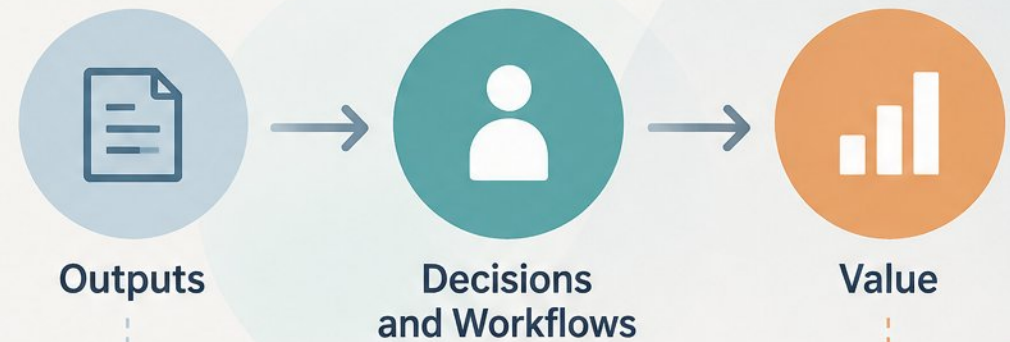


Measuring Data Science Outcomes: From Outputs to Decisions, Workflows, and Value

- Outcome measurement: hardest, least standardized part of the lifecycle
- High AI adoption $\neq$ improved decisions — "is it used?" is weak
- Outputs: models, dashboards, pipelines. Outcomes: better, faster choices
- Three lenses: decision quality, workflow efficiency, organizational value
- Goals: prove impact without overclaiming causality; comparable results
- Good outcomes: meaningful, attributable, triangulated, time-bound, equity-sensitive
- You'll build: measurement plan, trace spec, governance cadence



From Model Metrics to Decision Metrics

Why AUC and RMSE rarely convince decision-makers



Model quality (predictive) $\neq$
decision quality (better choice) $\neq$
outcome quality (real result)



Map scores to decision relevance:
expected value, asymmetric error cost,
hit rate at threshold



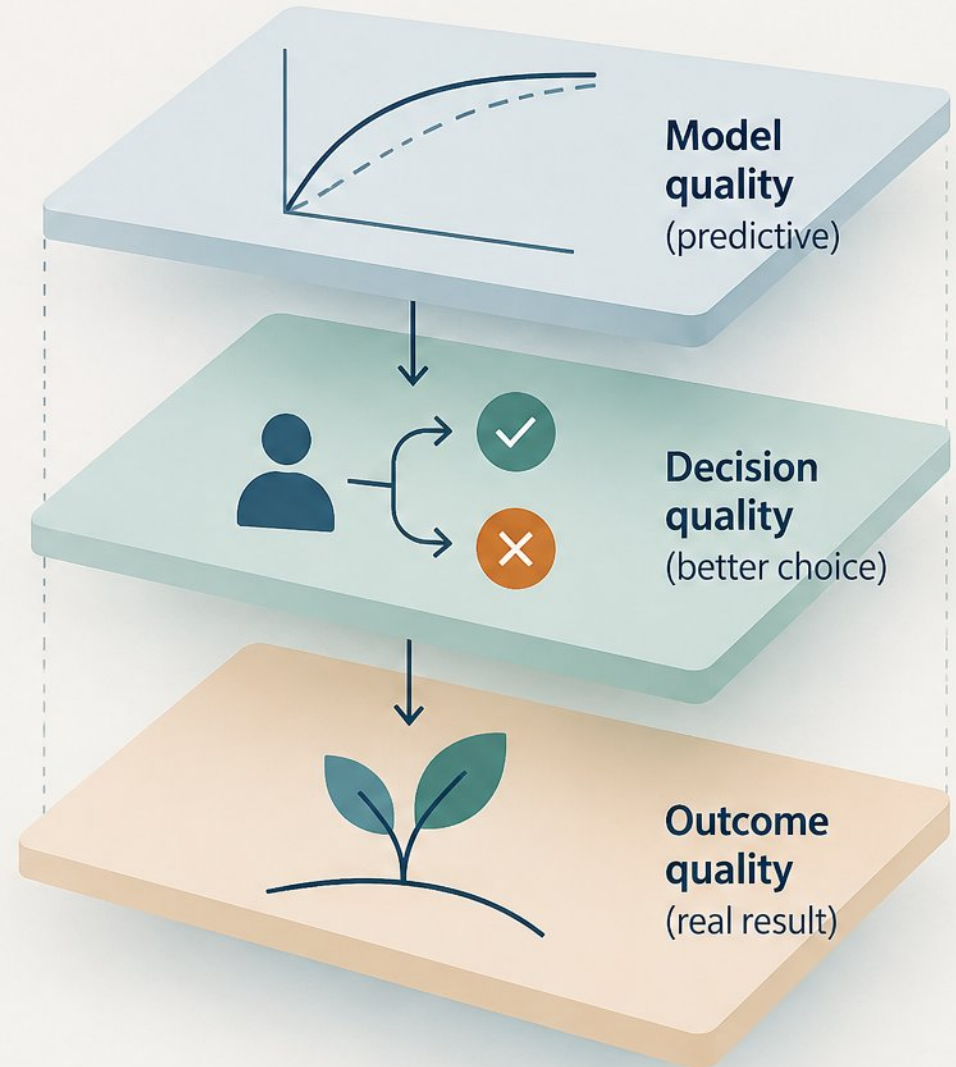
Decision metrics: reliance,
override rate and timing, escalation rate,
reversal rate, decision cycle time



Adoption is an exposure measure,
not performance — high use can
coexist with poor outcomes



Team gain framing: compare
human-only, AI-only, and human+AI
on final decisions



Instrumenting Workflows and Human-in-the-Loop Signals

PEOPLE
CONTEXT
BETTER OUTCOMES



Decision traces: what was recommended, shown, chosen, and modified



Log human actions as event attributes: approved, denied, modified, expired, auto-approved



Timing metrics: review latency, intervention latency, time-to-calibration, wait time



Workflow measures: cycle time, queue time, rework, handoffs, automation rate, touchpoints



Instrument interventions to capture who, why, how long, and outcome



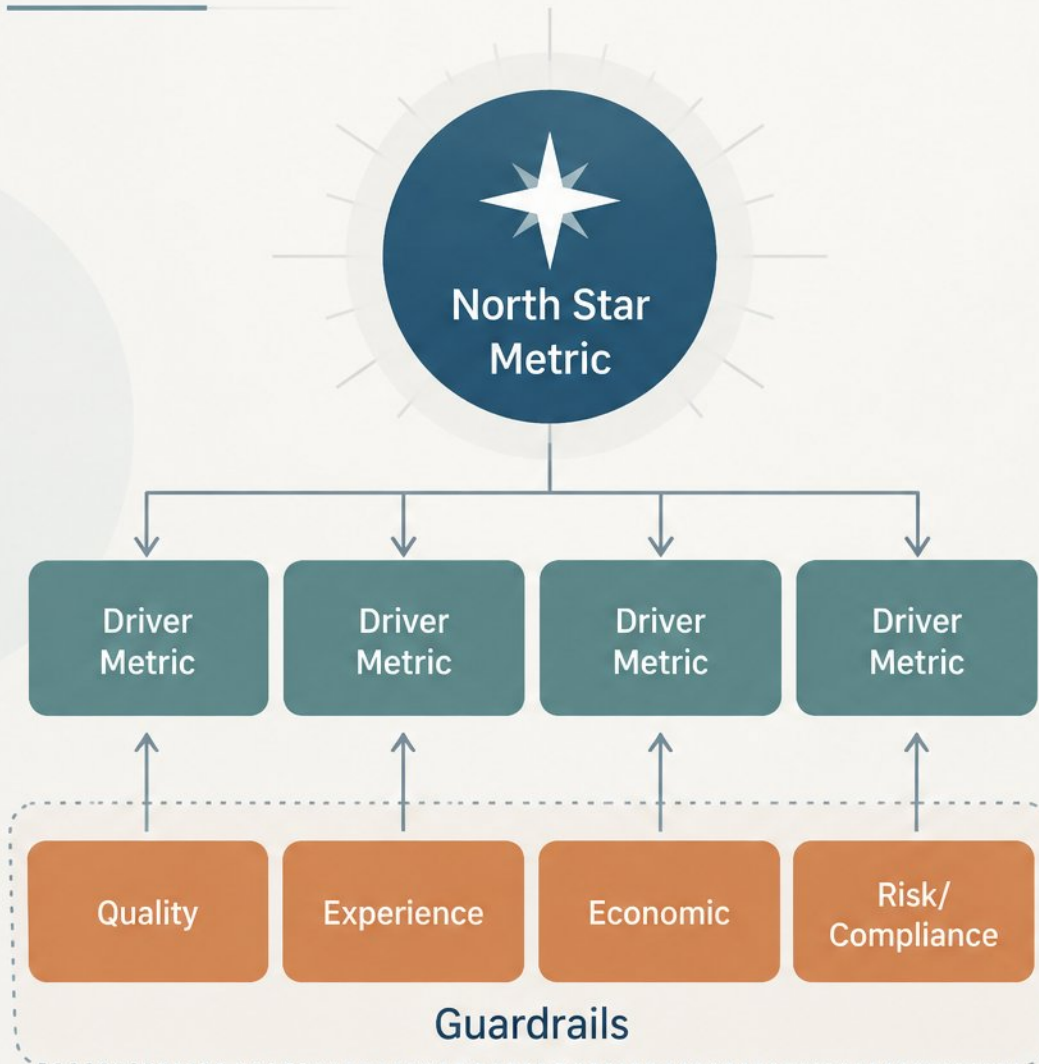
Pitfalls: no context, missing counterfactuals, fabricated panels, approval-flow confusion



Constraints: reviewer privacy, redaction, tamper-evident chains, ethical experiment design

Choosing and Governing Metrics

MEASURE • FOCUS • MAKE PROGRESS



- Working size: one North Star, 3–5 drivers, 2–4 guardrails (6–10 numbers)
- 30–40 metric dashboards usually mean nobody decided what matters
- Leading vs. lagging: activation diagnoses this week, revenue lags furthest
- Guardrails pre-committed: quality, experience, economic, risk/compliance
- Metric contract: meaning, formula, grain, owner, checks, consumers
- Definition changes need proposal, impact review, approver, version bump
- Goodhart's law: North Star bonuses invite quiet guardrail breaches

CLEARER METRICS
BRIGHTER DECISIONS

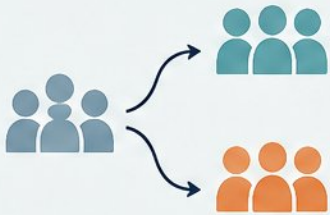
FROM MEASUREMENT
TO MEANINGFUL IMPACT

Causal Inference for Outcome Measurement

BETTER
QUESTIONS
LEAD TO
CLEARER
ANSWERS

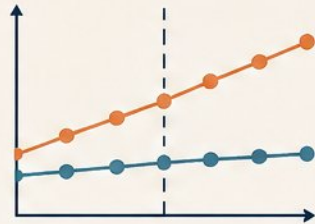
- Many data science questions are genuinely causal; prediction alone overclaims impact
- Randomized designs anchor attribution: A/B, switchback, holdouts, cluster randomization
- Quasi-experimental options: diff-in-diff, synthetic control, regression discontinuity, IV
- Name threats to validity: interference, novelty effects, selection bias, drift
- Report the identifying assumption plainly; use prior RCTs to validate updated models
- Case evidence: geo holdouts plus Bayesian synthetic control estimated ad lift when A/B was infeasible

A/B Test



Randomized comparison
between treatment
and control

Difference-in-Differences



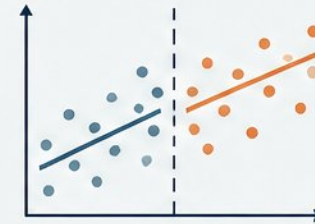
Compare changes
over time between
groups

Synthetic Control



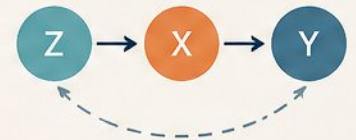
Use a weighted
combination of controls
to estimate the counterfactual

Regression Discontinuity



Compare outcomes
just above and below
a threshold

Instrumental Variables



Use an instrument that
affects treatment but
not the outcome directly

Designing Measurement Plans Before Deployment

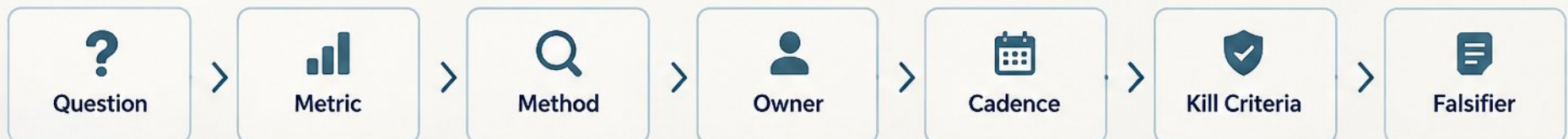
- Pre-register the falsifiable hypothesis and primary metric before launch
- Define success thresholds and minimum detectable effect upfront
- Build a logic model: resources → activities → outputs → outcomes
- Capture baseline, run power analysis, power guardrails for harmful effects
- Specify instrumentation, template, owners, cadence, and kill criteria early



Logic Model



Measurement Plan Template



Quantifying Business and Organizational Value



Convert outcomes to money:

revenue lift, cost avoided, risk reduced, time saved



Example: a 2-point drop in misrouted claims often beats total time saved



Value model: V_d direct savings, V_q error-rate gains, V_k better decisions



TCO, not license cost: tech is only 10–20% of year-one spend

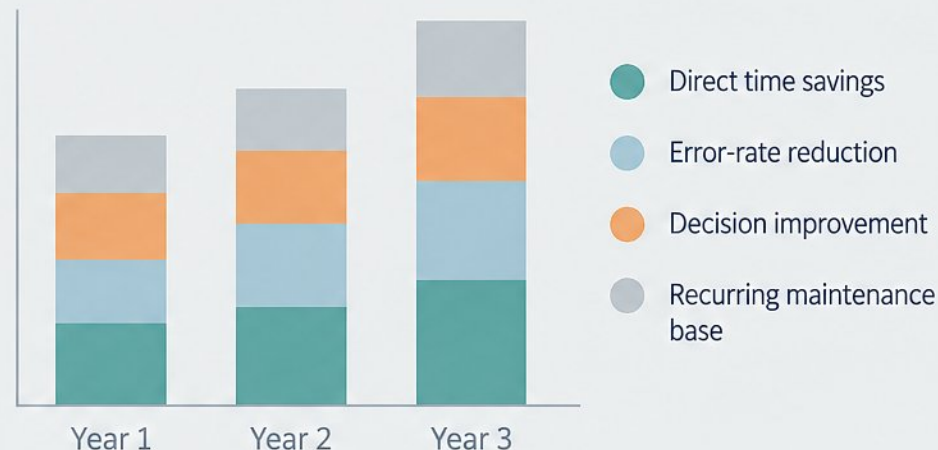


Year two base runs 30–50% of year one; savings are one-off, monitoring perpetual

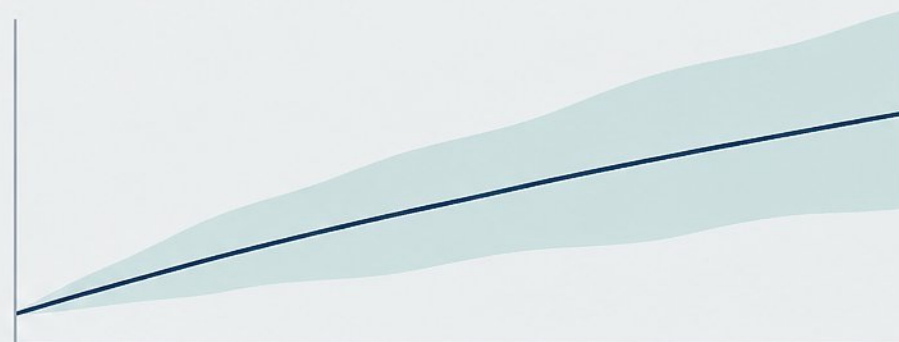


Report ranges, confidence intervals, and risk-adjusted NPV — never a single number

Three-Year Value and Cost Model

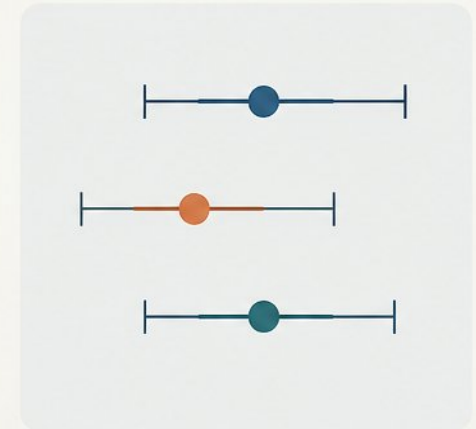
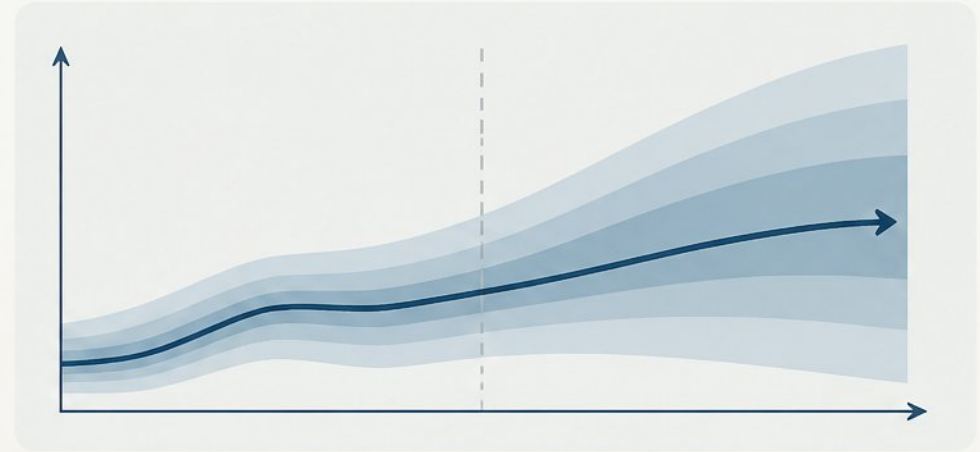


Three-Year NPV with Uncertainty



Stakeholder Reporting and Decision Narratives

- Executives want the bottom line, key trends, and one clear action
- Analysts want methodology, sources, significance, and confidence intervals
- Structure results: question → method → findings → caveats → recommendation
- Put uncertainty on the page — confidence fans, tornado charts, error bars
- Name the confounder and alternative explanation before the audience does
- Unattributed numbers and cluttered dashboards destroy credibility
- Close with an implication banner: action, owner, timing



Action



Owner

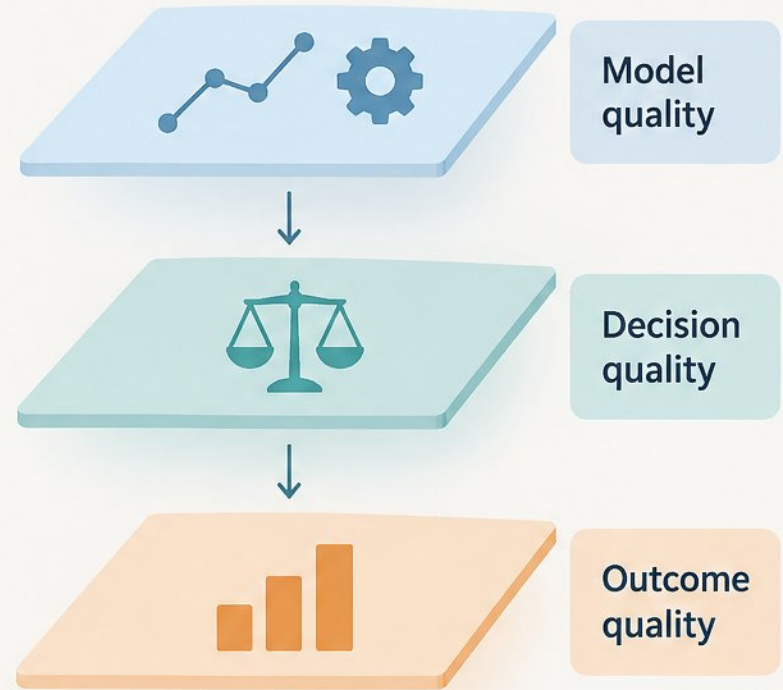


Timing



Monitoring, Drift, and Lifecycle Evaluation

- Watch data drift, concept drift, label shift, and pipeline failures
- Label-free proxies: disagreement rates, suitability signals, distribution stats
- Treat degradation detection as two-sample hypothesis tests, not ad-hoc thresholds
- Scan subgroups hierarchically to find where and why decay lands
- Monthly ops table: agreement, escalation, reversal, latency, cost per decision
- Cadence: monthly blind sample, calibration on each version, quarterly threshold review



Monthly ops table

Agreement	Escalation	Reversal	Latency	Cost per decision	Confidence-band share

Auditing Human Review:

Is the Human Step Substantive or a Formality?



Audit question: does human review reduce risk, or just ratify AI output?



Audit fields per system: reviewer information, logged time per case, override mechanism, documented reasoning



Verdict every system: Substantive, Formality, or Mixed; Formality opens remediation



Reviewer flags: near-zero overrides, instant reviews, high reversal rates



Act on flags: workload and tooling review before performance conversations

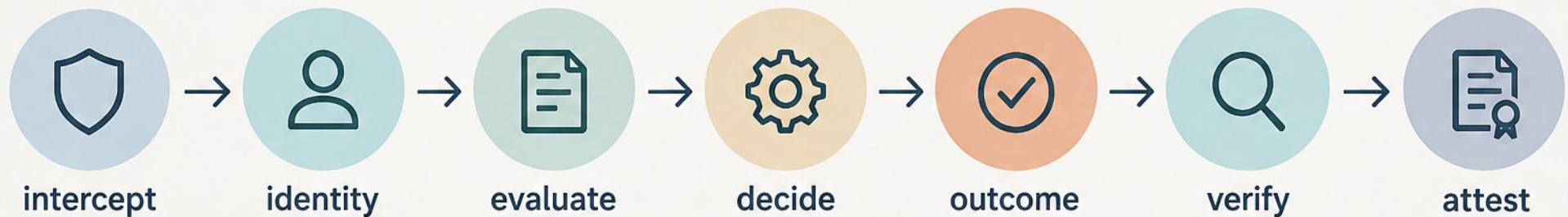


Feed second-review outcomes into coaching, thresholds, and next audit cycle

	—	—	—	—	—
	—	—	—	—	—
	—	—	—	—	—
	—	—	—	—	—
	—	—	—	—	—



Tracing and Evidencing Individual Decisions



- Trace captures: intent, policy block fired, outcome, tamper-evident proof
- Read seven stages in order: intercept → identity → evaluate → decide → outcome → verify → attest
- Policy verdict ≠ outcome: denial means action never ran — that is success
- Escalation stays pending until a human resolves it
- Keep integrity separate: chain untampered vs. target state verified
- Identity gaps are findings: unattributable actions fail the identity stage
- Triage surface: filter by severity, search by request ID, tool, or note



**Chain
untampered**



**Target state
verified**

Capstone: Building Your Measurement Plan and Governance Cadence

FROM QUESTIONS TO BETTER DECISIONS



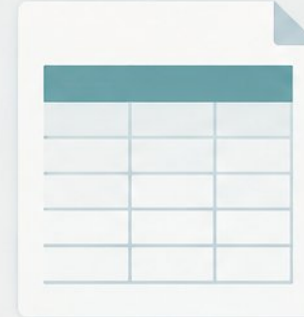
1

Decision inventory: pause dashboards 30 days — which five decisions still matter?

Decision Inventory



One-Page Plan



Guardrail Contract



2

Draft a one-page plan: question, metrics, thresholds, causal method, owner, kill criteria

Instrumentation Minimum



Governance Rhythm



Peer Critique



4

Write guardrail contracts: range, alert band, automatic response across four categories

5

Governance rhythm: monthly decisions review, quarterly threshold review, named metric owners

90-Day Rollout



6

Peer critique: falsifiable hypothesis, isolating assumption, benefit confidence, falsifier, stop criteria

7

90-day rollout: measure what decisions depend on, publish a narrow board-grade set

PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/c53e6862/public