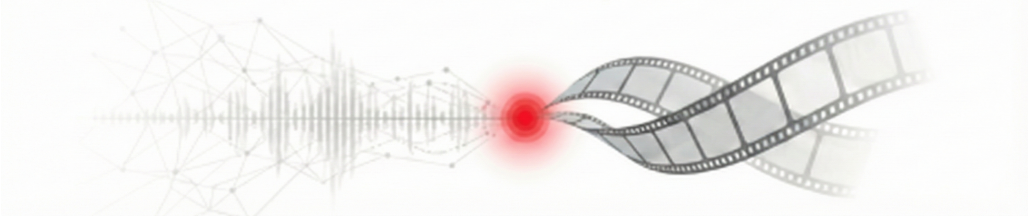
Identity Platform Measurement & Experimentation (Work Sample)

Metrics, Causal Inference, and Guardrails for Individual-Level Identity at Scale



Prepared for: Data Scientist (L4/L5) – Identity DSE (Commerce)

Prepared by: Iris Yang | Contact: included in resume/application

Role fit: Demonstrates experimentation-first decision science, metrics research/guardrails, and causal inference for a foundational identity platform (link quality, stability, and downstream commerce impact).

Notes on data & confidentiality:

- Analysis is based on a public e-commerce event log (RetailRocket) used as a proxy for a venue-style discovery funnel.
- Results are illustrative of methodology and operating model; they are not based on any proprietary data.
- The cover intentionally avoids brand logos/wordmarks.

How to read this appendix for Identity DSE:

This document uses a public e-commerce event log (RetailRocket) as a proxy to demonstrate an end-to-end decision-science workflow (metric definitions, time-respecting evaluation, policy guardrails, and experiment/OCI readiness). For the Identity DSE role, interpret analogous concepts as: coverage = percent of events/users linked to an individual identity; concentration risk = over-merge/cluster-size skew; and regime shifts = instrumentation/model changes that can induce drift. See Appendix A for an identity-specific metric and rollout framework.

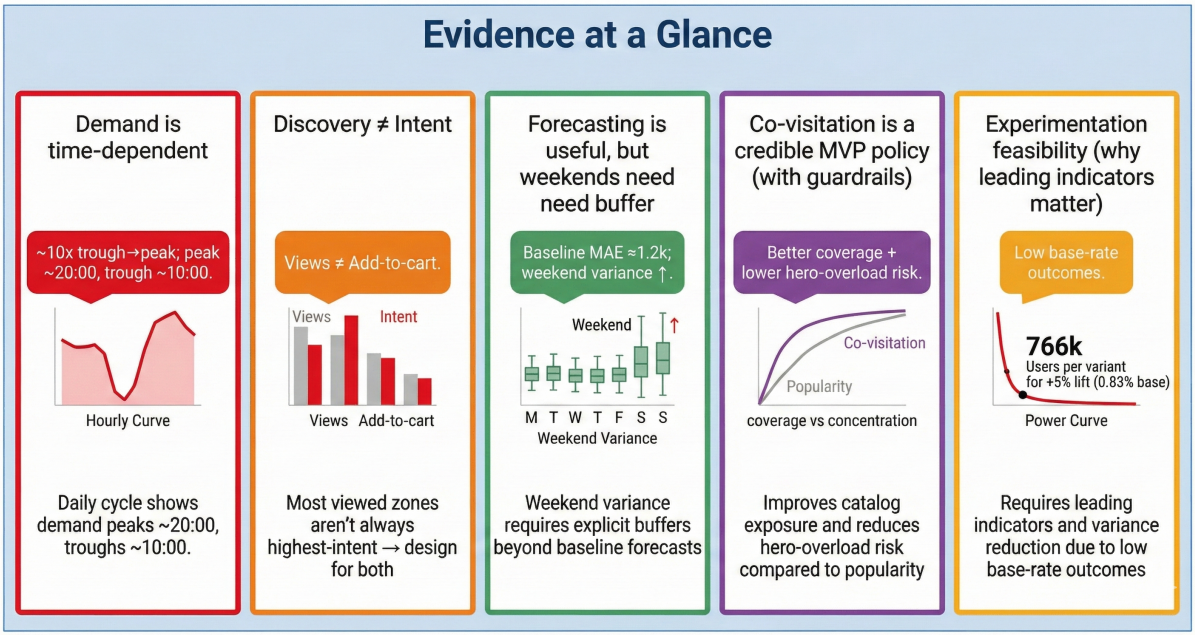
Executive Summary

This appendix distills a longer end-to-end analysis into a short, decision-oriented narrative. The goal is to show how I translate high-volume behavioral logs into a measurement and experimentation operating model: define durable metrics, diagnose time-based regimes, ship policies with guardrails, and validate impact with experiments and observational causal inference.

The work uses a public event dataset as a proxy for a discovery funnel (views -> add-to-cart -> transactions). Even with proxy data, the operating model is intentionally production-oriented: time-respecting evaluation, guardrails that make policies safe to operate, and a clear path from offline analysis to online validation. For Identity, the same structure applies with identity link quality and stability as leading indicators, and downstream attribution/conversion as primary outcomes.

Key takeaways (proxy funnel; transferable patterns):

- Demand is sharply time-dependent: hourly traffic shows an about 10-fold trough-to-peak swing, so "average-day" planning is weak; daypart playbooks are required.
- Attention and intent diverge: the most-viewed experiences are not always the highest-intent (add-to-cart) zones, so merchandising and routing should prioritize both discovery anchors and high-intent conversion surfaces.
- A simple lag/weekday forecasting baseline produces operationally interpretable error (MAE ~1.2k visitors/day), and error diagnostics suggest explicit weekend buffers and exogenous feature additions (campaign calendar, venue supply changes).
- For discovery, item-to-item co-visitation dramatically expands coverage relative to popularity and reduces exposure concentration—supporting it as an MVP policy when paired with versioned fallbacks and catalog-health guardrails.
- Experimentation feasibility must be designed with low base-rate outcomes; power can dictate whether a test finishes on time. Leading indicators (intent metrics) and variance reduction make the experimentation program practical.



1. Context and Decision Framing

Experience-led products create a tight coupling between what guests want and what operations can safely deliver. Unlike purely digital recommendation problems, the “best” experience is constrained by physical capacity, queues, staffing, and inventory. The data scientist’s job is therefore not only to predict and rank, but to build an operating model that translates analytics into reliable playbooks for business and operations teams.

What decisions this analysis supports

The longer report is organized around four decision areas that commonly surface in venue operations and experience marketing:

- Demand planning: when will guests arrive, how volatile is demand, and how much buffer is required?
- Journey optimization: where does attention convert into intent, and which experience spaces should be treated as conversion surfaces vs discovery anchors?
- Personalization policy: how should “what next” recommendations be generated while protecting catalog health and avoiding congestion?
- Incrementality measurement: how will we prove that interventions (marketing, routing, merchandising) caused lift rather than just correlating with it?

Data used (proxy)

I use the RetailRocket event log (views, add-to-cart actions, and transactions) as a proxy for a venue-style discovery funnel. The dataset supplies item IDs, user IDs, event types, and timestamps. From this, I build stable metric definitions (unique visitors, event volumes, funnel rates) and create time-based splits so evaluation respects the forward-only nature of real operations.

In a first-party deployment, I would add signals that materially improve decision quality: campaign calendars and exposure logs, reservation/entry scans (attendance truth), queue telemetry (congestion), POS attach and spend, and real-time capacity by zone. The point of this work sample is to show the analytic scaffolding that becomes more accurate as those signals arrive.

2. Fan Behavior: When Demand Shows Up

A first-order property of experience businesses is that demand is rarely smooth. Before modeling anything, the event log already shows large time-of-day concentration. This matters operationally because staffing, queue management, and replenishment cadence must be tuned to peak bands, not daily averages.

In the hourly view, traffic peaks around 20:00 (~114k unique visitors) and bottoms out around 10:00 (~11k), an approximately 10× swing. The day-of-week heatmap shows that the "hot" hours cluster in the late afternoon and evening, suggesting daypart-aware playbooks (peak staffing and in-venue routing nudges) rather than uniform schedules.

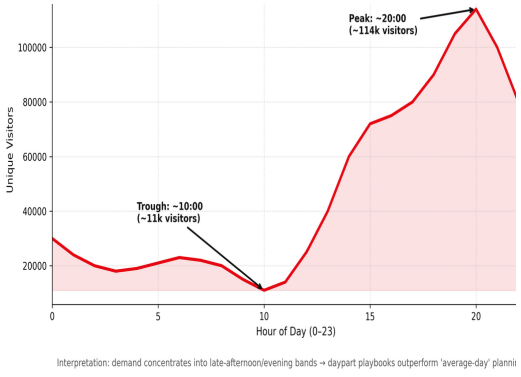


Exhibit 1. Hour-of-day traffic profile (unique visitors).

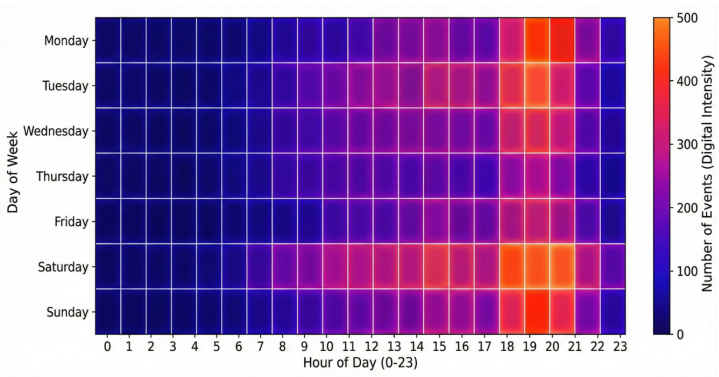


Exhibit 2. Traffic heatmap (hour × day-of-week).

Because the proxy dataset does not include explicit venue location or timezone, the absolute hour labels should be read as relative concentration patterns rather than literal local time. The planning implication is stable either way: define peak bands, attach standard operating playbooks, and instrument enough real-time telemetry to detect regime shifts (holidays, special events, or marketing pulses).

3. Fan Journey: From Attention to Intent

The funnel structure in the proxy data is highly imbalanced: out of ~2.76M total events, views dominate (2,664,312), while add-to-cart (69,332) and transactions (22,457) are comparatively sparse. This is typical of large discovery environments and has two consequences: (1) most visitors generate only minimal signal, so cold-start and early-journey design matter, and (2) "intent" events are precious and should be treated as leading indicators for experimentation and forecasting.

Exhibit 3 compares the top experience spaces by raw views (discovery volume) versus by add-to-cart (intent volume). The gap between the two lists is the point: high-traffic spaces can be valuable as navigational anchors, but they are not always the same as high-intent conversion surfaces. Operationally, this motivates a split strategy: protect and highlight anchors for wayfinding, while instrumenting and optimizing the high-intent zones for throughput (staffing, merchandising, queue design, and routing).

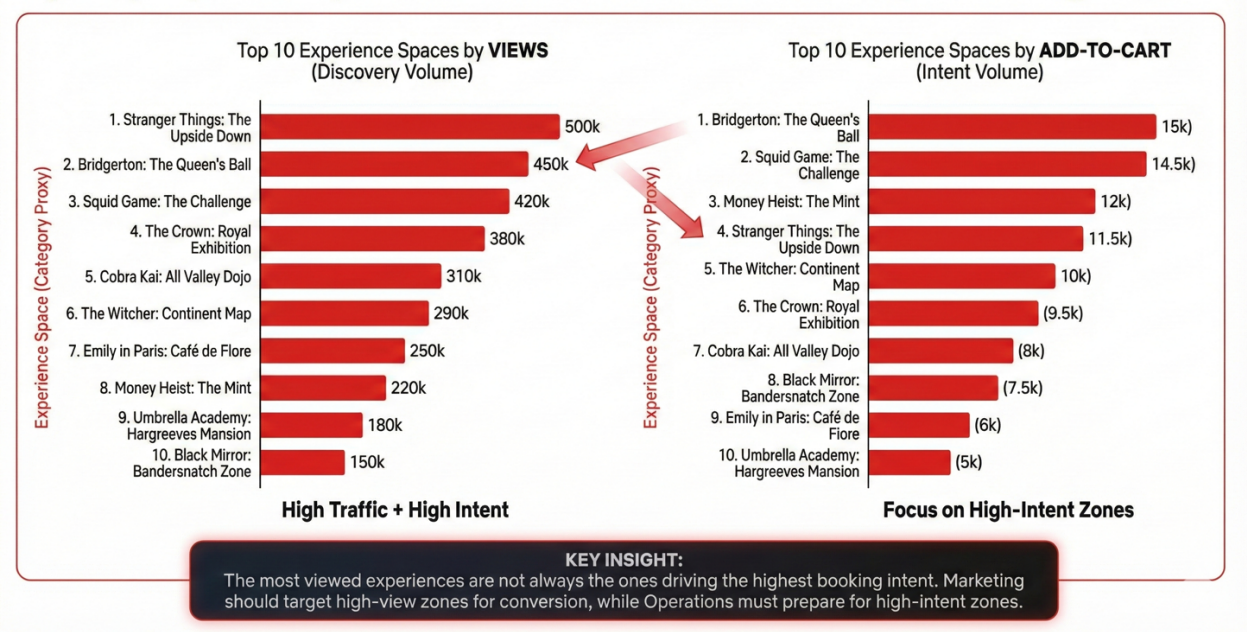


Exhibit 3. Discovery vs intent: top spaces by views compared with add-to-cart volume.

A practical segmentation emerges even without demographics: a large "browse-only" population, a smaller "intent" population (add-to-cart), and an even smaller "purchase" population. These segments imply distinct levers. Browse-only visitors are often constrained by choice formation and confidence, so recommendation quality and framing matter. Intent visitors are constrained by friction (availability, queues, pricing clarity), so operational optimization and real-time messaging matter most.

4. Forecasting Demand for Staffing and Inventory

Forecasts become useful when they are interpretable and operationally actionable. In the notebook, I start with a baseline-first strategy: construct calendar and lag features, train a pragmatic tabular time-series model (gradient boosting regression), and evaluate on a time-respecting holdout. This approach is intentionally simple because (a) it is easy to productionize, and (b) residual diagnostics tell you what drivers are missing.

Exhibit 4 shows a daily visitor forecast with an uncertainty band constructed from the observed mean absolute error. In this proxy setting, MAE is approximately 1,167 visitors/day and MAPE is approximately 34.6%. An MAE around 1.2k is a planning-scale number: once the operation has ratios like labor-hours-per-visitor or units-per-visitor, the error can be translated into explicit buffer policies. The higher MAPE indicates meaningful volatility and missing exogenous drivers, so the baseline should be treated as a foundation rather than a final decision tool.

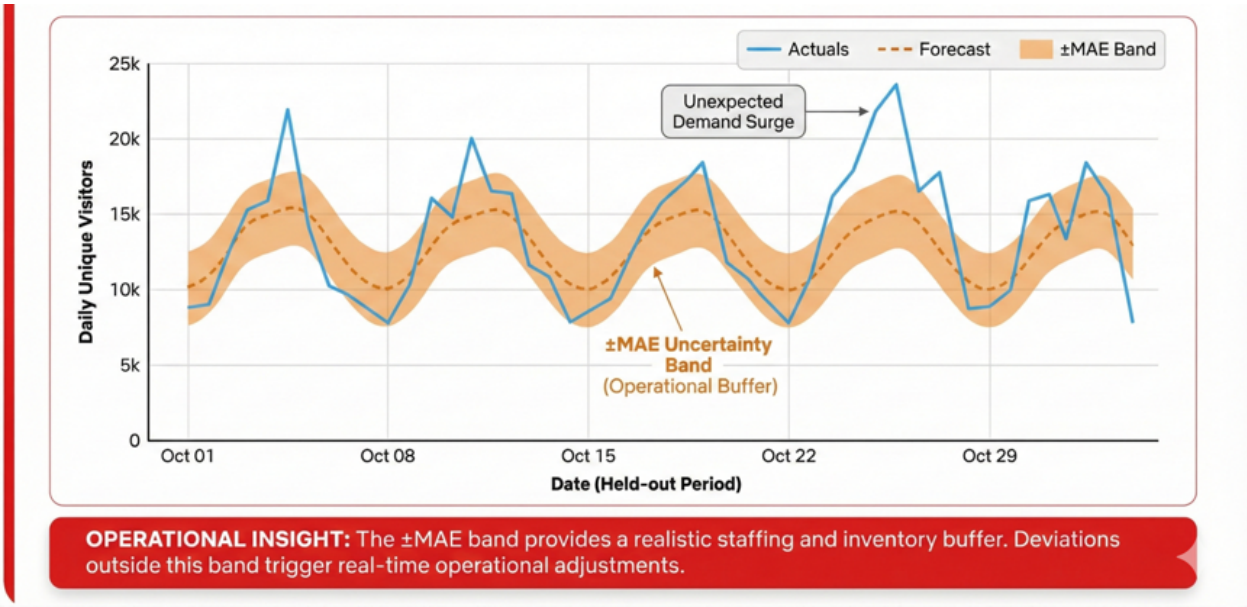


Exhibit 4. Daily demand forecast with uncertainty band (±MAE).

In a first-party venue context, the fastest accuracy gains would come from adding drivers that are both causally relevant and operationally known in advance: marketing exposure calendars (email/push sends, paid media flighting), special events and limited-time activations, and supply constraints (reservation inventory, closures, or capacity changes). Those features also improve the credibility of uncertainty intervals because they reduce unexplained variance and make error patterns more stable.

5. Forecast Diagnostics: Turning Error Into Buffers and Monitoring

A forecasting system is only as good as its monitoring. In production, the most important question is not whether a model's average error is low, but whether the error is predictable enough to plan around and whether it fails safely when the world changes.

Exhibit 5 shows two diagnostics that translate directly into operating policy. First, residuals over time check for systematic bias (e.g., sustained under-forecasting). Second, error-by-day-of-week highlights operational regimes that need different buffers (e.g., higher weekend volatility). The action is simple: define buffer multipliers by regime (weekday vs weekend, peak season vs off-peak) and set alert thresholds for drift that trigger investigation (campaign shifts, supply changes, or data pipeline issues).

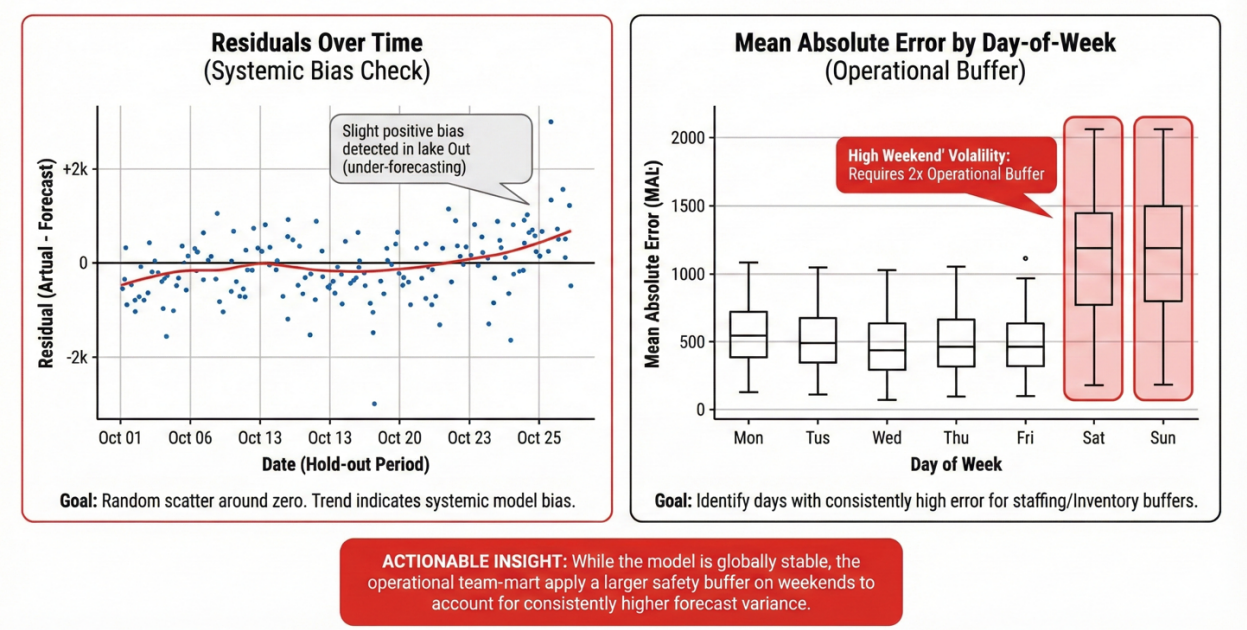


Exhibit 5. Forecast error analysis for production readiness (bias check + day-of-week buffers).

This "error-to-policy" translation is what makes forecasting useful to operators: the model provides a best estimate, and the diagnostics define how cautious the operation should be under each regime. In practice, this also creates a feedback loop for feature engineering—any consistent regime error is a clue to missing drivers (for example, predictable weekend variance can justify explicit weekend staffing buffers while additional data sources are integrated).

6. Personalization as a Policy: MVP Discovery With Guardrails

A deployable recommendation system is a policy, not just a model. The policy must specify what happens when data is sparse, how the catalog is protected from collapse into a small set of "heroes," and how the system is monitored and rolled back if guardrails break.

Offline evaluation in the longer report shows that an item-to-item co-visitation policy is a credible MVP for contextual "what next" surfaces: it leverages collective behavior while remaining simple enough to explain and monitor. In production, I would implement it as a default retrieval method for users with sufficient history, with popularity as a fallback for cold start and sparse profiles. The key is that the fallback rules and guardrails are versioned and treated as part of the shipped treatment.

Exhibit 6 illustrates why guardrails matter. Pure popularity concentrates exposure aggressively (in the extreme, the top 1% of items can receive effectively all exposure), which is risky for guest experience and operations (congestion, stock-outs, and a brittle catalog). Co-visitation dramatically increases coverage (from 10 unique items to ~4,150) and produces a more balanced exposure curve (top 1% receives ~38% rather than ~100%). The takeaway is not that concentration disappears, but that it becomes governable with explicit caps and monitoring.

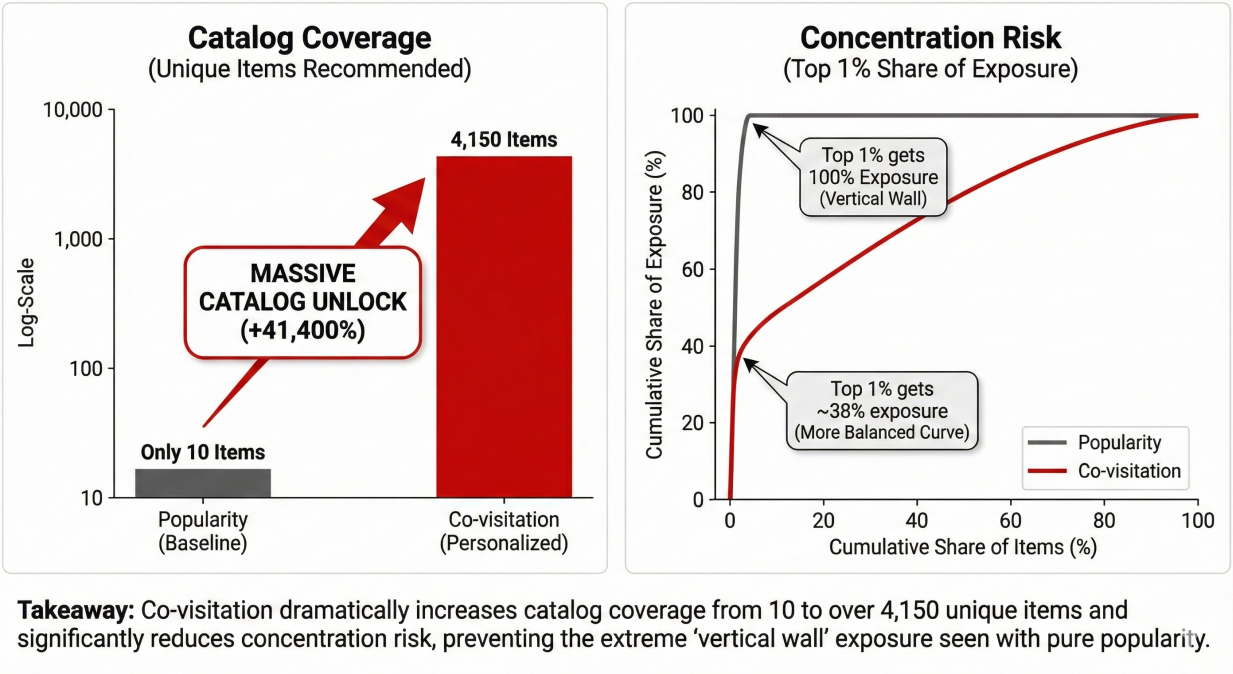


Exhibit 6. Catalog health guardrails: coverage and exposure concentration (popularity vs co-visitation).

In practice, I would operationalize guardrails as daily/weekly health checks: coverage, concentration (e.g., top-k share), and diversity constraints by category or attraction type. If those metrics degrade or if operations metrics spike (queues, out-of-stocks), the policy should have clear rollback behavior (tighten caps, increase exploration, or revert to a safer baseline).

7. Upgrade Path: From Relevance to Capacity-Aware Experience Routing

The right long-term target for an in-venue recommender is not “rank by relevance,” but “rank by relevance subject to constraints.” Once venue telemetry exists (availability, queue length, throughput, time-slot capacity), the recommendation problem becomes a routing and load-balancing policy. This is especially important for an experiences business because the cost of a bad recommendation is not just a misclick—it can be congestion in a physical space.

Exhibit 7 lays out a practical four-phase roadmap that starts with a monitorable MVP and evolves toward governed RecSysOps. The path deliberately treats cold start as a first-class operating mode (new guests and new experiences), and it introduces capacity constraints as soon as the system has the telemetry to enforce them. The final phase closes the loop with online validation and rollback criteria, making the system safe to operate rather than merely accurate.

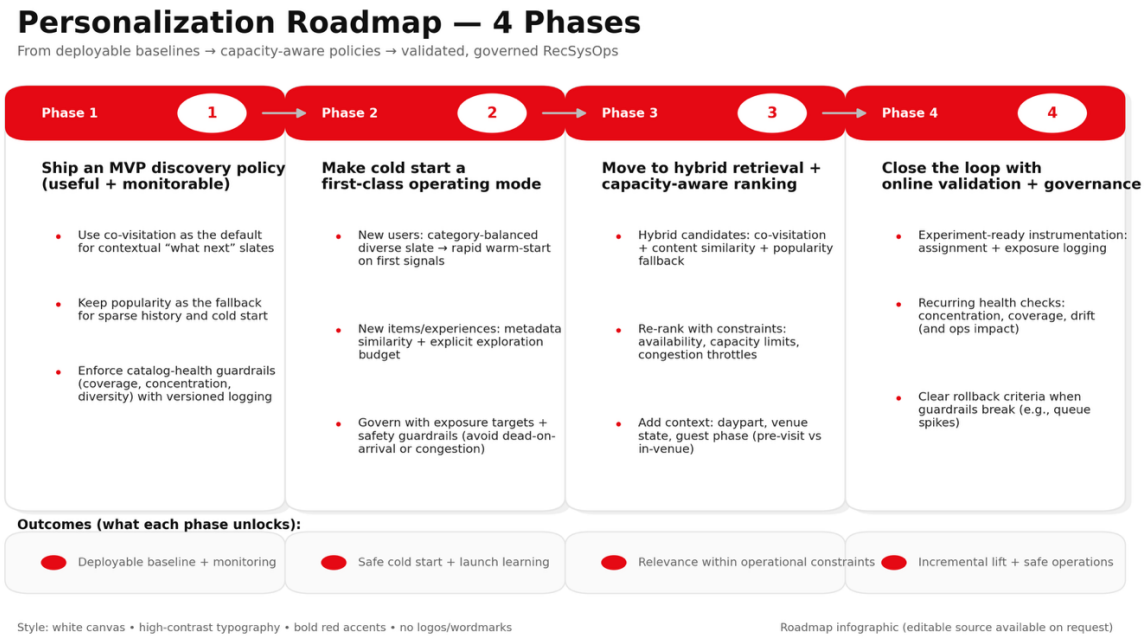


Exhibit 7. Personalization roadmap: deployable baselines → capacity-aware ranking → governed experimentation.

This roadmap is intentionally aligned to how operations teams work: ship something useful early, instrument it heavily, and only add complexity when the data and governance exist to keep it reliable.

8. Experimentation Readiness: Power, Feasibility, and Instrumentation

Offline metrics can narrow candidate solutions, but only experimentation (and disciplined causal inference) can establish incrementality. For experience operations, the additional constraint is feasibility: if the primary outcome is rare, even well-designed tests may take too long to reach a reliable conclusion.

Using the proxy funnel's purchase baseline (buyers/visitors ≈ 0.83%), a 5% relative lift requires on the order of ~766k users per variant for a two-sided test at α=0.05 with 80% power. Exhibit 8 visualizes the broader reality: small lifts imply extremely large sample sizes, which is why leading indicators and variance reduction are not "nice-to-haves" but gating design choices.

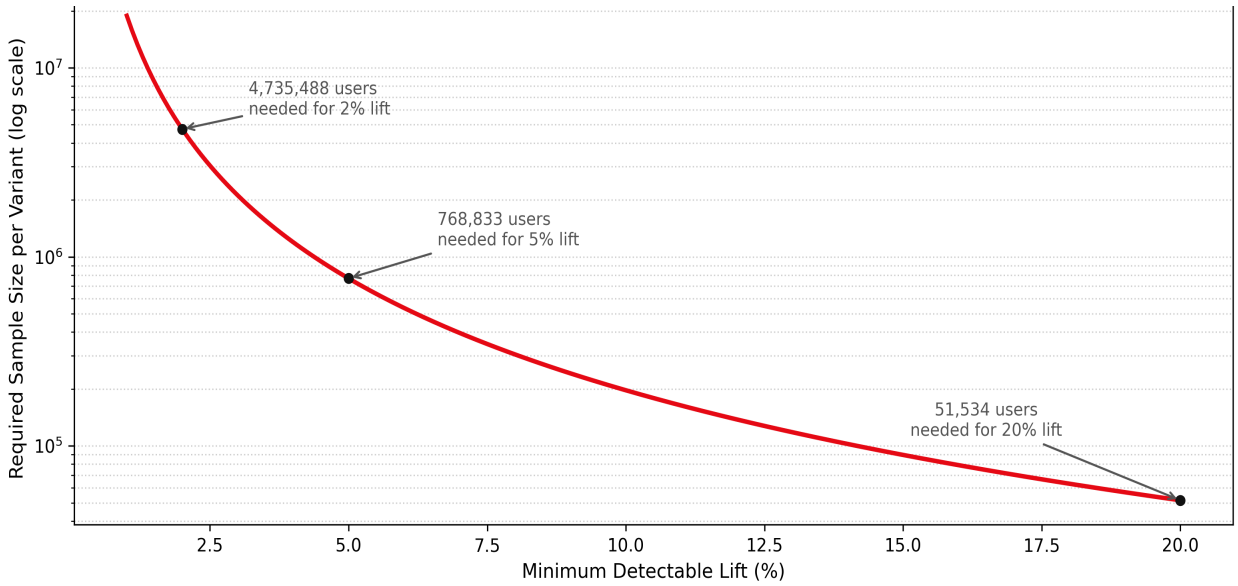


Exhibit 8. Power curve: required sample size per variant vs minimum detectable lift (baseline conversion ≈ 0.83%).

In a venue context, I would typically use a ladder of metrics: intent signals (adds-to-plan, reservations started, slot checks) as primary metrics for early-journey surfaces, and attendance/spend as confirmatory outcomes. This preserves alignment to business value while keeping tests feasible. I would also apply standard variance reduction and design techniques (pre-period covariates/CUPED, stratification by daypart and venue, and cluster designs when spillovers are likely).

9. Causal Inference as a Complement (When Randomization Isn't Available)

Not every business question can be randomized quickly. Marketing exposure, operational interventions, and experience launches often have selection effects (high-intent guests are more likely to be exposed, or ops teams prioritize interventions where they expect the biggest payoff). When that happens, naïve before/after comparisons can materially overstate impact.

Exhibit 9 shows a synthetic illustration from the notebook: a naïve comparison overestimates lift because it fails to correct for selection bias. Inverse propensity weighting (IPTW) adjusts for observable confounders and pulls the estimate closer to the ground truth. In a production environment, I would treat causal inference as a complement to experimentation, not a substitute—use it to prioritize hypotheses, to learn from operational rollouts where randomization is impractical, and to sanity-check surprising experimental outcomes.

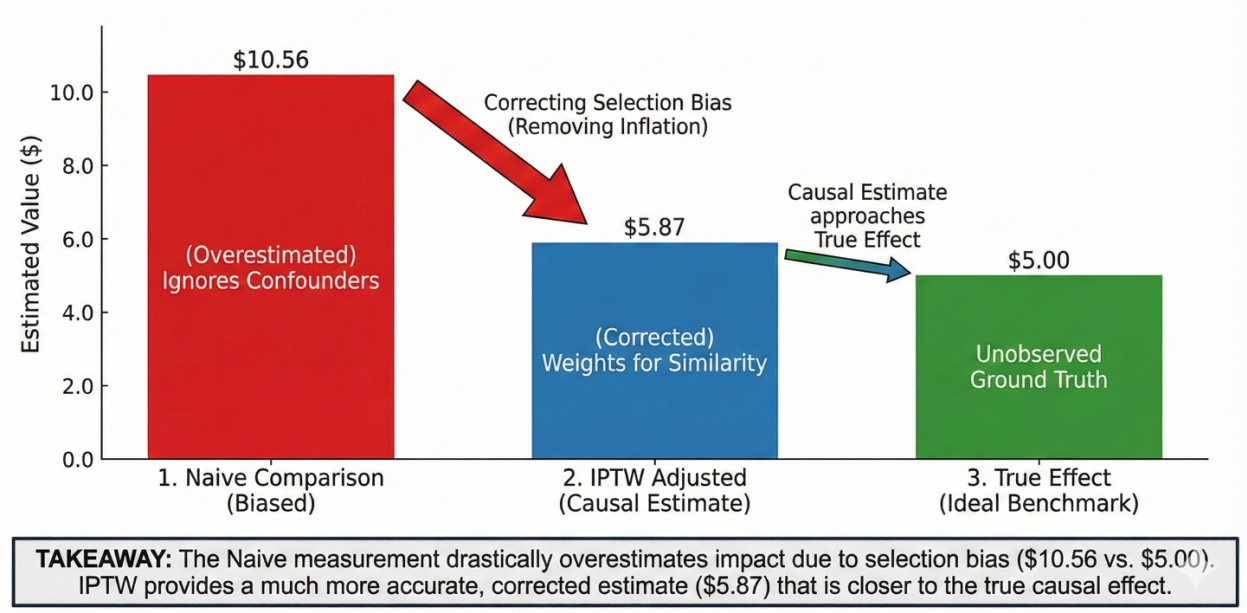


Exhibit 9. Causal inference demo: naïve lift vs IPTW-adjusted estimate (selection bias correction).

Causal approaches only work as well as the instrumentation. The minimum production requirements are: (1) assignment or exposure logging, (2) clear outcome definitions with no leakage (features must precede outcomes in time), and (3) recurring balance/diagnostic checks. Those elements also support governance: if guardrails break (e.g., queue spikes after a routing change), the system must support rollback and postmortem analysis.

Appendix A: Identity DSE Addendum

This addendum makes the Identity mapping explicit. The main body uses a public event log as a proxy to demonstrate decision-science mechanics; in an identity platform, the same mechanics apply to link quality, stability, and downstream commerce measurement.

A1. Core identity quality metrics (leading indicators)

- Coverage: percent of events/sessions attributed to an individual identity; share of unknown/unlinked traffic.
- Continuity: cross-device continuity and fragmentation (e.g., identities per account/profile; session stitching rates).
- Accuracy proxies: agreement with high-confidence anchors (login/verification/payment), collision rate checks, and sampled review sets.
- Stability: week-over-week reassignment rate; merge/split volume; drift in cluster-size distribution.

A2. Guardrails (do-no-harm constraints)

- Over-merge risk: sudden drop in unique identities, spike in top-1% cluster share, or unusual cross-household collisions.
- Over-split risk: inflated uniques, reduced continuity, and increased duplicate identities across devices/surfaces.
- Measurement integrity: exposure/assignment logging completeness, SRM checks, and stable denominators for key funnels.
- Privacy/governance: consent boundaries, retention windows, schema/versioning, and auditability of identity changes.

A3. Validation strategy (experiments + observational causal inference)

- Prefer staged rollouts with holdouts; run A/A tests and monitor SRM and logging before interpreting effects.
- Use a metric ladder: leading indicators (coverage/continuity/stability) for fast iteration; primary outcomes (attribution stability, conversion/LTV) for confirmation.
- When randomization is infeasible, use observational methods (propensity/weighting, negative controls, sensitivity checks) as a complement, not a substitute.

A4. Minimum instrumentation requirements

- Versioned identity mapping (ability to replay/compare identity versions) and a clear backfill policy.
- Exposure/assignment logs for experiments and algorithmic rollouts; timestamp hygiene to prevent leakage.
- Segment keys for monitoring (geo, device, market, surface) and alert thresholds with on-call response and rollback playbooks.

A5. 30/60/90 outline (foundational platform cadence): 0-30 days define metrics/guardrails, stand up dashboards, and validate experimentation integrity (A/A, SRM); 31-60 days ship the first change behind a staged ramp + holdout and quantify impact with segment cuts; 61-90 days operationalize release governance (versioning + rollback), automate drift detection, and connect identity improvements to commerce outcomes. Note: thresholds and success criteria should be calibrated with first-party telemetry and cross-functional definitions.