and writes the log entry before the count is made, with the ranked candidates, the falsifier and the date. No hire and no new campaign follow from a count that chance produces 14.9% of the time. The entry of Year 1 November stays open: it expects 22.2 wins from December to May, with an 80% range of 16 to 29, and it is graded in June. At low volume, diagnosis rests on cases and mechanisms, and a written falsifier keeps the story honest.

Birch, Year 2 January: the close rate falls after the outbound pilot

Birch, Year 2 January: about 44 new customers a window, 0.8 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 81 days, and the close rate reads a 20% change in about 7 months.

Since December, five agency sales development staff have run an outbound pilot, planned to end with January. In the January window, sellers contacted 261 inbound SQLs and 39 outbound SQLs. Read at the same age as earlier cohorts, the December and January cohorts are on course for a blended close rate of 14.7%, against 16.0% before. That is a fall of 8.3%. The head of sales proposes call coaching for the account executives. The analytics lead asks for the split first.

A close rate needs about 7 months to read a 20% change, so the detector cannot confirm a fall of 8.3% as a rate. But the split is arithmetic on counted shares, so it needs no wait (Table 3). It uses close rates by cohort, so age does not distort them (see *Forecast conversion and timing*). The inbound rate of 16.0% is the rate of the 3,544 SQLs of Year 1 before the pilot. Whether the January cohort keeps it is what the falsifier below tests. The outbound route has no earlier rate, so the split gives it its later rate.

Table 3. Birch, midpoint split of the blended close rate, before the pilot and in Year 2 January (Arithmetic on the company facts).

Route	Share of contacted SQLs, before	Share, January	Close rate, before	Close rate, January	Mix effect (points)	Rate effect (points)
Inbound	100.0%	86.8%	16.00%	16.00%	−2.11	0.00
Outbound pilot	0.0%	13.2%	no history	5.95%	+0.78	0.00
Blended			16.00%	14.68%	−1.32	0.00

The mix effect is the whole change. The inbound close rate is 16.0% in both periods, so no rate effect appears. Had the analytics lead given the outbound route the blended rate of 16.0% as its earlier rate, the split would have put half of the fall into rate. That would be an invented rate effect.

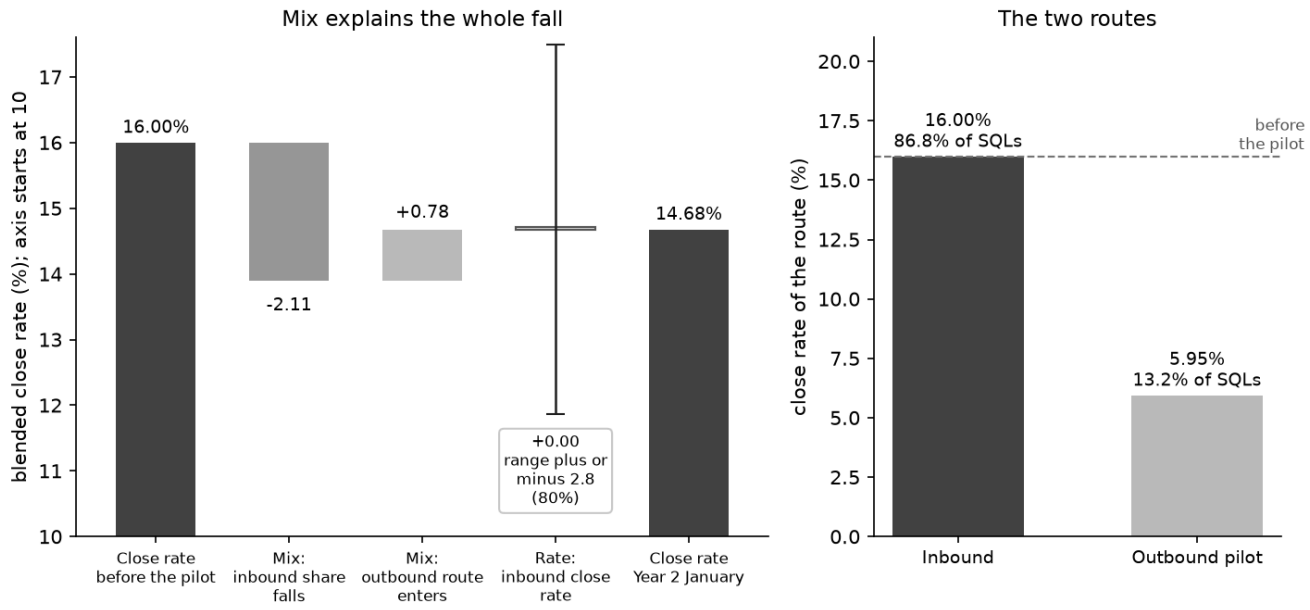


Figure 1. Birch, the blended close rate before the pilot and in Year 2 January. The mix effect is the whole fall. The bar for the rate effect shows the 80% range that the counts allow.

The rates carry noise. The outbound rate rests on 39 SQLs, so its standard error is 3.8 points. That makes the size of the mix effect uncertain, from about 0.7 to 2.0 points at 80%, but not its direction. The inbound rate rests on 261 SQLs, so its standard error is 2.3 points. Together with the noise of the earlier year, the rate effect has an 80% range of plus or minus 2.8 points, which is wider than the fall. The data show that mix is enough to explain the fall. They cannot rule out an inbound fall of 2 points, which would be a real rate story. So the diagnosis carries a falsifier.

If mix is the whole story, the January inbound cohort should win the usual number of deals by day 56. By day 56, about 80% of final wins have come in. With 261 SQLs at 16.0%, that is 33.3 wins. Allowing for overdispersion, the 80% range is 25 to 42. The falsifier reads: "Fewer than 25 inbound wins by day 56 means the inbound rate fell, and mix is not the whole story." The analytics lead also writes down what the falsifier cannot see. It has an even chance of firing only if the inbound rate fell by about a quarter. A smaller fall stays hidden until more cohorts are in. The last January SQL reaches day 56 at the end of March.

The split also opens a second question. Sellers are full. Their load is 123%, with 300 contacts possible in a window against 368 SQLs that arrive. Every outbound SQL that a seller contacts takes the place of an inbound SQL that would close at 16.0%. When sellers are full, the blended close rate is customers per seller contact, so its fall of 8.3% is also a fall of 8.3% in customers. Inbound alone brings 320 SQLs a window.

Without the pilot, sellers would have contacted 300 of them and won 48.0 customers a window. With it, they win 41.7 inbound and 2.3 outbound, which is 44.0. The 4.0 customers lost are worth about 125,000 euro a window, at 31,538 euro a customer on the value basis. At the end of January, 171 SQLs waited for contact.

If the falsifier holds, call coaching recovers none of this, because the loss comes from where seller time goes, not from how sellers sell. The pilot ends with January as planned, so this loss stops. The backlog does not stop, because inbound demand alone exceeds capacity. The recoverable effect sits in how seller time is used: more seller capacity or tighter qualification. *Commit with uncertainty* puts these options on one card.

The growth owner writes the finding at the observation tier, with the exact split as its evidence. The claim "mix, not skill" is firm. The claim "the inbound rate did not move" waits for day 56, and even then it is checked only for a large fall. The head of sales puts call coaching on hold. Both entries go into the log with their dates, without a meeting. At medium volume, mix often explains most of a fall. Split before you retrain anyone.

Cedar, Year 2 January: one country falls

Cedar, Year 2 January: about 1,900 new paying accounts a window, 1.3 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 28 days; the assisted motion, with 79 opportunities a window, needs about 7 months to read a 30% change.

Cedar sells in 24 countries. In Poland, trial to paid fell from 3.42% to 2.60%, a fall of 24.0%. Poland brings 4,416 sign-ups a window, 8.0% of all sign-ups. Shortly before, a change in how the pricing page shows the currency went live in Poland. The product owner proposes to roll it back. The analytics lead asks whether the display is the cause.

High volume helps less than it seems, because Poland has only 8.0% of the volume. The smallest change that Poland can show in a window against the window before is 32.0%. The fall is 24.0%, which is 2.2 standard errors. That would be notable for one metric chosen in advance. But Cedar watches 24 countries, and with many units one large fall in some country is expected now and then (see *Decide when a change needs action* on multiplicity). So one window cannot read it with confidence. The analytics lead turns to cheap checks and writes a falsifier for each before pulling the data.

Three causes go first. The gates rule out tracking: tracked sign-ups agree with the accounts created in the product database. The expectation rules out the season: the January peak is already in it. The split rules out mix: the channel mix of Polish sign-ups did not shift, so the fall sits inside the rate. Two causes remain: a company-wide cause and the display change.

Placebo. The 23 other countries did not get the display change. Pooled, they have 50,784 sign-ups a window, and one window can read a change of 9.4% there. They show no fall of that size. A single country cannot serve as the placebo. If the other countries split their sign-ups equally (Illustration only), one of them has 2,208 sign-ups and reads only 45%. By chance alone, about one of the 23 would fall by 24% or more in a window (simulated, Illustration only). So the placebo must be pooled. Figure 2 shows both views.

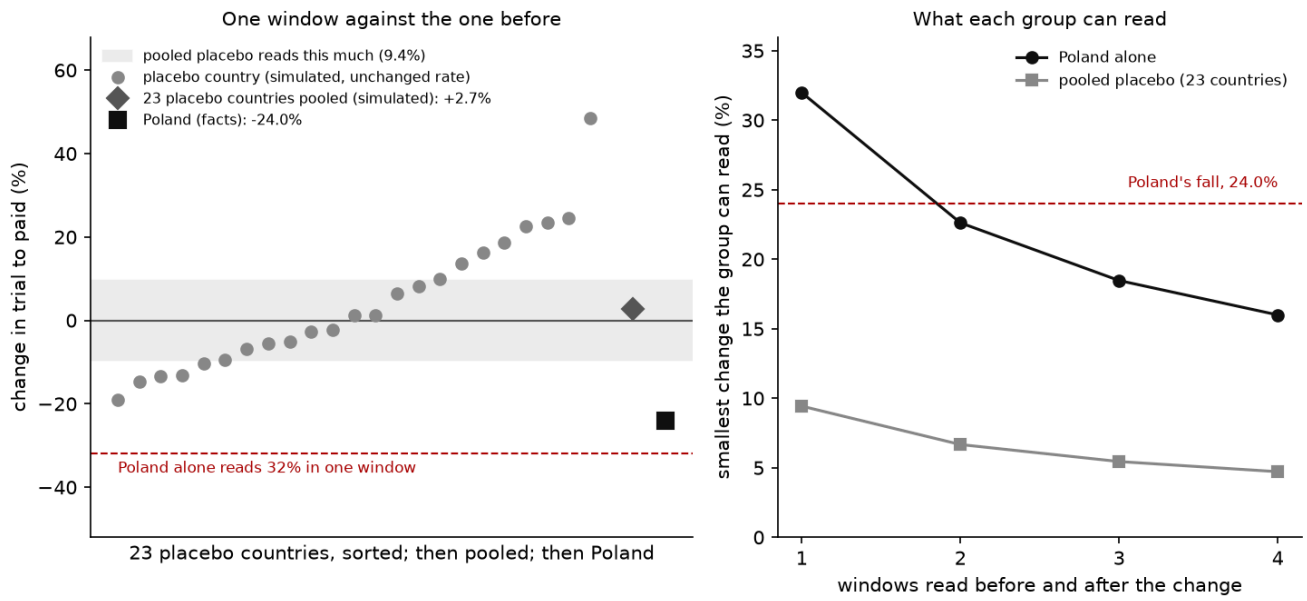


Figure 2. Cedar, trial to paid in Poland against the placebo countries. Left: one window against the one before, for each placebo country, for the 23 pooled, and for Poland. The placebo countries are simulated at an unchanged rate with an equal split of sign-ups (Illustration only). Right: the smallest change that each group can read.

Timing. The change log gives the date of the display change. The fall starts on that date and not before. It shows in every trial that reached the payment decision after the date, including trials that started a few days earlier. That is the diagonal of a period effect, which fits a change on a page that trial users meet when they decide to pay. **Specificity.** The display sits on the pricing page, close to the payment step. Sign-ups and activation do not move. Only trial to paid does.

Dose. Trial users who view the pricing page meet the new display. Trial users who upgrade inside the product see the price there, without the new display. If the display causes the fall, the first group falls more. The falsifier reads: "If both groups fall by the same amount, the display is not the cause." The pull is under way.

Placebo, timing and specificity have passed, and the dose check is still open. The tier is checks passed. Now the cost of waiting. Poland's fall costs 36.2 paying accounts a window. That is 2.0% of the window's new paying accounts. At 1,683 euro a paying account on the value basis, it is 60,900 euro a window, or about 2,200 euro a day. Poland alone reads the fall once two windows on each side are in: the readable change is then 22.6% and the fall is 3.2 standard errors. That needs one more window, which costs about 60,900 euro if the display is the cause. A randomised roll-back for half of the visitors needs 6,815 sign-ups per arm, which is 3.1 windows or 86 days. That is slower still.

If the display is not the cause, a roll-back costs one release, and the team can restore the new display later. That cost is small against 2,200 euro a day of waiting. So the check pays only if the chance of the wrong cause is high, and after three passed checks the owner judges it low. The product owner puts the roll-back for all Polish visitors on a card, and the checks-passed tier allows it. The roll-back carries its own falsifier: if Poland's rate stays near 2.60% for two windows after it, the display was not the cause. A climb back alone proves less, because a fall by chance also tends to return. The log entry holds the checks, the open dose check and the falsifiers. At high volume, cheap checks have power. A small unit, such as one country, still needs time.

The decision to carry forward

Keep competing explanations until evidence distinguishes them. Separate a mix change, a measurement fault and a behavioural change. Act on the mechanism your evidence supports, with the remaining uncertainty carried into the decision.

23 · Learn from outcomes

Did the action improve the business, and what should you believe next? Compare its outcome with the baseline path recorded before the action. A return to normal after a change may be natural reversion, seasonality or another action. The decision log preserves the comparison that hindsight would otherwise erase.

Evaluate two things separately. A treatment can have a useful effect even when a particular outcome is disappointing. A forecast can be well calibrated without the chosen treatment causing any gain. Neither evaluation replaces the other.

This chapter connects rollout comparisons, probability calibration and updates to future decisions. Learning changes the assumptions used to acquire, convert, retain and expand customers. It also changes how much information to buy before the next commitment.

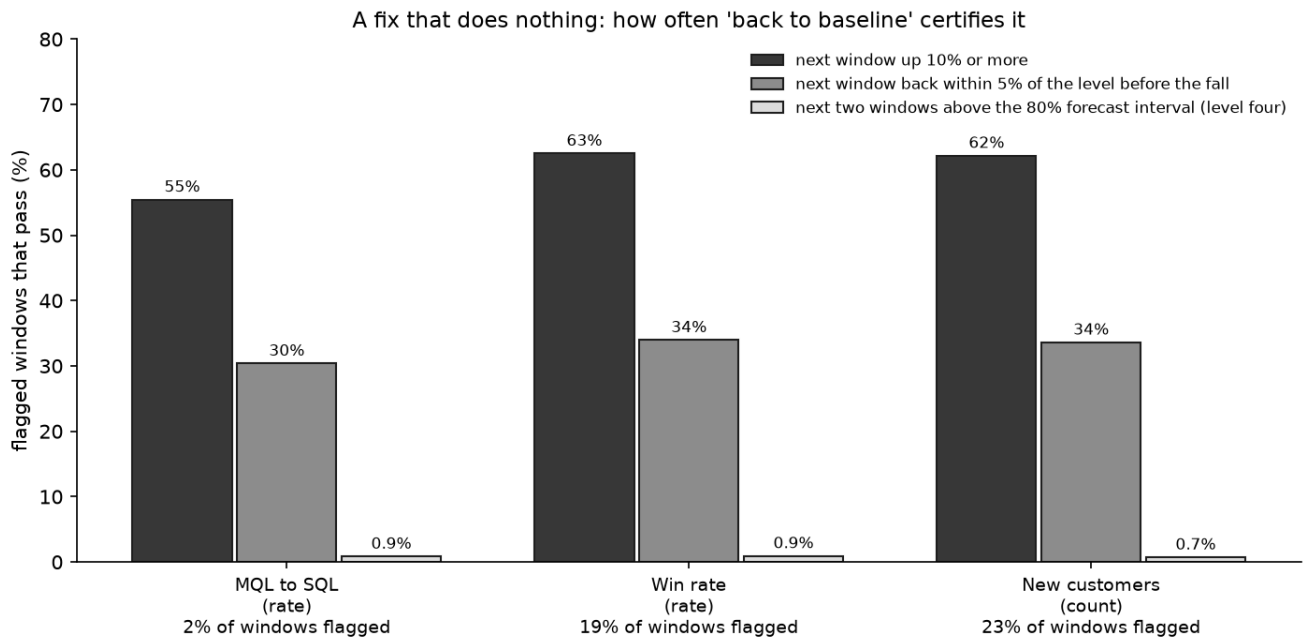
Why "back to baseline" is not proof

A metric falls, a team ships a fix, and a few weeks later the metric is back near its old level. The team writes "confirmed". A metric that fell by chance also tends to come back by chance. *Decide what the evidence can resolve* named this effect regression to the mean. *Decide when a change needs action* showed that the fixed rule "falls more than 15% against the prior window" fires mostly on lucky and unlucky windows.

A simulation of nothing measures the problem. Birch's rates stay at their usual level, the fixed 15% rule flags a window, and a fix that does nothing ships. How often does the next window look like success? The simulation runs 10,000 years of 14 windows each, with the overdispersion start value of 1.2 (Illustration only; fixed seed).

Three tests of success are common.

- The next window is at least 10% above the flagged window.
- The next window is back within 5% of the level before the fall.
- The next two windows both lie above the top of the 80% forecast interval, the level-four test of this chapter.



Noise only: of the windows that the fixed rule flags, the share that pass each test after a fix that does nothing. Birch rates and counts; 10,000 simulated years (Illustration only).
The first test passes 55% of the flagged windows for the MQL-to-SQL rate, 63% for the win rate and 62% for the new-customer count. So an inert fix is graded effective more than half the time. The second test passes about one window in three, and the third less than 1%. The rule flags 1.8% of windows for the MQL-to-SQL rate, 19% for the win rate and 23% for the new-customer count. At the bottom of the funnel, a team that reacts to every flag will "confirm" many fixes that did nothing.

The cure is to compare the outcome with what would have happened without the fix, the **counterfactual**. We never see it, so we estimate it from a control group, a holdback or a forecast. A result is proof only if it beats that estimate by more than the noise.

Levels of evidence for a shipped action

Proof has degrees. Fix the level of evidence when you decide, not when you read the result.

Table 1. The levels of evidence for a shipped action, and what each may say.

Level	What it is	May say
One: opinion	The owners think it worked.	"We believe it worked."
Two: case study	One story, with no comparison.	"It worked here."
Three: before and after	The metric after against the metric before.	"Moved; cause not shown."
Four: before and after against a forecast interval	The metric leaves a forecast interval written in advance and beats the reversion path.	"Confirmed, with the interval."
Five: randomised test or holdback	A random group stays on the old version.	"Confirmed, with the effect and its interval."

Only levels four and five may print "confirmed". The ladder grades an **effect claim**: an action or a cause changed a metric. A **forecast entry** (a count, a range or a signpost) makes no such claim, so it is graded against its falsifier alone. This ladder is Standard practice: the growth literature ranks a plain before-and-after

read below a randomised read (*Growth Data Analytics Playbook*, the section on randomised control trials).

A natural comparison group sits between levels three and four. Examples are a placebo group, a dose check or a roll-out by region. *Diagnose what changed* called these cheap checks, and its evidence tier of "checks passed" is the right label. Such a comparison may say "supported" but not "confirmed", because nobody randomised the groups. The level that the design can reach goes into the log entry before the result is known.

Grading against the forecast interval and the reversion path

Every entry that expects an effect holds two forecasts: one with the action, and one with no action, the **reversion path** that the metric would follow back to its usual level if you did nothing. A fix that only matches this path has shown no effect.

The simplest path moves a share of the way back to the level. The share is the reliability of the window, as in *Decide when a change needs action*.

> **Calculation**
>
> reversion path for the next window = level + reliability × (now − level).
>
> Evidence label: Arithmetic. The reliability is the share of the spread between windows that is real.

Take Birch's MQL-to-SQL rate. Its level is 25.0% and its standard error in one window is 1.37 points. A window two standard errors low reads 22.3%. With a reliability of 0.10 (Illustration only), the path puts the next window at 24.7%. A fix that is followed by 25.0% has beaten the path by 0.27 points. That is 0.2 standard errors, which is nothing.

A shipped action is confirmed at level four when two conditions hold. First, the metric leaves the 80% forecast interval in the expected direction in two consecutive windows, or pooled over three. Second, the metric beats the reversion path. The numbers of windows are start values (Plausible but untested). On the noise-only simulation above, the rule confirms an inert fix in 0.9% of flagged windows for the win rate, against 63% for "back to baseline".

The rule has a price, and the price is power. A fix that really lifts the rate by 20% passes the rule in 98% of simulated cases for the MQL-to-SQL rate. It passes in 31% for the win rate and in 19% for the new-customer count. So at the bottom of the funnel, most real fixes will end at level three: moved, cause not shown. Plan a holdback when the decision matters.

One extension, written in advance. Some reads are not ready on the confirms-by date, because a cohort has not matured. The entry may then carry one maximum extension, with its length written at decision time. The reason must be maturity, not a disappointing result. After the extension the grade is "inconclusive, stopped". A second extension needs a new entry with a new belief. Birch's January cohort shows the clean form: the falsifier of *Diagnose what changed* waits, from the start, until the last January SQL reaches day 56. That wait is part of the design, not an extension.

Holdbacks for rolled-out changes

A holdback is the strongest cheap proof for a shipped change. A share of units stays on the old version, chosen at random, and the two groups are compared. *Choose whether to test or act* sized holdbacks for decisions; here we use them to grade.

A link between an action and a good result can be learned only if you stop the action somewhere and watch (Denrell, in *The Oxford Handbook of Organizational Decision Making*, the chapter on superstitious behaviour). A holdback stops it for a small share of units, on purpose.

Size and duration. A small holdback reads less than an even split would. The standard error of a difference is proportional to the square root of one over the share times one minus the share. A 10% holdback therefore has 1.67 times the standard error of an even split with the same traffic. Time buys the difference back, because a long holdback collects more traffic.

Cost. The units that stay on the old version forgo the gain. The cost is the held share times the expected gain. Table 2 uses Cedar's pricing-page change of *Commit with uncertainty*. The shrunk lift is 4.1%, 17,592 activated users arrive in a window, 9.5% of them pay, and a paying account is worth 1,683 euro on the value basis, with a quality ratio of 0.90. The change earns about 1,345,000 euro over one year of accounts.

Table 2. A holdback on the pricing-page change, read after one year (Illustration only).

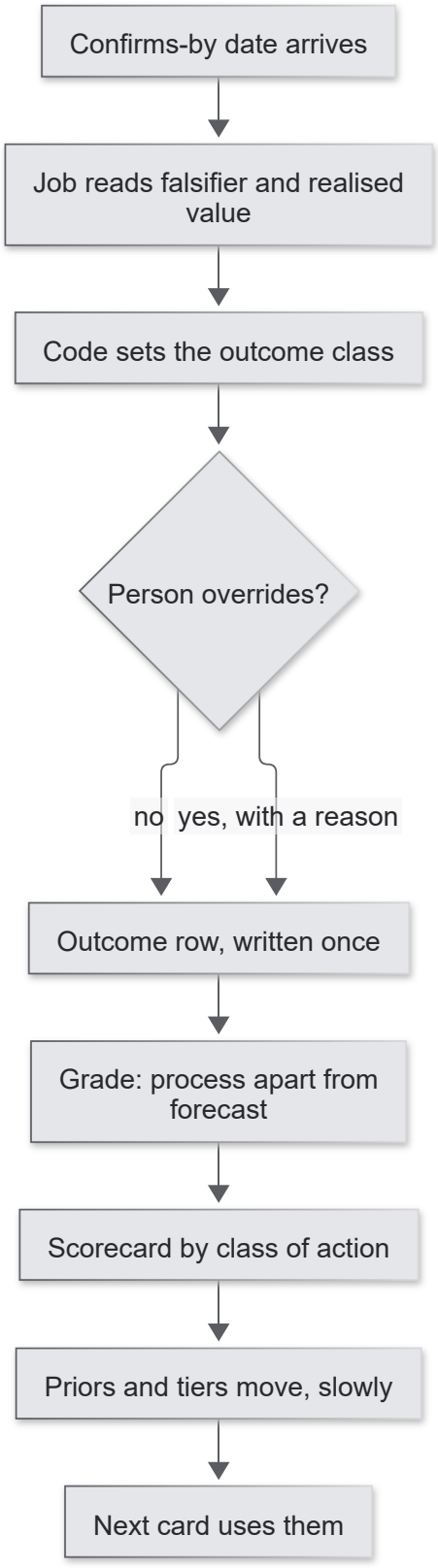
Share held back	Smallest lift readable	Gain forgone in a year, euro
5%	8.5%	67,000
10%	6.1%	134,000
20%	4.6%	269,000
50% (an even split)	3.6%	672,000

Even at 20%, the holdback cannot read a lift of 4.1%. It still guards against harm, because a large loss shows quickly, and it feeds the pooled record of the register. But it cannot confirm this change, so the log entry says so, and the grade will be "not confirmable".

Never hold back a safety fix or a legal change. Withholding a price or an offer from customers raises fairness and consumer-law questions, so ask counsel first. This is not legal advice.

The decision log at the read

On the confirms-by date a job reads each due entry, compares the realised value with the falsifier written in advance, and sets the outcome class by code. The four classes of *Commit with uncertainty* are confirmed, refuted, inconclusive and void. Code sets the class because a person who knows the story will bend it.



From the confirms-by date to the grade, the scorecard and the prior update.

The grade has two parts: the forecast against its falsifier, and the process apart, with the checks of the section on scarce outcomes below. A person may override the class only with a written reason. The scorecard counts overrides by function, never by person, and a high rate is a finding about the rules.

The stored class and the grade words of this chapter relate as follows. A falsifier that did not fire is "held". For a forecast entry the forecast is the claim, so a held forecast is stored as confirmed. For an effect entry a held falsifier is stored as confirmed only at level four or five, and only when the design could read the claim by its date. Otherwise it is stored as inconclusive, with the note that the falsifier did not fire. A falsifier that fired, or a missed signpost, is stored as refuted. A design that could not read the claim by the date is "not confirmable", stored as inconclusive. Void is for an entry whose measurement broke. An entry whose date has not come is "open" and has no outcome row yet.

Table 3 shows one graded entry in full: Birch's backlog belief of Year 2 February.

Table 3. A graded log entry: Birch, the uncontacted SQL backlog (Illustration only).

Field	Entry
Question	Will qualification alone keep the backlog at or below 65 SQLs at every month-end until July?
Belief, written in Year 2 February	Yes, with 97% probability, "with demand on its trend and no other change to the funnel".
Interval	Highest month-end backlog by July: P10 0, median 8, P90 38 SQLs.
Falsifier and confirms-by date	A backlog above 65 at any month-end. First week of Year 2 August.
Realised	Month-end backlog: May 22.4, June 52.2, July 79.7.
Outcome class, set by code	Refuted: the falsifier fired in July.
Kind of entry	A forecast, with a probability and an interval; the levels of Table 1 do not apply.
Cause named at the read	The shorter form shipped in May and raised demand. The card did not hold it.
Process	Belief, falsifier and date were written first. The planned response, a hire approved within five working days, ran in the month of the crossing.
Override	None.

The belief carried a condition, "no other change to the funnel", and the owners then shipped the form. At a company that runs tests, that condition never holds. The grade does not rescue the belief by calling it void: it stays refuted. Write the planned changes into the belief, or widen the interval.

Calibration

A forecaster is calibrated when 70% forecasts come true about 70% of the time. The log makes this checkable, because each entry holds a probability. Three scores summarise a pool of resolved predictions.

The **hit rate** is the share of predictions that came true. The **Brier score** is the average squared distance between the stated probability and the outcome, with the outcome set to 1 if the event happened and 0 if not (Brier, 1950, *Verification of forecasts expressed in terms of probability*, Monthly Weather Review). Lower is better. A forecast of 50% for every case scores 0.25.

> **Calculation**
>
> Brier score = average of (stated probability − outcome)².

> Evidence label: Arithmetic.

The **calibration slope** is the slope of realised outcomes on stated probabilities. A slope of 1 is perfect; below 1 means too much confidence. We also read the gap between the average stated probability and the hit rate, for one pool: a class of actions or one company's log.

Three limits apply.

- A pool needs about 50 resolved entries before it reads anything. A test on 50 entries, with 70% claimed against 55% true, finds the gap in 71% of cases at an exact 5% level. At 100 entries the figure is 90%. At 10 entries it is 26%. To find a gap of 15 points in 80% of cases, the test needs about 70 entries. A gap of 9 points needs about 160 (Arithmetic). Until the pool can read the gap, write "the check cannot show over-confidence yet", never "calibrated".
- A calibration slope is the noisiest of the scores. It needs the stated probabilities to spread widely, and owners state them in a narrow band.
- A person never gets a score. One owner with 12 resolved entries and 8 hits has a hit rate with a 95% interval of 39% to 86%. This is a Wilson interval: an interval for a share that stays sound when n is small (Wilson, 1927, *Probable inference, the law of succession, and statistical inference*, Journal of the American Statistical Association 22(158), 209 to 212). The *Formula sheet* gives the formula. In a forecasting tournament, the Brier scores of the same people on two sets of 94 questions correlated at only 0.36 (Atanasov and Himmelstein, in Seifert (ed.), 2023, *Judgment in Predictive Analytics*, the chapter on talent spotting). Report by class of action and by role. Scoring workers is also a legal matter: ask counsel and follow works-council steps where they exist. This is not legal advice.

Updating priors and tiers

A score is worth nothing unless it changes the next belief, and the change must be slow, because the data are few.

A company may give each class of claim a standard confidence, which some call a tier, for example 70%. Do not confuse it with the evidence tier of a finding in *Diagnose what changed*, which is a text label and not a probability. After outcomes arrive, the standard confidence moves towards the record of its class.

> **Calculation**
>
> new confidence = (k × old confidence + hits) ÷ (k + resolved outcomes).
>
> k is the prior weight: the number of outcomes that the old value is worth. Start value k = 10, with the result kept between 10% and 95% (Plausible but untested).

A worked case: a tier of 70% with 5 hits in 12 outcomes becomes 54.5%. With k equal to 10, a tier needs about 10 outcomes to move halfway to its own record. So little changes in the first year. A person approves every change. A proposal is dropped when the 95% interval of the record contains the current confidence, because then the data cannot support a move. When several tiers are read at once, a correct tier will now and then miss its interval by chance. So a person may hold a tier whose interval misses by a hair, with a written reason. When no tier moves, the right output is "no change, too few outcomes".

The same weighting applies to the prior of an effect. *Commit with uncertainty* estimated the real spread of effects from the register once it held about 30 tests. Re-estimate it every quarter. Be careful with small registers, as the Cedar example shows: the spread read 3.6 points at 30 tests and 3.0 at 140. If an effect cannot be read for years, no outcome will move its prior. Say so in the log, and keep the prior as a belief.

Testing the system

Test the detector, the rules and the models with known cases before anyone relies on them, and again after every change. *Decide when a change needs action* listed these tests for one detector. An **evaluation suite** makes them a standing set for every rule, detector and model.

A suite has four parts.

1. **Noise-only runs.** Run the detector on data in which nothing changes. Every alarm is false. Count the alarms in each run and compare their mean with the false-alarm budget.
2. **Seeded faults.** A **seeded fault** is a known change that you plant in a replay of history or in simulated data, with a known start and size. The suite records whether the system finds the fault, and its detection delay: the time from the start of the fault to the first alarm. A fault in tracking, such as a tag that fires twice, tests the data gates; a fall in a rate or a count tests the detector. Replay each fault many times with fresh noise, because one replay is one draw.
3. **Severity sweeps.** A **severity sweep** repeats one seeded fault at growing sizes, for example falls of 5%, 10%, 20%, 30% and 40%. For each size it records the share of runs in which the system finds the fault within a stated time. The resulting curve shows the smallest change that the system finds reliably, and the size below which it is blind. Run the sweep at the grain of the real case, because a fall in one country is a different case from a fall in the total.
4. **Stress.** Make the world differ from the model. At least vary the overdispersion, because false alarms depend on it more than on volume.

The suite needs acceptance limits, written before the run. Three start values serve (Plausible but untested). The mean number of false alerts per run must not exceed the budget by more than two standard errors of that mean. A planted tracking fault must be caught by the gate in every run. A planted fall of the size that the company acts on must be found within a stated time in at least 80% of runs. When a test passes with no failure in n runs, the 95% upper bound on the failure rate is about 3 divided by n, the **rule of three** (Hanley and Lippman-Hand, 1983, *If nothing goes wrong, is everything all right? Interpreting zero numerators*, JAMA 249(13), 1743 to 1745). So 1,000 clean replays bound the miss rate of the simulated replays at 0.3%. The bound holds for the simulated world, not for the real one.

A text labeller or a model needs a **golden set**: a set of cases with agreed correct answers. A person compares the system's labels with the answers, counting the agreement and the cases where the system declines to answer. *Run the growth system* sets the promotion limits for the labellers.

A suite must be able to fail: a scenario that only shows that a rule fires proves nothing. The Cedar example shows a rule that passed every test of a large fall and was blind to the fall that mattered.

Learning when outcomes are scarce

Where outcomes are few, judge the decision process instead of the forecasts. A process is good when five things hold.

- The belief was written before the outcome, with a probability or an interval.
- The falsifier was a number with a date, and it matched the size of the claim.
- The entry was read on its date, by code, with the class set before anybody told the story.
- The planned response ran as written.
- The lesson was written down, including when the outcome was bad luck.

A good decision can have a bad outcome, and a likely miss is weak evidence. Grade the decision on what was known then. Between outcomes, use outside base rates as the prior, as *Commit with uncertainty* does. A company that never reaches 30 tests keeps them for good. The log still pays, because it forces the written falsifier and the read on the date.

What changes with volume

Regime	What learning can read	What to do
Low volume	About a dozen resolved entries. No calibration, no register slope.	Grade the process. Use outside base rates. Keep the log anyway.
Medium volume	A pooled calibration after about two years, for large gaps only: a gap of 9 points needs about 160 resolved entries. A slope only with a wide interval.	Read the pool by class of action, and say what it cannot show yet. Move standard confidences slowly. Plan a holdback where the decision is large.
High volume	A register of 100 tests or more: the slope of realised on predicted effects, and the optimism of raw lifts. A suite with many noise-only runs.	Update the priors of effects. Run the suite after every change of a rule, at the grain of the real case.

Alder, Year 3 February: ten resolved entries

Alder, Year 3 February: about 3.9 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year.

The log holds 16 entries, and 10 are resolved; the book has told six of them. The chief executive asks whether the company decides well. A pooled test on 10 entries finds a gap of 15 points between a claimed 70% and a true 55% in only 26% of cases. A hit rate of 7 in 10 would read from 40% to 89%. Alder's register holds 5.7 well-recorded tests, and it will never reach the 30 that own priors need. The analyst grades the six told entries.

Table 4. Alder, the six graded entries (Illustration only: realised values follow the median rule of the facts).

Entry	Kind and level of evidence	Result	Stored as
Selling has not changed: 22.2 wins from December to May, 80% range 16 to 29. Realised 22.	Forecast: interval written in advance	Held	Confirmed
A thin pipeline caused the 7 wins: more than 92 opportunities would refute it. Realised 118.	Forecast: count against a falsifier	Refuted	Refuted
Second vertical, six months: at least 12 engaged accounts and 3 opportunities. Realised 16 and 5.	Forecast: dated signpost	Held	Confirmed

Entry	Kind and level of evidence	Result	Stored as
Second vertical, twelve months: at least 2 new customers. Realised 1.	Forecast: dated signpost	Missed signpost	Refuted
Account research raises meetings booked. Realised 6 treated and 6 held back.	Effect: five, but the quarter is too small	Not confirmable	Inconclusive
Account-based programme in two tranches, with a 20% account holdout.	Effect: five, with a power of 13%	Not confirmable	Inconclusive

The wins forecast held, so the expectation model passed its first check. The pipeline falsifier fired against the candidate cause. A thin pipeline was not the main cause, and chance explains the rest. Seven wins or fewer would occur in 14.9% of quarters if selling had not changed, and four deals ended without a decision.

The card gave a chance of 36% that two customers would arrive by Year 3 February, with 1.3 expected; the simulation's facts give 1.43 and 40%. So a miss was the likelier result, at about 60%. The two signpost entries have a Brier score of 0.113 against 0.25 for always saying 50%, but two entries tell nothing about skill. A second number is also 36%: the card's belief that the full move pays, after a missed second signpost. It is far below the bar of 95% for a move that is hard to reverse. Stage two rests on the owners' judgement of the market beyond 24 months, as the entry said at the start.

The research pilot and the programme holdout were built at level five, and neither could read its claim. The pilot's counts were both 6, and *Choose whether to test or act* had said in advance that a quarter could not show whether research helps. So both are stored as inconclusive. The programme was graded on process: its reading rule was written first and followed.

The chief executive decides not to start stage two now. The analyst writes a new entry for the next signpost, with a belief and a date, before anybody reads new data. At low volume learning is qualitative and borrows outside base rates. Keep the log anyway: it forced six falsifiers with dates, and four of them gave a reading.

Birch, Year 3 February to March: did the fix work?

Birch, Year 3 February: about 62 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate.

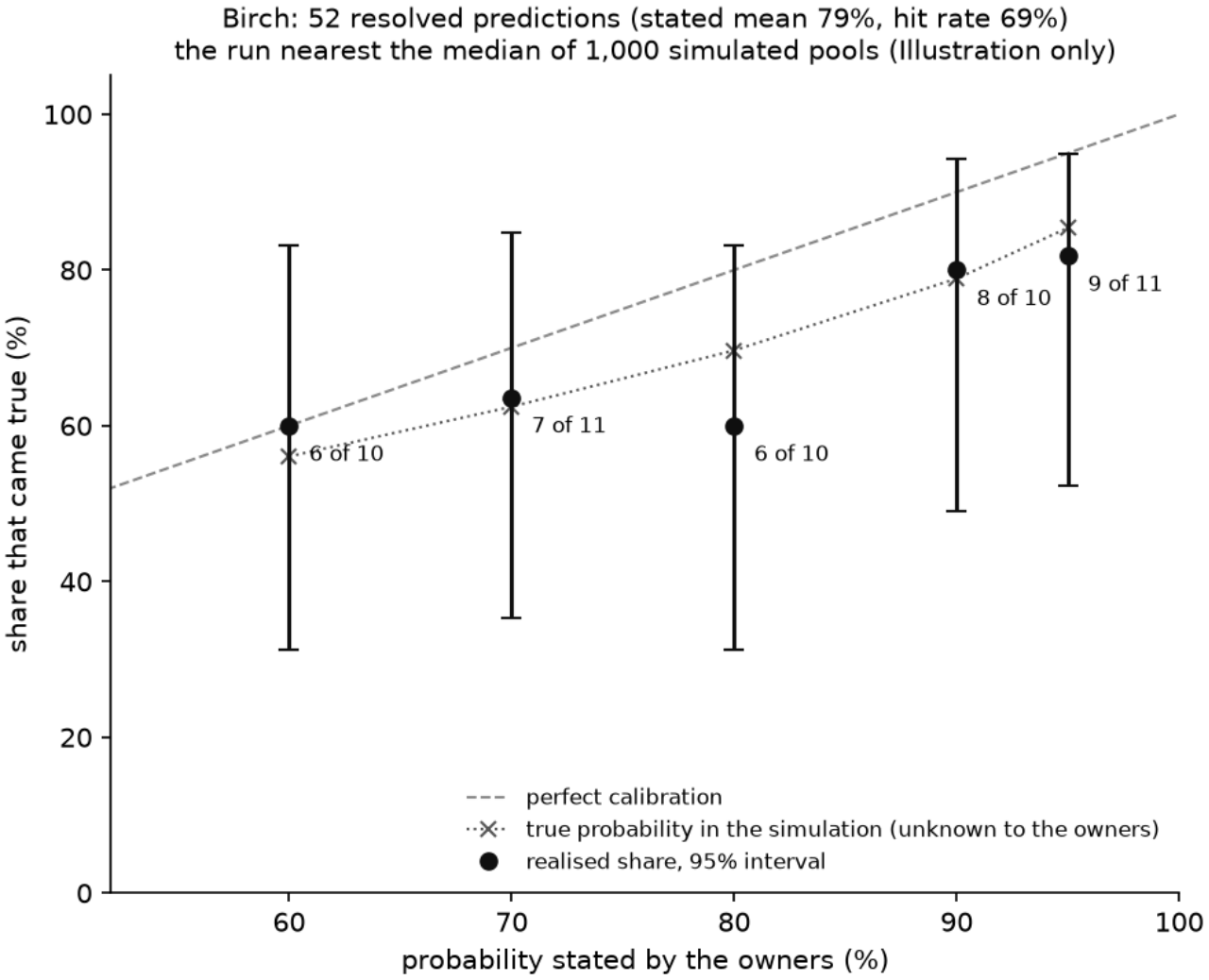
The log holds 64 entries, and 52 are resolved. The growth owner asks whether the qualification fix worked and how well calibrated the owners are.

The fix. It stopped 6.0% of SQLs, with an 80% range of 5.5% to 6.5% from the replay, as designed. The claim that matters is the gain of 7.0% in the close rate per contacted SQL, and the close rate needs 1,228 days, about 3.4 years, to read it. Before and after is level three here, and it is confounded. Since Year 2 February Birch also changed eight other things, from the shorter form and six budget pulses to the rebuilt lead score and the price rise.

Two partial reads exist, and neither settles it. The sales-accepted rate rose from 80% to 83%, a rise of 3.75%, below the 6.2% that a quarter against a quarter reads. The backlog ended July at 79.7 SQLs, about 200 below the card's do-nothing path. But 171 of those 200 were the old queue, which the fix sent back to nurture. The shorter form's extra demand sits in the rest.

So the mechanism is supported, the effect on the close rate is not confirmable, and the entry reads "moved, cause not shown". A holdback of qualification would need the same 3.4 years. The sales owner keeps the rule and its two signposts, the stopped share and the time that opportunities spend in their stage.

The pool. Table 5 groups the 52 resolved predictions, which are simulated, by the probability that the owners stated. They include the two probability entries that the book told: no false alarm in the first quarter (73%, it held) and no backlog crossing (97%, it did not).



Birch: realised share against stated probability for 52 resolved predictions, with 95% intervals; the run nearest the median of 1,000 simulated pools (Illustration only).

Table 5. Birch, the first pooled calibration (Illustration only).

Stated level	Entries	Came true	Share	95% interval
60%	10	6	60%	31% to 83%
70%	11	7	64%	35% to 85%
80%	10	6	60%	31% to 83%
90%	10	8	80%	49% to 94%
95%	11	9	82%	52% to 94.9%

The owners stated an average of 79.2%, and 69.2% came true (36 of 52), with a 95% interval of 56% to 80%. The gap is 10.0 points, and the interval still holds 79%. The Brier score is 0.218, against 0.250 for always saying 50% and 0.213 for always saying the pool's own hit rate, a rate known only afterwards. So the owners add nothing over a plain base rate. The calibration slope is 0.65 with a standard error of 0.50: below 1, which points to over-confidence, but too noisy to say so by itself.

So after 52 resolved entries the check cannot show over-confidence yet. That is a finding about the size of the pool, not a pass for the owners. In this simulation they are over-confident by 8.5 points. The exact test of the section on calibration finds a true gap of 9 points (79% claimed against 70% true) in only 39% of pools of 52. From 158 resolved entries on, it finds that gap in at least 80% of pools (Arithmetic). At four resolved entries a month, Birch's log reaches 158 at the end of Year 5 May. The analytics lead writes "cannot show over-confidence yet; read again at 158 entries".

One pool is also one draw. The shown pool is the run nearest the median of 1,000 pools built from the same truth. Of those pools, 39% read a lower hit rate, 12% the same and 49% a higher one. A typical pool reads 69%, with P10 to P90 of 62% to 77%. The slope runs from 0.04 to 1.20 at P10 to P90, so a pool of 52 does not read it.

One match proves nothing either. The card's gain of 7% per contacted SQL equals the simulation's true gain, because the book built it so, to show the method. The honest test is in *Commit with uncertainty*: the card fails if the belief falls by 2.2 points.

Tiers. Four of the five records hold their tier inside the 95% interval, so those proposals drop. The 95% record, 9 of 11, has an interval that ends at 94.9%: it misses its tier by a hair. If all five tiers were correct, at least one of the five intervals would still miss its tier in 24% of pools (Arithmetic). So the growth owner holds all five tiers, logs "no change, too few outcomes" with that reason, and reads the 95% tier again next quarter. The weighted update with a prior weight of 10 would give 60.0%, 66.7%, 70.0%, 85.0% and 88.1%; the log records it unused.

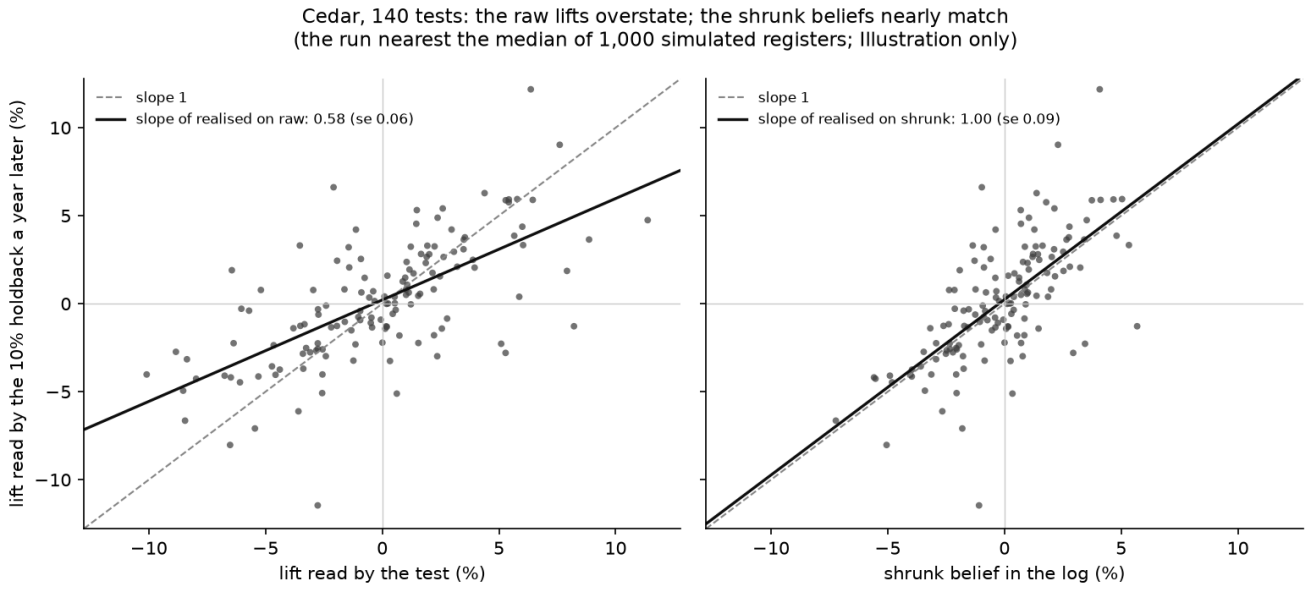
The second trip. The hiring rule tripped again at the end of Year 3 February, with a backlog of 103.6 against the rewritten limit of 100. The sales owner approved a twelfth seller within the week: the process working a second time. It is not a refuted forecast, because the rewrite of Year 2 December logged no probability; the lesson is to log one. The half media shift entry stays open until its conversion ratio is read in Year 3 May.

At medium volume a pooled calibration becomes readable after about two years, but only for large gaps, and it reads the gap before the slope.

Cedar, Year 3 March: a register of 140 tests

Cedar, Year 3 March: about 2,100 new paying accounts a window, 1.2 times the derived typical; high volume: every self-serve rate reads a 20% change within a window.

The register holds 140 tests. Each leaves two records: the lift that the test read, and the lift that a 10% holdback read a year later. The belief in the log is the shrunk lift. The register prior has a mean of 0 and a spread of 3 points. The simulation draws the standard error of each test between 0.95% and 4.7%, and assumes that every test leaves a holdback read. Real registers hold such reads only for shipped changes.



Cedar: realised lift against the lift that the test read (left) and against the shrunk belief (right), 140 tests (Illustration only).

The analytics lead fits the slope of realised on predicted. Against the raw lifts the slope is 0.58, with a standard error of 0.06. About 58% of what tests read is real. Against the shrunk beliefs the slope is 1.00, with a standard error of 0.09. The shrinkage works. The shown run is the run nearest the median of 1,000 runs (the 51st percentile on both slopes). Across the runs, P10 to P90 is 0.90 to 1.10 on shrunk beliefs and 0.49 to 0.66 on raw lifts.

The winners show the curse most clearly: at the one-sided 5% level, 24 of the 140 tests count as winners. They read 5.3% on average, and the holdback read 3.5%. The raw lift overstates by a factor of 1.5. The shrunk belief averaged 3.4%.

The register also updates the prior. The estimate of the real spread was 3.6 points at 30 tests, 3.3 at 62 and 3.0 at 140. At 30 tests the estimate sat well above the true 3.0: a small register misleads. The register mean, with a prior weight of 10 tests, moves from 0.0 to minus 0.1 points. The growth lead leaves the prior at 0 and 3 points.

Table 6. Cedar, the four graded entries (Illustration only).

Entry	Level of evidence	Result	Stored as
The currency display caused Poland's fall from 3.42% to 2.60%.	Natural comparison: dose check and roll-back	Held, "supported". Pricing-page viewers fell more; the rate was back at 3.42% from February.	Inconclusive: a natural comparison does not confirm
Pricing-page winner: shrunk lift 4.1%.	Five, but a 10% holdback reads only lifts above 6.1% in a year	Not confirmable; graded through the register	Inconclusive
Onboarding combination: shrunk lift 4.3%, 80% range 2.2% to 6.3%.	Five, holdback after about 75 days	Held. The realised lift was 6.0%.	Confirmed
Paid social: 0.94 euro per euro.	Geo test	Open. No second test has a date.	No row yet

The dose check, not the roll-back, carries the weight (*Diagnose what changed*). The onboarding lift fell above its shrunk belief and inside its range: one test where shrinkage was too cautious.

Now the suite. Cedar monitors 120 metrics every day. The daily rule alerts when a standardised count falls more than 3.04 standard errors. That limit spends a budget of one false alert a week across 120 metrics and 7 days.

Noise only. In 1,000 simulated weeks the rule raises 0.85 false alerts a week on average. The limit for acceptance is 1.06, so it passes. With a true overdispersion of 1.5 it raises 2.4 alerts a week, and with 2.0 it raises 6.8. Both fail, so the suite would catch a wrong overdispersion.

Seeded fault, the tag release. Sign-ups fall by 38%, which is 13.9 standard errors. The rule alarms on the first day in 1,000 of 1,000 replays. By the rule of three, that bounds the miss rate of the simulated replays at 0.3%, with 95% confidence. The sweep of sign-ups, day by day, finds a fall of 10% on the first day in 74% of runs, and within seven days in all runs. A fall of 5% is found within seven days in 56% of runs.

Seeded fault, the Polish fall. The same daily rule finds a fall of 24% in Poland within 28 days in 0% of runs. It finds no fall at any size of the sweep, up to 40%. The reason is arithmetic. Poland has about 5.4 paying accounts a day. Even a day with none gives a z-score of only minus 2.16, so the limit of 3.04 can never be crossed. The rule was blind by design. *Decide when a change needs action* said that small counts need exact limits.

The analytics lead proposes a rule that reads each country once a window. Its budget is one false alert a quarter across 24 countries and 3.25 windows, a limit of 2.23 standard errors. In noise-only runs it raises 0.80 alerts a quarter, which passes. It finds the Polish fall in one window in 74% of runs, and within two windows in 92%, above the acceptance limit of 80%. The sweep shows its range: 20% in two windows in 76% of runs, 10% in 25%, and 5% in 8%. The growth lead approves the rule, and the suite reruns after every change.

At high volume the system can measure its own optimism. The suite must test at the grain of the case that matters.

The decision to carry forward

Evaluate the action against its counterfactual and the forecast against its recorded probabilities. Update beliefs without erasing the original decision. Revisit future targeting, funding and test choices when the learning changes their economics.

24 · Run the growth system

How do you make this practice work every week? Assign ownership of commercial decisions and of the evidence that supports them. The same written record must connect the action, assumptions, computation and later outcome. A hand-off between people need not split the reasoning into separate disciplines.

Work runs asynchronously unless a decision needs a meeting. The owner decides by a deadline, records the commitment and returns to its evidence on the planned date. Code computes reproducible quantities. People judge the objective, assumptions and acceptable exposure. Automation operates inside those boundaries.

The system below joins cadence, roles and learning. Its success is the quality and speed of economically useful decisions, including decisions to stop. The first rollout establishes the minimum history needed to judge that success without waiting for a perfect data platform.

The rule: code computes, models draft, people decide

One rule governs the whole operating system. Table 1 shows it.

Table 1. Three kinds of worker in the loop.

Worker	What it does	What it never does
Code (a scheduled job)	Computes every number: rates, intervals, cards, scores, grades. Runs on a timetable. The same input gives the same output.	Weighs a trade-off between options. Reads free text.
AI model (inside an agent)	Reads text. Drafts prose: a memo, a log entry, a label with a quote.	Computes a number of its own. Decides. Writes to a system without a gate.
Person (the owner)	Decides. Accepts a proposal, adjusts it inside a stated range, or rejects it.	Edits a logged belief after the outcome is known.

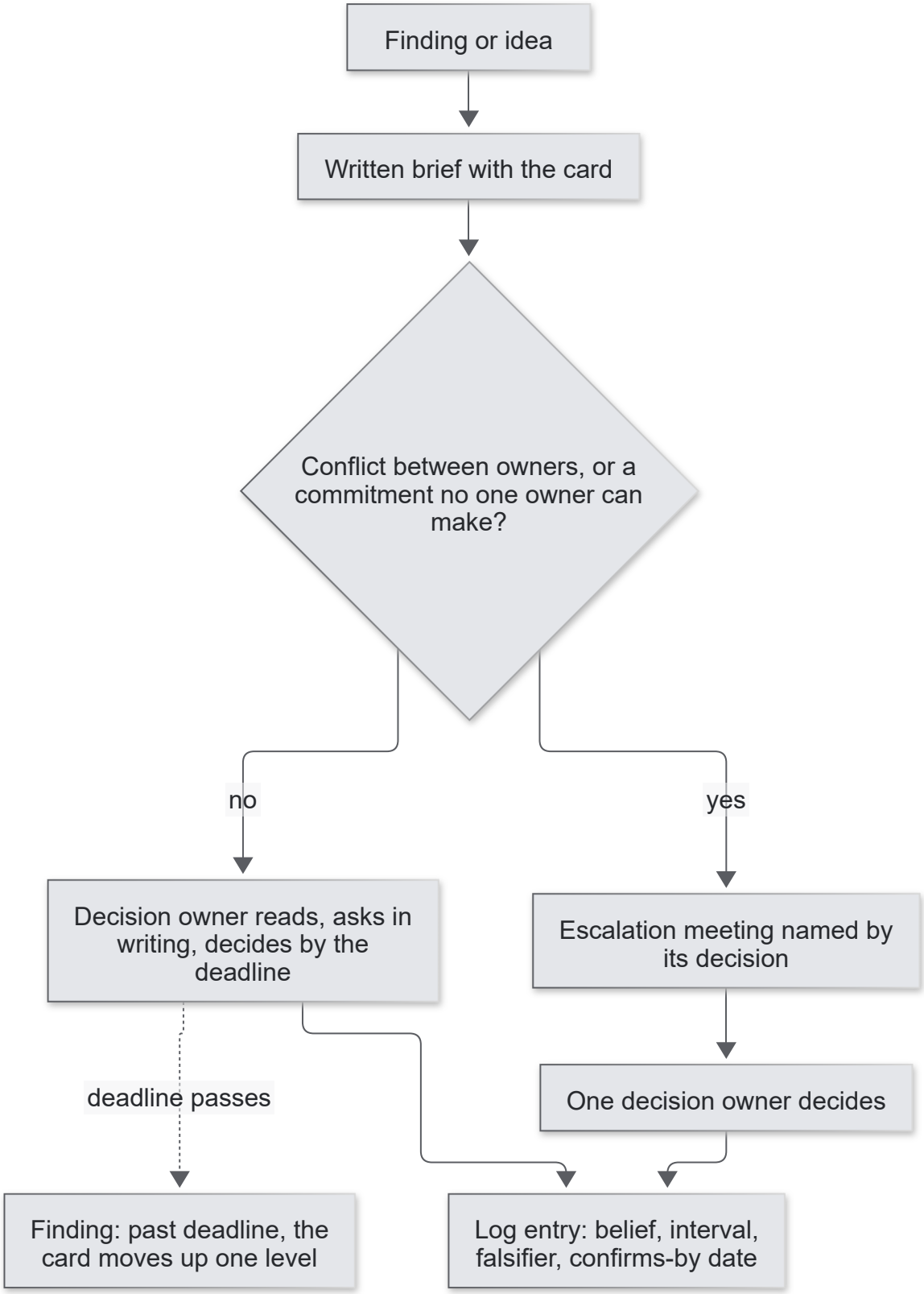
The split follows the strengths of each worker. Code is exact and repeatable, but it cannot weigh a trade-off between the two wings of the bowtie. A person can weigh it, but cannot read 120 metrics every morning, which is the number that Cedar monitors. A **large language model** is an AI model trained on large amounts of text to read and write text. It reads well and writes fluent prose. It can also write a fluent, wrong number. So no number comes from a model. Code makes the number, and the model receives a placeholder to quote.

Agents advise. They never decide on money, price, people, commitments or consent. These five stay with a person at every step of the autonomy ladder below. We also never ask an agent for output about one person. Every result goes by role, route or stage.

Asynchronous by default

A meeting is a poor way to make most decisions. It adds a wait: a fortnightly meeting makes a card wait one week on average. The material is old when the meeting reads it. And a vote replaces the number on the card with an opinion (see *Fund the next best action*). So the standard path has no meeting and no vote.

An **asynchronous decision** is a decision made without a meeting. A written brief goes to one owner, who decides by a deadline. The decision log records it.



The asynchronous decision path. The meeting is the exception, drawn as a branch.

The written brief

The **written brief** starts the path. It is one page, and it holds six things.

- The question that the decision answers.
- The options, always including do nothing.
- The decision card of *Commit with uncertainty*: value, range, chance of paying and the input that would flip the choice.
- What the owner needs from others, and by when.
- The **decision owner**: the one named role that decides. It is always one of the roles in Table 3.
- The deadline. It is a date, not a meeting. The start value is five working days (Plausible but untested).

Code computes the card. An agent may draft the brief from the card and the findings. The analytics lead checks that every number comes from a job. The owner may ask questions in writing and then decides alone. Silence is not consent. A card that passes its deadline becomes a finding. It goes up one level, in the end to the bowtie owner, with the date it stood open.

Choosing experiments and projects

The card ranks the work. A project or an experiment goes forward when its expected net value clears the hurdle of *Fund the next best action*. For a test, the value of information on the card is part of that value (see *Choose whether to test or act*). The owner of the card decides by its deadline. Work in progress stays inside its limit. Nobody votes. A company that retires a council that chose by score does not set up a new council. A number and an owner take its place.

Escalation

Escalation is a meeting called for one decision. Two cases justify it.

1. Two owners have a real conflict that the card cannot settle. Each owns a part of the result, and the disagreement rests on an input that nobody can measure in time.
2. The commitment is larger than the authority of any one owner. An example is a promise to customers that binds both growth and finance.

The bowtie owner calls the meeting. The meeting is named by the decision it makes, for example "whether assisted sellers may commit to the annual-plan discount". The brief goes out before the meeting. Only the owners in conflict and the decision owner attend. The bowtie owner is the decision owner unless the brief names another. The meeting ends when the log entry is written. A meeting that ends without a decision is a finding too.

A company may cap the number of escalation meetings. Alder allows at most one a quarter. The cap is a rule and not a target. It forces the owners to settle most conflicts on paper.

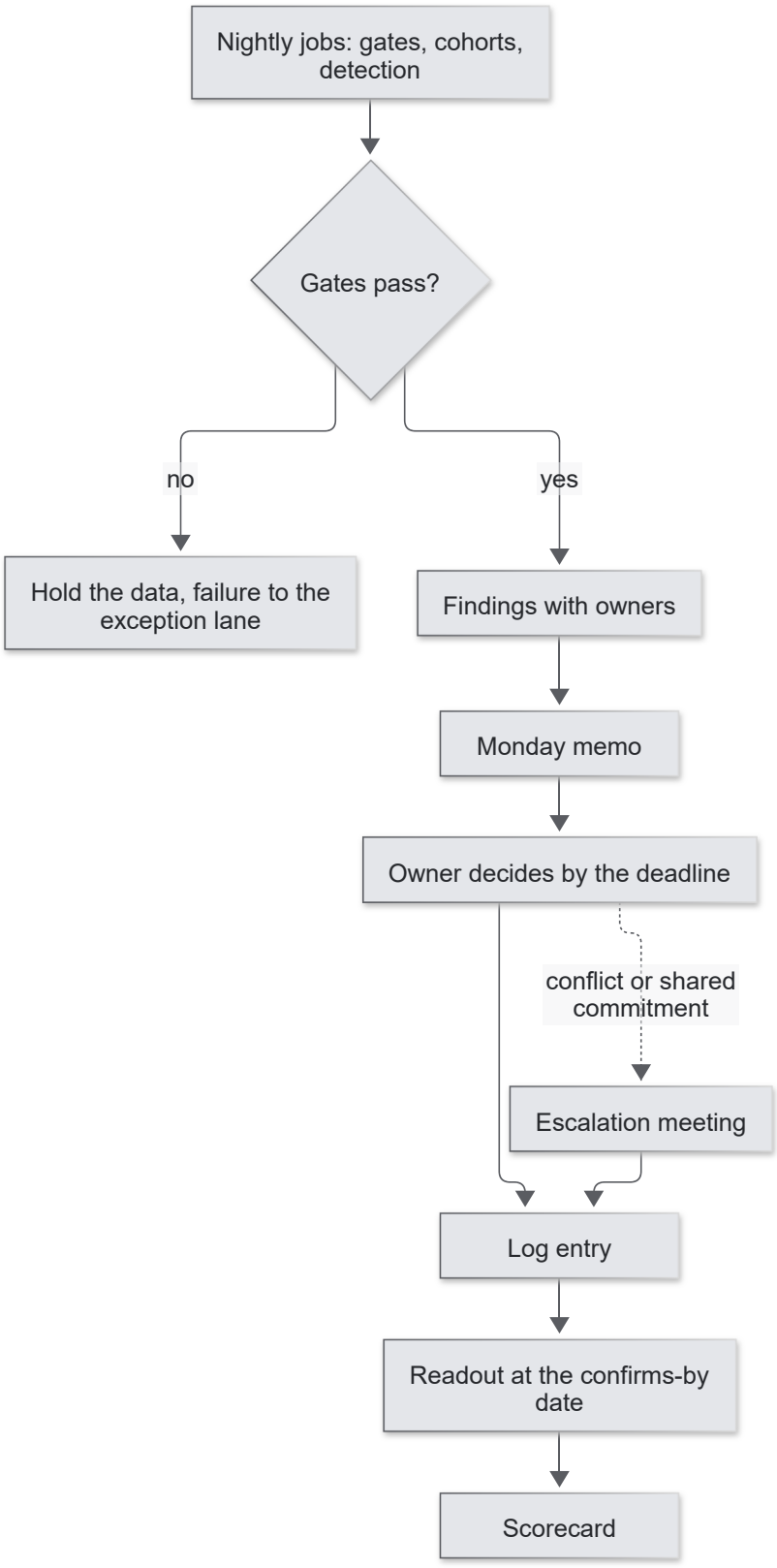
The rhythm of the loop, without meetings

The rhythm follows the data and not the calendar. A rate that matures in a quarter gains nothing from a daily read. *Decide what to measure* gave each stage a loop delay. *Decide when a change needs action* showed how long a detector needs to see a change. Reading faster than the data arrive only adds noise.

Five beats make the rhythm: a daily check, a weekly memo, owner decisions at their deadlines, a monthly bowtie note and a quarterly refresh. A grade also arrives at the confirms-by date of each log entry. Table 2 lists them, with the readout and the escalation meeting. Only the last row is a meeting.

Table 2. The rhythm of the loop, with the variant for low volume.

Beat	What it decides or informs	Owner	Timing	Output	At low volume
Data gates and tests	Whether today's data can be used	Analytics lead	Daily, after each load	Pass or hold. Failures only go to the exception lane.	The same
Detection run	Which metrics moved beyond noise	Analytics lead runs it. Stage owners receive findings.	Nightly. An act-level finding reaches its owner at once.	Findings, each with an owner and a deadline	Counts and activities in time. Rates as trailing-year figures with intervals.
Weekly memo	Findings, proposed decisions, last week's grades, questions only people can answer	Analytics lead owns it. Stage owners read it.	Monday	One page	Monthly
Owner decisions	Each open card	The decision owner	By the deadline on the brief	A log entry	The same
Readout and grade	Whether the belief of a decision held	Analytics lead runs it. The decision owner reads it.	At the confirms-by date	Outcome class, grade and scorecard row (see *Learn from outcomes*)	Judge the process, because items are few
Bowtie note	Trade-offs between the wings: where to look, what to stop	Bowtie owner	Monthly	One written page from the executive page	The same
Quarterly refresh	Models, priors, calibration, specifications, the tile list and the risk register	Analytics lead with the finance owner. Each change goes to its decision owner in Table 4.	Every 91 days	Refreshed parameters, one log entry for each change	Outside base rates stay the priors
Escalation meeting	One named decision	Bowtie owner	Only when a case arises	A decision and a log entry	Capped by rule



One week of the loop. The jobs run at night. The owner decides inside the week. The meeting is the dotted exception.

Most weeks hold no card at all. "No action is supported" is a valid memo. The memo shows each rate with its readability from *Decide what the evidence can resolve*. So nobody acts on a lagging rate that the volume cannot read. It proposes a decision only for an act-level finding or a due signpost.

Keep two things apart. A gate is a fixed rule that checks the data. For example, it checks that a load arrived and that the CRM amount matches the first invoice. A gate does not learn. Detection is a statistical run on data that passed its gates. A failed gate holds the data. A detection finding asks for a look or a decision. Detection does not decide whether a gate passes.

The reading speed rises with volume, and the deciding speed does not. At high volume every rate has power, so gates and detection run daily. At low volume a weekly count is mostly chance, as the Alder example shows. The deadline of a decision is the same in both cases.

Roles and ownership

A **role** is a set of decisions and duties that the loop needs. A person holds a role. One person may hold several. A **stage owner** is the person who owns the metrics and the levers of one funnel stage: the marketing owner, the sales owner or the customer success owner. Mechanics in this book name the role. Examples may use the company's title for the person, but each chapter uses one name for each person. Table 3 is the one role list of the book. It also shows who holds each role at the three companies.

Table 3. The role list: what each role owns and who holds it (holders: Illustration only).

Role	What it owns	Alder	Birch	Cedar
Bowtie owner	Settles conflicts between the left wing (acquisition) and the right wing (retention and expansion). Reads the executive page each month. Owns the rhythm of the loop and the escalation path. Is the one decision owner of an escalation meeting.	The chief executive	The growth owner	The chief executive
Growth owner	The left wing as a whole, when marketing and sales have separate owners: the detection rules of the funnel, the diagnosis of left-wing findings and the trade between demand and seller capacity.	Not a separate role (the chief executive)	The growth owner (also the bowtie owner)	The growth lead (the self-serve funnel end to end)
Marketing owner (stage owner)	The stages up to the MQL or the sign-up: channels, paid spend, creative, landing pages, forms, the lead definition with sales and the audience contract.	The marketer	The head of marketing	The growth lead
Sales owner (stage owner)	From the SQL or the product-qualified account to closed-won: qualification with marketing, the hand-off contract, seller capacity, the hiring trigger and the sales process with its signposts.	The head of sales	The head of sales	The assisted-sales manager (assisted motion)

Role	What it owns	Alder	Birch	Cedar
Customer success owner (stage owner)	Onboarding, adoption, renewal and expansion. The save plays, the after-the-close page and the renewal calendar.	The head of customer success	The head of customer success	The growth lead for the self-serve book (cancellation flow, save offers). The assisted-sales manager, with the customer success managers, for assisted accounts.
Product owner	One product surface inside the growth team, such as the pricing page, onboarding or the cancellation flow. Proposes changes and roll-backs on a card.	None	None	A product owner in the growth team
Finance owner	The value basis (discount rate and horizon), the ARR bridge with revenue operations, CAC and payback, budgets and tranches, and the match of billing with the CRM.	The head of finance	The chief financial officer	The chief financial officer
Analytics lead	The measurement system and the daily read of data health. The jobs that compute every number: cards, detection, readouts and scorecards. Test design. Advises, and never decides an option.	The analyst (half a post)	The analytics lead	The analytics lead (one of two analysts)
Revenue operations	The CRM, the billing map and the stage definitions. Checks that the ARR bridge closes at each month end.	The head of finance with the analyst	Sales leadership and operations (four people)	The analytics lead with finance
Decision owner	Not a separate role. It is the one named owner of a given decision, always one of the roles above.			

Agents and jobs are not roles. The memo agent at Birch, the deal-note labeller at Alder and the support-ticket labeller at Cedar draft and label. A person in one of the roles above decides.

Two names need a note. At Birch the head of sales is the sales owner, and the growth owner also holds the bowtie role. At Cedar the finance owner is the chief financial officer, and the chief executive holds the bowtie role.

A person who holds several roles can write the brief, weigh it and decide it. Two guards protect the decision. First, the analytics lead computes the card and decides only on measurement: definitions, rules and limits. It never decides a business option. Second, the finance owner is consulted in writing on any commitment of money, and the log shows the consultation. Table 4 sets the decision rights that follow from the roles.

Table 4. Decision rights (Illustration only).

Decision	Decision owner	Consulted in writing
A change to a metric definition, a rule or a limit	Analytics lead	The stage owner. The finance owner, if money is involved.

Decision	Decision owner	Consulted in writing
Spend and creative inside the budget	Marketing owner	Analytics lead, on the card
A hire under a written trigger	Sales owner	Finance owner is told
A hire or a budget with no trigger	Bowtie owner	Finance owner and the stage owner
The discount rate and the horizon of the value basis	Finance owner	Analytics lead
Moving an agent to its next step	The owner of the work the agent touches	Analytics lead gives the evidence
Anything that touches workers, recordings, ad-platform data or AI chat with customers	Bowtie owner, after counsel	Counsel. The works council, where one exists.

The bowtie owner earns the role in conflicts between the wings. Take a plain case. The marketing owner wants more demand. The customer success owner says that new customers would land on an onboarding team that is already full. Each owner is right about their own stage. Only the value basis lets one person compare the two options: the expected gross profit of each over the horizon, after the capacity limit (see *Allocate sales attention*). This is why the bowtie owner reads the executive page each month. The bowtie note then says where to look and what to stop.

Every report, scorecard and agent output goes by role. Nothing ranks a person.

Decision discipline in writing

A fast process loses the friction that silos used to supply by accident. The finance person who asked about payback and the analyst who doubted the marketer both caught errors. Put that friction back on purpose, in writing and at low cost. Five habits do it.

- **Independent estimates.** Each consulted role writes its own estimate of the key belief on the card before it sees any other view. The belief is the effect, the chance or the date. The analytics lead then shows the spread. Teams that work apart before a manager combines their views keep more of their different knowledge (National Research Council, 2011, *Intelligence Analysis for Tomorrow*, on collaboration). The estimates are beliefs, and the card labels them so.
- **A pre-mortem.** A **pre-mortem** is a written list of the ways a large bet could fail, made before the bet, as if it had already failed (Klein, 2007, *Performing a project premortem*, Harvard Business Review 85(9), 18 to 19). Write one for every large bet and every irreversible bet, which is every bet held to the 95% bar. The list feeds the falsifier and the signposts.
- **Bounded overrides.** An owner may change an input or reject a proposal inside a stated range, with a reason in the log. People trust a decision aid more when they can adjust its output. Adjustments limited to a stated range also gave more accurate decisions than free ones (Dietvorst and others, 2016, reported by Burton, Stein and Jensen in Seifert (ed.), 2023, *Judgment in Predictive Analytics*). An override outside the range goes to the bowtie owner. The log counts overrides by function and never by person.
- **The exception lane.** The **exception lane** shows each person only what needs them. That means gate failures, findings at the act level, briefs addressed to them and items past their deadline. All other items run without their attention. This is how a small team can work with 120 monitored metrics.
- **The yield test.** Apply the **yield test** to every recurring meeting that survives. If the meeting changed no decision in four sessions, retire it (a start value, Plausible but untested).

Jobs and agents

A **scheduled job** is code that runs on a timetable and computes numbers. An **AI agent** is a program that uses an AI model and a few defined tools. It reads or drafts text inside set limits, and then it hands over to a person. Most of the loop is jobs. A job cannot invent a number, and it has tests. Only two text tasks need a model in the core of the loop: turning findings into a page, and labelling free text.

Table 5 lists the roster. A **human gate** is a step where a person must approve before anything else happens.

Table 5. The roster of jobs and agents.

Job or agent	Trigger	Input	Output	Human gate	Evaluation
Data checks	After each load, daily	Source tables, tests, reconciliations, consent ratio	Pass or hold. Failures as findings.	Analytics lead acknowledges the same working day	Seeded faults replayed
Cohort and economics engine	Nightly	Stage events, renewals, costs	Rates by cohort age, snapshots, cohort economics	Analytics lead locks baselines	Hand-worked fixture. No join sees future data.
Rule engine (detection)	Nightly	Metric windows, expectation model, limits	Findings with z-score, decomposition and readability	Analytics lead, for any rule or limit	Noise-only runs. Severity sweep.
Evaluation suite	At every change to a rule, card, parameter or prompt	Scenarios, seeds, golden sets	Pass or fail table, false alarms, delay	Owners write the expected outcomes	A deliberately broken rule must fail
Readout	At the confirms-by date	Fields written before the decision	Outcome class with interval	The decision owner reads it	Runs on data with no effect. Planted lift.
Scorecard	Monthly	The decision log	Hit rate, Brier score, calibration, proposed prior changes	The bowtie owner approves any change	A log with no real effect must give no proposed change
Scoring	Daily	Outcomes, account traits, usage	Lead and account scores with a version	The stage owners set the cut-offs: marketing and sales for lead scores, customer success for account scores	Out-of-time lift and calibration
Lint	At each change to a specification	Metric cards, lineage, glossary	A report of drift and gaps	The analytics lead	Seeded defects
Weekly memo writer (agent)	Monday, after the rule engine	Findings and computed tables only	A one-page memo with number tokens. Draft log rows.	The owner reads and decides	Golden weeks. Noise-only weeks. Number lint.
Text labeller (agent)	Nightly batch	Redacted notes or tickets, a closed label list	A label, an exact quote and an abstain flag	The taxonomy belongs to the analytics lead. Low confidence goes to a person.	Golden set. Kappa. Abstain rate.
Account research (agent, optional, later)	Weekly	Allow-listed public pages for the target list	Dated trigger events with source and quote	A person, before any outreach	Precision audit of a sample
Inbound first responder (agent, optional, last)	A new inbound lead	The lead event and an approved facts file	A reply and a booked slot	A person reviews every reply	Randomised against the human queue

Five layers hold the work. The rules are written as data, so that a person can read and version them. A deterministic engine applies the rules. A statistical layer adds intervals and false-alarm budgets. A language layer drafts text from the results. A human gate decides. Each layer has one job and its own test. A fault in the language layer cannot change a number, because the layers below have already fixed the numbers.

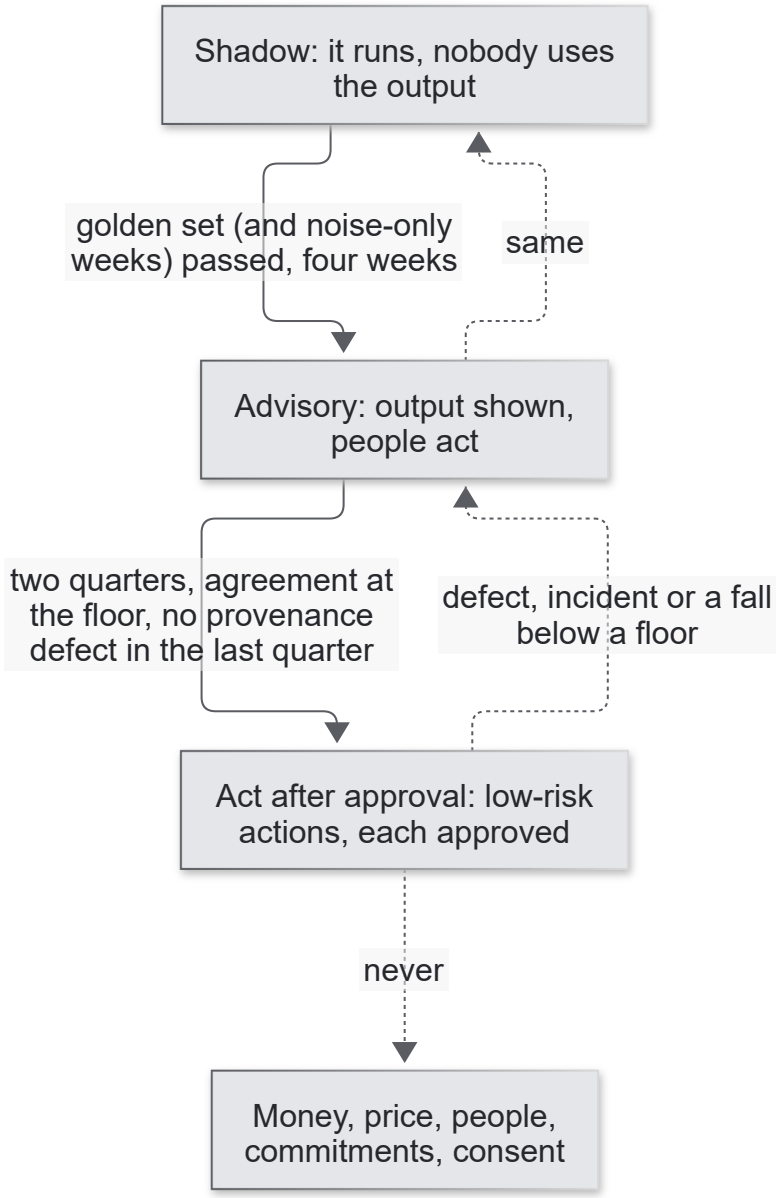
Build in this order: jobs first, then the memo writer and the labeller, then the optional agents. The account researcher reads the open web, so it carries the risk of injected text and of personal data. The first responder combines private data, untrusted input and outbound messages, so it is built last. We do not promote it beyond advisory in this book.

The **weekly memo writer** reads findings and computed tables, never raw data. It writes in a fixed order. First come the gates, the lead finding and the items set aside. Then come the causes with the next check, and one proposal for each stage or "no action". Last come last week's grades and the questions only people can answer. It may say "no action is supported". Its golden weeks are past weeks with a diagnosis that the owners agreed. The **text labeller** assigns a label from a closed list to a note, a ticket or a consented transcript. It must quote the words that support the label, or it must abstain. *Diagnose what changed* showed how to correct a share for the errors of a labeller.

Implementation note: a workflow scheduler, or the task scheduler of the data warehouse (a database built for analysis), runs the jobs. A version-controlled repository holds the rules, the prompts and the golden sets. A ticket queue holds the exception lane.

The autonomy ladder

An agent earns freedom in steps. The **autonomy ladder** has three steps and no fourth.



The autonomy ladder. Promotion needs evidence. Demotion is immediate.

In **shadow**, the agent runs and people compare its output with their own work. Its output informs no decision. It leaves shadow when it passes its golden set, the cases with agreed correct answers of *Learn from outcomes*. A memo writer must also pass the noise-only weeks. And it must have run at least four weeks (Plausible but untested).

In **advisory**, the agent shows its output and people may act on it. It stays here for at least two quarters, which is 26 weeks (Plausible but untested). It then needs agreement with the owners of at least 80%, and no provenance defect in the last quarter.

In **act after approval**, the agent takes low-risk actions only, and a person approves each action. A low-risk action is reversible in one step, stays inside the company, touches no personal data beyond its task and leaves a log entry. Examples are opening a data-fix ticket, drafting a log entry for an owner to sign, and re-running a failed data job. There is no step beyond. An agent never decides on money, price, people, commitments or consent, whatever its record.

Demotion is immediate. A provenance defect, an incident, or a fall below a floor returns the agent to the step below, and the log records why.

Agreement is a weak proof. It measures how often owners accepted the output, and an owner may accept a memo because it reads well. With few memos the rate is also uncertain. So read agreement beside the override reasons and a hand audit of the claims. The floors are gates and not proofs.

A labeller's golden-set test has two floors of its own. **Cohen's kappa** measures agreement between two coders beyond what chance alone would give.

> **Calculation**
>
> kappa = (observed agreement − chance agreement) ÷ (1 − chance agreement)

A kappa of 1 is perfect agreement, and 0 is no better than chance. Landis and Koch (Landis and Koch, 1977, *The measurement of observer agreement for categorical data*, Biometrics 33(1), 159 to 174) call a kappa of 0.61 to 0.80 "substantial". That is a description and not a rule. Our floors are a kappa of at least 0.60 and an **abstain rate** of at most 15%. The abstain rate is the share of items that the labeller leaves to people. Both are start values (Plausible but untested). Re-read both on a fresh golden sample, monthly where the volume allows. A kappa from a few hundred items has noise, so print its n. A value just above a floor is not a clear pass. Kappa also falls when one label is rare, even at high raw agreement, because chance agreement is then high (Arithmetic). So show the raw agreement beside it.

Guardrails for AI

Every agent follows one standard. It has seven parts.

1. **Numbers are tokens.** The model writes a placeholder, and code fills it from the job that computed the value. A **number token** is such a placeholder. For example, the model writes "the win rate fell by `{{win_rate_change}}`" and code fills in the value. A **lint** is an automatic check against rules. This lint rejects any digit in model output that no job produced. The lint fails closed: a doubtful memo does not go out.
2. **Every claim has a source.** The record of where a claim came from is its **provenance**. A sentence without a pointer to a record is dropped. A quote must be exact text from the source. A **provenance defect** is a number or a claim in agent output that no job or source supports.
3. **Untrusted text is data. Prompt injection** is text in the data that tries to give instructions to a model. A note, a ticket or a web page may hold it. The context that reads untrusted text holds no write tool and no private data beyond its task. Its output has a fixed form.
4. **Privacy by design.** Send the least data. Remove names before a call. Process in a region that counsel accepts. Do not allow training on your inputs. Report by role.
5. **Versions are pinned.** Pin the model and the prompt. Every run can be replayed from its input. A change to either goes through the evaluation suite first.
6. **A spend cap and a kill switch.** A **kill switch** stops an agent and its schedule at once. Drill it, so you know that it works.
7. **Evaluation before trust.** Use golden sets, noise-only weeks and a red-team set, which is a set of notes in which a tester has hidden instructions. On noise-only weeks, the share of memos that call for action must not exceed the false-alarm rate of the detector (see *Decide when a change needs action*). To audit

provenance, check a random sample of claims. If 59 claims hold no defect, the defect rate is below 5% with 95% confidence (Standard practice).

Legal and risk register

Counsel must review anything that touches workers, call recordings, data sent to ad platforms, or AI chat with customers before you build it. This is not legal advice. The register in Table 6 names each risk and its control. Review it every quarter. Log counsel's answers with their date.

Table 6. The risk register. Verify every legal row with counsel.

Risk	Where it arises	Control or open question	Owner
Worker monitoring	Seller funnel plots, events per seller, usage records	Aggregate by role. Keep person-level output out of scope. Involve the works council early where one exists. Ask counsel whether the EU AI Act lists the system as high risk.	Bowtie owner
Call recording and transcripts	The text labeller	Consent of all parties. A **data protection impact assessment**, which is a written assessment of the risk to people that the GDPR asks for before high-risk processing. Retention limits. Redaction before the model call.	The owner of the team that is recorded
Data sent to ad platforms	Offline conversion events, audiences	A lawful basis for each field. Hashing is pseudonymisation, not anonymisation (see *Choose acquisition channels*).	Marketing owner
AI chat with customers	The inbound first responder	Tell the person that they talk to an AI. Review every reply.	Sales owner
Linking an analytics id to a person	The identity map	Consented sessions only. Erasure must reach every layer (see *Decide what to measure*).	Analytics lead
Web-sourced personal data	Account research	A notice to the person. No automated collection of members of a social network.	Marketing owner
Prompt injection	Every agent that reads text	No write tool in a context that reads untrusted text. A red-team set.	Analytics lead
Fabricated numbers	The memo writer	Number tokens and a failing-closed lint	Analytics lead
Automation bias	Owners who accept a memo because it reads well	Independent estimates first. Logged override reasons. A hand audit.	Bowtie owner
Cost	The whole layer	A monthly spend cap. People cost for upkeep and labelling dominates.	Finance owner

The **EU AI Act** sets duties for AI systems by their level of risk. Ask counsel which duties apply to each agent. In Austria and Germany the law gives a works council a say in systems that can monitor staff. Ask counsel which systems this covers, and involve the works council before you build any view that could point at one person.

Rolling out

Build the loop as thin vertical slices and not as layers. A **vertical slice** is a thin piece of the system that runs end to end at low fidelity, from data to decision. A layered plan builds all gates, then all rules, then all dashboards. Nothing runs from end to end for months, and nobody learns what the loop needs. A sliced plan runs the whole loop once in the first weeks, then deepens each part.

One trap needs care. A gate that needs data which arrives later cannot report in week one. Show it as "not yet measurable", and do not show it as "pass".

Table 7. The first 90 days. A quarter is 91 days, which is 13 weeks.

Weeks	The slice	What runs	What waits
1 to 2	One stage, end to end	One metric specification. Gates on its source. A hand-run detection. One written brief with a card. One log entry. The roles written down.	Every agent
3 to 6	The nightly spine	Nightly gates, cohort engine and rule engine. The exception lane. The first Monday memo, written by the analytics lead by hand. The first confirms-by dates.	The evaluation suite
7 to 13	The learning side	The evaluation suite with noise-only runs. The memo writer in shadow. The first readouts and grades. The first quarterly refresh. The risk register goes to counsel.	The labeller and the optional agents, until their data steps are done

A company with one analyst can still do this. In the first slice the analyst is the analytics lead. Roles combine, as they do at Alder. A slice is done when a decision made inside it is in the log with a belief and a confirms-by date. Its grade must also be scheduled. Without that date the learning move cannot start.

The analytics lead writes the first memos by hand. This shows what the agent will later draft, and it gives the golden weeks that test the agent. The 90 days end with the first quarterly refresh. The first step up the ladder comes after that, when the evidence exists.

Measuring how the loop runs

The operating system is itself something to measure. Table 8 lists what to count, how, and why. Every measure comes from the time stamps of the decision log, the exception lane and the agent runs. The analytics lead reads the page each quarter, and the bowtie owner uses it to decide what to change.

Table 8. The operating page: how the loop itself runs (start values: Plausible but untested).

Measure	How it is counted	Why	Start limit or floor
Time from finding to decision	Stamp of the finding to stamp of the log entry. Median by role.	Shows where decisions wait	Inside the deadline of the brief
Share of decisions taken asynchronously	Log entries with no meeting ÷ all entries	Shows if the default holds	All entries except escalations
Decisions past their deadline	Cards open beyond the deadline	Each one is a finding	None

Measure	How it is counted	Why	Start limit or floor
Meeting hours per decision	Hours of escalation meetings ÷ decisions they made	Prices the meetings that remain	Reviewed each quarter
Alert acknowledgement time	Gate failure to acknowledgement	Shows if exceptions are seen	The same working day
Agent agreement with owners	Outputs taken as written ÷ outputs (for a memo writer: memos)	Gate for promotion	At least 80%
Provenance defects	Claims in agent output with no source	Gate for promotion	None in the last quarter
Labeller agreement (kappa)	Kappa on a golden sample	Gate for the labeller	At least 0.60
Abstain rate	Items left to people ÷ items	Shows the workload	At most 15%

Do not claim hours saved. Nobody recorded the old meetings, so no baseline exists. Measure the new system against its own start values. Report every measure by role or function.

What changes with volume

Regime	Rhythm of reading	Agents	Meetings	Main risk
Low volume (Alder)	One written note a month. Gates run daily and show only failures.	A labeller on text. The memo writer has little to draft.	Capped by rule. At most one escalation a quarter.	One person holds many roles
Medium volume (Birch)	A weekly memo. Owners decide by deadline.	Memo writer and labeller. Steps follow evidence.	Escalation only	Agreement mistaken for accuracy
High volume (Cedar)	Daily gates and detection on every rate. A weekly memo.	Many jobs. A labeller on thousands of items.	Escalation only. The yield test retires the rest.	Scale of text and legal review

Alder, Year 3 April: a monthly rhythm, and a labeller for deal notes

Alder, Year 3 April: about 4.1 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year. The win rate needs about 2.7 years to read even a 30% change, and the lead count reads a 20% change in 23 days.

The chief executive holds the bowtie-owner role. Alder has no separate growth owner, so the chief executive holds that role too. Five other people fill the remaining roles. They are the marketer, the head of sales, the head of customer success, the head of finance and the analyst, who works half time. In Year 3 April the chief executive writes the rhythm down.

The first question is how fast to read. Alder's trend is about 19.5 opportunities a month, which is about 4.5 a week. By chance alone, a count of 4.5 has a standard error of 47% of its size. A count of 19.5 has 23%. A weekly memo would show a swing that no owner could use. So the rhythm is monthly. The nightly gates run, and the analyst reads failures only. Once a month the analyst sends one written note. It holds the counts of engaged accounts and opportunities, the signposts that fall due and the log entries that fall due. Counts carry their ranges. Rates that the volume cannot read do not appear.

Table 9. The layout of Alder's monthly note (Illustration only).

Block	What it holds
Counts	Engaged accounts and opportunities for the window, each with its range
Signposts due	Each signpost that falls due this month, with its planned response
Log entries due	Entries whose confirms-by date falls this month, with the grade when the readout is ready
Decisions open	Each card with its owner and its deadline
Questions only people can answer	A short list

Each owner replies in writing within five working days, with a decision or "no change". The chief executive writes the bowtie note from the replies and the executive page. Many of Alder's decisions sit in the right wing, where renewals and expansions come as cases and not as rates (see *Expand accounts and generate referrals*).

The log holds 18 entries, and 12 are resolved. The register holds 6.3 well-recorded tests, far short of 30, so Alder keeps outside base rates as priors. The chief executive sets one more rule: at most one escalation meeting a quarter. The head of finance is consulted in writing on every commitment of money. The analyst computes every card and decides no option. These two guards matter because one person holds the bowtie role and the growth role.

In the same month the deal-note labeller reaches advisory. Two steps came first: the consent step, with counsel's review of notes that name customer contacts and sellers, and the works-council step. The scope that came out of them is labels by theme and stage, never by seller. Personal names are removed before the model reads a note. On a golden set of 200 notes the labeller abstains on 10%. On the 180 notes that it labels, it agrees with a person at a kappa of 0.70. That clears both floors of the golden-set test, a kappa of at least 0.60 and an abstain rate of at most 15%. Landis and Koch would call 0.70 "substantial", and that remains a description.

The labeller reads text and computes nothing. With few deals it saves reading time but yields few labelled items. Shares built on small counts keep their intervals, as *Diagnose what changed* showed.

The head of sales owns the work that the labeller touches, so the head of sales decides the promotion. The analyst supplies the evidence. The log entry asks whether the labeller stays at or above both floors on a fresh golden set at the next quarterly refresh. The falsifier is a kappa below 0.60 or an abstain rate above 15%. The planned response is a return to shadow. The lesson at low volume is that the rhythm is monthly and agents help with text and not with numbers.

Birch, Year 3 April: the memo agent is promoted

Birch, Year 3 April: about 65 new customers a window, 0.8 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate. The win rate reads a 30% change in 54 days, and the close rate reads a 20% change in about 5 months.

Birch has run without a standing council since Year 2 July. Owners decide on cards by their deadlines. The weekly memo agent started in shadow in that month and moved to advisory in Year 2 September. By Year 3 April it has written about 30 weekly memos in advisory. That is 30.3 weeks, more than the 26 weeks that two quarters require.

The growth owner holds the bowtie role and owns the rhythm, so the growth owner decides whether the agent moves up. The analytics lead brings the evidence that the ladder asks for.

- Time in advisory: 30 memos over 30.3 weeks.
- Agreement: the owners took the decision proposed in 26 of the 30 memos as written, which is 87%. A memo counts only when the owners took all its proposals without change. The floor is 80%.
- Provenance: no defect in the last quarter. The lint and the hand audit found no number that a job did not produce.

The log now holds 60 resolved predictions, so the memo can also report the pooled calibration of *Learn from outcomes*. That is context for the owners, not a gate of the ladder.

Read the 87% with care. It is 26 of 30. The 95% Wilson interval (see *Learn from outcomes*) runs from 70% to 95%, so the true agreement could sit below the floor. Agreement is also not accuracy. So what protects Birch? The next step is small. The agent may open a data-fix ticket, draft a log entry for an owner to sign, or re-run a failed data job. A person approves each one. The step is reversible. A single provenance defect returns the agent to advisory the same day. And the analytics lead reads the four memos that the owners did not take as written. The lead looks for a pattern in the override reasons, by type of proposal and never by person.

In the meantime the loop ran without a meeting. At the end of Year 3 February the month-end backlog stood at 103.6 uncontacted SQLs. The hiring rule of *Commit with uncertainty* tripped for the second time. Birch had rewritten its trigger in Year 2 December at 100 SQLs. The sales owner approved a twelfth seller within the week and logged the decision. No meeting was held. The rule paused until the new seller works at full capacity. The twelfth seller starts this month, ramps at half capacity until June and works at full capacity from Year 3 July. The backlog is 93.2 at the end of April. *Set targets and staff capacity* tells the story. Here it counts as evidence for the default path. The rule, the owner and the deadline were written before the backlog crossed.

The growth owner promotes the agent and writes the entry. The question is whether the agent keeps an agreement of at least 80% and shows no provenance defect in the next quarter. Either failure is the falsifier. The planned response is a return to advisory, and the confirms-by date is the next quarterly refresh. The agent never acts on money, price, people, commitments or consent. The lesson at medium volume is that a weekly written memo replaces the reading phase and most meetings. Trust grows one step at a time, and each step needs evidence.

Cedar, Year 3 May to June: many automations, one meeting

Cedar, Year 3 May: about 2,000 new paying accounts a window, 1.2 times the derived typical; high volume: every self-serve rate reads a 20% change within a window. Trial to paid reads a 10% change in 21 days; the assisted motion, with 75 opportunities a window, needs about 7 months to read a 30% change.

Cedar's register holds 152 tests and its log holds 126 entries. Cedar runs the most automation of the three, so the question is not whether the jobs run. It is whether anyone can trust them at this scale. Take the support-ticket labeller. It has run in advisory since Year 2 July, which is ten months.

Cedar has about 50,100 paying accounts at the start of Year 3 May. At 0.25 tickets for each paying account each month, that is about 12,500 tickets a month. Nobody can check them all. So each month the analytics lead labels a golden sample of 300 tickets by hand, which is 2.4% of the flow. The lead then compares the

sample with the labeller. This month the labeller abstains on 8% of the sample. On the 276 tickets that it labels, the kappa is 0.78. Both clear the floors. At this scale the abstain rate is a workload. Eight per cent is about 1,000 tickets a month for the support team. At the ceiling of 15% it would be about 1,900. The labels feed the loss-reason and save-play analyses, with the correction and the intervals of *Diagnose what changed*.

In Year 3 May and June Cedar also reviews its risk register with counsel. Ticket text holds personal data, so it is redacted before any model call. Support staff are workers, so labels go by theme and never by the person who handled the ticket. Conversion data sent to ad platforms is an open row for counsel. The row for AI chat with customers stays closed. No agent talks to customers, and none will before counsel has reviewed the plan. Counsel's answers go into the log with their date. The yield test then retires the standing meetings that changed no decision. Only the escalation path stays.

One meeting remains, and it is a real one. The growth lead proposes a standing commitment: assisted sellers may offer up to the annual-plan discount on any annual assisted contract. That discount is 25%. Assisted deals now average 8.57% off list. The chief financial officer objects. The break-even discount for the annual plan is 28.1% on the real hazard (see *Model recurring growth*). That leaves 3.1 points of room, but the figure comes from the hazard of self-serve monthly customers in their first year. It says nothing about assisted accounts. A standing commitment also binds the price of every contract signed under it.

The case meets both tests. The commitment crosses growth and finance, and no single owner can make it. The assisted motion is a low-volume business inside a high-volume company. Its win rate needs about seven months to read a 30% change, so no test can settle the question in time. Each owner writes an independent estimate of the share of assisted deals that would take the full discount. The brief shows both estimates, the card and the do-nothing option.

The meeting is named "whether assisted sellers may commit to the annual-plan discount". The chief executive is the one decision owner. In this illustration the chief executive declines the standing commitment and keeps the case-by-case path. Any assisted discount above the current average now needs the written approval of the chief financial officer inside the five-working-day deadline. The signpost is the average assisted discount, read each month against 8.57%. The log entry holds it, and the meeting ends when the entry is written. The lesson at high volume is that automation needs evaluation discipline and legal review, and that meetings shrink to the true conflicts.

The decision to carry forward

Give each recurring choice an owner, a decision date, an evidence record and a later read. Use automation to maintain that record and surface exceptions. Keep the economic objective and the authority to commit explicit.

Glossary

This glossary defines each technical term of the book in plain words, with the formula where one exists. Every entry names the chapter that develops the term.

The funnel and the bowtie

Funnel

The stages a stranger passes through to become a customer. Counts at different stages have different units.

First explained in: Choose the growth objective.

Silo

Work in separate teams with separate metrics, tools and definitions. Each team can do well by its own measure while the whole company loses.

First explained in: Choose the growth objective.

Bowtie

Also called: revenue bowtie.

One map of the whole revenue system. The left wing is awareness, education and selection; the knot is the commit; the right wing is onboarding, adoption and impact, renewal and expansion.

First explained in: Choose the growth objective.

Awareness

First left-wing stage: a buyer learns that your company may solve a problem (top of the funnel).

First explained in: Choose the growth objective.

Education

Second left-wing stage: the buyer learns how to solve the problem (middle of the funnel).

First explained in: Choose the growth objective.

Selection

Third left-wing stage: the buyer compares options and picks one (bottom of the funnel).

First explained in: Choose the growth objective.

Commit

Also called: mutual commit.

The knot of the bowtie. The company owes a result by a date; the customer owes time, data and a named person.

First explained in: Choose the growth objective.

Onboarding

The work between signature and a working set-up. It is the first right-wing stage.

First explained in: Choose the growth objective.

Adoption

The stage where the product becomes part of the customer's routine.

First explained in: Choose the growth objective.

Impact

The measurable result that the customer gets, and pays for at each renewal.

First explained in: Choose the growth objective.

Renewal

A customer's decision to extend a contract for another term.

First explained in: Choose the growth objective.

Expansion

Growth in the recurring revenue of a customer that stays.

First explained in: Expand accounts and generate referrals.

Top of the funnel (TOFU)

The stages that create awareness and attract visitors and leads.

First explained in: Choose the growth objective.

Middle of the funnel (MOFU)

The stages where leads show intent and are qualified.

First explained in: Choose the growth objective.

Bottom of the funnel (BOFU)

The stages where sales qualifies, proposes and closes.

First explained in: Choose the growth objective.

Pirate metrics (AAARRR)

Also called: AAARRR, AARRR.

Six stages of a customer's path:

- Awareness: accounts learn the company exists.
- Acquisition: a stranger becomes a lead.
- Revenue: the deal closes and is billed.
- Activation: the customer reaches first value.
- Retention: use continues and the customer renews.
- Referral: customers bring new customers.

First explained in: Choose the growth objective.

Lead

A person or company with contact details, or identified as a prospect, not yet judged worth a sales touch.

First explained in: Choose the growth objective.

Marketing-qualified lead (MQL)

A lead that meets marketing's written definition of worth a sales touch.

First explained in: Choose the growth objective.

Sales-accepted lead (SAL)

An MQL that sales accepts to work under the hand-off contract. Some companies skip this stage or use the term for the SQL. Birch puts acceptance after the SQL, as its own choice.

First explained in: Choose the growth objective.

Sales-qualified lead (SQL)

A lead that sales has contacted and judged to have a real problem and a fit. It is counted once, on the date of first qualification.

First explained in: Choose the growth objective.

Opportunity

Also called: deal.

A potential sale with an expected value and a stage in the sales process.

First explained in: Choose the growth objective.

Closed-won and closed-lost

An opportunity that ended in a signed contract (won) or with no sale (lost).

First explained in: Choose the growth objective.

Customer relationship management system (CRM)

Also called: CRM.

The system where sales records leads, opportunities and deals.

First explained in: Choose the growth objective.

Activation

The point where a new customer first gets value, defined as one visible event.

First explained in: Activate customers.

Ideal customer profile (ICP)

The written description of the companies that gain most from the product.

First explained in: Choose the growth objective.

Go-to-market

How a company reaches, wins and keeps customers.

First explained in: Choose the growth objective.

Loop

A path by which the result of an action returns to influence the next action.

First explained in: Choose the growth objective.

Open loop

A loop in which nobody compares the output with the intention.

First explained in: Choose the growth objective.

Closed loop

Also called: feedback loop.

A loop that measures the output, compares it with the intention and changes the input.

First explained in: Choose the growth objective.

Correcting loop

A closed loop that shrinks a gap between output and intention.

First explained in: Choose the growth objective.

Compounding loop

A loop that feeds its own growth; g is the extra customers per customer, below 1.

Formula: total customers = first round ÷ (1 - g).

First explained in: Choose the growth objective.

Loop delay

The time from an action until its result can be read honestly.

First explained in: Choose the growth objective.

Data-informed and data-centric

Data-informed: people decide, with data as one input. Data-centric: stage events are dated and code computes the numbers.

First explained in: Choose the growth objective.

Top-down, bottom-up and hybrid method

Top-down splits a total into parts; bottom-up adds up units; hybrid compares both.

First explained in: Choose the growth objective.

Window (28-day)

The 28-day period for every comparison, against the 28 days before.

First explained in: Choose the growth objective.

Display grain

The time step of a chart or table, never a second comparison window.

First explained in: Choose the growth objective.

Volume regime

A label for company volume, read on the last stage rate before revenue in its main motion. High is 28 days or fewer to read a 30% change, Medium up to 91 days, Low more. The 30% and the limits are start values (Plausible but untested).

Formula: days = 28 × n required ÷ n per window, with n required = 16 × (1 - p) ÷ (p × 0.30 × 0.30).

First explained in: Choose the growth objective. Full method: Decide what the evidence can resolve.

Barometer

One sentence of at most 30 words: the company, the date, new customers per window against the derived typical, the volume regime, and the deepest stage rate that reads a 20% change in the period that fits the regime.

First explained in: Choose the growth objective. Full method: Decide what the evidence can resolve.

Derived typical

New customers per 28-day window that a company of a given ARR and contract size would win at published medians; an order of magnitude only.

Formula: derived typical new customers per window = 0.234 × ARR ÷ (13 × yearly contract value).

First explained in: Choose the growth objective.

Evidence label

A tag on a claim that is not arithmetic, one of five:

- Arithmetic: follows from the numbers.
- Standard practice: many use it.
- Plausible but untested: not tested on your data.
- Assertion: stated without support.
- Illustration only: a made-up number.

First explained in: Choose the growth objective.

Stage and entry criterion

A stage is one funnel step. Its entry criterion is the timestamped event that moves an item in.

First explained in: Choose markets and accounts.

Volume, rate and time (the three metric families)

A volume is a count or euro amount; a rate is the share reaching the next stage; a time is the wait between timestamps.

First explained in: Choose markets and accounts.

Conversion rate

The share of items at one stage that reach the next.

Formula: conversion rate = count that reach the next stage ÷ count that entered the stage.

First explained in: Choose markets and accounts.

Step rate and cumulative rate

A step rate compares neighbouring stages; a cumulative rate compares a later stage with an earlier one.

Formula: cumulative rate = step rate 1 × step rate 2 × … × step rate n.

First explained in: Choose markets and accounts.

MQL rate, SAL rate and SQL rate

Shares of leads that become MQLs, of MQLs that become SQLs, and of SQLs that sales accepts.

Formula: MQL rate = MQLs ÷ leads; SQL rate = SQLs ÷ MQLs; SAL rate = sales-accepted leads ÷ SQLs.

First explained in: Choose markets and accounts.

Lead volume

The count of leads created in a window.

First explained in: Model recurring growth.

Impression

One showing of an ad. Impressions are not people.

First explained in: Model recurring growth.

Session

One visit to the website.

First explained in: Choose markets and accounts.

Click-through rate (CTR)

The share of ad impressions that produce a click.

Formula: CTR = clicks ÷ impressions.

First explained in: Model recurring growth.

Click-to-lead rate

The share of clicks that become leads.

Formula: click-to-lead rate = leads ÷ clicks.

First explained in: Choose markets and accounts.

Click-to-session ratio

The share of ad-platform clicks that appear as sessions in analytics. It is a tracking ratio, not a lever.

Formula: click-to-session ratio = sessions in analytics ÷ clicks in the ad platform.

First explained in: Choose markets and accounts.

Opportunity rate

The share of sales-qualified leads that become opportunities.

Formula: opportunity rate = opportunities ÷ SQLs.

First explained in: Model recurring growth.

Win rate

The share of decided opportunities that end closed-won.

Formula: win rate = closed-won ÷ (closed-won + closed-lost).

First explained in: Choose the growth objective. Full treatment: Choose markets and accounts.

Close rate

The share of sales-qualified leads that become customers.

Formula: close rate = opportunity rate × win rate.

First explained in: Choose the growth objective. Full treatment: Choose markets and accounts.

Product of rates

Customers equal the top volume times every rate in the chain.

Formula: customers = volume × rate 1 × rate 2 × … × rate n.

First explained in: Choose markets and accounts.

Improvable rate

A step rate between two counted stages that a team can change by its work.

First explained in: Model recurring growth.

Attrition

The loss of items at each step of a sequence, with a share u lost per step.

Formula: share left = (1 - u)^n.

First explained in: Model recurring growth.

Interaction of improvements

Gains at different rates multiply; they do not add.

Formula: combined gain = (1 + a) × (1 + b) - 1.

First explained in: Choose markets and accounts.

Motion

How a company sells to and serves a customer group: self-serve, sales-assisted or sales-led.

First explained in: Choose markets and accounts.

Touch level

The amount of human contact in selling and serving: none, low, medium or high.

First explained in: Choose markets and accounts.

Self-serve motion

The customer buys and starts alone.

First explained in: Choose markets and accounts.

Sales-assisted motion

A person joins a product-led or inbound path to qualify, demonstrate or close.

First explained in: Choose markets and accounts.

Product-led growth (PLG)

Growth in which the product brings in, converts and expands customers; users try it before they talk to a person.

First explained in: Choose markets and accounts.

Sales-led growth

Growth in which the customer meets a person before the product.

First explained in: Choose markets and accounts.

Product-led sales

Usage tells sales which accounts to contact; a person joins after usage starts.

First explained in: Choose markets and accounts.

Revenue line

One product sold through one go-to-market motion, with its own rates.

First explained in: Model recurring growth. Full treatment: Choose markets and accounts.

Buying group

The people at one customer who take part in a purchase decision.

First explained in: Choose markets and accounts.

Buying-group coverage

The share of expected buying-group roles with an engaged contact; a candidate leading signal.

Formula: buying-group coverage = roles with an engaged contact ÷ expected roles.

First explained in: Choose markets and accounts.

Account-based marketing (ABM)

Marketing and sales that treat chosen target accounts, not leads, as the unit.

First explained in: Choose markets and accounts.

Engaged account

An account that meets a written engagement rule, for example two or more of its people acting in a window.

First explained in: Choose markets and accounts.

Unit of the funnel

The thing counted as it moves through the stages: a lead, contact, account or workspace.

First explained in: Choose markets and accounts.

Contact

One person in the records, belonging to one account.

First explained in: Choose markets and accounts.

Account

One company in the records.

First explained in: Choose markets and accounts.

Workspace

A space in the product where a team of users works; the unit of a self-serve funnel.

First explained in: Choose markets and accounts.

Subscription (record)

The billing record of a recurring contract.

First explained in: Choose markets and accounts.

Product-qualified lead and account (PQL, PQA)

A user or company whose product use signals buying intent.

First explained in: Choose markets and accounts.

Product-qualified account rate

The share of active accounts that reach the defined buying-intent level of product use.

Formula: product-qualified account rate = product-qualified accounts ÷ active accounts in the same period.

First explained in: Choose markets and accounts.

Sign-up

The moment someone creates an account, on a free plan or a trial.

First explained in: Choose markets and accounts.

Trial-to-paid rate

The share of trial starts that later pay, for one cohort.

Formula: trial-to-paid rate = paying accounts ÷ trial starts, for the same cohort.

First explained in: Choose markets and accounts.

Average contract value (ACV)

The yearly recurring value of one contract, without one-off fees.

Formula: ACV = total yearly recurring value won ÷ number of contracts won.

First explained in: Choose markets and accounts.

Pipeline

The set of open opportunities: volume is their count, value is the sum of their expected contract values.

First explained in: Choose markets and accounts.

Recurring revenue

Annual recurring revenue (ARR)

The yearly value of all active recurring contracts on one date.

Formula: ARR = sum of the annualised recurring value of all active contracts.

First explained in: Choose the growth objective. Full treatment: Model recurring growth.

Monthly recurring revenue (MRR)

ARR expressed per month.

Formula: MRR = ARR ÷ 12.

First explained in: Model recurring growth.

Subscription pricing

The buyer pays to use the product for a term, never to own it.

First explained in: Model recurring growth.

Seat-based pricing

A subscription price that grows with named users.

First explained in: Model recurring growth.

Usage-based pricing

A price that grows with use. Usage above the committed minimum is re-occurring revenue.

First explained in: Model recurring growth.

Hybrid pricing

A fixed platform fee plus a variable part. Only the fixed fee and a committed minimum recur.

First explained in: Model recurring growth.

Billing term

How often the customer pays.

First explained in: Model recurring growth.

Committed minimum

The fixed amount a usage-based contract obliges the customer to pay. It is recurring revenue.

First explained in: Model recurring growth.

Billed in advance

The customer pays for the whole term at the start.

First explained in: Choose the growth objective. Full treatment: Model recurring growth.

Billing line

One product on one contract, the unit that the recurring-revenue test classifies.

First explained in: Model recurring growth.

Recurring revenue and re-occurring revenue

Recurring: the same customer pays again, access stops when payment stops, and the contract fixes date and amount. Re-occurring repeats without such a contract.

First explained in: Model recurring growth.

Recurring-revenue share

The share of total revenue that is recurring. Read changes against 3 percentage points as a start value, then 2 standard deviations of your own change.

Formula: recurring-revenue share = recurring revenue ÷ total revenue.

First explained in: Model recurring growth.

New ARR

ARR from customers that were not customers before.

First explained in: Model recurring growth.

Returning ARR

ARR from a former customer that comes back.

First explained in: Model recurring growth.

Expansion ARR

Extra ARR from customers that stay.

First explained in: Model recurring growth.

Contraction ARR

ARR lost from customers that stay but downgrade.

First explained in: Model recurring growth.

Churned ARR

ARR lost from customers that leave.

First explained in: Model recurring growth.

Net new ARR

The change in ARR over a period.

Formula: net new ARR = ending ARR - starting ARR = new + returning + expansion - contraction - churned.

First explained in: Choose the growth objective. Full treatment: Model recurring growth.

ARR bridge

Carries starting ARR to ending ARR through the five movements and must close to the euro.

Formula: ending ARR = starting ARR + new + returning + expansion - contraction - churned.

First explained in: Model recurring growth.

Logo

One customer, counted as one unit.

First explained in: Model recurring growth.

Logo retention

The share of starting customers still customers at the end.

Formula: logo retention = customers at the end that were present at the start ÷ customers at the start.

First explained in: Model recurring growth.

Gross revenue retention (GRR)

The share of one cohort's starting ARR still paid at the end, counting losses only.

Formula: GRR = (starting ARR - churned ARR - contraction ARR) ÷ starting ARR.

First explained in: Choose the growth objective. Full treatment: Model recurring growth.

Net revenue retention (NRR)

One cohort's ending ARR as a share of its starting ARR, so it can exceed 100%.

Formula: NRR = GRR + expansion ARR ÷ starting ARR.

First explained in: Model recurring growth.

Renewal rate

The share of contracts due in a period that renewed.

Formula: renewal rate = contracts renewed ÷ contracts due.

First explained in: Model recurring growth.

Renewal number

The count of a renewal decision in a contract's life. Renewal 1 is the first decision.

First explained in: Model recurring growth.

Renewal calendar

Every contract with its renewal date, expected outcome and named owner.

First explained in: Model recurring growth.

Churn

The loss of customers or ARR by non-renewal, downgrade or non-payment.

First explained in: Choose the growth objective. Full treatment: Model recurring growth.

Leaky bucket

The loss of ARR at renewal, which new ARR must replace.

Formula: loss share of new ARR = (churned ARR + contraction ARR) ÷ (new ARR + returning ARR).

First explained in: Model recurring growth.

Survival curve

The share of a customer group still active after each renewal.

Formula: S(k) = share still customers after renewal k.

First explained in: Model recurring growth.

Hazard

The share of customers at risk at one renewal that leave at it. Deals use the same ratio per age step.

Formula: hazard(k) = customers lost at renewal k ÷ customers at risk at renewal k. For any step, hazard = items that end in the step ÷ items at risk at the start of the step.

First explained in: Model recurring growth.

Censoring

Observation ends before the event of interest is observed. Administrative censoring occurs at the reporting cut-off. An event that prevents the target event is a competing outcome, not automatically independent censoring.

First explained in: Model recurring growth.

Product-limit estimate (Kaplan-Meier)

Also called: Kaplan-Meier estimate.

The standard survival curve when some cases are censored.

Formula: S(k) = (1 - hazard 1) × (1 - hazard 2) × … × (1 - hazard k).

First explained in: Model recurring growth.

Half-width

The distance from an estimate to either end of its interval.

Formula: half-width of a rate = 1.96 × square root of (p × (1 - p) ÷ n), for a 95% interval.

First explained in: Model recurring growth.

Growth streams

The three sources of recurring-revenue growth: new customers, expansion and renewal.

First explained in: Model recurring growth.

ARR concentration

The share of ARR in the largest accounts.

Formula: ARR concentration = ARR of the ten largest accounts ÷ total ARR.

First explained in: Choose markets and accounts.

Unit economics and value

Gross margin

The share of revenue left after the cost of revenue.

Formula: gross margin = (revenue - cost of revenue) ÷ revenue.

First explained in: Value customers and cash.

Gross profit

The euro left after the cost of revenue.

Formula: gross profit = revenue - cost of revenue.

First explained in: Choose the growth objective. Full treatment: Value customers and cash.

Cost of revenue

The cost of delivering the product: hosting, third-party software, support, onboarding and delivery. Renewal and upsell work are not in it.

First explained in: Value customers and cash.

Cost to serve

The cost of keeping and growing a customer that is not in the cost of revenue, such as renewal work.

First explained in: Value customers and cash.

Customer acquisition cost (CAC)

The cost of winning one customer. Cost and customers must come from the same population.

Formula: CAC = acquisition cost ÷ customers won from the same population.

First explained in: Value customers and cash.

Paid CAC

Paid media spend per customer won from paid sources. Never call it CAC alone.

Formula: paid CAC = paid media spend ÷ customers won from paid sources.

First explained in: Value customers and cash.

Fully loaded CAC

Paid media plus the people, tools and overhead of marketing and sales, per new customer.

Formula: fully loaded CAC = (paid spend + marketing and sales people, tools and overhead) ÷ all new customers.

First explained in: Value customers and cash.

New-customer CAC ratio

The cost of each euro of new ARR, using the new-customer share of sales and marketing cost. Published median: 1.76 in US dollars (Benchmarkit and Pavilion, 2024).

Formula: new-customer CAC ratio = new-customer sales and marketing cost ÷ new ARR from new customers. Accrual payback in months = 12 × ratio ÷ gross margin.

First explained in: Value customers and cash.

Cohort paid CAC

The paid spend of one acquisition cohort per customer it produced.

First explained in: Value customers and cash.

Population rule

A CAC takes its cost and its customers from the same population.

First explained in: Value customers and cash.

Lifetime value (LTV)

Also called: customer lifetime value (CLV).

The expected gross profit of one customer over the horizon, discounted, using survival by renewal.

Formula: LTV = sum over years k = 0 to H − 1 of [S(k) × GP(k) ÷ (1 + r)^k], with S(k) the share of customers still active in contract year k (S(0) = 1), GP(k) the gross profit of an active customer in year k, r the discount rate and H the horizon in years.

First explained in: Value customers and cash.

Net expansion

Expansion less contraction of customers who stay.

First explained in: Value customers and cash.

Discount rate

The yearly rate at which a euro received later counts for less. The researched default is 12%, with a range of 8% to 20%.

First explained in: Choose the growth objective. Full treatment: Value customers and cash.

Horizon

The years of customer life that the value counts. The default is five years.

First explained in: Choose the growth objective. Full treatment: Value customers and cash.

Value basis

Expected gross profit over the horizon, discounted at the company's rate. The rate and the horizon are parameters that each company sets.

First explained in: Choose the growth objective. Full treatment: Value customers and cash.

Expected value

The average outcome of an uncertain quantity, weighted by probability.

Formula: expected value = sum of (each outcome × its probability).

First explained in: Value customers and cash.

Present value

What a future amount is worth today.

Formula: present value = future amount ÷ (1 + r)^k, for an amount received in year k.

First explained in: Value customers and cash.

Value ladder

The euro value of one unit at each stage, from customer to sign-up.

Formula: value of a unit = probability it becomes a customer × value of one customer.

First explained in: Value customers and cash.

Accrual payback and cash payback

Accrual payback counts gross profit as earned. Cash payback counts it when each invoice is billed.

Formula: accrual payback in months = fully loaded CAC ÷ (contract value × gross margin ÷ 12).

First explained in: Value customers and cash.

Payback period

The time the gross profit of a customer needs to repay its fully loaded CAC.

Formula: accrual payback in months = fully loaded CAC ÷ (contract value × gross margin ÷ 12).

First explained in: Value customers and cash.

Payback ceiling

The longest payback a segment accepts. Start values: 12 months for small business, 18 for mid-market, 24 for enterprise (Standard practice for the ceilings).

First explained in: Value customers and cash.

LTV to CAC ratio

Lifetime value divided by CAC for the same customer. Always name the CAC.

Formula: LTV to CAC = LTV ÷ CAC.

First explained in: Value customers and cash.

SaaS magic number

A quick check of recurring-revenue growth against the previous quarter's sales and marketing spend. It is Standard practice with limits.

Formula: magic number = (recurring revenue this quarter - recurring revenue last quarter) × 4 ÷ sales and marketing spend last quarter.

First explained in: Value customers and cash.

Parameter and start value

A parameter is an input each company sets. A start value is the first value you use, labelled "Plausible but untested".

First explained in: Value customers and cash.

Distribution

The values a quantity can take, with the probability of each.

First explained in: Value customers and cash.

Flaw of averages

Average inputs return the average case, not the average output.

First explained in: Value customers and cash.

Monte Carlo simulation

Drawing many random input sets from their distributions to show the spread of the output.

First explained in: Value customers and cash.

Seed

The starting number of a random-number generator. A fixed seed makes a run repeat.

First explained in: Value customers and cash.

P10, P50 and P90

The values below which 10%, 50% and 90% of simulated outcomes fall.

First explained in: Value customers and cash.

Switching value

The amount one input must move, others held at their means, before the conclusion changes.

First explained in: Value customers and cash.

Time and cohorts

Cohort

A group of items that entered at the same time or share an origin.

First explained in: Choose the growth objective. Full treatment: Forecast conversion and timing.

Cohort table

Also called: cohort-age table.

One row for each cohort, one column for each age.

First explained in: Forecast conversion and timing.

Partial cell

A cohort-table cell where some members are too young to judge.

First explained in: Forecast conversion and timing.

Cohort age

The time since a cohort entered.

First explained in: Forecast conversion and timing.

Same-period rate

A rate counted within one calendar period.

First explained in: Choose markets and accounts. Full treatment: Forecast conversion and timing.

Cohort rate

A rate for one cohort at a stated age.

First explained in: Choose markets and accounts. Full treatment: Forecast conversion and timing.

Cohort conversion at age

The share of a cohort that reached the outcome by a stated age.

Formula: cohort conversion at age a = items of the cohort that reached the outcome within a days ÷ items of the cohort that are at least a days old.

First explained in: Forecast conversion and timing.

Maturity

The state when most of a cohort's outcomes have happened, set at the 90th percentile of the cycle (Plausible but untested).

First explained in: Forecast conversion and timing.

Age, period and cohort effects

An age effect shapes every row. A period effect shows along a diagonal. A cohort effect shows as one row. A real change shows down a column.

First explained in: Forecast conversion and timing.

Time to event

The time from entry to an outcome, such as a deal won or a customer lost.

First explained in: Forecast conversion and timing.

At risk

Items still open at the start of an age step and watched through it.

First explained in: Forecast conversion and timing.

Cumulative conversion

The share of a cohort that reached an outcome by a given age, built from step hazards.

Formula: cumulative conversion at age a = sum over the steps up to a of (share still open at the start of the step × hazard of the outcome in the step).

First explained in: Forecast conversion and timing.

Competing outcomes

Terminal outcomes that exclude each other on a defined journey, such as won and lost, or activation and departure before activation. Still open is the state before either terminal outcome.

First explained in: Forecast conversion and timing.

Projected eventual conversion

An estimate of the final rate of a cohort not yet mature.

Formula: projected eventual conversion = cohort conversion at age a ÷ share of final outcomes in by age a (from complete cohorts).

First explained in: Forecast conversion and timing.

Cycle percentile

The time within which a given share of deals finish. Set deadlines on P80, not the median.

First explained in: Forecast conversion and timing.

Time in stage

The days a deal spends in one stage.

First explained in: Forecast conversion and timing.

Deal age

The days since an opportunity was opened.

First explained in: Forecast conversion and timing.

Stalled deal

An open deal older than P90 of the cycle.

First explained in: Forecast conversion and timing.

Open pipeline

The deals still open, grouped by age band and stage.

First explained in: Forecast conversion and timing.

Conversion by age

The chance that an open deal of a given age and stage wins within a stated horizon.

First explained in: Forecast conversion and timing.

Snapshot and as-of view

A snapshot is a stored copy of the data on one date. An as-of view rebuilds a past date.

First explained in: Forecast conversion and timing.

Bookings forecast

A forecast of new contracts from the open pipeline, reported as a range.

Formula: expected wins = wins booked + sum over open deals of (conversion by age) + expected wins from deals not yet opened.

First explained in: Forecast conversion and timing.

Little's law

Average open items equal the arrival rate times the average time open.

Formula: open items = arrivals per day × average days open.

First explained in: Forecast conversion and timing.

Growth and capacity

Growth rate

The relative change in a quantity over time.

Formula: year-on-year growth = ARR now ÷ ARR a year ago - 1; compound monthly growth = (end ÷ start)^(1 ÷ months) - 1.

First explained in: Set targets and staff capacity.

Growth stage

A phase of company life: start-up, scale-up or mature. The ARR bands are conventions (Assertion).

First explained in: Model recurring growth.

S-curve and ceiling

Growth that starts slowly, speeds up and levels off at a ceiling. Never fit the ceiling to two points.

Formula: ARR(t) = K ÷ (1 + e^(-r × (t - t∅))); K is the ceiling.

First explained in: Set targets and staff capacity.

Attainable share

The share of reachable accounts that can realistically pay. It is a judgement, written with a range.

Formula: ceiling = reachable accounts × attainable share.

First explained in: Set targets and staff capacity.

Cohort-based forecast

A forecast of ARR built from units, with many random runs that give a distribution.

Formula: ARR at the end = ARR now - churned - contraction + expansion + new.

First explained in: Set targets and staff capacity.

Dispersion

The extra spread of a count beyond pure chance, written ϕ. If ϕ is 1, the variance equals the mean. Estimate it from at least 13 windows; until then show forecasts at 1 and at 2. A detector needs one value, so *Detection* starts at 1.2, inside that bracket (Plausible but untested).

Formula: ϕ = average over windows of $(\text{actual} - \text{expected})^2 \div \text{expected}$.

First explained in: Set targets and staff capacity.

Seasonality

Regular, calendar-driven movement in the numbers.

Formula: seasonality index = usual level of the month ÷ average level.

First explained in: Set targets and staff capacity.

Prediction interval

The range in which a new value falls with a stated probability.

Formula: count n, relative half-width $= z \times \sqrt{(\phi \times (1 + 1 \div k) \div n)}$; rate p over n trials, relative half-width $= z \times \sqrt{(\phi \times (1 - p) \times (1 + 1 \div k) \div (p \times n))}$; k is the number of baseline windows.

First explained in: Set targets and staff capacity.

Probability of reaching a target

The share of simulated years that end at or above the target level.

First explained in: Set targets and staff capacity.

Probabilistic target (stretch and commit)

A target stated as a probability. Stretch is the median outcome (50%). Commit is reached with 80% probability, so it is the 20th percentile.

First explained in: Set targets and staff capacity.

Capacity

The amount of work that a team can do in a period.

Formula: heads = required volume ÷ (productivity per head at full ramp × attainment).

First explained in: Choose the growth objective. Full treatment: Set targets and staff capacity.

Utilisation

The share of available time that people spend on work. Above 100% the queue grows without limit.

Formula: utilisation = work time ÷ available time.

First explained in: Set targets and staff capacity.

First-touch queue

The new SQLs that wait for a seller to make first contact.

Formula: mean wait = backlog ÷ SQLs contacted a day (Little's law).

First explained in: Set targets and staff capacity.

Ramp time

The time that a new hire needs to reach full productivity.

Formula: months from hire to steady wins = ramp months + cycle months.

First explained in: Set targets and staff capacity.

Staffing ratio

The number of people in one role per unit of volume or ARR. Published ratios are start values.

First explained in: Set targets and staff capacity.

Scalability, sustainability and durability tests

Three tests of a go-to-market model. Scalability: the next customer repays within the payback ceiling (24, 18 or 12 months). Sustainability: the chance of that. Durability: new customer value exceeds its fully loaded CAC.

First explained in: Choose markets and accounts.

Rule of 40

An investor check: the growth rate plus a profit margin, both in percent, should be at least 40.

Formula: growth rate % + profit margin % ≥ 40.

First explained in: Set targets and staff capacity.

Burn multiple

Net cash burned for each euro of net new ARR (Sacks, Craft Ventures, 2020, *The Burn Multiple*).

Formula: burn multiple = net cash burned ÷ net new ARR.

First explained in: Set targets and staff capacity.

Leverage of improvements

Lever

One change that a team can make to the funnel or to customers after the sale.

First explained in: Find the most valuable constraint.

Leverage

How much the value of the whole business changes when one lever works, measured as expected profit.

First explained in: Find the most valuable constraint.

Marginal value

The expected discounted gross profit that a gain of one percent brings to one rate.

Formula: marginal value of a rate = 1% × customers a year × value of a customer.

First explained in: Find the most valuable constraint.

Headroom

The gap between a rate today and a ceiling or benchmark.

Formula: headroom = ceiling or benchmark value ÷ current value - 1.

First explained in: Model recurring growth. Full treatment: Find the most valuable constraint.

Expected profit of a lever

The discounted gross profit of the extra customers that a lever brings during its life, minus its cost.

Formula: expected profit = sum over months of (extra customers × quality ratio × value of a customer × discount factor of the month) - cost.

First explained in: Find the most valuable constraint.

Lever life

The months of extra flow counted for a lever. Start value: 12 months (Plausible but untested).

First explained in: Find the most valuable constraint.

Quality dilution

The fall in average unit value when volume rises, because added units convert less often.

Formula: quality dilution ratio = value of the marginal unit ÷ value of the average unit.

First explained in: Find the most valuable constraint.

Spend elasticity

The percent rise in results from a 1% rise in channel spend. Below 1 means each extra euro buys less.

Formula: spend elasticity = percent change in leads ÷ percent change in spend.

First explained in: Find the most valuable constraint.

Capacity constraint

A limit on how many items a team can process in a window.

First explained in: Allocate sales attention.

Shadow price

The value of one more unit of a scarce resource, such as a seller slot.

Formula: shadow price of a slot = value of the item that the slot would serve, if the constraint binds; otherwise 0.

First explained in: Allocate sales attention.

Leverage order heuristic

The rule of thumb "start at the top of the funnel". It fails when downstream capacity binds, added volume is of lower quality, the market is small and known, payback breaches its ceiling, the leak after the close is larger, or a data gate fails.

First explained in: Find the most valuable constraint.

Order of work

The sequence in which a team does levers.

First explained in: Find the most valuable constraint.

Statistics and readability

Binomial and Poisson noise

Random variation in counts: binomial for successes out of fixed tries, Poisson for events in a period.

First explained in: Decide what the evidence can resolve.

Standard error

The typical size of the random error in a sample estimate.

Formula: standard error of a rate = square root of (p × (1 - p) ÷ n).

First explained in: Forecast conversion and timing. Full treatment: Decide what the evidence can resolve.

Confidence interval

The range of values that the data cannot rule out.

First explained in: Model recurring growth.

Significance level

The accepted chance of calling a difference real when it is chance. Usual value: 5%.

First explained in: Decide what the evidence can resolve.

Statistical power

The chance that a test finds a real effect of the minimum detectable size. Usual value: 80%.

First explained in: Decide what the evidence can resolve.

Minimum detectable effect (MDE)

The smallest change that a test or window can detect at the stated power and significance.

First explained in: Decide what the evidence can resolve.

Exact two-proportion formula

The sample size per arm for a test of two rates. It uses 5% two-sided significance and 80% power; $\bar{p}$ is the mean of the two rates.

Formula: n per arm = $(1.96 \times \sqrt{(2 \times \bar{p} \times (1 - \bar{p})} + 0.84 \times \sqrt{(p_1 \times (1 - p_1) + p_2 \times (1 - p_2))})^2 \div (p_1 - p_2)^2$.

First explained in: Decide what the evidence can resolve.

Rule of 16

Also called: readability rule.

A quick approximation of the items that a stage needs to show a relative change in its rate. It under-sizes a test by about 2% to 9%.

Formula: n per period or per arm = 16 × (1 − p) ÷ (p × r × r), for a rate p and a relative change r.

First explained in: Choose the growth objective. Full treatment: Decide what the evidence can resolve.

Sample size

The number of items needed to detect an effect of a given size.

Formula: weeks = 2 × n per arm ÷ weekly flow.

First explained in: Decide what the evidence can resolve.

Weeks of a test

Also called: weeks to read.

The weeks that a test needs to fill its arms.

Formula: weeks = 2 × n per arm ÷ weekly flow of eligible units.

First explained in: Decide what the evidence can resolve.

Readability

How large a change must be before your volume can show it.

First explained in: Decide what the evidence can resolve.

Readability map

A table of the smallest change that each funnel stage can show in a window, a quarter and a year.

First explained in: Decide what the evidence can resolve.

Monitoring mode

Sizing for a metric compared with a baseline of k earlier windows, not a control group.

Formula: variance multiplier = 1 + 1 ÷ k.

First explained in: Decide what the evidence can resolve.

Overdispersion

More variation than pure chance predicts, because the true rate itself wobbles between windows. Multiply sample sizes by ϕ and interval widths by $\sqrt{\phi}$. The same quantity as dispersion in Set targets and staff capacity.

Formula: overdispersion factor $\phi = 1 + n \times \tau^2 \div (p \times (1 - p))$, with τ the real window-to-window spread of the true rate.

First explained in: Decide what the evidence can resolve.

Design effect

The factor by which clustered units inflate the variance.

Formula: design effect $= 1 + (m - 1) \times \rho$, with m the cluster size and ρ the intra-cluster correlation.

First explained in: Decide what the evidence can resolve.

Intra-cluster correlation

How alike the members of one cluster are, from 0 (independent) to 1 (act as one).

First explained in: Decide what the evidence can resolve.

Covariate adjustment

Using a known baseline variable to remove part of the noise in an outcome.

Formula: variance multiplier $= 1 - r^2$, with r the correlation between covariate and outcome.

First explained in: Decide what the evidence can resolve.

Regression to the mean

Extreme values tend to be followed by less extreme ones, by chance alone.

First explained in: Decide what the evidence can resolve.

Partial pooling

A method that pulls the rates of small cells towards the rate of the whole group.

First explained in: Choose markets and accounts. Full treatment: Decide what the evidence can resolve.

Shrinkage

The pull of a raw rate towards the pooled rate. A small cell is pulled more.

Formula: shrunk rate = reliability × raw rate + (1 - reliability) × pooled rate.

First explained in: Decide what the evidence can resolve.

Shrunk rate

The rate of a cell after partial pooling.

Formula: shrunk rate = pooled rate + reliability × (raw rate - pooled rate).

First explained in: Decide what the evidence can resolve.

Reliability

The share of the spread between cells that is real, not noise. Applied to test results, the same number is the shrinkage factor of Commit with uncertainty.

Formula: reliability = $\tau^2 \div (\tau^2 + \sigma^2)$, with τ^2 the real spread between cells and σ^2 the noise within a cell.

First explained in: Decide what the evidence can resolve.

Funnel plot

A chart of each cell's rate against its size, with limits at 2 and 3 standard errors.

First explained in: Decide what the evidence can resolve.

Omnibus test of equal rates

One test of whether all cells could share one true rate.

Formula: chi-square = sum over cells of $(\text{wins} - n \times \text{pooled rate})^2 \div (n \times \text{pooled rate} \times (1 - \text{pooled rate}))$.

First explained in: Decide what the evidence can resolve.

Wilson interval

An interval for a share of n cases that stays sound when n is small, where the simple interval can pass 0% or 100%.

Formula: centre = $(p + z^2 \div 2n) \div (1 + z^2 \div n)$; half-width = $z \times \sqrt{(p(1 - p) \div n + z^2 \div 4n^2)} \div (1 + z^2 \div n)$; z = 1.96 for 95%.

First explained in: Learn from outcomes.

Measurement and data

Decision metric

A metric that a named person uses to make a decision.

First explained in: Decide what to measure.

Reference metric

A metric kept for context that drives no decision by itself.

First explained in: Decide what to measure.

Leading and lagging metric

A leading metric moves first and you can act on it this week. A lagging metric records the result later.

First explained in: Decide what to measure.

Evidence grade of a leading metric

A grade for how well a leading metric predicts the outcome: predictive at unit level, screened in a series, or shown causal.

First explained in: Decide what to measure.

Early warning

A leading metric that has no evidence grade. It may prompt a look, not a decision.

First explained in: Decide what to measure.

Metric specification

The written definition of one metric: name, numerator, denominator, population, window, grain, owner, source, minimum sample and limit model.

First explained in: Decide what to measure.

Population

The set of items that a metric counts. Never mix populations in one rate.

First explained in: Value customers and cash. Full treatment: Decide what to measure.

Source of truth

Also called: primary source.

The one system that decides the value of a metric.

First explained in: Decide what to measure.

Triangulation

Comparing one quantity across sources to explain differences.

First explained in: Decide what to measure.

Account key

A stable identifier that links all records of one customer or prospect across systems.

First explained in: Choose markets and accounts. Full treatment: Decide what to measure.

Identity resolution

Matching records from different systems to the same person or account.

First explained in: Decide what to measure.

Stage event

A time-stamped record that an item entered a stage.

First explained in: Decide what to measure.

Data contract

A written agreement between the producers and the users of data.

First explained in: Decide what to measure.

Data-quality test

An automatic check on the data. While a test error is open, no rate is read.

First explained in: Decide what to measure.

Test severity

Also called: severity.

The class of a data-quality test. An error blocks dependent tiles. A warning shows beside the tile.

First explained in: Decide what to measure.

Test pass rate

The share of data-quality tests that pass in the latest load.

Formula: test pass rate = tests passed ÷ tests run.

First explained in: Decide what to measure.

Freshness

How recent the newest data is.

First explained in: Decide what to measure.

Reconciliation

Checking that two sources agree on a total, within a tolerance set for each pair.

Formula: reconciliation gap = |primary total - other total| ÷ primary total.

First explained in: Decide what to measure.

Tracking coverage

The share of real actions that analytics saw. A value above 1 means that a tag fires twice.

Formula: tracking coverage = events in analytics ÷ matching records in the system of record.

First explained in: Decide what to measure.

Tracking plan

The events and properties that the website or product must send.

First explained in: Decide what to measure.

UTM parameters

Tags on a link that record the source, medium and campaign of a visit.

First explained in: Decide what to measure.

Engaged session

A session with real interaction: more than 10 seconds, a key event, or at least two page views.

First explained in: Decide what to measure.

Key event

An action marked as important in a web-analytics tool, such as a demo request.

First explained in: Decide what to measure.

Cookieless ping

An event that a tag sends while consent is denied. It is not a user or a session.

First explained in: Decide what to measure.

Consent capture ratio

The share of ad clicks that appear as sessions in your analytics. It is not the consent rate.

Formula: consent capture ratio = sessions recorded in analytics ÷ clicks billed by the ad platforms.

First explained in: Decide what to measure.

Attribution model

A rule that shares credit for a conversion among the touches before it.

First explained in: Decide what to measure.

Attribution coverage

The share of leads or revenue that carry a recorded source.

Formula: attribution coverage = leads with a source ÷ all leads.

First explained in: Decide what to measure.

Attribution range

The credit of a channel written as a lower and an upper bound. If two ranges overlap, the data cannot rank the channels.

Formula: lower = credited conversions ÷ all conversions; upper = (credited + unattributed) ÷ all conversions.

First explained in: Decide what to measure.

Decision dashboard

A dashboard built for named decisions, with few tiles per page.

First explained in: Decide what to measure.

Tile and tile limit

A tile is one metric on a dashboard. Its limit is the value at which its status changes.

First explained in: Decide what to measure.

Pace tile

A tile that compares progress in the current period with the plan.

First explained in: Decide what to measure.

Metric detail pane

The view that opens from a tile, with the specification and goal line.

First explained in: Decide what to measure.

Retirement rule

The rule that removes unused tiles and pages. Start values: four quarters inside limits, or a page unopened for two quarters (Plausible but untested).

First explained in: Decide what to measure.

Semantic model

A shared layer over the data that holds metric definitions.

First explained in: Decide what to measure.

Detection and diagnosis

Gate

A check on data quality, tracking and volume that must pass before you read any stage metric.

First explained in: Find the most valuable constraint. Full treatment: Decide when a change needs action.

Expectation model

A model of the value a metric should have, built from level, season and known changes.

Formula: expected value = level × season × known changes.

First explained in: Decide when a change needs action.

Change event

A dated record of a known change in the business or in tracking. It moves the expectation.

First explained in: Decide what to measure.

Z-score

The distance between the observed and the expected value, in standard errors.

Formula: z = (observed - expected) ÷ standard error.

First explained in: Decide when a change needs action.

Control chart

A plot of a measure over time, with limits set by the noise of the process.

First explained in: Decide when a change needs action.

Control limit

The edge of the band around the expected value, in standard errors, set by simulation.

First explained in: Decide when a change needs action.

Exponentially weighted moving average (EWMA)

A running average of the z-score that weights the newest window by λ.

Formula: EWMA(t) = λ × z(t) + (1 - λ) × EWMA(t-1), scaled to standard-normal units, with λ between 0 and 1.

First explained in: Decide when a change needs action.

EWMA weight (λ)

Also called: smoothing weight.

The weight an EWMA gives to the newest window. Start value 0.2 (Plausible but untested).

First explained in: Decide when a change needs action.

Cumulative sum (CUSUM)

A running total of falls below the expected value, less an allowance each window, never below zero.

Formula: CUSUM(t) = max(0, CUSUM(t-1) - z(t) - k), to watch for falls.

First explained in: Decide when a change needs action.

Allowance (k)

Also called: reference value.

The amount a CUSUM subtracts each window: half the smallest fall worth catching. Start value 0.5 (Plausible but untested).

First explained in: Decide when a change needs action.

False-alarm budget

The chance that noise alone alarms one rate in one year (default 10%, Plausible but untested), and the cap for the whole list.

First explained in: Decide when a change needs action.

Detection delay

The time from a real change to its alert; a tighter false-alarm budget lengthens it.

First explained in: Decide when a change needs action.

Maturity lag

The time from a real change in a slow rate to its alarm, including the cohort wait to reach reading age.

First explained in: Decide when a change needs action.

Multiplicity

The rise in false alarms when you watch many metrics at once.

First explained in: Decide when a change needs action.

False discovery rate

The expected share of false findings among all findings reported (Benjamini and Hochberg, 1995).

First explained in: Decide when a change needs action.

Alert

A signal that has crossed a limit and awaits review.

First explained in: Set targets and staff capacity. Full treatment: Decide when a change needs action.

Alert list

The page that lists alerts by severity, with their values, level and owner.

First explained in: Decide when a change needs action.

Finding

The record of one reviewed alert: what moved, evidence, likely cause and next step.

First explained in: Decide when a change needs action.

Watch and act levels

Watch asks for a look. Act asks for a decision. A lagging metric alone reaches watch only.

First explained in: Decide when a change needs action.

Leading pair

A leading metric and the lagging metric it predicts, read together. Act on the leading side.

First explained in: Decide when a change needs action.

Tracking break

A change in what is measured, not in what happens.

First explained in: Decide when a change needs action.

Mix effect and rate effect

A blended rate changes through group shares (mix effect) or rates inside groups (rate effect). While the mix effect exceeds half of the total movement (Plausible but untested), tell no story about rates.

Formula: mix effect = sum of (new share − old share) × (old rate + new rate) ÷ 2; rate effect = sum of (old share + new share) ÷ 2 × (new rate − old rate). Source: Kitagawa, 1955.

First explained in: Diagnose what changed.

Decomposition

Splitting a change in a total into parts that add up to it.

First explained in: Diagnose what changed.

Differential diagnosis

The ordered list of causes that can produce one symptom, with a "none of these" outcome.

First explained in: Diagnose what changed.

Separating metric

A metric on which two candidate causes predict different values. Write each prediction before you check.

First explained in: Diagnose what changed.

Falsifier

The result that would prove a diagnosis wrong, written before you open the data.

First explained in: Diagnose what changed.

Dose-response check

A cause check. If more exposure to the cause gives a larger effect, the cause is more likely.

First explained in: Diagnose what changed.

Placebo check

A cause check on units the cause could not have touched. The same change there rules the cause out.

First explained in: Diagnose what changed.

Timing check

A cause check on dates. The effect must start after the cause.

First explained in: Diagnose what changed.

Specificity check

A cause check on where the effect shows: a cause moves only the metrics it can touch.

First explained in: Diagnose what changed.

Evidence tier of a finding

The support for a finding. Observation: no check run. Checks passed: at least two different cheap checks, none failed (Plausible but untested). Experiment: randomisation or quasi-experiment. A check not run is "not run", never "passed".

First explained in: Diagnose what changed.

Recoverable effect

The effect an action can win back. Rank actions by it divided by cost.

Formula: recoverable effect = size of the change × share explained by the cause × share of that part the action recovers.

First explained in: Diagnose what changed.

Loss reason

The reason a lost deal is coded with, from deal notes or interviews. Correct the shares for labelling error.

First explained in: Diagnose what changed.

Text labelling

Assigning codes to text, such as loss reasons. A model drafts them and a person samples them to measure error rates.

First explained in: Diagnose what changed.

True-positive rate and false-positive rate

The share of true cases a labeller finds, and the share of non-cases it flags.

First explained in: Diagnose what changed.

Corrected share (of a coded label)

Also called: true share.

A coded share corrected for the errors of the labeller (Standard practice).

Formula: true share = (coded share − false-positive rate) ÷ (true-positive rate − false-positive rate). Source: Rogan and Gladen, 1978.

First explained in: Diagnose what changed.

Discovery sampling

A hand check that proves an error rate is below a tolerable rate when the sample holds no error.

Formula: n = ln(1 − confidence) ÷ ln(1 − tolerable rate), rounded up.

First explained in: Diagnose what changed.

Decisions and prioritisation

Decision card

One page that prices one decision on the value basis, with the net value of each option, including "do nothing", as a distribution.

Formula: net value = v × l × D − K − r. Here v is the value of one unit, l the extra units, D the effective windows, K the cost and r the expected harm.

First explained in: Commit with uncertainty.

Expected net value

The mean net value of an option against doing nothing, in euro of discounted gross profit.

First explained in: Commit with uncertainty.

Break-even effect

The smallest effect that pays for an option, with the other inputs at their means.

Formula: break-even effect = (K + r) ÷ (value of one unit × units exposed × D).

First explained in: Commit with uncertainty.

Reversible and irreversible decision

A reversible decision is cheap to undo. An irreversible one is not, so its shipping threshold is higher.

First explained in: Commit with uncertainty.

Staged move

A cheap, reversible first step with signposts, then a later step that waits for them.

First explained in: Commit with uncertainty.

Do-nothing option

The option to change nothing. Every decision card includes it as the status quo.

First explained in: Commit with uncertainty.

Risk limit

The most cash one decision may commit before its first signpost. Start value 2% of yearly gross profit (Plausible but untested).

First explained in: Commit with uncertainty.

Change-impact review

A list of what each option does on the other wing of the bowtie, priced into the harm term or set as a guardrail.

First explained in: Commit with uncertainty.

Prior

What you believe about an effect before you see the new data.

First explained in: Commit with uncertainty.

Outside view

Judging a case by the results of similar past cases, not by its own details.

First explained in: Commit with uncertainty.

Base rate

How often an outcome happens in a class of similar cases, such as the share of tests that win.

First explained in: Commit with uncertainty.

Base rate of winning tests

The share of the tests of a company that end as significant winners. Use outside base rates until the register holds about 30 tests.

Formula: share of significant wins that are false = α × (1 − b) ÷ (α × (1 − b) + power × b), with b the share of ideas that truly work.

First explained in: Commit with uncertainty.

Reference class

The group of similar past cases that you use for a base rate.

First explained in: Commit with uncertainty.

Posterior

What you believe after you combine the prior with the new data.

First explained in: Commit with uncertainty.

Shrinkage factor

The share of the observed spread of effects that is real. A noisy test has a small factor, so its result moves far towards the prior. It is the reliability of Decide what the evidence can resolve, applied to test results.

Formula: shrinkage factor = $\tau^2 \div (\tau^2 + \sigma^2)$, with τ the real spread of effects in the register and σ the standard error of the test. Shrunk effect = prior mean + shrinkage factor × (observed lift − prior mean).

First explained in: Commit with uncertainty.

Winner's curse

The best-looking result is overstated because it was picked from many noisy results. Plan with the shrunk effect.

First explained in: Commit with uncertainty.

Probability of loss

Also called: chance of loss.

The share of simulated draws in which an option has a net value below zero.

First explained in: Commit with uncertainty.

Probability of positive net value

The share of simulated draws in which an option has a net value above zero.

First explained in: Commit with uncertainty.

Probability of being the best option

Also called: chance of best.

The share of simulated draws in which an option has the highest net value on the card.

First explained in: Commit with uncertainty.

Shipping threshold

The probability of positive net value at which you ship: at least 90% if reversible, 95% if irreversible (Plausible but untested start values).

First explained in: Find the most valuable constraint. Full treatment: Commit with uncertainty.

Decision log

The write-once record of decisions and of forecasts that carry a belief, written before the outcome. Beliefs are scored by class of action, never by person.

First explained in: Model recurring growth. Full treatment: Commit with uncertainty.

Confirms-by date

The date by which the data must show whether a decision, a forecast or a diagnosis was right.

First explained in: Diagnose what changed.

Signpost

A dated, observable event that tells you early whether a decision is on track.

First explained in: Set targets and staff capacity. Full treatment: Commit with uncertainty.

Scoring frameworks (ICE, RICE, PIE, PXL)

Retired methods that multiply small-scale scores for impact, confidence, ease, reach or effort. We rank by expected value instead.

First explained in: Fund the next best action.

Rank correlation (Kendall's tau)

A number from −1 to 1 that compares two orders of the same items.

First explained in: Fund the next best action.

Cost of delay

The gross profit that a change loses for each week that it starts later. It is zero for an item that cannot yet earn.

Formula: cost of delay per week = expected gross profit of the effect ÷ weeks of lever life. Cost of a delay of d weeks = cost of delay per week × d.

First explained in: Choose whether to test or act. Full method: Fund the next best action.

Cost of delay per week of work

Also called: CD3; cost of delay divided by duration; weighted shortest job first.

A ranking value: do first the item with the highest cost of delay per week of the scarce resource it needs.

Formula: cost of delay per week of work = cost of delay per week ÷ weeks of the scarce resource that the item needs.

First explained in: Fund the next best action.

Work in progress (WIP) and its limit

Work in progress is the number of items started and not finished. A limit caps it.

First explained in: Fund the next best action.

Throughput

The number of items finished per period.

First explained in: Fund the next best action.

Time in system

The weeks one item takes from start to finish.

Formula: time in system = work in progress ÷ throughput.

First explained in: Fund the next best action.

Slot-week

One week of the whole flow of eligible units, used by one test.

Formula: slot-weeks of a test = units that the test needs ÷ eligible units per week.

First explained in: Fund the next best action.

EVSI per slot-week

The expected value of sample information of a test divided by the slot-weeks it uses.

Formula: EVSI per slot-week = EVSI of the test ÷ slot-weeks of the test.

First explained in: Fund the next best action.

Hurdle

The minimum expected value per unit of the scarce resource that a card must clear to be funded.

First explained in: Fund the next best action.

Probability of covering the target

The share of simulated draws in which the summed net values of the funded cards reach the target.

First explained in: Fund the next best action.

Portfolio

All funded bets taken together, judged by total risk, such as the chance of at least one false confirm.

First explained in: Fund the next best action.

False confirm

A change that passes its rule although it has no effect. Among n independent null bets, the chance of at least one is 1 – (1 – α)^n.

First explained in: Fund the next best action.

Kill and scale triggers

Rules written before funding: a kill trigger stops a bet when a signpost fails, and a scale trigger adds funding when one passes.

First explained in: Fund the next best action.

Trigger sheet

The page written before funding with the load-bearing assumptions, signposts and one contingent action per trigger.

First explained in: Fund the next best action.

Load-bearing assumption

A belief whose failure ends the case for a decision.

First explained in: Fund the next best action.

Tranche

One step of staged funding. Use it only when the saving from an early stop, times the chance it is correct, exceeds the cost of delay plus the value false stops destroy.

First explained in: Fund the next best action.

Explore and exploit

Exploit means using what works. Explore means trying new options to learn.

First explained in: Fund the next best action.

Experiments

A/B test

Also called: randomised controlled test; randomised two-version test.

A test that randomly splits units between two versions and compares one primary metric.

First explained in: Choose whether to test or act.

Randomisation unit

The unit randomly assigned to a version: a visitor, an account or a workspace.

First explained in: Choose whether to test or act.

Assignment log

One line per unit in a test: unit, arm, time and test name.

First explained in: Choose whether to test or act.

Holdout

A group kept from a treatment, to compare its outcome with the treated group.

First explained in: Choose whether to test or act.

Holdback

A holdout inside a rollout: the share of units that stays on the old version.

First explained in: Commit with uncertainty. Full method: Choose whether to test or act.

Staged rollout

Releasing a change in steps, each with a holdback, a stop rule for harm and a next-step condition.

First explained in: Choose whether to test or act.

Value of information

The expected gain from learning before you decide, net of its cost and delay. Test only above zero.

First explained in: Find the most valuable constraint. Full treatment: Choose whether to test or act.

Expected value of perfect information (EVPI)

The gain from knowing the true effect before you decide. It caps the cost of a test.

Formula: EVPI = expected value of the best choice with full knowledge – expected value of the best choice now.

First explained in: Choose whether to test or act.

Expected value of sample information (EVSI)

What a test of a given size is worth on average. It never exceeds EVPI.

Formula: EVSI = expected value of the best choice after the test result – expected value of the best choice now.

First explained in: Choose whether to test or act.

Expected net gain of a test

What a test is worth after its cost and delay. Choose the size where it peaks.

Formula: expected net gain = EVSI – cost of the test – cost of delay a week × weeks until the result.

First explained in: Choose whether to test or act.

Test-and-roll

A small test on a known audience picks a version, and the winner goes to all remaining units.

Formula: gain over a coin flip = $(N - 2n) \times s \div \sqrt{\pi}$, where $s = \sigma^2 \div \sqrt{(\sigma^2 + p \times (1 - p) \div n)}$, for n units per version in an audience of N.

First explained in: Choose whether to test or act.

Sample-ratio mismatch

A split between versions that differs from plan by more than chance allows, voiding the test. Alarm at p below 0.001 (start value, Plausible but untested).

First explained in: Choose whether to test or act.

Balance check

Arms compared on facts known before the test. A large gap means broken assignment.

First explained in: Choose whether to test or act.

Test duration

The weeks of flow for both arms plus the time for the outcome to mature.

Formula: weeks of flow = 2 × units per arm ÷ weekly flow of eligible units.

First explained in: Choose whether to test or act.

Intention to treat

Analysing units by the version assigned, whatever they did afterwards.

First explained in: Choose whether to test or act.

Primary metric

The one metric on which a test is decided, chosen before the test starts.

First explained in: Choose whether to test or act.

Composite metric

A primary metric that multiplies a guardrail into the main result, so harm cancels the gain.

First explained in: Choose whether to test or act.

Guardrail metric

A metric that must not fall while the primary metric rises. It needs a non-inferiority margin.

First explained in: Choose whether to test or act.

Non-inferiority margin

The largest guardrail fall you accept, set from break-even. The one-sided 95% lower bound must lie above minus the margin.

Formula: units per arm = $2 \times (1.645 + 0.842)^2 \times p \times (1 - p) \div (\text{margin} \times p)^2$, with no true difference, 80% power and a one-sided 5% test.

First explained in: Choose whether to test or act.

Proxy and surrogate metric

A metric that stands in for a slow or rare outcome. A decision waits for the outcome unless validated.

First explained in: Choose whether to test or act.

Look

A planned reading of a test before it ends. Each look adds false-positive risk.

First explained in: Choose whether to test or act.

Sequential testing

Reading a test at planned points while keeping the false-positive rate under control.

First explained in: Choose whether to test or act.

Alpha spending

A rule that divides the allowed false-positive rate among the planned looks.

First explained in: Choose whether to test or act.

Factorial design

A test that varies several elements together. Screen first, then confirm the best combination.

First explained in: Choose whether to test or act.

Multi-armed bandit

A method that shifts traffic towards the better option during the test. It is not the default.

First explained in: Choose whether to test or act.

Quasi-experiment

A comparison around a change nobody randomised, to estimate its effect. Keep "confirmed" for randomised tests and holdbacks.

First explained in: Diagnose what changed. Full method: Choose whether to test or act.

Interrupted time series

A quasi-experiment. It compares the trend of one series before and after a change.

First explained in: Choose whether to test or act.

Difference in differences

A quasi-experiment. It compares the change in a treated group with the change in an untreated group.

Formula: effect = (treated after - treated before) - (control after - control before).

First explained in: Choose whether to test or act.

Regression discontinuity

A quasi-experiment. It compares units just above and below a cut-off that decides treatment.

First explained in: Choose whether to test or act.

Synthetic control

A quasi-experiment. It blends untreated units to match the treated unit before the change, then compares them after.

First explained in: Choose whether to test or act.

Randomisation inference

A way to judge a result with few units: compare the observed effect with every possible assignment.

First explained in: Choose whether to test or act.

Switchback design

A test in which the whole market switches between versions over time periods, in random order.

First explained in: Choose whether to test or act.

Novelty effect

A gain that appears because a change is new, and fades as users adapt.

First explained in: Choose whether to test or act.

Superior, equivalent within the margin, unresolved

Three readings of a result. Superior: the interval lies above break-even. Equivalent: the lower bound lies above minus the margin. Unresolved: the interval spans the deciding line.

First explained in: Choose whether to test or act.

Marketing measurement

Incrementality

The part of an outcome that an action caused, beyond what would have happened anyway.

Formula: incremental conversions = conversions with the ads − conversions without the ads.

First explained in: Choose acquisition channels.

Incremental conversions

The conversions that the ads caused: conversions with the ads minus conversions without them.

First explained in: Choose acquisition channels.

Incremental CAC

Spend divided by incremental conversions. It is higher than the credited CAC when some credited conversions would have happened anyway.

Formula: incremental CAC = spend ÷ incremental conversions.

First explained in: Choose acquisition channels.

Lift test

A test that compares an exposed group with a held-out group to measure the extra outcome.

First explained in: Choose acquisition channels.

Audience holdout

A random share of a target audience kept from seeing the ads, to show what happens without them.

First explained in: Choose acquisition channels.

Account holdout

A random share of target accounts that gets no programme touches in account-based marketing.

First explained in: Choose acquisition channels.

Geo experiment

Ads run in some regions and not others, assigned at random or in matched pairs.

First explained in: Choose acquisition channels.

Matched pairs

A design that pairs units by a variable such as size and treats one unit of each pair.

First explained in: Choose acquisition channels.

Budget pulse

A planned change in the spend of one channel: up to 30% up or down for two weeks.

Formula: standard error of the elasticity = weekly noise ÷ (standard deviation of log spend × square root of the weeks).

First explained in: Choose acquisition channels.

Spend response curve

The curve linking the spend on a channel to its results. It bends and holds only inside the tested range.

Formula: leads = L0 × (spend ÷ S0)^e.

First explained in: Choose acquisition channels.

Saturation

The bend of the spend response curve: each extra euro buys less. The saturation point is where marginal CAC hits the payback ceiling.

First explained in: Choose acquisition channels.

Tested range

The range of spend that a pulse or test has varied.

First explained in: Choose acquisition channels.

Carry-over

Also called: adstock.

The effect of spend that continues in later weeks. The half-life is when half remains.

First explained in: Choose acquisition channels.

Marginal CAC

The cost of the next customer, not the average one.

Formula: marginal cost per lead = average cost per lead ÷ spend elasticity; marginal CAC = marginal cost per lead ÷ customers per lead.

First explained in: Choose acquisition channels.

Customers per lead

The share of leads that eventually become customers.

Formula: customers per lead = product of the stage rates from lead to won.

First explained in: Choose acquisition channels.

Marginal gross profit per euro

The value of a customer divided by the marginal CAC. Move budget until it is equal across channels.

Formula: marginal gross profit per euro = value of a customer ÷ marginal CAC.

First explained in: Choose acquisition channels.

Return on ad spend (ROAS)

Revenue credited to ads divided by their spend. Do not act on it alone.

Formula: ROAS = credited revenue ÷ ad spend.

First explained in: Choose acquisition channels.

Incremental ROAS

The extra gross profit that an incrementality test shows, divided by the spend.

Formula: incremental ROAS = extra gross profit from the ads ÷ ad spend.

First explained in: Choose acquisition channels.

Marketing mix model (MMM)

A regression of total results on every channel spend over a long history, with controls. Experiments calibrate it.

First explained in: Choose acquisition channels.

Calibration of a model by experiments

Using a randomised test result as a prior or likelihood term of a model.

First explained in: Choose acquisition channels.

Bias allowance

The width added to a source because it may be wrong in a way more data cannot cure. Start values (Plausible but untested), as shares of the estimate: 5% randomised test, 20% calibrated mix model. Also 40% uncalibrated mix model and 50% attribution report or untested plan.

First explained in: Choose acquisition channels.

Precision weighting

Combining estimates from several sources with weights that fall as error and bias allowance grow.

Formula: weight = 1 ÷ (standard error² + bias allowance²); combined estimate = sum of (weight × estimate) ÷ sum of weights.

First explained in: Choose acquisition channels.

Levers before the sale

Total addressable market (TAM)

The companies or yearly revenue your product could serve. A ceiling, not a forecast.

First explained in: Choose markets and accounts.

TAM coverage

Two shares: how much of the market the list covers, and how much of the list you engaged.

Formula: list coverage = units on the target list ÷ units in the market outside your customers; reach coverage = target accounts engaged in the period ÷ target accounts on the list.

First explained in: Choose markets and accounts.

Target account list

The named part of the total addressable market that you work.

First explained in: Choose markets and accounts.

ICP match rate

The share of checkable leads that match the ideal customer profile.

Formula: ICP match rate = enriched leads that match ÷ enriched leads.

First explained in: Decide what to measure.

Buyer persona

The people in a target company, by what each tries to achieve.

First explained in: Choose markets and accounts.

Campaign structure

How you split ad budget into campaigns and ad sets, each with one budget and one audience.

First explained in: Choose acquisition channels.

Signal density

Conversion events per ad set each week. An unverified platform rule of thumb says about 50 events per ad set.

Formula: events per ad set per week = conversion events that the account sends in a week ÷ ad sets that share them.

First explained in: Choose acquisition channels.

Conversion event and offline conversion import

A conversion event is an action sent to an ad platform. An offline conversion import sends later outcomes, such as an MQL, from the CRM.

First explained in: Choose acquisition channels.

Value-based lead events

Lead events sent to an ad platform with a value equal to expected gross profit (Plausible but untested).

First explained in: Choose acquisition channels.

Audience and audience contract

An audience is a group you target. Its contract records owner, purpose, consent basis and expiry date.

First explained in: Improve offers and conversion.

Reach and frequency

Reach is the unique people who saw an ad; frequency is the average views each.

Formula: frequency = impressions ÷ reach.

First explained in: Improve offers and conversion.

Cost per thousand impressions

The price of 1,000 showings of an ad.

Formula: cost per thousand impressions = spend ÷ impressions × 1,000.

First explained in: Choose acquisition channels.

Cost per click

Spend divided by the clicks that the platform bills.

Formula: cost per click = spend ÷ billed clicks.

First explained in: Improve offers and conversion.

Cost per lead and value per lead

Cost per lead is paid spend per paid lead. Value per lead is its expected gross profit.

Formula: cost per lead = paid spend ÷ paid leads; value of a lead = value of a customer × customers per lead.

First explained in: Improve offers and conversion.

Break-even incremental share

The share of a channel's credited conversions that must be incremental to return one euro per euro spent.

Formula: break-even incremental share = 1 ÷ credited value per euro.

First explained in: Choose acquisition channels.

Creative fatigue

Falling response when the same people see the same ad often. A first-exposure rate c falls by a share d per exposure (Illustration only).

Formula: click-through rate at exposure k = c × (1 - d)^(k - 1); marginal cost of a click at exposure k = cost per thousand impressions ÷ (1,000 × click-through rate at exposure k).

First explained in: Improve offers and conversion.

Value proposition

What the customer gets and why it beats the alternatives.

First explained in: Improve offers and conversion.

Offer

What you give for an action, such as a demo.

First explained in: Improve offers and conversion.

Landing page

The page where a visitor arrives, built for one action.

First explained in: Improve offers and conversion.

Qualified session and qualified-session rate

An engaged session from a visitor in the target market, counted on identified sessions.

Formula: qualified-session rate = qualified sessions ÷ sessions.

First explained in: Improve offers and conversion.

Call to action

The one action a page asks the visitor to take.

First explained in: Improve offers and conversion.

Conversion rate optimisation (CRO)

Testing and improving the rates at which visitors act on pages and forms.

First explained in: Improve offers and conversion.

Form completion rate

The share of started forms that are submitted. It does not show lead quality.

Formula: form completion rate = forms submitted ÷ forms started.

First explained in: Improve offers and conversion.

Qualified completions

Submitted forms that later become MQLs, per session.

Formula: qualified completions per session = MQLs from the form ÷ sessions on the page.

First explained in: Improve offers and conversion.

Demand baseline

Demand that exists before your campaigns act, such as brand searches.

First explained in: Choose acquisition channels.

Branded search share

Searches naming your brand as a share of category brand searches.

Formula: branded search share = searches for your brand ÷ searches for all brands in the category.

First explained in: Choose acquisition channels.

Answer-engine visibility

How often AI answer tools name your company in your category.

First explained in: Choose acquisition channels.

Hand-off contract

A written agreement among marketing, sales development and sales on MQL acceptance and rejection.

First explained in: Remove buying friction.

MQL rejection rate

The share of passed MQLs that sales rejects.

Formula: MQL rejection rate = MQLs rejected by sales ÷ MQLs passed to sales.

First explained in: Remove buying friction.

Rejection reason

One reason from a short fixed list that a seller records on rejection.

First explained in: Remove buying friction.

Lead scoring

A score or probability for a specified outcome per lead. A purchase prediction does not by itself estimate the additional value caused by contact.

First explained in: Allocate sales attention.

Calibration of a score

A score is calibrated when leads with a predicted 20% chance convert about 20% of the time.

First explained in: Allocate sales attention.

Explore arm

A random sample below the score cut-off that receives contact. It informs outcomes under contact in that group. A comparable no-contact group is also needed to estimate the effect.

First explained in: Allocate sales attention.

Explore-arm rate

The share of explore arm leads that reach the label.

Formula: explore-arm rate = arm leads that reached the label ÷ arm leads contacted.

First explained in: Allocate sales attention.

Cut-off

A decision boundary for assigning scarce contact capacity. Set it from incremental value and the resource constraint; a purchase-probability threshold is valid only under stated simplifying assumptions.

Contact gain = expected discounted gross profit under contact minus that under no contact, less incremental contact expenditure. Allocate a feasible set with the highest total gain.

First explained in: Allocate sales attention.

Nurture

A planned message series for leads not ready for sales.

First explained in: Remove buying friction.

Suppression arm

A random group of leads that gets no nurture.

First explained in: Remove buying friction.

Nurture incrementality

What the nurture adds, measured against the suppression arm.

Formula: nurture incrementality = rate of nurtured leads - rate of suppressed leads, over the same period.

First explained in: Remove buying friction.

Deliverability and send health

Deliverability is the share of emails that reach the inbox. Send health is the email complaint rate and the email bounce rate.

Formula: complaint rate = complaints ÷ delivered emails; bounce rate = bounced emails ÷ sent emails.

First explained in: Remove buying friction.

Speed to lead

Minutes from lead creation to the first human touch.

First explained in: Remove buying friction.

Contact rate

The share of leads that reach a person at the first attempts.

Formula: contact rate = leads that reached a person ÷ leads touched.

First explained in: Remove buying friction.

SPICED

A discovery method that records Situation, Pain, Impact, Critical Event and Decision.

First explained in: Remove buying friction.

MEDDIC

A qualification method: Metrics, Economic buyer, Decision criteria, Decision process, Identify pain, Champion.

First explained in: Remove buying friction.

Critical event

A dated event in the customer's business with consequences if missed, confirmed by the customer.

First explained in: Remove buying friction.

Champion

A person at the customer with influence who works to make the deal happen.

First explained in: Remove buying friction.

Mutual action plan

A shared, dated plan of steps to a signed contract.

First explained in: Remove buying friction.

Critical-event coverage

The share of opportunities with a customer-confirmed critical event before the proposal (Plausible but untested as a closing predictor).

First explained in: Remove buying friction.

Champion presence rate

The share of opportunities with a champion who did a costly act by the proposal date (Plausible but untested as a closing predictor).

First explained in: Remove buying friction.

Pricing and packaging

The price levels and what each plan contains.

First explained in: Set prices and packages.

List price, discount depth and price realisation

List price is the published price; discount depth is the gap to the signed price; price realisation is the share kept.

Formula: discount depth = 1 - net price ÷ list price; price realisation = net price ÷ list price.

First explained in: Set prices and packages.

Willingness to pay

The highest price a buyer accepts, estimated by research.

First explained in: Set prices and packages.

Gross-profit price frontier

The close rate at which a new price earns the same gross profit as the old price.

Formula: break-even close rate, as a share of before = gross profit per customer-year before ÷ gross profit per customer-year after, with gross profit per customer-year = price × (1 - cost share that scales with price) - cost to serve per customer.

First explained in: Set prices and packages.

Pipeline coverage

Open pipeline held against the target still to win; at win rate w, at least 1 ÷ w.

Formula: pipeline coverage = open pipeline value ÷ new ARR still needed.

First explained in: Remove buying friction.

After the close

Impact journey

The path from commit through onboarding, activation, adoption and renewal to expansion.

First explained in: Activate customers.

Time to first value

The time from commit until the customer first sees a result from the product.

First explained in: Activate customers.

Time to activation

The time from a defined eligible start until activation. Departure that prevents activation is a competing outcome; reporting ending first creates administrative censoring.

First explained in: Activate customers.

Activation rate

The probability that an eligible entrant activates inside the window before departure. With incomplete follow-up, estimate the cumulative incidence and distinguish departures from administrative censoring.

For a fully observed cohort: activation rate = entrants activated before departure inside the window ÷ eligible entrants at the time origin.

First explained in: Activate customers. Full treatment: Activate customers.

Adoption depth

The share of intended users, roles or sites that use the product, and how many key features. It leads renewal.

First explained in: Retain customers profitably.

Health score

A model score per account that predicts later renewal from usage, support, sentiment and commercial signals.

First explained in: Retain customers profitably.

Save play

An action to keep an at-risk account, such as an offer or a call.

Formula: value of a play = change in staying probability × gross profit kept - cost of the play; break-even change = cost ÷ gross profit kept.

First explained in: Retain customers profitably.

Uplift

The change in an outcome that an action causes, against a control group.

Formula: uplift = outcome rate of treated - outcome rate of control.

First explained in: Retain customers profitably.

Persuadable customer

A customer whose outcome changes because of an action.

First explained in: Retain customers profitably.

Incremental save rate

The change that a save play causes in the share of at-risk accounts that stay.

Formula: incremental save rate = share staying among treated – share staying among control.

First explained in: Retain customers profitably.

Cut in cancellations

The share of would-be cancellers that a save play turns into stayers.

Formula: break-even cut = cost of the play per offered customer ÷ gross profit that a save keeps.

First explained in: Retain customers profitably.

Whitespace

The part of an account where the product could deliver impact and does not yet.

First explained in: Expand accounts and generate referrals.

Upsell and cross-sell

Upsell sells a higher tier or more of the same product. Cross-sell sells another product.

First explained in: Expand accounts and generate referrals.

Expansion rate

The flow of expansion ARR, against the opening ARR of the same customers.

Formula: expansion rate = expansion ARR ÷ opening ARR of the same customers.

First explained in: Expand accounts and generate referrals.

Cost of expansion

The fully loaded sales and marketing cost per 1.00 of expansion ARR. The book uses a starting ceiling of 0.50 and a starting objective at or below 0.40, both Plausible but untested. Survey quantiles provide context rather than validating those operating limits (KeyBanc Capital Markets, 2021, *Private SaaS Company Survey*).

Formula: cost of expansion = sales and marketing cost of expansion work ÷ expansion ARR won.

First explained in: Expand accounts and generate referrals.

Price-rise break-even

The extra loss of renewals at which a price rise earns the same discounted gross profit as no rise.

Formula: break-even extra loss = p ÷ (m + p) = p ÷ (1 + p − c), with p the rise as a fraction of the old price, m the gross margin and c = 1 − m the cost share.

First explained in: Set prices and packages.

Advocacy

Customers who recommend the company: references, reviews and referrals.

First explained in: Expand accounts and generate referrals.

Referral rate

The share of new customers that came from a referral.

Formula: referral rate = new customers from referrals ÷ all new customers.

First explained in: Expand accounts and generate referrals.

Customer success manager

A role that helps customers reach impact and renew.

First explained in: Expand accounts and generate referrals.

Account manager

A role that owns expansion in existing accounts.

First explained in: Expand accounts and generate referrals.

Renewal project

A large renewal at low volume, run as a project with a named owner and a written plan.

First explained in: Retain customers profitably.

Churn reason

The coded reason that a customer left. Correct the shares for coder error.

First explained in: Expand accounts and generate referrals.

Learning

Counterfactual

What would have happened if you had not acted, estimated for example from a control group.

First explained in: Learn from outcomes.

Back to baseline

The metric returns to its old level after a fix. This is not proof that the fix worked.

First explained in: Learn from outcomes.

Reversion path

The path a metric would follow back to its usual level if you did nothing.

Formula: reversion path for the next window = level + reliability × (now − level).

First explained in: Learn from outcomes.

Level of evidence for a shipped action

How strong the proof is that an action worked, from opinion up to a randomised test or holdback. Only the top two levels may say "confirmed".

First explained in: Learn from outcomes.

Pre-registered maximum extension

The one extension of a confirms-by date that an entry may carry, set at decision time.

First explained in: Learn from outcomes.

Outcome class

The class that code gives to an entry at its confirms-by date: confirmed, refuted, inconclusive or void.

First explained in: Learn from outcomes.

Forecast entry and effect entry

A forecast entry states a count, a range or a signpost. An effect entry claims that an action changed a metric.

First explained in: Learn from outcomes.

Grade of a log entry

The reading of an entry: held, refuted, missed signpost, not confirmable or open.

First explained in: Learn from outcomes.

Hit rate

The share of resolved predictions that came true. Read it for a pool, never for one person.

First explained in: Learn from outcomes.

Brier score

A measure of the accuracy of probability forecasts. Lower is better.

Formula: Brier score = average of (stated probability − outcome)², with outcome 1 if the event happened and 0 if not.

First explained in: Learn from outcomes.

Calibration and calibration slope

A forecaster is calibrated when 70% forecasts come true about 70% of the time. A slope below 1 means too much confidence.

First explained in: Learn from outcomes.

Slope of realised on predicted effect

The slope of the lift that a holdback read later, on the lift that was predicted.

First explained in: Learn from outcomes.

Prior weight

The number of outcomes that an old value is worth in a weighted update (Plausible but untested).

Formula: new confidence = (prior weight × old confidence + hits) ÷ (prior weight + resolved outcomes).

First explained in: Learn from outcomes.

Confidence tier

The standard confidence that a class of claims carries, for example 70%.

First explained in: Learn from outcomes.

Evaluation suite

A standing set of tests for a rule, detector or model, with limits set before the run.

First explained in: Learn from outcomes.

Acceptance limit

The result that a suite test must reach for the rule to pass, written before the run.

First explained in: Learn from outcomes.

Noise-only run

A run of a detector on data in which nothing changes, so every alarm in it is false.

First explained in: Decide when a change needs action.

Seeded fault

A known change planted in a replay of history or in simulated data.

First explained in: Decide when a change needs action. Full treatment: Learn from outcomes.

Severity sweep

One seeded fault repeated at growing sizes, to find the smallest change found reliably.

First explained in: Decide when a change needs action. Full treatment: Learn from outcomes.

Rule of three

When n clean runs show no failure, the 95% upper bound on the failure rate is about 3 ÷ n (Hanley and Lippman-Hand, 1983).

Formula: upper bound ≈ 3 ÷ n.

First explained in: Learn from outcomes.

Golden set

A set of cases with agreed correct answers, used to test a text labeller or a model.

First explained in: Learn from outcomes.

Operating system, AI and automation

Asynchronous decision

A decision made without a meeting: a written brief goes to one owner, who decides by a deadline.

First explained in: Run the growth system.

Written brief

The one-page document that starts an asynchronous decision.

First explained in: Choose the growth objective. Full treatment: Run the growth system.

Decision owner

The one named role that decides a question by its deadline.

First explained in: Commit with uncertainty. Full treatment: Run the growth system.

Escalation

A meeting called for one decision, only for a conflict between owners or a commitment no single owner can make.

First explained in: Run the growth system.

Role

A set of decisions and duties that the loop needs. A person holds it. Agents are not roles.

First explained in: Run the growth system.

Stage owner

The role that owns the metrics and levers of one funnel stage: the marketing owner, the sales owner or the customer success owner.

First explained in: Run the growth system.

Analytics lead

The role that owns the measurement system. It decides only on measurement.

First explained in: Run the growth system.

Bowtie owner

The one role that settles conflicts between the acquisition wing and the retention and expansion wing.

First explained in: Choose the growth objective. Full treatment: Run the growth system.

Revenue operations (RevOps)

The role that keeps the CRM, billing map and stage definitions, and checks that the ARR bridge closes.

First explained in: Run the growth system.

Weekly memo

A one-page written summary, ready on Monday, of findings and proposed decisions.

First explained in: Run the growth system.

Bowtie note

A monthly page from the bowtie owner on the trade-offs between the two wings.

First explained in: Run the growth system.

Quarterly refresh

The review every 91 days of models, priors, calibration, specifications, tiles and the risk register.

First explained in: Run the growth system.

Exception lane

A route that shows each person only the items that need them.

First explained in: Run the growth system.

Independent estimate

A belief that a person writes down before seeing anyone else's.

First explained in: Run the growth system.

Pre-mortem

A written list of the ways a large bet could fail, made before the bet.

First explained in: Run the growth system.

Bounded override

A change by an owner to an input or proposal, inside a stated range and with a reason in the log.

First explained in: Run the growth system.

Yield test

If a recurring meeting changed no decision in four sessions, retire it (Plausible but untested).

First explained in: Run the growth system.

Scheduled job

An automatic task that runs on a timetable and computes numbers.

First explained in: Run the growth system.

AI agent

A program that uses an AI model and a few defined tools. It drafts text, then hands over to a person.

First explained in: Run the growth system.

Large language model (LLM)

An AI model trained on large amounts of text to read and write text.

First explained in: Run the growth system.

Autonomy ladder

The steps by which an agent earns freedom: shadow, then advisory for at least two quarters, then low-risk actions that a person approves one by one. Agents never decide on money, price, people, commitments or consent.

First explained in: Run the growth system.

Human gate

A step where a person must approve before anything else happens.

First explained in: Run the growth system.

Cohen's kappa

A measure of agreement between two coders beyond chance: 1 is perfect, 0 is chance. The start floor for a labeller is 0.60 (Plausible but untested). Chance agreement is high when one label is rare, so the same raw agreement gives a lower kappa (Arithmetic).

Formula: kappa = (observed agreement − chance agreement) ÷ (1 − chance agreement).

First explained in: Run the growth system.

Abstain rate

The share of items that a labeller leaves to people. The start ceiling is 15% (Plausible but untested).

Formula: abstain rate = items left to people ÷ all items.

First explained in: Run the growth system.

Number token

A placeholder that an AI model writes in a draft. Code fills in the value.

First explained in: Run the growth system.

Lint

An automatic check of text or code against rules.

First explained in: Run the growth system.

Provenance

The record of where a claim or a number came from: the source, the query and the time.

First explained in: Run the growth system.

Provenance defect

A number or claim in agent output that no job or source supports.

First explained in: Run the growth system.

Prompt injection

Text in the data that tries to give instructions to an AI model.

First explained in: Run the growth system.

Kill switch

A control that stops an agent and its schedule at once.

First explained in: Run the growth system.

Risk register

A table of the legal, security and statistical risks, each with a control and an owner.

First explained in: Run the growth system.

Vertical slice

A thin piece of the system that runs end to end, from data to decision.

First explained in: Run the growth system.

Operating page

The table of measures of how the loop itself runs, reported by role.

First explained in: Run the growth system.

Law and privacy (not legal advice)

Works council

An elected body that represents employees. In Austria and Germany it often has a say on staff monitoring.

First explained in: Decide what the evidence can resolve.

Personal data

Any information about an identified or identifiable person.

First explained in: Decide what to measure.

General Data Protection Regulation (GDPR)

The EU law on the processing of personal data.

First explained in: Decide what to measure.

Lawful basis

A legal ground that the GDPR lists for processing personal data, such as consent or contract.

First explained in: Decide what to measure.

Consent

A free, specific, informed and clear agreement by a person to the processing of their data.

First explained in: Decide what to measure.

Pseudonymisation

Replacing direct identifiers with codes, for example by hashing. The data is still personal data.

First explained in: Choose acquisition channels.

Profiling

Automated processing of personal data to judge or predict aspects of a person.

First explained in: Choose acquisition channels.

Data protection impact assessment (DPIA)

A written assessment of risks to people that the GDPR asks for before high-risk processing.

First explained in: Run the growth system.

EU AI Act

The EU law that sets duties for AI systems by their level of risk.

First explained in: Run the growth system.

Implementation

Data warehouse

A database built for analysis, into which every source system is copied.

First explained in: One possible implementation.

Raw layer

The first layer of the warehouse. It holds each source exactly as delivered.

First explained in: One possible implementation.

Clean layer

The second layer of the warehouse. It holds typed and deduplicated tables.

First explained in: One possible implementation.

Serving layer

The third layer of the warehouse. It holds the views that count.

First explained in: One possible implementation.

Load log

A table with one row for each source and each load.

First explained in: One possible implementation.

Parity test

A check that a tile gives the same number in the view and in the semantic model.

First explained in: One possible implementation.

Rule record

A rule written as data, so that a person can read and version it.

First explained in: One possible implementation.

Finding record

The stored form of a finding. Nobody edits it.

First explained in: One possible implementation.

Quantitative growth decisions

Quantitative growth

The practice of carrying a commercial choice through customer behaviour, observation, modelling, action and economic evaluation as one argument.

First explained in: Choose the growth objective.

Administrative censoring

Follow-up ends at a reporting cut-off before the event or competing outcome is observed. It is an observation limit rather than a terminal customer outcome.

First explained in: Activate customers.

Incremental contact gain

The expected discounted gross profit under contact minus that under no contact, less incremental contact expenditure. Its estimation needs causal evidence or explicit assumptions.

First explained in: Allocate sales attention.

Appendix A: One possible implementation

The chapters of this book describe a loop without naming a product. A loop still needs machines. This appendix shows one way to build them. It follows a prototype, the Funnel Diagnostic Tool, and the stack around it.

Read it as one set of choices, not as the right set. A company at low volume may run the same design in a spreadsheet and a few saved queries. The tables use generic roles. Lines that start "Implementation note:" name the products of the prototype. Notes that say "checked" hold facts from vendor documentation that we read on 15 October 2026. Vendor facts change, so check them again before you rely on them.

The appendix moves from the data to the decision, and it ends with the limits of the prototype. The start values that its rules load are in Appendix B, *Parameters and start values*, and its formulas are in Appendix D, *Formula sheet*.

The assumed stack

Table 1 lists the systems. Six are sources, and each decides the value of some metrics, as *Decide what to measure* explains. The last three rows decide nothing.

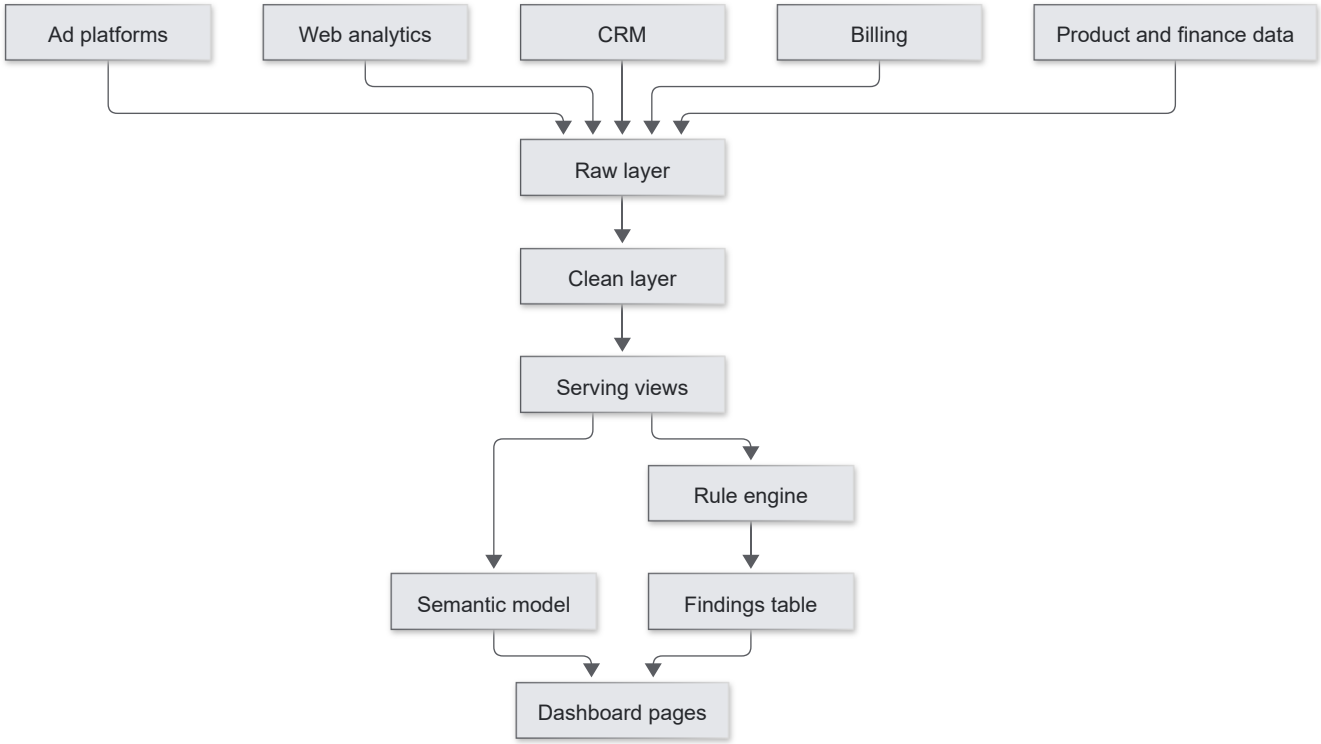
Table 1. The stack by role.

Role	It decides	How it lands
Ad platforms	Spend, impressions, clicks	One row per day and ad
Web analytics, with a raw export	Sessions and events of the visitors it recorded	One row per event
CRM	Leads, stage dates, deals, the amount at signing	Lead, deal and activity tables
Billing	Invoices, ARR and its movements, renewals	Billing lines and daily ARR snapshots
Product analytics	Use of the product after the sale	One row per customer, day and feature
Finance cost table	Every cost line, by function and motion	One row per month, function and motion
Warehouse	Nothing. It stores the copies and joins them.	All tables and views
Semantic model	Nothing. It divides sums and never originates a number.	Measures
Dashboards	Nothing. They show tiles.	Pages

A **data warehouse** is a database built for analysis. Every source is copied into it, so that the sources can be joined. A funnel is small data, and only the web event table grows fast. So choose the warehouse by the skills of the team, not by volume.

Implementation note: this design assumes Microsoft Azure and Power BI. The warehouse is Azure SQL Database, Microsoft Fabric or Azure Databricks. The semantic model and the dashboards are Power BI. The ad platforms are Google Ads, Meta Ads and LinkedIn Campaign Manager. Web analytics is Google Analytics 4,

which exports raw events only to Google BigQuery, so BigQuery stays as the landing zone and a copy job moves the flat rows onward. The CRM is HubSpot, Salesforce or Dynamics 365. Billing is Stripe or Chargebee. Product analytics is Amplitude or Mixpanel.



The data path. Counts flow through the views. The semantic model divides. The rule engine writes findings that the pages display.

Layers, the daily run and time

Three layers organise the warehouse. The **raw layer** holds each source exactly as delivered. It is append-only and nobody edits it, so any number can be rebuilt. The **clean layer** holds the tables of the next sections, typed and deduplicated, with times in UTC and tag values in lower case. The **serving layer** holds the views that count. Personal data stays in the clean layer and never enters the semantic model.

One pipeline runs the day in a fixed order. Each step starts only when the one before it succeeded.

1. Pull the ad platforms, the CRM and the web export. Re-pull recent days, because sources restate past days and late events arrive. A start is 28 days of ad spend and three days of web events (Plausible but untested).
2. Rebuild the clean tables. Write one row per source to the **load log**: source, load time, row count and latest date loaded.
3. Run the data tests of *Decide what to measure*. An error holds the data, and the tiles that depend on it go grey.
4. Build the serving views and the cohort tables from *Forecast conversion and timing*.
5. Run the rule engine. It writes metric values and findings.
6. Refresh the semantic model from the pipeline, never on its own clock. The model then cannot refresh on half-loaded data.

7. Send each finding to its owner through the exception lane of *Run the growth system*.

A source that is more than a day old greys out every tile that depends on it. A daily run also pays: a weekly read waits 3.5 days on average for a fault, and a daily run waits 0.5 days (Arithmetic).

Time zones

Store every event time in UTC and convert only when you read. Platforms set account zones in different ways (see the note below), so keep a small reference table, `source_account`, with the zone and the day boundary of each account. The views map each day to the reporting zone of the company.

Compare daily numbers across tools only through this mapping. Weekly and 28-day windows hide most of the effect.

Implementation note: checked 15 October 2026. A Google Ads standard account no longer supports time zone changes. The body of the Meta article on changes could not be read, so confirm in Ads Manager that a change opens a new account. LinkedIn states that its ad reports use UTC and is silent on an account zone. Google Analytics 4 chooses the zone at property creation, applies a change going forward only, and exports `event_date` in the property zone and `event_timestamp` in UTC.

The tracking plan

The **tracking plan** of *Decide what to measure* becomes a short table. Table 2 lists the events that the lower funnel needs. No tool reserves any of these names.

Table 2. Events the prototype requires.

Event	Fires when	Parameters	Feeds
`cta_click`	A call to action is clicked	button, page type	Click rate of calls to action
`form_first_input`	The visitor first touches a form field	form	Form start and abandonment
`form_success`	A form submits without error	form	Form completion
`pricing_view`	The pricing page or section is viewed	plan, optional	Intent signal
`scroll_depth`	The visitor passes a depth threshold	percent	Content depth
`lead_confirmed`	The CRM confirms the new lead	offer, lead type	Leads from the web, attribution coverage
`experiment_exposure`	A visitor meets the difference under test	experiment, variant	Counts per arm of a test

Fire the lead event only after the CRM confirms the record. Fire an exposure event at exposure, not at assignment. Register each parameter in the web analytics tool, or it cannot be reported. Store the consent state with every event, so that you can split the rows.

Implementation note: checked 15 October 2026. Google Analytics 4 sends `form_start` and `form_submit` as automatic form events, and its list of reserved names holds both. Its pages do not state what a custom event with a reserved name does, so the prototype uses distinct names. The BigQuery export holds cookieless pings

and never holds modelled data, and the consent fields sit in `privacy_info`. One point is open: Google says each cookieless session gets a different `user_pseudo_id`, and third parties say the field is null. Count the null rows in your own export.

The core tables

Six tables are enough to compute the funnel up to the first invoice. Times end in `_utc`. Every ratio divides two stored counts. A zero denominator returns a blank (NULL), never a zero. A zero reads as a broken funnel, and a blank as a broken tag.

Table 3. The six core tables.

Table	Grain	Main columns
`ad_spend_daily`	Day, platform, account, campaign, ad set, ad	Spend, impressions, clicks, reach and video views (nullable), planned spend, campaign type
`web_events`	One event	Time, user key, session key, source, medium, campaign, content, term, page, event name, parameters, engagement time, consent state
`crm_leads`	One lead	Created at, source, tag values, lifecycle stage, lead score, MQL time, rejection time and reason, disqualification time and reason, ICP match, motion
`crm_deals`	One deal	Lead key, account key, created at, SQL time, opportunity time, won and lost times, amount, list amount, stage, owning team
`crm_activities`	One activity	Lead key, type (call, email, meeting), created at, human flag
`billing_lines`	One contract line and billing period	Contract, customer, product, period start and end, earned amount, revenue class, billing term, committed minimum

Four columns need a rule. Count MQLs on the MQL time, and never infer it from the current stage. Count opportunities on the time the deal passed the opportunity test, not on the creation time. Most CRMs create the deal row at the SQL. Flag automated email as not human, so that it does not count as contact. Classify each billing line as recurring, re-occurring or one-off, as *Model recurring growth* sets out.

The deal table carries a team or a role, never the key of one seller. Results go out by role. Counsel and the works council must review before any seller-level view exists. This is not legal advice.

Extension tables

Add an extension table when a chapter needs it. The first four are not optional. Build them first, because history that was never recorded cannot be rebuilt.

Table 4. Extension tables.

Table	Grain	Source role	Read by
stage_event	One item entering a stage; append-only	CRM events	All cohort rates
deal_snapshot_daily	One open deal and one day	CRM	Pipeline age, as-of views
account_map	One account key with its CRM, billing and workspace records, and a match status	Built by the analytics lead	Customer counts, traceable ARR
change_event	One dated change, the metrics it touches, its expected size	Tags, ad accounts, CRM, price list	Detection
load_log	One source and one load	Pipeline	Freshness, greying
source_account	One account, its zone and day boundary	Reference	The day mapping
arr_snapshot	One customer and one date: ARR, price part, motion	Billing	ARR bridge, retention
arr_movement	One movement: new, returning, expansion, contraction or churn, with reason and price part	Billing	The bridge and its reasons
subscription	One subscription: signed, go-live, first impact, notice, renewal date, term, renewed	Billing, CRM, product	Activation, renewal
cost_allocation_monthly	Month, function, motion: people cost, tools cost, allocation rule	Finance cost table	Fully loaded CAC, cost to serve
motion_dim	One motion: touch level, segment, product	Finance with the data team	Every card split by motion
product_usage_daily	Customer, day, feature	Product analytics	Active accounts, adoption
account_health_weekly	Customer and week: family scores, band, families scored	Joined sources	Accounts at risk
customer_owner	Customer, manager role, start and end dates	Success records	ARR per manager

Two rules guard the billing tables. The sum of ARR by motion on a date equals total ARR on that date, and a missing motion is a data fault. Starting ARR plus the five movements equals ending ARR to the euro. When the identity fails, raise the reconciliation flag and do not insert a balancing figure.

Implementation note: small reference tables, such as planned spend and the cost table, are Excel files that the pipeline loads. Each has a named owner and a last-edited date.

The views

Views count, and the semantic model divides. Fifteen views serve the funnel up to the first invoice, in four layers: the base tables, staging without counting, counting at a fixed grain, and combining. Table 5 groups the views.

Table 5. The fifteen views.

View	Layer	What it does
Flat web events	Staging	Turns the nested raw export into flat rows
Paid traffic rules	Staging	One list of paid media and sources, edited in one place, medium first
Sessions	Staging	One row per session, keyed on user and session so that a session across midnight stays one
Paid media daily	Counting	Delivery and cost from the platforms
Paid and qualified sessions	Counting	Sessions from paid traffic, with one session rule set once
Leads, MQLs, SQLs and customers daily	Counting	Four views. Each counts on its own event time. Customers carry revenue at signing.
Junk leads daily	Counting	Leads disqualified inside the junk window for a junk reason
Speed to lead daily	Counting	Time to the first human touch, and the contact rate
Unified daily funnel	Combining	Every daily count and cost in one row, and the target of the parity test
Stage feeds	Combining	Three views, one for each third of the funnel, with only the metrics of that stage

The three stage feeds are thin wrappers that keep each metric off the wrong page. Same-period rates in the feeds serve volume monitoring. The rules read the cohort versions.

Three more views serve the page after the close: the movement of each customer in each period; the five movements of each month, with a gap column that must be zero; and retention by month, with starting ARR, ARR kept after capping each customer at its starting value, ending ARR of the starting customers, and logo counts. None of them divides. The measures compute gross revenue retention, net revenue retention and logo retention from sums on the selection of the reader.

Implementation note: the prototype names these objects `vw_ga4_events_flat`, `vw_dashboard_tof_daily` and so on, in a schema called `funnel`.

The semantic model

The semantic model loads the views and divides. A ratio stored per day cannot be added up again. A reader who picks 28 days and one campaign needs spend and leads summed over that selection, then divided. Only a measure runs on the selection.

Three rules keep the model honest. Every measure traces to columns of the views, and every column traces to one source. No table typed into the model holds a metric value. The loading step holds no business logic, because a filter hidden there is a definition that nobody reviews.

A window selector and a comparison switch give the current window, the earlier window and the change. Anchor the window on the last loaded date from the load log, never on today.

One reconciliation page lists a few measures three ways: the measure for the last 28 days, the same figure from the stored metric values, and the difference. This is the parity test. Counts must match exactly. Ratios must match to the last decimal shown. Usual causes of a gap are dates converted in one place only, a relationship that filters too much, and a refresh that ran before the load ended.

Implementation note: in Power BI the formulas are DAX measures, the window switch is a calculation group, and the loading step is Power Query. Import mode suits this volume. Direct Lake reads Delta tables and not SQL views, so on Fabric the serving views become tables first. We did not re-check storage modes in October 2026, so read the current documentation before you choose.

The six dashboard pages

Each page has a reader, a rhythm and a decision. Every tile prints four things: the value, its interval, the n behind it and its limit status, as *Decide what to measure* sets out. A tile that fails a gate goes grey and states why. Never hide a tile, because a missing tile reads as a metric that does not exist.

Table 6. The pages.

Page	Decision it serves	Reader and rhythm	Main tiles
Executive	Where to look and what to stop	Bowtie owner, monthly	One row per bowtie stage, gate status, readability, ARR bridge with reconciliation flag, recurring share
Top of the funnel	Spend, creative, landing pages	Marketing owner, weekly	Spend and clicks by channel, click-to-lead on one population, click-through spread by creative, consent capture ratio beside every web number
Middle of the funnel	Qualification and nurture	Marketing and sales owners, weekly	MQLs, MQL-to-SQL on a cohort basis, lead score distribution, junk-lead share, rejection reasons
Bottom of the funnel	Capacity and deal review	Sales owner, weekly	SQLs, uncontacted backlog, seller load, speed to lead, opportunity and win rate on a cohort basis, sales cycle at the 80th percentile
Data and sources	Release or hold the data	Analytics lead, daily	Freshness, test pass rate, reconciliation gaps, tracking and attribution coverage, change events, the metric list with primary and secondary sources
After the close	Onboarding, renewal, expansion	Customer success owner, weekly and monthly	Activation, time to activation, active accounts by touch level, accounts at risk, renewal calendar, expansion by reason

Each tile opens a detail pane with the definition, the numerator, the denominator, the source, the owner and a link to the rows. A change-event strip sits under every chart that spans a change. A findings panel lists each finding with its evidence, sorted by level. It keeps suppressed findings visible and greyed, with the reason. A panel with only the top finding hides the logic that earns trust.

Alerts belong in the pipeline. A message from the findings table carries the diagnosis, not only a number that crossed a line.

Implementation note: in Power BI the pages are one report, each with its own display folder of measures. The change-event strip is a combined line and column chart, because Power BI has no native annotation layer on a time axis. A Fabric Activator rule or Power Automate sends the message from the findings table. The prototype draws the same pages in a web page and not in Power BI.

The records

Four records carry the loop. A person reads each as a table, and a program reads each as a row.

The metric specification

The specification of *Decide what to measure* is stored as one row per version. Its fields are the name, numerator, denominator, population, window, grain, owner, source, minimum sample, limit model, version and the date from which the version holds. Versions are never overwritten. A change adds a row.

The rule record

A **rule record** is a rule written as data, so that a person can read and version it. Table 7 lists the fields.

Table 7. The fields of a rule record.

Field	Meaning
Name	The rule, in words, and its stage
Metric	The specification it watches, with its version
Kind	Gate or monitor
Expectation	Level, season and the change events that move it
Monitor	Smoothing weight, overdispersion factor and the limit tuned to the false-alarm budget
Levels	The watch limit and the act limit
Minimum sample	Below it, the metric reads "too few to read"
Depends on	The gates that must pass first
Suppressed by	Rules and gates that make this rule meaningless
Check to run next	The cheap check that tells causes apart
Owner, deadline	A role and the days allowed
Version, reason, review date	Who may change it, why, and when to look again

The prototype holds a fixed percentage change per rule, which fails for the reasons that *Decide when a change needs action* gives. A rule record holds an expectation and a calibrated limit instead, set from the false-alarm budget and the overdispersion factor.

The finding record

A **finding record** is one alert after review, as in *Decide when a change needs action* and *Diagnose what changed*. The engine writes it, and nobody edits it. Table 8 lists the fields.

Table 8. The fields of a finding record.

Field	Meaning
Identifier	A hash of the rule version, the window and the inputs, so a replay gives the same identifier
Run date and window	When it fired and which window it reads
Metric and size	The direction, the size, the z-score and the monitor score
Level	None, watch or act
Evidence	Gate states, change events nearby, other metrics that moved
Suppressed by	Empty, or the rule or gate that explains it away
Likely cause	The leading candidate, with the next check
Evidence tier	Observation, checks passed or experiment
Owner, deadline	A role and a date
Status	Open, acknowledged, decided or past deadline

One shared cause makes one finding, not one for each metric. A decision past its deadline becomes a finding of its own and moves up one level.

The decision card and the log

The decision card is the output of the Monte Carlo of *Commit with uncertainty*. Store it with its inputs, so that you can rerun it.

Table 9. The fields of a stored decision card.

Field	Meaning
Question and options	Always including "do nothing"
Inputs	Each with a distribution, an evidence label, a source and an owner
Seed and draws	So that the card reproduces
Outputs	Mean, P10 and P90, chance of loss, chance of best, break-even effect, switching values, value of perfect information
Constraints	Payback, capacity and the risk limit, with the chance of breach
Change-impact lines	One line for each option
Threshold	90% for a change that can be reversed, 95% for one that cannot
Owner, deadline	A role and a date

The decision log holds two tables that nobody edits. The decision table holds the question, the belief, its interval, the options, the falsifier, the confirms-by date and the owner as a role. The store stamps the time. The outcome table is written once, later, with the class confirmed, refuted, inconclusive or void. The companion code holds both as SQL, with views for the decisions that are due and for the graded ones.

What executes the rules: five layers

A dashboard cannot be the engine. A rule has an order, gates first and suppression after, which a formula in a visual expresses badly. An alert must fire when nobody opens the report. A stored finding can be audited months later, but a measure leaves no record. A procedure can be tested on known cases.

So the work has five layers, each with one job and its own test.

1. **Rules as data.** Cards, rules, parameters and actions sit in a versioned store, each with a version, an owner, a reason and a review date.
2. **A deterministic engine.** A scheduled job applies the rules to the metric values of each window and writes hashed findings. The same input gives the same output.
3. **A statistical layer.** One library computes intervals, pooling, minimum samples and grades. The engine calls it and never reimplements it.
4. **A language layer.** An AI model drafts the weekly memo and the text of a finding from stored results. It computes nothing, as *Run the growth system* requires.
5. **A human gate.** A person approves, and the approval writes the log.

Formulas live once, in the views and the semantic model. The parity test asserts on every load that each tile gives the same number in the view and in the measure.

Implementation note: this design assumes tables in Azure SQL Database, an engine written in Python that runs as a Fabric pipeline or on a timer, Power BI measures as the parity oracle, and the rule files in a Git repository.

One pass per window

The engine runs the same pass on every window of 28 days. It compares the window with the expectation, not with the window before.

1. Load the parameters. Compute each metric for the window and its expectation. Mark any rate under its minimum sample as too few to read.
2. Evaluate the gates in order: data quality, tracking, volume. For each failed gate, mark the dependent metrics unreadable and suppress their rules.
3. Evaluate the monitors, stage by stage. Apply change events, so that a known change moves the expectation and raises no alert.
4. Merge findings that share a cause. Apply suppression. Keep the suppressed findings, with the reason.
5. Attach to each open finding the check that tells its causes apart, as *Diagnose what changed* teaches.
6. Write the metric values and the hashed findings.
7. Hand the findings to the memo writer and the exception lane. A person decides.

8. On the next window, read every decision that is due and record its outcome, as *Learn from outcomes* describes.

The prototype lists findings in stage order. Treat that as a display order only: the card ranks the work by expected profit and value of information, as *Fund the next best action* shows.

Dialect notes

SQL differs between engines in a few places that break queries silently. Write the views in plain SQL where you can. Table 10 lists the traps.

Table 10. SQL constructs that differ between engines.

Construct	The trap
Safe division	Without a guard, a zero denominator raises an error. Without a decimal factor, two integers divide to an integer.
Local date from UTC	Declare the stored value as UTC first, then convert. A missing first step shifts nothing and gives UTC dates.
Difference in minutes	A boundary-counting function counts clock boundaries, not elapsed minutes. Count seconds and divide.
Alias in grouping	Some engines reject a select alias in a group clause. Compute it in a common table expression first.
Text comparison	Collation decides case. Lower-case both sides of every tag comparison.
Nested fields	Flat tables cannot unnest. Flatten before the data reaches the engine.

Implementation note: for the Azure engines, use `a * 1.0 / NULLIF(b, 0)` for safe division, in Databricks SQL `try_divide(a, b)`. In T-SQL the local date is `CAST(ts AT TIME ZONE 'UTC' AT TIME ZONE 'W. Europe Standard Time' AS date)`, and in Databricks SQL it is `to_date(from_utc_timestamp(ts, 'Europe/Vienna'))`. T-SQL uses Windows zone names, and Databricks uses IANA names. A Fabric warehouse cannot store `nvarchar`, `datetime` or `datetimeoffset` columns, so store UTC in `datetime2` (checked 15 October 2026). It also does not enforce keys (from the earlier manual; not re-checked), so the data tests check uniqueness themselves. The companion code marks its two dialect-specific expressions with macros.

What the prototype does and does not do

The Funnel Diagnostic Tool is a working web page. It models a fictional mid-market software company and plants one fault at a time. It carries 92 metric cards, 30 findings (3 gates and 27 stage findings), 20 scenarios and 15 views. A severity control scales a fault from none to full, so that a reader can watch a limit trip. A diagnosis rail lists the findings, gates first, and keeps overruled findings visible. A failed tracking gate turns the tiles below it grey.

It does not yet do the following.

- **Read real data.** It runs on a simulated company. Several rules read simulated fields that real data lacks: creative history, audience overlap, form drop-off and retargeting conversions by week. A real build needs those feeds, and no feed for one seller.

- **Store findings.** The page computes them in the browser and has no store. No finding can be audited later, and no alert fires when nobody looks.
- **Detect against an expectation.** It compares the window with the window before, with fixed limits. It has no change events, no seasons and no tuned false-alarm budget.
- **Read cohorts.** Some rates divide same-period counts, because the page holds no stage history.
- **Value an action.** It has no decision card, no value of information, no decision log and no grading.

Use the prototype as a teaching instrument and a layout model. Build in the order of *Run the growth system*. Start with one stage end to end: a specification, a gate, a hand-run detection, one card and one log entry. Add the nightly engine next, and the memo writer last.

Key points

- Each source system decides some metrics. The warehouse, the semantic model and the dashboards decide none.
- Store times in UTC, keep a table of account time zones, and test the clock changes.
- Build the stage history and the account map first. Nobody can rebuild them.
- Views count, the semantic model divides, and a parity test proves they agree.
- Rules are data. A deterministic engine applies them and stores hashed findings. A model drafts text. A person decides.
- The prototype shows the layout. It does not yet read real data, store findings or value an action.

Appendix B: Parameters and start values

A parameter is an input that the company sets and the data do not fix. This appendix lists the start values of the chapters, with their labels. Each company tunes its own values. A change goes to its owner, named in the decision rights of *Run the growth system*, and into the decision log. The last column names the role that sets the value. Appendix D, *Formula sheet*, holds the formulas. Appendix E, *Sources and further reading*, lists the sources.

Value and money

Parameter	Start value	Label	Chapter	Set by
Discount rate	12% a year; range 8% to 20%	Standard practice for the method; Plausible but untested for the number	Value customers and cash	Finance owner
Horizon	5 years; sensitivity 3 to 5	Plausible but untested	Value customers and cash	Finance owner
Payback ceiling	12 months small business, 18 mid-market, 24 enterprise; also the scalability limit in *Set targets and staff capacity*	Standard practice; Plausible but untested at the margin	Value customers and cash	Finance owner
Expansion cost ceiling	0.50 per 1.00 of expansion ARR; good level 0.40 or below	Plausible but untested; Standard practice for the good level	Expand accounts and generate referrals	Finance owner
Recurring-share band	3 points; after 12 periods, 2 standard deviations of own change, floor 1 point	Plausible but untested	Model recurring growth	Analytics lead
Cost of one work day	Yearly analyst cost ÷ 220 working days; a test takes 10 analyst days	Plausible but untested	Choose whether to test or act	Finance owner
Benefit period of a change	1 year (52 weeks)	Plausible but untested	Choose whether to test or act	Analytics lead
Hurdle	10 times the pay of one person-week (yearly pay ÷ 52)	Plausible but untested	Fund the next best action	Decision owner
Risk limit	2% of yearly gross profit, per decision, before its first signpost	Plausible but untested	Commit with uncertainty	Finance owner

Decisions and evidence

Parameter	Start value	Label	Chapter	Set by
Shipping thresholds	90% chance of positive net value if reversible; 95% if not	Plausible but untested	Commit with uncertainty	Analytics lead, with the finance owner

Parameter	Start value	Label	Chapter	Set by
Own prior	Outside base rates until the register holds about 30 tests	Plausible but untested	Commit with uncertainty	Analytics lead
Decision deadline	5 working days	Plausible but untested	Commit with uncertainty	Decision owner
Holdback share	10%	Illustration only	Choose whether to test or act	Decision owner
Budget pulse	Up to 30% per channel, for 2 weeks	Plausible but untested	Choose acquisition channels	Marketing owner
Work in progress limit	2 items at once for Alder's pair	Illustration only	Fund the next best action	Decision owner

Decision owner: the role that owns the card (*The decision card and the decision log*).

Time, volume and detection

Parameter	Start value	Label	Chapter	Set by
Window, quarter, year	Window 28 days; quarter 91 days (13 weeks, 3.25 windows); year 13 windows (364 days)	Arithmetic	Decide what the evidence can resolve	Analytics lead
Volume regimes	Days to read a 30% change: high 28 or fewer, medium up to 91, low more than 91	Arithmetic for the days; Plausible but untested for 30% and the limits	Decide what the evidence can resolve	Analytics lead
Overdispersion factor	1.2, inside 1 to 2, never below 1; estimate from 13 or more windows	Plausible but untested	Decide when a change needs action	Analytics lead
False-alarm budget	10% a year per act limit; 50% per watch limit; one false alarm a quarter for the list	Plausible but untested	Decide when a change needs action	Analytics lead
Alert limit	Not typed: the value that the yearly maximum of noise-only years passes in 10% of years	Plausible but untested, through the budget	Decide when a change needs action	Analytics lead
Detector weights	EWMA weight 0.2; CUSUM allowance 0.5	Standard practice for the methods; Plausible but untested for the values	Decide when a change needs action	Analytics lead

Checks and agents

Parameter	Start value	Label	Chapter	Set by
Calibration check	About 50 resolved entries read a large gap; 70 read 15 points, 160 read 9 points (80% power)	Arithmetic	Learn from outcomes	Analytics lead

Parameter	Start value	Label	Chapter	Set by
Agent gates	At least two quarters in advisory; agreement at least 80%; kappa at least 0.60; abstain rate at most 15%	Plausible but untested	Run the growth system	Analytics lead
Yield test	Retire a meeting that changed no decision in 4 sessions	Plausible but untested	Run the growth system	Bowtie owner

Appendix C: The three companies: facts and timeline

Three invented companies run through this book: Alder, Birch and Cedar. They sit at low, medium and high volume, so every chapter can test its method at each volume and say where it holds. *How to read this book* introduces them. This appendix holds their facts and the timeline of every dated event.

A value with a published source carries that source in the chapter that uses it; a value without one is an illustration. All money is in euro. Benchmarks in US dollars are compared without conversion, so each comparison is approximate.

One source for every number

Every number about the companies comes from one shared set of facts, written as code. A script for each chapter recomputes every number that the chapter prints, and a script writes the tables below. Table 1 lists the shared conventions.

Table 1. The conventions of the shared facts.

Convention	Value
Window	28 days
Year	13 windows
Quarter	91 days, 3.25 windows
Month	1.08 windows of flow
Calendar	30 months, Year 1 January to Year 3 June
Calendar month in the facts engine	365.25 / 12 days; a window is 28 days, so a year of 13 windows is 364 days (Arithmetic)
Regime test: change to read	30% relative change in the last rate before revenue
Regime test: power and significance	80% power, 5% two-sided
High volume	28 days of flow or fewer
Medium volume	more than 28 and up to 91 days
Low volume	more than 91 days
Barometer: change to read	20% relative change
Barometer: periods	a window = 28 days, a quarter = 91 days, a year = 364 days
Barometer: derived typical	23.4% of opening ARR as new-logo ARR a year
Value basis: discount rate	12% a year (range 8% to 20%)
Value basis: horizon	5 years

How to read the barometer

The barometer is one sentence. It names the company and the date, and gives the new customers in a window against a derived typical. It names the volume regime and the deepest stage rate that reads a 20% change in the period that fits the regime: a window, a quarter or a year.

The derived typical is a reference, not a target: the new customers that a company of this size and contract value would win if its new-logo ARR were the share of opening ARR in Table 1, a share from published reports. The sentence prints the count for that month, season included. The regime, the ratio and the deepest rate use the trend level, so that a label does not change with the season.

The volume regime of each company

Count the days of flow that a company needs to read a 30% change in its last rate before revenue, by the rule of *Decide what the evidence can resolve*. Volume is high at 28 days or fewer, medium up to one quarter and low above it.

Table 2. The volume regime of each company at Year 1 January, at the trend level.

Company	Regime	Last rate before revenue	Days to read a 30% change	New customers a window	Derived typical	Times the typical	Medium from (new customers a window)	High from
Alder	Low	win rate (opportunity to won)	1,161	3.3	3.6	0.93	42.7	139
Birch	Medium	win rate (opportunity to won)	75	45.1	41.2	1.09	37.2	121
Cedar	High	activated-to-paid rate	3	1,383	811	1.70	49.5	161

Table 3. The deepest rate that reads a 20% change, and the barometer detail, at Year 1 January.

Company	Within a window	Within a quarter	Within a year	Detail
Alder	none	none	none	The win rate needs about 3.2 years to read even a 30% change, and the lead count reads a 20% change in 27 days.
Birch	lead-to-MQL rate	SQL-to-opportunity rate	win rate (opportunity to won)	The win rate reads a 30% change in 75 days, and the close rate reads a 20% change in about 7 months.
Cedar	activated-to-paid rate	activated-to-paid rate	activated-to-paid rate	Trial to paid reads a 10% change in 32 days; the assisted motion, with 72 opportunities a window, needs about 8 months to read a 30% change.

Alder is low volume, Birch medium and Cedar high, so one method gives different answers for each. Table 3 also shows that Cedar holds two regimes: its assisted motion reads far more slowly than its self-serve motion.

The profiles

Each profile describes the company at Year 1 January, unless a row says otherwise. Retention follows the customers open at the start for the next 12 months. Unit economics follow *Value customers and cash*.

Table 4. Profile of Alder, low volume.

Item	Value
Arc of the story	learning to decide well without statistical significance
Sells	quality-management software for manufacturing plants: inspection plans, deviations, audits and supplier quality
Segment	enterprise and upper mid-market manufacturers
Motion	sales-led and high touch: account-based marketing, trade fairs and events, implementation partners
Contract and billing	platform fee plus site licences; annual contracts (75% of customers) and three-year contracts (25%); billed annually in advance
Countries	6: Germany, Austria, Czechia, Poland, Italy, Slovenia
ARR at the start (euro)	12,000,000
Paying customers at the start	190
Largest account	1,320,000 euro ARR (11%): a plant group with 24 licensed sites, Alder's largest account
Motion: account-based, core vertical	engaged accounts, opportunities, new customers; new contract 60,000 euro; cycle 120 days (median), 191 days (80th percentile)
New customers a window, trend	3.3
Times the derived typical	0.93
Volume regime	Low: 1,161 days of flow to read a 30% change in the win rate (opportunity to won)
Gross revenue retention	92.8%
Net revenue retention	107.5%
Logo retention	93.3%
Gross margin	79.9%
Sales and marketing, share of revenue	33.3%
New-customer CAC ratio	1.33
CAC payback	20.0 months
Expansion cost per euro of expansion ARR (Year 1 January)	0.24
Value of a new customer (euro)	220,768

Item	Value
ARR growth, Year 1	30.9%
ARR growth, Year 2	26.6%
ARR growth, Year 3 January to June (not annualised)	8.0%
ARR at the end of Year 3 June (euro)	21,479,921
Customers at the end of Year 3 June	271
Team at the start	account executives 4, head of sales 1, solution engineers 3, marketers 1, analysts 0.50, customer success managers 3, support staff 4, onboarding consultants 5
Sales capacity in use (trend)	69%

Table 5. Profile of Birch, medium volume.

Item	Value
Arc of the story	from reports and arguments to a priced decision loop
Sells	finance-team software for the monthly close and spend control
Segment	mid-market finance teams (small-deal sizes, inside sales)
Motion	sales-assisted inbound: demo request, sales development, account executive; paid media through professional-network ads, search ads and social ads
Contract and billing	annual contracts, billed annually in advance, auto-renewing
Countries	9: Germany, Austria, Netherlands, Belgium, Ireland, France, Spain, Italy, Nordic countries
ARR at the start (euro)	22,000,000
Paying customers at the start	2,250
Motion: sales-assisted inbound	leads, MQLs, SQLs, opportunities, new customers; new contract 9,600 euro; cycle 34 days (median), 56 days (80th percentile)
New customers a window, trend	45.1
Times the derived typical	1.09
Volume regime	Medium: 75 days of flow to read a 30% change in the win rate (opportunity to won)
Gross revenue retention	89.9%
Net revenue retention	102.1%
Logo retention	90.0%
Gross margin	80.5%
Sales and marketing, share of revenue	39.8%
New-customer CAC ratio	1.29
CAC payback	19.2 months

Item	Value
Expansion cost per euro of expansion ARR (Year 1 January)	0.28
Value of a new customer (euro)	31,538
ARR growth, Year 1	31.4%
ARR growth, Year 2	26.3%
ARR growth, Year 3 January to June (not annualised)	14.0%
ARR at the end of Year 3 June (euro)	41,635,836
Customers at the end of Year 3 June	3,378
Team at the start	account executives 10, sales development representatives 4, sales leadership and operations 4, solution engineers 3, marketers 11, customer success managers 15, account managers 4, analytics lead 1, support staff 10, onboarding specialists 6
Sales capacity in use (trend)	94%

Table 6. Profile of Cedar, high volume.

Item	Value
Arc of the story	high volume is not high certainty
Sells	a team scheduling and collaboration tool for small and medium businesses
Segment	small and medium businesses, self-serve; larger accounts with sales assistance
Motion	product-led: freemium plus a 14-day trial of paid features that every sign-up starts, self-serve monthly and annual plans; a sales-assist team for larger accounts
Contract and billing	self-serve monthly and annual plans paid by card; assisted deals on annual contracts billed in advance
Countries	24 European countries (the list is in the shared facts)
ARR at the start (euro)	37,000,000
Paying customers at the start	26,640
Motion: self-serve, product-led	visitors, sign-ups, activated sign-ups, new paying accounts; new contract 821 euro; cycle 9 days (median), 19 days (80th percentile)
Motion: sales-assisted	product-qualified accounts, opportunities, new customers; new contract 16,000 euro; cycle 45 days (median), 71 days (80th percentile)
New customers a window, trend	1,383
Times the derived typical	1.70
Volume regime	High: 3 days of flow to read a 30% change in the activated-to-paid rate
Gross revenue retention	76.1%
Net revenue retention	92.1%
Logo retention	75.0%

Item	Value
Gross margin	81.3%
Sales and marketing, share of revenue	34.8%
New-customer CAC ratio	0.69
CAC payback	10.2 months
Expansion cost per euro of expansion ARR (Year 1 January)	0.10
Value of a new customer (euro)	1,683
ARR growth, Year 1	35.7%
ARR growth, Year 2	30.3%
ARR growth, Year 3 January to June (not annualised)	14.0%
ARR at the end of Year 3 June (euro)	74,566,833
Customers at the end of Year 3 June	51,937
Team at the start	growth team (product, marketing, data) 12, account executives 5, assisted-sales manager 1, solution engineers 1, analysts 2, support staff 22, customer success managers 3
Sales capacity in use (trend)	80%

The barometer strip

Table 7 shows the barometer of each company at the five months that all three share and at the last month that a chapter uses. For Cedar, a self-serve company, a new customer is a new paying account. No regime changes. The position against the typical falls for all three, because the typical grows with ARR faster than the count of new customers.

Table 7. The barometer sentence of each company at six months that the chapters use.

Barometer sentence	Chapters with an example at that month
Alder, Year 1 January: about 3.2 new customers a window, 0.9 times the derived typical; low volume: no rate below the lead reads a 20% change within a year.	Choose the growth objective
Alder, Year 1 September: about 3.6 new customers a window, 0.8 times the derived typical; low volume: no rate below the lead reads a 20% change within a year.	Decide what the evidence can resolve
Alder, Year 2 February: about 3.5 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year.	Commit with uncertainty
Alder, Year 2 July: about 3.5 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year.	Fund the next best action
Alder, Year 2 December: about 3.1 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year.	Retain customers profitably

Barometer sentence	Chapters with an example at that month
Alder, Year 3 April: about 4.1 new customers a window, 0.7 times the derived typical; low volume: no rate below the lead reads a 20% change within a year.	Run the growth system
Birch, Year 1 January: about 47 new customers a window, 1.1 times the derived typical; medium volume: a quarter reads a 20% change down to the SQL to opportunity rate.	Choose the growth objective
Birch, Year 1 September: about 48 new customers a window, 1.0 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate.	Decide what the evidence can resolve
Birch, Year 2 February: about 51 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate.	Commit with uncertainty
Birch, Year 2 July: about 51 new customers a window, 0.8 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate.	Fund the next best action
Birch, Year 2 December: about 62 new customers a window, 0.9 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate.	Set prices and packages
Birch, Year 3 April: about 65 new customers a window, 0.8 times the derived typical; medium volume: a quarter reads a 20% change down to the sales-accepted to opportunity rate.	Run the growth system
Cedar, Year 1 January: about 1,600 new paying accounts a window, 1.7 times the derived typical; high volume: every self-serve rate reads a 20% change within a window.	Choose the growth objective
Cedar, Year 1 September: about 1,600 new paying accounts a window, 1.4 times the derived typical; high volume: every self-serve rate reads a 20% change within a window.	Decide what the evidence can resolve
Cedar, Year 2 February: about 1,700 new paying accounts a window, 1.3 times the derived typical; high volume: every self-serve rate reads a 20% change within a window.	Commit with uncertainty
Cedar, Year 2 July: about 1,500 new paying accounts a window, 1.4 times the derived typical; high volume: every self-serve rate reads a 20% change within a window.	Fund the next best action
Cedar, Year 2 December: about 1,800 new paying accounts a window, 1.3 times the derived typical; high volume: every self-serve rate reads a 20% change within a window.	Retain customers profitably
Cedar, Year 3 May: about 2,000 new paying accounts a window, 1.2 times the derived typical; high volume: every self-serve rate reads a 20% change within a window.	Run the growth system

The timeline

Table 8 lists every dated event in date order, with the chapter that tells it. Some events have no example of their own: they are the quiet results of an earlier decision, such as a seller who ramps up after a hire. Read the timeline as three threads.

Alder. Three customer counts open the story. One lost deal looks like a fall in the win rate. A second vertical opens in steps and is graded later. A champion leaves, and a written rhythm closes the story.

Birch. Sales and marketing blame each other. A decision card puts qualification first, with a written hiring trigger. Six budget pulses read the response of leads to spend. The trigger trips twice, and the owners grade their own forecasts.

Cedar. Sign-ups rise while net revenue stays flat. One country falls after a display change. A pricing-page winner leads to an onboarding test and a geo test. Save offers go only where they change the outcome, and standing meetings retire.

Table 8. The shared timeline: every dated event, in date order.

Date	Company	Event	Chapter
Year 1 January	Alder	**Three customer counts.** Each team reports its own count.	Choose the growth objective
Year 1 January	Birch	**Sales is slow, leads are bad.** The MQL differs between systems.	Choose the growth objective
Year 1 January	Cedar	**Sign-ups up, net revenue flat.** Many experiments a month.	Choose the growth objective
Year 1 February	Alder	**Three accounts drive the expansion.**	Model recurring growth
Year 1 March	Birch	**The 10-year lifetime.** Retention becomes survival by renewal.	Model recurring growth
Year 1 March	Cedar	**Annual discount re-priced.** Priced on the real hazard shape.	Model recurring growth
Year 1 April	Alder	**Payback on a handful of deals.**	Value customers and cash
Year 1 April	Birch	**Paid efficiency looks good.** Paid spend over all customers.	Value customers and cash
Year 1 April	Cedar	**What belongs in gross margin.** Support and free-user hosting.	Value customers and cash
Year 1 May	Alder	**The account is the unit.**	Choose markets and accounts
Year 1 May	Birch	**A sales-accepted stage.**	Choose markets and accounts
Year 1 May	Cedar	**Two motions.**	Choose markets and accounts
Year 1 June	Alder	**Deals slip.**	Forecast conversion and timing
Year 1 June to July	Birch	**A campaign lifts SQLs.** The same-period close rate falls.	Forecast conversion and timing
Year 1 June	Cedar	**An outage week.** A cohort effect in the grid.	Forecast conversion and timing
Year 1 July	Alder	**Plus 40% and two more sellers.**	Set targets and staff capacity
Year 1 July	Birch	**A fixed band raises alarms.**	Set targets and staff capacity
Year 1 July	Cedar	**Home market slows.**	Set targets and staff capacity
Year 1 August	Alder	**Leads or expansion.** Expansion is worth more.	Find the most valuable constraint
Year 1 August	Birch	**Top-of-funnel levers lead.** Seller capacity is the next limit.	Find the most valuable constraint

Date	Company	Event	Chapter
Year 1 August	Cedar	**Cheap sign-ups.** Worth far less than average.	Find the most valuable constraint
Year 1 September	Alder	**Ranking four sellers.**	Decide what the evidence can resolve
Year 1 September	Birch	**A demo script test.**	Decide what the evidence can resolve
Year 1 September to December	Cedar	**Sixth assisted seller in ramp.**	Set targets and staff capacity
Year 1 September	Cedar	**Intervals too narrow.** Users cluster in workspaces.	Decide what the evidence can resolve
Year 1 October	Alder	**Records go live.**	Decide what to measure
Year 1 October	Birch	**Consent banner redesign.** Web analytics and CRM disagree.	Decide what to measure
Year 1 October	Cedar	**Nine dashboards to three.**	Decide what to measure
Year 1 November	Alder	**One lost deal.** The win rate looks lower.	Decide when a change needs action
Year 1 November	Birch	**Calibrated detection.**	Decide when a change needs action
Year 1 November	Cedar	**A tag release breaks tracking.** The daily gate catches it.	Decide when a change needs action
Year 1 December to Year 2 January	Birch	**Outbound pilot.**	Diagnose what changed
Year 2 January	Alder	**Fewer wins last quarter.**	Diagnose what changed
Year 2 January	Birch	**The close rate falls.** Split into mix and rate.	Diagnose what changed
Year 2 January	Cedar	**Sixth assisted seller at full capacity.**	Set targets and staff capacity
Year 2 January	Cedar	**One country falls.** Placebo countries do not move.	Diagnose what changed
Year 2 February	Alder	**A second vertical.** Food and beverage, a staged entry.	Commit with uncertainty
Year 2 February	Birch	**An eleventh seller or better qualification.** Qualification first, with a hiring trigger.	Commit with uncertainty
Year 2 February	Birch	**Backlog re-qualified.**	Commit with uncertainty
Year 2 February	Cedar	**A pricing-page winner.** Won after repeated looks.	Commit with uncertainty
Year 2 March	Alder	**Second vertical opens.**	Choose whether to test or act
Year 2 March to May	Alder	**Account-research pilot.** Randomised holdback of target accounts.	Choose whether to test or act
Year 2 March to April	Birch	**Shorter demo form test.**	Choose whether to test or act
Year 2 March to April	Cedar	**Onboarding factorial.**	Choose whether to test or act

Date	Company	Event	Chapter
Year 2 April	Cedar	**Support labeller in shadow.**	Run the growth system
Year 2 May to October	Alder	**Account holdout.**	Choose acquisition channels
Year 2 May	Birch	**Shorter form shipped.** More leads, lower lead-to-MQL rate.	Choose whether to test or act
Year 2 May	Birch	**Pulse: professional-network ads up.**	Choose acquisition channels
Year 2 May	Cedar	**Onboarding winner shipped.**	Choose whether to test or act
Year 2 May to June	Cedar	**Geo test of paid social.** Calibrates the mix model.	Choose acquisition channels
Year 2 June	Birch	**Pulse: search ads up.**	Choose acquisition channels
Year 2 July	Alder	**Many ideas, few people.**	Fund the next best action
Year 2 July	Birch	**Pulse: social ads up.**	Choose acquisition channels
Year 2 July	Birch	**Council retired, trigger trips.** Owners decide on cards.	Fund the next best action
Year 2 July	Cedar	**Support labeller advisory.**	Run the growth system
Year 2 July	Cedar	**Sixty ideas a quarter.**	Fund the next best action
Year 2 August	Alder	**Target list and channels.** Reviewed against the ideal customer profile.	Choose markets and accounts
Year 2 August	Birch	**Pulse: professional-network ads down.**	Choose acquisition channels
Year 2 August to November	Birch	**Trigger paused.**	Fund the next best action
Year 2 August to September	Birch	**Ad sets consolidated.** Value-based lead events sent.	Choose acquisition channels
Year 2 September	Alder	**Six-month signpost passed.**	Choose whether to test or act
Year 2 September	Birch	**Pulse: search ads down.**	Choose acquisition channels
Year 2 September to November	Birch	**New seller in ramp.**	Fund the next best action
Year 2 September	Birch	**Memo agent advisory.**	Run the growth system
Year 2 September	Cedar	**Creative and channel quality.** Value per sign-up by channel.	Improve offers and conversion
Year 2 October	Alder	**New packaging.** Discovery checks; packaging for new deals.	Set prices and packages
Year 2 October	Birch	**Pulse: social ads down.**	Choose acquisition channels
Year 2 October	Birch	**Platforms learn from better events.** More leads for the same spend.	Choose acquisition channels
Year 2 October to November	Birch	**Lead score rebuilt.** Rebuilt on a downstream label.	Allocate sales attention
Year 2 October to November	Cedar	**Randomised sales touch.** Among product-qualified accounts.	Allocate sales attention

Date	Company	Event	Chapter
Year 2 December to Year 3 January	Alder	**The champion leaves.** Renewal with reduced scope.	Retain customers profitably
Year 2 December	Birch	**Half the media shift.**	Choose acquisition channels
Year 2 December	Birch	**Trigger rewritten.**	Fund the next best action
Year 2 December	Birch	**New seller at full capacity.**	Fund the next best action
Year 2 December	Birch	**Rebuilt score in use.** Fewer SQLs, more accepted.	Allocate sales attention
Year 2 December	Birch	**A list price rise at renewal.** With a held-out share.	Set prices and packages
Year 2 December to Year 3 January	Cedar	**Save offers.** With a time-boxed control group.	Retain customers profitably
Year 3 January	Birch	**Price rise applies.**	Set prices and packages
Year 3 February	Alder	**Grading the log.**	Learn from outcomes
Year 3 February	Birch	**Trigger trips again.** Approved within the week, no meeting.	Run the growth system
Year 3 February to March	Birch	**First pooled calibration.** Did the qualification fix work?	Learn from outcomes
Year 3 February	Cedar	**Targeted save offers.** To the later tenure bands.	Retain customers profitably
Year 3 February to March	Cedar	**Learning arm for the first three months.**	Retain customers profitably
Year 3 March to June	Birch	**Trigger paused again.**	Run the growth system
Year 3 March	Cedar	**About 140 tests.**	Learn from outcomes
Year 3 April	Alder	**A monthly written rhythm.** Text labeller reaches advisory.	Run the growth system
Year 3 April to June	Birch	**Twelfth seller in ramp.**	Run the growth system
Year 3 April	Birch	**Memo agent promoted.** Low-risk actions after approval.	Run the growth system
Year 3 May to June	Cedar	**Meetings retired.**	Run the growth system

Using the facts

Do not copy these numbers into your own plan. Run a method on the company nearest to your volume, check that you reach the chapter's number, then use your own data. If two places disagree, the shared facts hold the right number. Appendix B, *Parameters and start values*, lists the start values that the methods use.

Appendix D: Formula sheet

This appendix holds the core formulas in one place. Each formula appears as its home chapter writes it, with the same symbols. Each entry says in one line what the formula gives and links to the chapter that explains it. Read that chapter, with its evidence labels and limits, before you rely on a formula. Appendix B, *Parameters and start values*, lists the start values that the formulas use.

Noise and sample size

Standard error and interval of a rate. It gives the noise of a rate p measured on n units. See *Decide what the evidence can resolve*.

> **Calculation**
>
> standard error of a rate = $\sqrt{(p \times (1 - p) \div n)}$
>
> 95% interval = rate ± 1.96 × standard error

Sample size, exact form. It gives the units needed in each arm of a two-arm test.

> **Calculation**
>
> n per arm = $(1.96 \times \sqrt{(2 \times \bar{p} \times (1 - \bar{p})} + 0.84 \times \sqrt{(p_1 \times (1 - p_1) + p_2 \times (1 - p_2))})^2 \div (p_1 - p_2)^2$
>
> p_1 is the baseline rate, p_2 is the rate you hope for, and $\bar{p}$ is their average.
>
> weeks = 2 × n per arm ÷ weekly flow of eligible units

Rule of 16, the book's convention for readability. It gives the smallest relative change that a period can show, and the days that a stage needs to read a 30% change. It is Standard practice, and a convention, not the exact form above.

> **Calculation**
>
> smallest relative change = $4 \times \sqrt{((1 - p) \div (p \times n))}$ for a rate p with n units in each period
>
> smallest relative change = $4 \div \sqrt{n}$ for a count of n events in each period
>
> n required = $16 \times (1 - p) \div (p \times 0.30 \times 0.30)$, the units needed in each of two equal periods
>
> days to read = 28 × n required ÷ n per window

Overdispersion. It gives the factor by which real wobble between windows widens the noise.

> **Calculation**
>
> variance of a window rate = $p \times (1 - p) \div n + \tau^2$
>
> overdispersion factor $\phi = 1 + n \times \tau^2 \div (p \times (1 - p))$

Design effect. It gives the factor by which clustered units, such as contacts in one account, raise the sample size.

> **Calculation**
>
> design effect = 1 + (m - 1) × ρ
>
> m is the average number of units per cluster. Multiply the sample size by the design effect.

Shrinkage. It pulls a noisy rate towards the pooled rate. The more noise, the stronger the pull.

> **Calculation**
>
> shrunk rate = pooled rate + reliability × (raw rate - pooled rate)
>
> reliability = $\tau^2 \div (\tau^2 + \sigma^2)$
>
> τ^2 is the real spread between cells, and σ^2 is the noise within a cell, which equals p × (1 - p) ÷ n.

The window, the quarter and the year. A window is 28 days. A quarter is 91 days, which is 3.25 windows, not 3. A year is 364 days, which is 13 windows. The 28-day window is a convention of this book (Arithmetic). The facts of the three companies count calendar months of 365.25 ÷ 12 days. See *Decide what the evidence can resolve*.

Value

Gross profit and customer value. All money in the book is expected gross profit over a horizon, discounted. LTV is the value of one customer. See *Value customers and cash*.

> **Calculation**
>
> gross profit = revenue – cost of revenue
>
> gross margin = gross profit ÷ revenue
>
> LTV = sum over years k = Ø to H – 1 of [S(k) × GP(k) ÷ (1 + r)^k]
>
> S(k) is the share of customers still active in contract year k, with S(Ø) = 1. GP(k) is the gross profit in year k of a customer that is active. r is the discount rate and H the horizon in years.

Value ladder. See *Value customers and cash*. It gives the value of a unit at any stage.

> **Calculation**
>
> value of a unit at a stage = chance that the unit becomes a customer × value of one customer
>
> chance = product of the step rates from that stage to the customer

Accrual payback. It gives the months that the gross profit of a customer needs to repay the CAC.

> **Calculation**
>
> accrual payback in months = fully loaded CAC ÷ (contract value × gross margin ÷ 12)

Decisions

Net value. The card draws it many times. See *Commit with uncertainty*.

> **Calculation**
>
> net value = v × l × D − K − r
>
> v is the value of one unit. l is the extra units in each window. D is the number of windows in which the effect counts. K is the cost. r is the expected harm on the other wing of the bowtie.
>
> break-even effect = (K + r) ÷ (value of one unit × units exposed × D)

Chance of clearing the bar. It is the share of draws with a net value above zero. Ship when it is at least 90% for a reversible change and at least 95% for an irreversible one. Both are start values (Plausible but untested).

Winner's curse. *Commit with uncertainty* calls the reliability the shrinkage factor.

> **Calculation**
>
> shrinkage factor = $\tau^2 \div (\tau^2 + \sigma^2)$
>
> shrunk effect = prior mean + shrinkage factor × (observed lift − prior mean)

Value of information. It gives what a test is worth before you run it. See *Choose whether to test or act*.

> **Calculation**
>
> EVPI = expected value of the best choice with full knowledge - expected value of the best choice now
>
> EVSI = expected value of the best choice after the test result - expected value of the best choice now
>
> expected net gain of a test = EVSI - cost of the test - cost of delay a week × weeks until the result

Cost of delay. It gives the value that a week of waiting loses. See *Fund the next best action*.

> **Calculation**
>
> cost of delay per week = expected gross profit of the effect ÷ weeks of lever life
>
> cost of a delay of d weeks = cost of delay per week × d

Detection and diagnosis

Standardised signal, EWMA and CUSUM. They turn a gap from the expectation into one signal, and then watch it over time. See *Decide when a change needs action*.

> **Calculation**
>
> z = (observed − expected) ÷ standard error.
>
> EWMA(t) = λ × z(t) + (1 − λ) × EWMA(t − 1), scaled to standard-normal units.
>
> CUSUM(t) = max(0, CUSUM(t − 1) + (−z(t)) − k), to watch for falls.

Mix and rate split. It splits a change in a blended rate into two parts that add up to it. See *Diagnose what changed*.

> **Calculation**
>
> mix effect = sum over groups of (new share − old share) × (old rate + new rate) ÷ 2.
>
> rate effect = sum over groups of (old share + new share) ÷ 2 × (new rate − old rate).
>
> mix effect + rate effect = new blended rate − old blended rate, with nothing left over.

Break-even tests

Break-even price rise. It gives the extra loss of renewing customers at which a rise of p stops paying. See *Set prices and packages*. Here m is the gross margin and c = 1 − m is the cost share.

> **Calculation**
>
> break-even extra loss of renewing customers = p ÷ (m + p) = p ÷ (1 + p − c)

The revenue version p ÷ (1 + p) is a trap. It forgets that a lost customer takes the margin and not the price. It sets the limit too low. For new business the same logic gives the break-even close rate in *Set prices and packages*.

Save-play break-even. See *Retain customers profitably*. It gives the smallest gain in the chance of staying that pays for a play.

> **Calculation**
>
> value of a play = change in the chance of staying × gross profit kept − cost of the play
>
> break-even change = cost of the play ÷ gross profit kept

Break-even incremental share. It gives the share of credited conversions that must be incremental for a channel to return its spend. See *Choose acquisition channels*.

> **Calculation**
>
> break-even incremental share = 1 ÷ credited value per euro

Grading and small counts

All four entries below are explained in *Learn from outcomes*.

Brier score. It gives the average squared miss of stated probabilities. A forecast of 50% for every case scores 0.25.

> **Calculation**
>
> Brier score = average of (stated probability − outcome)²

Calibration slope. It is the slope of realised outcomes on stated probabilities. A slope of 1 is perfect. A slope below 1 means too much confidence.

Wilson interval. It gives an interval for a share of n cases that stays sound when n is small. See *Learn from outcomes*. *Run the growth system* also uses it. Standard practice.

> **Calculation**
>
> centre = $(p + z^2 \div 2n) \div (1 + z^2 \div n)$; half-width = $z \times \sqrt{(p(1 - p) \div n + z^2 \div 4n^2)} \div (1 + z^2 \div n)$; $z = 1.96$ for 95%
>
> interval = centre ± half-width

Rule of three. When n clean runs show no failure, it bounds the failure rate (Hanley and Lippman-Hand, 1983).

> **Calculation**
>
> 95% upper bound on the failure rate $\approx 3 \div n$

Appendix E: Sources and further reading

This catalogue groups the manuscript's sources by commercial decision. Sources used across decisions can appear more than once. A listing identifies the source of a method or context; it does not establish a causal effect in the example companies.

Practitioner essays motivate mechanisms and hypotheses. Statistical sources support methods under their assumptions. Benchmark reports describe their own samples. All company results remain Illustration only. The evidence labels in the chapters state the support for the claim being made.

Some bibliographic details and external claims remain unverified, as listed below. Local summaries and incomplete extracts are inspiration rather than substitutes for checked primary methods. The reference library also includes *Lean Analytics* and The Business Engineer's *Revenue Modeling*. Their inspected local files are a summary and a potentially truncated extract respectively; neither supplies a new quantitative estimate in this edition.

Sources by decision

Choose the growth objective

- Winning by Design, 2023, *Revenue Architecture* (the bowtie).
- McClure, 2007, *Startup Metrics for Pirates*, talk (pirate metrics).
- Croll and Yoskovitz, *Lean Analytics* (limits of a rule that decides alone).
- Gupta, Lehmann and Stuart, 2004, *Valuing Customers* (the 12% discount rate). Label: Plausible but untested.
- van Belle, 2008, *Statistical Rules of Thumb* (items needed to read a change). Label: Standard practice.
- Kohavi and others, 2014, *Seven Rules of Thumb for Web Site Experimenters*, KDD (items needed to read a change). Label: Standard practice.
- Benchmarkit and Pavilion, 2024, *2024 B2B SaaS Performance Metrics Benchmark Report* (derived typical new customers).
- SaaS Capital, 2023, *SaaS Retention Benchmarks for Private B2B Companies* (retention medians).
- First Page Sage, 2025, *B2B SaaS funnel conversion benchmarks* (stage conversion ranges).
- ChartMogul and ProductLed, 2026, *The SaaS Conversion Report* (sign-up to paid range).
- Balfour, *The First Step To Building A Growth Machine* (process before tactics). Local essay; year not verified.
- Balfour, Winters, Kwok and Chen, 2018, *Growth Loops are the New Funnels*, Reforge (growth outputs returning as inputs). Practitioner framework.

Value customers and cash

- KeyBanc Capital Markets, 2021 (cost of revenue split). Label: Standard practice for the cost split.
- Benchmarkit and Pavilion, 2024 (gross margin and CAC ratio).

- Gupta, Lehmann and Stuart, 2004, *Valuing Customers*, Journal of Marketing Research (12%, from a range of 8% to 16%). Label: Plausible but untested.
- Damodaran, 2026, *Cost of Equity and Capital (US)* (software cost of capital, January 2026). Label: Plausible but untested.
- Skok and Reiss, undated, *Discount rate for DCF*, Matrix Partners (rates for private firms). Label: Plausible but untested.
- Bessemer Venture Partners, 2021, *Scaling to $100 Million* (payback ceilings by segment). Label: Standard practice.
- Skok, undated, *SaaS Metrics 2.0*, a practitioner essay (the ratio of 3).
- Evans, 2017, *Business Analytics: Methods, Models, and Decisions*, Pearson, Global Edition, 2nd edition (the flaw of averages, in the chapter on risk analysis).

Model recurring growth

- KeyBanc Capital Markets, 2021, *Private SaaS Company Survey* (professional services share). Label: Plausible but untested.
- SaaS Capital, 2023 (median gross retention). Label: Standard practice for the definitions.
- Benchmarkit and Pavilion, 2024 (median net retention). Label: Standard practice for the definitions.
- Kaplan and Meier, 1958, the product-limit estimate. Label: Standard practice.
- Schmittlein, Morrison and Colombo, 1987, the Pareto/NBD model. Label: Standard practice.
- Fader, Hardie and Lee, 2005, the BG/NBD model. Label: Standard practice.
- Fader and Hardie, 2007, the shifted beta-geometric model. Label: Standard practice.
- Linoff and Berry, *Data Mining Techniques* (hazards and tenure).
- Winning by Design, 2023 (first principles).
- Winning by Design, 2023 (data model, deal model, touch levels, buying group).
- Benchmarkit and Pavilion, 2024 (CAC ratio).
- Winning by Design, 2023 (growth stages, quota ranges, three tests). Label: Assertion, no sample.
- Benchmarkit and Pavilion, 2024 (growth and rule of 40 medians).
- KeyBanc Capital Markets, 2021 (payback and expansion cost medians). Label: Plausible but untested.

Choose markets and accounts

- KeyBanc Capital Markets, 2021, *Private SaaS Company Survey* (professional services share). Label: Plausible but untested.
- Benchmarkit and Pavilion, 2024 (median net retention). Label: Standard practice for the definitions.
- Winning by Design, 2023 (first principles).
- KeyBanc Capital Markets, 2021 (cost of revenue split). Label: Standard practice for the cost split.
- Benchmarkit and Pavilion, 2024 (gross margin and CAC ratio).
- Skok and Reiss, undated, *Discount rate for DCF*, Matrix Partners (rates for private firms). Label: Plausible but untested.
- Bessemer Venture Partners, 2021, *Scaling to $100 Million* (payback ceilings by segment). Label: Standard practice.

- Skok, undated, *SaaS Metrics 2.0*, a practitioner essay (the ratio of 3).
- Winning by Design, 2023 (data model, deal model, touch levels, buying group).
- Skok, undated (touch levels).
- ChartMogul and ProductLed, 2026 (trial conversion).
- Benchmarkit and Pavilion, 2024 (CAC ratio).
- Ebsta and Pavilion, 2023, *2023 B2B Sales Benchmark Report* (deal timing by size).
- First Page Sage, 2025 (funnel conversion spans).
- Winning by Design, 2023 (growth stages, quota ranges, three tests). Label: Assertion, no sample.
- Benchmarkit and Pavilion, 2024 (growth and rule of 40 medians).
- Bessemer Venture Partners, 2021 (payback ceilings). Label: Standard practice for the ceilings; Plausible but untested at the margin.
- KeyBanc Capital Markets, 2021 (payback and expansion cost medians). Label: Plausible but untested.
- Winning by Design, 2023 (tiers of target accounts).
- Balfour, 2017, *Get Out of the ARPU-CAC Danger Zone with Channel Model Fit* (channel and model fit). Practitioner framework.

Choose acquisition channels

- Blake, Nosko and Tadelis, 2015, *Consumer Heterogeneity and Paid Search Effectiveness*, Econometrica (brand search).
- Gordon, Zettelmeyer, Bhargava and Chapsky, 2019, *A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook*, Marketing Science (observational methods against experiments).
- Lewis and Rao, 2015, *The Unfavorable Economics of Measuring the Returns to Advertising*, Quarterly Journal of Economics (small ad effects need large samples).
- Johnson, Lewis and Nubbemeyer, 2017, *Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness*, Journal of Marketing Research (ghost ads).
- Vaver and Koehler, 2011, *Measuring Ad Effectiveness Using Geo Experiments*, Google (geo experiments).
- Jin and others, 2017, *Bayesian Methods for Media Mix Modeling with Carryover and Shape Effects*, Google (saturation and carry-over).
- Hanssens, Parsons and Schultz, 2001, *Market Response Models* (power curve). Label: Standard practice.
- *Marketing Measurement and Analytics*, chapter on multi-channel measurement (limits of mix models). Title only, not verified.
- Regulation (EU) 2016/679 (Article 4(5) and Recital 26).
- European Data Protection Board, 2025, *Guidelines 01/2025 on Pseudonymisation*, version for public consultation (hashing is pseudonymisation).
- Blake, Nosko and Tadelis, 2015 (demand that exists before the ads).

Improve offers and conversion

- Regulation (EU) 2016/679 (Article 6, consent basis).
- Hauser and others, in *From Little's Law to Marketing Science*, on consideration sets (the must-have claim).

- Thomke, 2020 (one third of experiments help).

Allocate sales attention

- Provost and Fawcett, *Data Science for Business* (decision payoff and the expected value of an action). The purchase-probability threshold is restricted to the assumptions stated in the chapter.

Remove buying friction

- Winning by Design, 2023 (hand-off contract, SPICED, critical event). Label: Standard practice for the hand-off contract.

Set prices and packages

- Gupta, Lehmann and Stuart, 2004, *Valuing Customers* (discounted customer-value basis). The price break-even expressions are Arithmetic under the chapter's stated cost and survival assumptions.

Forecast conversion and timing

- Linoff and Berry (cohorts defined at entry).
- Ebsta and Pavilion, 2023 (stalled and slipped deals).
- Little, 1961, Operations Research (Little's law). Label: Arithmetic.

Set targets and staff capacity

- Winning by Design, 2023 (growth stages, quota ranges, three tests). Label: Assertion, no sample.
- Benchmarkit and Pavilion, 2024 (growth and rule of 40 medians).
- Bridge Group, 2025, *SDR Models and Metrics Report* (reps per development rep).
- Gainsight, 2022, *What Is The Right CSM to Customer Ratio?* (accounts and ARR per success manager).
- Bhunia and others, *Advanced Optimization and Operations Research* (queueing theory). Label: Standard practice.
- Sacks, Craft Ventures, 2020, *The Burn Multiple* (the burn multiple).
- Winning by Design, 2023 (hand-off contract, SPICED, critical event). Label: Standard practice for the hand-off contract.

Activate customers

- Userpilot, 2025, *Benchmark Report* (time to first value).

Retain customers profitably

- van Smeden and others, 2016, *No rationale for 1 variable per 10 events criterion for binary logistic regression analysis*, BMC Medical Research Methodology 16, 163 (events per variable).
- Provost and Fawcett (uplift).
- Rodrigues (uplift models).

Expand accounts and generate referrals

- KeyBanc Capital Markets, 2021 (reported expansion-cost distribution). The book's operating limits are Plausible but untested.

Find the most valuable constraint

- Bessemer Venture Partners, 2021 (mid-market payback ceiling). Label: Standard practice for average payback; Plausible but untested at the margin.

Decide what to measure

- Linoff and Berry (one place for each definition).
- Kihn and O'Hara, *Customer Data Platforms* (account matching, first-party data).
- Wexler, Shaffer and Cotgreave, *The Big Book of Dashboards* (decision dashboards).
- Thomke, 2020 (tiered data flags).
- European Union, 2016, *Regulation (EU) 2016/679 (General Data Protection Regulation)* (consent, Articles 4, 6 and 7).
- Google, 2025, *Privacy Sandbox next steps* (Chrome cookie policy; browser pages checked 15 October 2026).
- Google, 2025, *Update on plans for Privacy Sandbox technologies*.
- Mozilla, 2022, *Firefox rolls out Total Cookie Protection by default* (Firefox).
- WebKit, undated, *Tracking Prevention* (Safari).
- Google, undated, Analytics Help, *BigQuery Export*, *Behavioral modeling* and *Engaged sessions* (export, consent fields and engaged sessions; checked 15 October 2026).

Decide what the evidence can resolve

- Deng, Xu, Kohavi and Walker, 2013, *Improving the sensitivity of online controlled experiments by utilizing pre-experiment data* (covariate adjustment).
- Kohavi and others, 2014 (money metrics need larger samples).
- Thomke, 2020, *Experimentation Works*, Harvard Business Review Press (reports the Kohavi point).
- Benchmarkit and Pavilion, 2024 (derived typical new customers).
- First Page Sage, 2025 (published spans).
- McCullough, 2021, *Business Experiments with R*, Wiley (no power after the result).
- Spiegelhalter, 2005, *Funnel plots for comparing institutional performance*, Statistics in Medicine, volume 24 (funnel plots).
- Berger, *Statistical Decision Theory and Bayesian Analysis* (empirical Bayes). Label: Standard practice.

Choose whether to test or act

- Berger (value of information with a normal prior). Label: Arithmetic, with the normal approximation (Standard practice).
- *Bayesian Networks and Influence Diagrams* (information never has a negative value).

- Kohavi and others, 2014 (money metrics are skewed).
- Lan and DeMets, 1983, alpha spending (O'Brien-Fleming looks).
- McCullough, 2021 (test-and-roll). Label: Arithmetic.
- Li and others, 2006, as reported in McCullough, 2021 (few active interactions).

Commit with uncertainty

- Camm and others, *Business Analytics* (chapter on decision analysis). Label: Arithmetic.
- Berger (empirical Bayes). Label: Standard practice.
- Thomke, 2020 (base rates of winning tests). Label: Standard practice.
- Arkes and others, 1988, reported in National Research Council, 2011, *Intelligence Analysis for Tomorrow* (hindsight bias).

Fund the next best action

- Reinertsen, 2009, *The Principles of Product Development Flow* (weighted shortest job first).
- Walker, Marchau and Kwakkel, in *Decision Making under Deep Uncertainty* (trigger sheets).
- Thomke, 2020 (queues in test teams).
- Samuel-Cahn, 1984, as summarised by Chawla and others, 2024, *Algorithmic Game Theory*, Springer (the prophet inequality: one threshold for one slot).

Decide when a change needs action

- *Better Business Decisions from Data*, on control charts (control chart limits).
- Roberts, 1959, *Control chart tests based on geometric moving averages*, Technometrics (EWMA). Label: Standard practice.
- Page, 1954, *Continuous inspection schemes*, Biometrika (CUSUM). Label: Standard practice.
- Benjamini and Hochberg, 1995, *Controlling the false discovery rate*, Journal of the Royal Statistical Society, Series B (false discovery rate).
- Thomke, 2020 (Twyman's law).
- *Growth Data Analytics Playbook* (replace metrics that read too slowly).

Diagnose what changed

- Kitagawa, 1955, *Components of a difference between two rates*, Journal of the American Statistical Association (mix and rate split). Label: Arithmetic.
- Simpson, 1951, *The interpretation of interaction in contingency tables*, Journal of the Royal Statistical Society, Series B (a trend that reverses in the total).
- Hill, 1965, *The environment and disease: association or causation?*, Proceedings of the Royal Society of Medicine (checks for a cause).
- Rodrigues, *Product Analytics* (the same checks for product data).
- Rogan and Gladen, 1978, *Estimating prevalence from the results of a screening test*, American Journal of Epidemiology (correcting a coded share). Label: Standard practice.

Learn from outcomes

- *Growth Data Analytics Playbook* (ladder of proof). Label: Standard practice.
- Denrell, in *The Oxford Handbook of Organizational Decision Making*, on superstitious behaviour (learning needs a holdback).
- Brier, 1950, *Verification of forecasts expressed in terms of probability*, Monthly Weather Review (Brier score). Label: Arithmetic.
- Wilson, 1927, *Probable inference, the law of succession, and statistical inference*, Journal of the American Statistical Association 22(158), 209 to 212 (the Wilson interval).
- Atanasov and Himmelstein, in Seifert (ed.), 2023, *Judgment in Predictive Analytics* (low reliability of scores).
- Hanley and Lippman-Hand, 1983, *If nothing goes wrong, is everything all right? Interpreting zero numerators*, JAMA 249(13), 1743 to 1745 (rule of three).

Run the growth system

- National Research Council, 2011 (independent estimates).
- Klein, 2007, *Performing a project premortem*, Harvard Business Review 85(9), 18 to 19 (pre-mortem).
- Dietvorst and others, 2016, reported by Burton, Stein and Jensen in Seifert (ed.), 2023 (bounded overrides).
- Landis and Koch, 1977, *The measurement of observer agreement for categorical data*, Biometrics 33(1), 159 to 174 (kappa bands).

Details not yet verified

- *Marketing Measurement and Analytics*: author, publisher and year. The chapter gives the title only, and the title is unverified. Rafael will read the title page.
- Camm and others, *Business Analytics*: edition, publisher and year.
- Berger, Linoff and Berry, Rodrigues, Wexler, Shaffer and Cotgreave, Kihn and O'Hara, Croll and Yoskovitz, Provost and Fawcett, and Bhunia and others: publisher and year.
- Hauser and others, Walker, Marchau and Kwakkel, and Denrell: the year of the book that holds the chapter.
- *Better Business Decisions from Data*, *Bayesian Networks and Influence Diagrams* and *Growth Data Analytics Playbook*: authors, publisher and year.
- Userpilot, 2025: the full title of the report.
- Little, 1961, and Lan and DeMets, 1983: the titles of the articles.
- The WebKit page and the three Analytics Help pages: year.
- *Revenue Architecture*: authors and publisher.
- The EU AI Act and the General Data Protection Regulation, which *Run the growth system* mentions with no article: the official site could not be read.

Further reading

All of these appear in the lists above.

- Winning by Design, *Revenue Architecture*: the whole revenue system and the bowtie.
- Thomke, *Experimentation Works*: how companies run many tests and learn from them.
- Kohavi and others, *Seven Rules of Thumb for Web Site Experimenters*: short rules for sample size and metrics.
- McCullough, *Business Experiments with R*: power, test-and-roll and factorial designs.
- Provost and Fawcett, *Data Science for Business*: expected value as the rule for acting on a prediction.
- Gupta, Lehmann and Stuart, *Valuing Customers*: customer value as discounted profit.
- Gordon and others, *A Comparison of Approaches to Advertising Measurement*: why observational methods can mislead.
- Jin and others, *Bayesian Methods for Media Mix Modeling with Carryover and Shape Effects*: the base of the media mix model.
- Reinertsen, *The Principles of Product Development Flow*: cost of delay and queues.
- Seifert (ed.), *Judgment in Predictive Analytics*: how people and models work together.