ADKDD 2026 · JEJU, KOREA

Estimating Heterogeneous Treatment Effects in Online Advertising Experiments: A Hierarchical Bayesian Approach

A production framework for aggregating treatment effects across thousands of noisy ad campaigns.

Aljoša Vodopija · Anže Alič · Tim Poštuvan · Martin Jakomin · Blaž Škrlj — Teads, Ljubljana

teads

MOTIVATION

Sparse conversions, heterogeneous effects

Judged on conversions, not clicks

Performance advertising is bought on CPA and conversion rate — the metrics experiments must move.

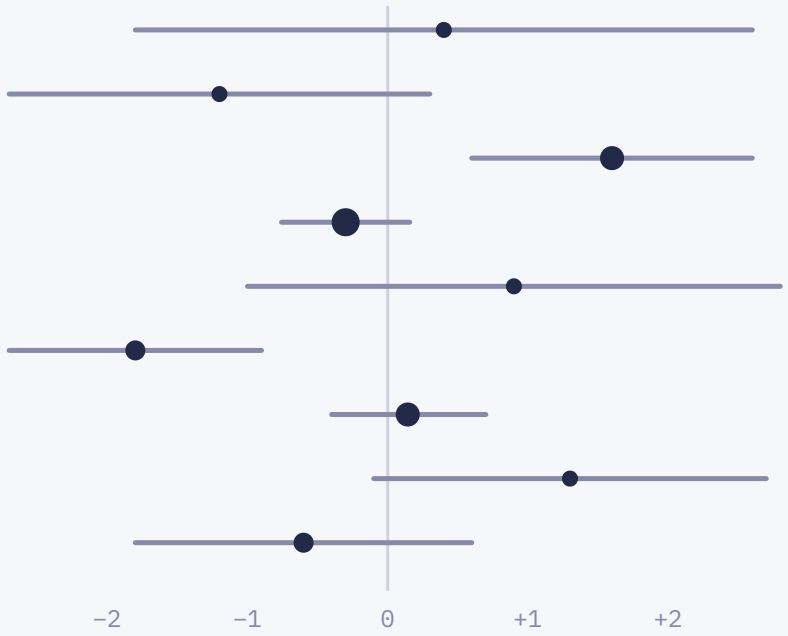
Conversions are sparse and diverse

From low-intent page visits to high-value purchases — inflating the variance of observed CPA.

Effects genuinely vary by campaign

Targeting, creatives, and auction dynamics differ — treatment-effect heterogeneity is real, not noise.

ONE EXPERIMENT, CAMPAIGN BY CAMPAIGN



Campaign-level effect estimates $\hat{\theta}_i$ — wide intervals from sparse counts, scattered centers from real heterogeneity.

THE GAP

The limits of inverse-variance weighting

THE STANDARD BASELINE

Treat each campaign as an independent unit: estimate a log rate-ratio per campaign, then aggregate with inverse-variance weights.

$$\hat{\theta}_i = \log\left(\frac{(c_{i1} + 0.5) \,/\, s_{i1}}{(c_{i0} + 0.5) \,/\, s_{i0}}\right)$$

$$\hat{\theta} = \frac{\sum_i w_i \hat{\theta}_i}{\sum_i w_i}, \qquad w_i = \frac{1}{\mathrm{SE}_i^2 + \hat{\tau}^2}$$

Random-effects form — $\hat{\tau}$ estimated with Paule-Mandel, a moment-based estimator.

WHERE IT BREAKS

$\hat{\tau}$ comes from a single noisy experiment. Underestimate it, and standard errors shrink — false positives inflate. Sparse counts and high heterogeneity is exactly the regime CPA experiments live in.

And the academic literature on CPA experimentation is thin — conversions treated as equivalent, or results pooled at coarse levels like advertiser vertical.

CONTRIBUTIONS

A framework built for production

Three contributions, validated offline and online.

1

Hierarchical Bayesian model

Campaign-level effects and cross-campaign heterogeneity modeled jointly — stable under sparse conversions.

2

Bayesian sequential testing

HDI + ROPE decisions recomputed hourly as data accumulate — principled early stopping in production.

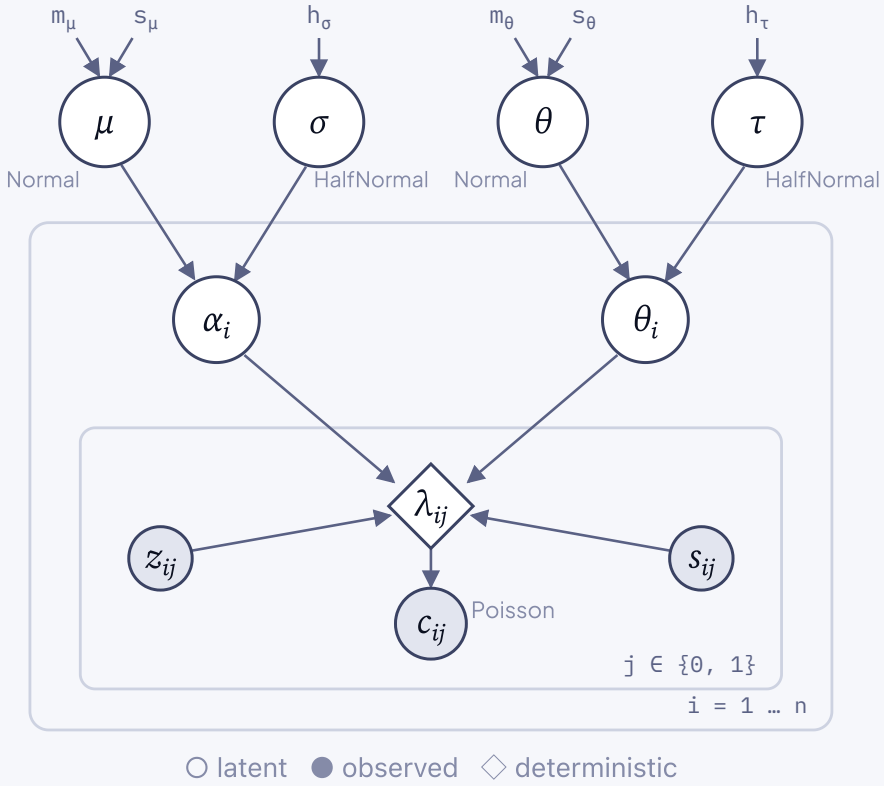
3

Deployed at scale

Live on a DSP handling 5M+ requests/sec — 35,000+ campaigns, 3,000+ advertisers, three case studies.

METHOD · 1 OF 2

The hierarchical Bayesian model



LIKELIHOOD — SPEND AS EXPOSURE

$$c_{ij} \sim \text{Poisson}(\lambda_{ij})$$

$$\log \lambda_{ij} = \alpha_i + \theta_i z_{ij} + \log s_{ij}$$

PRIORS — PARTIAL POOLING

$$\alpha_i \sim \mathcal{N}(\mu, \sigma^2), \qquad \theta_i \sim \mathcal{N}(\theta, \tau^2)$$

Low-volume campaigns shrink toward the population mean; high-volume campaigns keep their own signal.

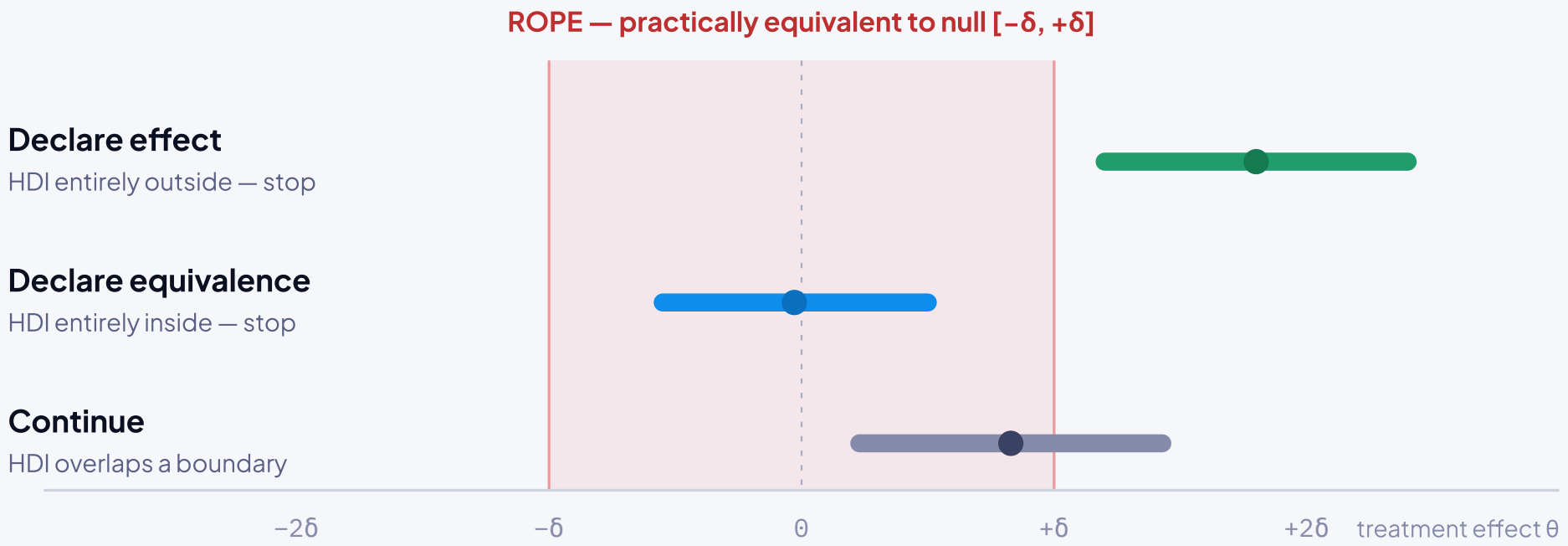
HYPERPRIORS — WEAKLY INFORMATIVE

Scaled in units of the reference effect δ, so the data dominate.

Unlike IVW, θ and τ are inferred *jointly* — uncertainty about heterogeneity flows into the posterior of the global effect.

METHOD · 2 OF 2

Sequential testing with HDI and ROPE



Posterior refreshed hourly as data arrive

δ calibrated from historical A/A tests

Precision stop: HDI width drops below a threshold

OFFLINE EVALUATION

Calibrated simulations

DATA-GENERATING PROCESS

Mirrors production — calibrated to five days of DSP data.

- 3,000 advertisers, ~9 campaigns each.
- Campaign CPA and spend drawn log-normally.
- Per-campaign effects $\theta_i \sim \mathcal{N}(\theta, \tau^2)$.

DELIBERATELY MISSPECIFIED TO TEST ROBUSTNESS

- Conversions drawn NegBin (overdispersed) — the model assumes Poisson.
- Advertiser hierarchy induces correlation the model does not see.

Protocol

θ	0 · 1.5δ · 2δ · 3δ
τ	1δ · 2δ · … · 10δ
200 runs	per setting · α = 5%
Measures	type I error · power · coverage · RMSE · bias

HBM fit with PyMC / NUTS — 7–9 min per 30k-campaign snapshot on 4 cores.

OFFLINE EVALUATION

Reliability under heterogeneity

35%

IVW type I error (τ = 10δ) — 6% for HBM.

80%

HBM power at typical production heterogeneity (τ = 5δ) — 79% for IVW.

0.02δ vs. **0.32δ**

Bias under the null — HBM close to unbiased, IVW positive.

READ

IVW is fine up to τ ≈ 4δ, then fails exactly where CPA experiments live. Its high-τ power is miscalibration, not sensitivity.

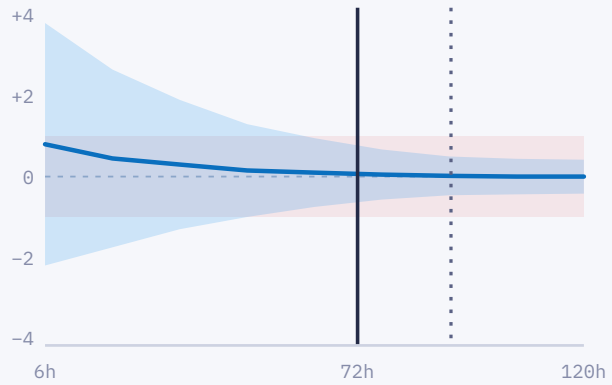
ONLINE VALIDATION

Three production case studies

95% HDI — mean — ROPE ±δ | stop ⋮ precision

A/A validation

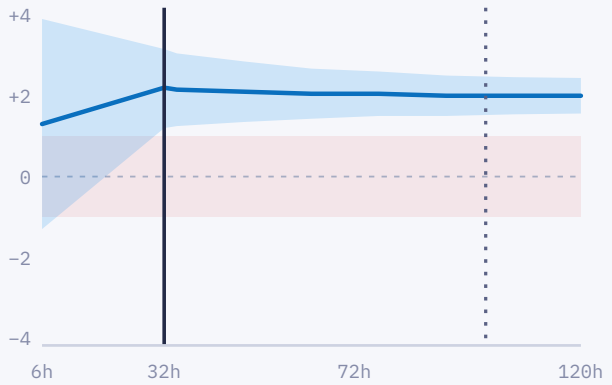
Equivalence · 72 h



Posterior converges to zero; HDI falls fully inside the ROPE. Pipeline validated end to end.

New CVR model

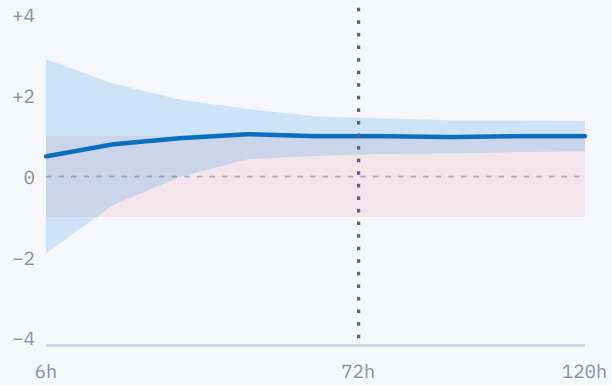
Effect · 32 h



+2δ lift; HDI clears the ROPE at ~32 h — stopped well before the 5-day horizon.

Creative retrieval

Precision stop · 72 h



+δ; HDI above zero — conventional tests ship it. ROPE says the effect's too small to justify rollout.

CONCLUSION

Takeaways

01 Choose by regime — HBM matches IVW at moderate heterogeneity and is markedly more reliable when heterogeneity is high.

02 HDI + ROPE turns posteriors into decisions in business units — statistical detection ≠ practical significance.

03 Running in production today, informing platform-wide rollout decisions.

↗ `github.com/outbrain-inc/ab-tea-lab`