Return-Aware Experimentation: Three Rules & Three Tricks

Alex Deng @2025 NABE Tech

About Me

- 15 years in measurement for large-scale, algorithm-centric products, experiencing both sides: designing A/B testing platforms and leading feature teams.
- Been on both sides of the fence: as an experimentation platform architect and gatekeeper, and as a leader of feature team with real skin in the game.

To kick off this session, I thought it would be insightful to share **three fundamental rules of thumb** and illustrate them through **three powerful technical tricks** that have proven tremendously applicable in return-aware experimentation research.

Rule #1: It's About Business People

Organizational Goals

Every organization aims to deliver the **biggest cumulative win with limited resources**. Our role as "gatekeepers" is designing policies that optimize for this reality.

Data-Driven Culture

Organizations and individual contributors define **impact** based on metrics they can land. In this environment, **certified measurement of return** is everything that truly matters.

Human Dynamics

People deeply care about metrics because they're incentivized by them.

- This creates confirmation bias—we scrutinize rules that go against us, but rarely question those that benefit us.
- Incentivized to deeply understand the rules
- Incentivized to game the rule

Rule #2: Think Generally & Be Future-Proof

Scrutinize Assumptions

If you don't scrutinize them, your stakeholders certainly will (Rule #1).

Be transparent and self-critical

Build for Reality

True innovation comes from making theory applicable to real-world complexity.

Build for Future

Building models for highly specific scenarios yields low ROI.

Today's specific model likely won't work tomorrow. Build for future.

Rule #3: Keep Things Simple

Easy to Explain & Sell

People resist what they can't understand, especially when their goals depend on it. Remember Rule 1: if they can't grasp it, they won't trust or adopt it.

Easier Implementation

Simpler systems have fewer failure modes, making them more robust and easier to debug. This reduces development time and maintenance overhead.

Better Scalability

Simple solutions scale better. Prefer "embarrassingly parallel" approaches that enable massive parallelization across large datasets.

With these three rules in mind, let's explore three tricks that embody these principles.

Trick #1: Efficiency Augmentation — The CUPED Story

2011: The Challenge

At Bing/Microsoft, we lacked traffic for high statistical power on sessions per user—our key metric.

1

Reality Check

It failed A/A tests—consistently underestimating variance. Real user behavior is far more complex than simple Poisson.

2

Breakthrough Moment

My colleague Ya achieved 10x variance reduction using Poisson distribution and log-linear models. We were extremely excited!

3

The Revelation

We needed model-agnostic efficiency augmentation, leading to a more robust, general approach.

4

> **CUPED:** Controlled Experiment Using Pre-Experiment Data—a widely adopted technique for variance reduction.

Trick #1: Bare Minimum Assumptions & Power of Simplicity

- I.i.d. Observations

- Population Mean Shift

  Treatment groups experience a population shift in the mean. Heterogeneity is important but often a secondary-order effect.

- Central Limit Theorem

  Relying on large samples and CLT allows robust inferences even with complex underlying distributions.

The key: Δ_0 can be anything with mean 0. Adding any multiple of it to estimator Δ won't add any bias.

We call it a zero-mean augmentation. Using pre-experiment data to construct plenty of zero mean augmentation term.

$$\Delta^* = \Delta + \Delta_0$$

$$\Delta^*(\theta) = \Delta + \theta\Delta_0$$

$$Var(\Delta^*) < Var(\Delta)$$

$$\theta^* = -\frac{Cov(\Delta_0, \Delta)}{Var(\Delta_0)}$$

Bonus: these techniques appears in Reinforcement Learning methods like subtracting baseline in your reward signal, and DeepSeek's GRPO for KL estimation

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s).$$

$$\mathbb{D}_{KL}\left[\pi_\theta || \pi_{ref}\right] = \frac{\pi_{ref}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - \log\frac{\pi_{ref}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - 1,$$

Trick #2: Mixture Prior & EM Algorithm

1 Noisy Observations

We only observe a noisy version of true effects. Signal-to-noise ratios are sizable, meaning we often operate in low to mid statistical power regimes.

2 Winner's Curse

We constantly battle the tendency to overestimate treatment effects due to selection bias—we only highlight the "winners."

The Silver Lining

We have a wealth of experiment results to leverage.

Tweedie's Formula

$$\mu \sim \pi$$

$$\Delta \sim Normal(\mu, \sigma^2)$$

$$\mathbb{E}(\mu|\Delta) = \Delta + \frac{\sigma^2}{N} l'(\Delta), \quad \text{Var}(\mu|\Delta) = \frac{\sigma^2}{N}\left\{1 + \frac{\sigma^2}{N} l''(\Delta)\right\}$$

We can estimate the posterior mean directly from observed marginal distribution using log-likelihood derivative—without explicitly modeling the prior!

> This elegant result allows us to **de-bias our estimates** without strong assumptions about underlying distributions.
>
> Yet.. Homoscedasticity (constant variance) of Δ is too strong — rarely holds with varying sample sizes which is subject to heavy human influence!

Trick #2: Mixture Prior & EM Algorithm

The Pragmatic Solution: Ghidorah prior (Deng et al. 2021)

Real treatment effect distributions exhibit **fat tails**, as shown in "A/B Testing with Fat Tails (2020)" research.

Use a mixture of distributions of **0 (no effect), Gaussian (incremental idea) and Laplace (fat tail, breakthrough ideas).**

We use **EM algorithm** to iteratively fit mixture weights, prior parameters, and posterior probabilities.

> In practice, a large number of Gaussian mixtures can achieve similar results

Trick #3: Calibration with Experiment Splitting

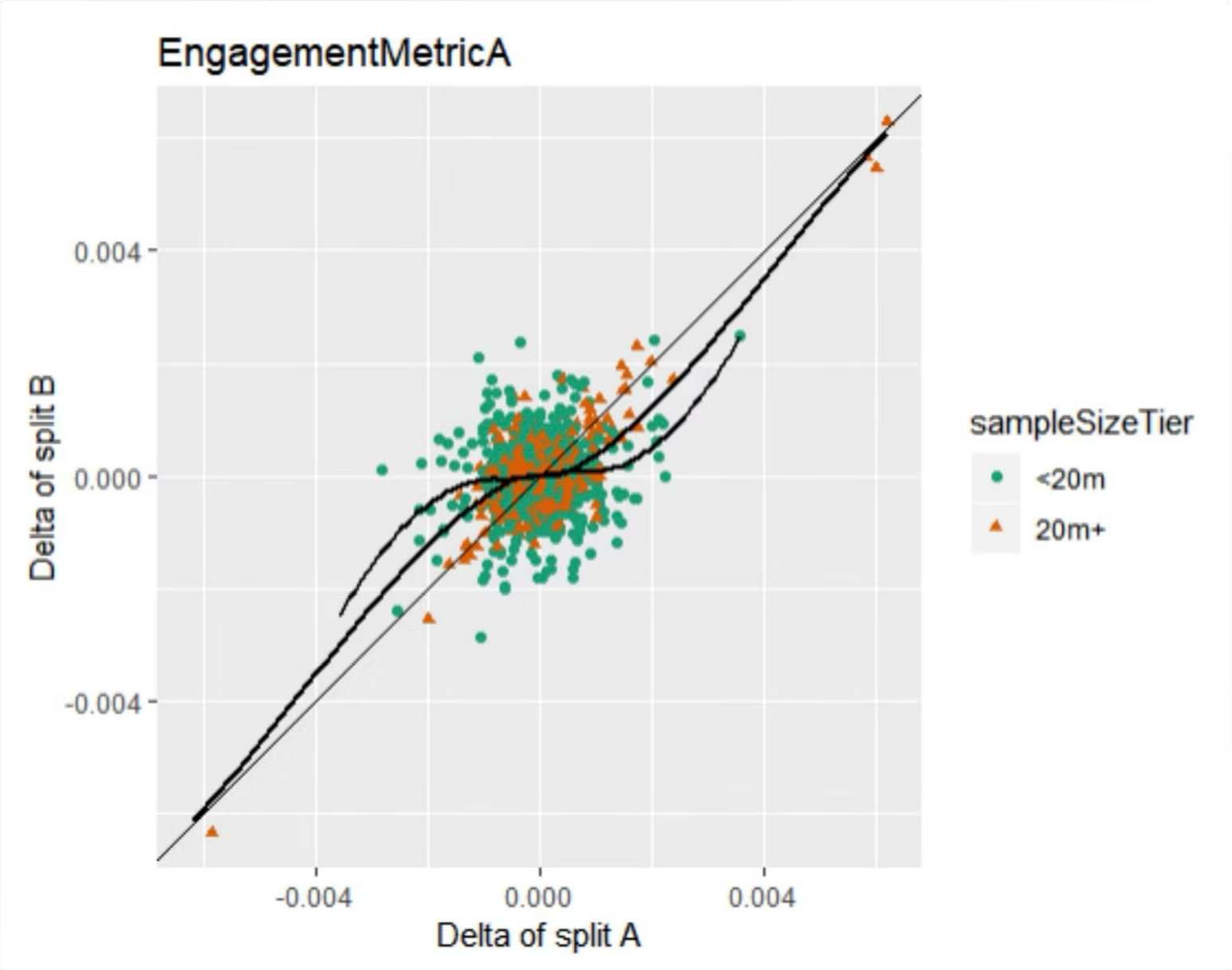
Calibration

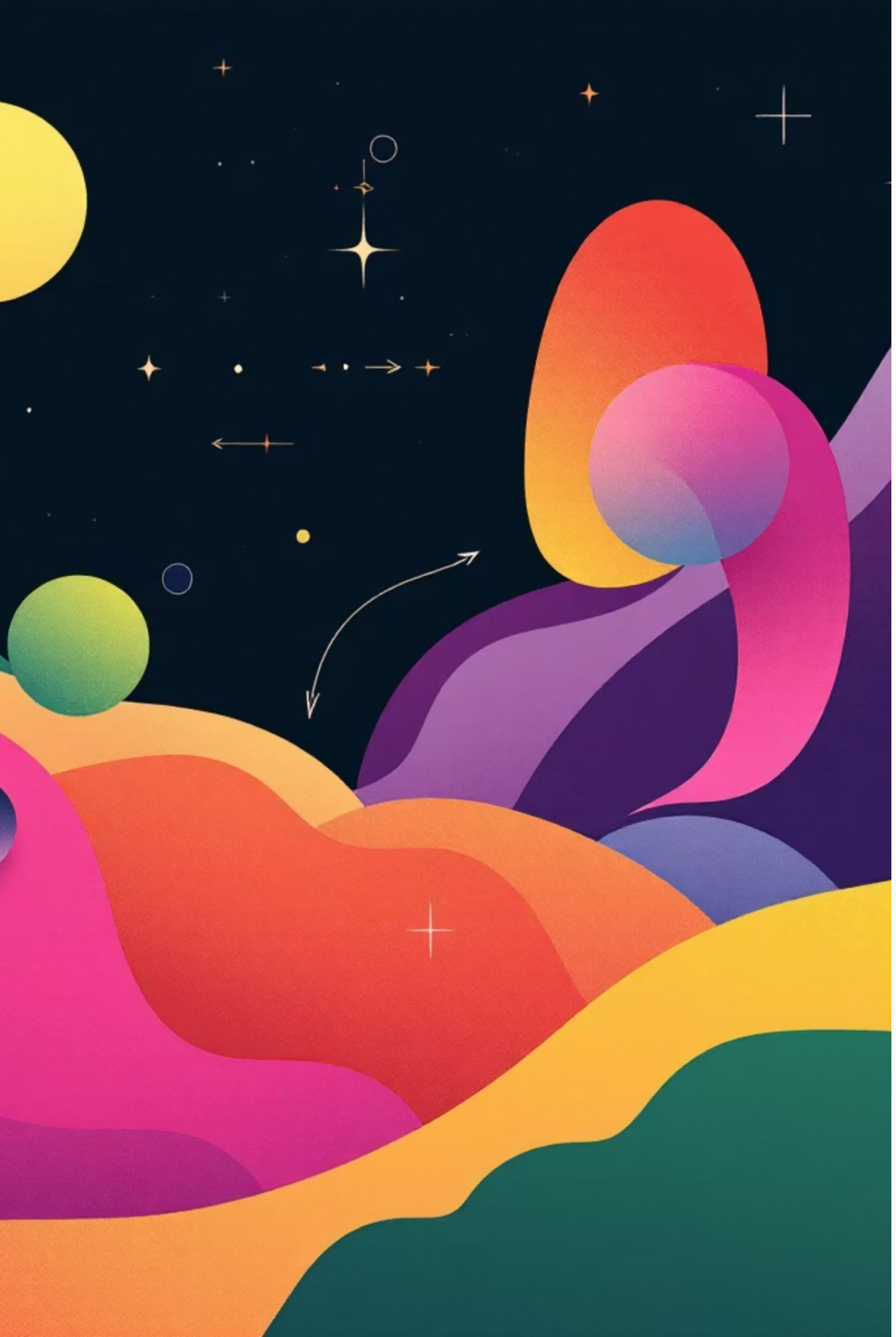
Bayesian conditional statement matches empirical data

The Breakthrough

Coey & Cunningham, 2019

Experiment splitting provides **synthetic ground truth** to enable robust model evaluation and validation.





Trick #3: Experiment Splitting

Coey & Cunningham 2019

Experiment splitting allows fitting regression models as shrinkage estimator

Deng et al. 2021

Theory-Assisted Regression with Experiment Splitting **(TARwES)** uses posterior under various prior distributions as predictors in regression

Chou et al. 2025

KDD Best Paper with deeper applications—authors present today!

Future Directions in Return-Aware Experimentation

Multi-Metric Models

Move beyond single metrics to multi-dimensional metric movement. Develop generative models that represent ideas and predict their impact across a scorecard of interconnected metrics, including noise.

Dynamic Incentives

Recognize that user incentives evolve as they understand how historical results shape priors and affect return estimates. This feedback loop can alter treatment effect distributions, influencing the mix of incremental vs. breakthrough ideas.

Metadata Governance

Emphasize the critical role of high-quality, consistent metadata. Robust governance and quality control are essential for the reliability and utility of experimentation systems and are rich areas for future research.

Thank you for your attention!

Thank You!