tatari

AdKDD 2026

Causal Ad Incrementality without Experiments

Addressable-Audience CEM + DML at Billion-Impression Scale

Jaeyeong Kim · Alexander J. Rasmussen · Jake P. Reschke · Abdullah M. Spall

Tatari

A measurement gap worth explaining

One B2B SaaS campaign, two internal estimators — a measurement disagreement, not a causal chain.

Incumbent estimate — SBI

6.5K

28-day incremental visits

Proposed pipeline — CEM + DML

33.8×

220K

28-day incremental visits

SBI = Simulation-Based Incrementality (the incumbent estimator) · CEM = Coarsened Exact Matching · DML = Double Machine Learning

Why experiments are not the routine answer

Gold-standard causal tools are valuable, but hard to run continuously for every advertiser.

Ghost ads

Require platform cooperation and user-level randomization.

Key Limits

Cost Privacy Speed

PSA controls

Can lose randomization under optimizer routing.

Key Limits

Cost Speed Privacy

Geo lift tests

Slow, noisy, and not scalable to hundreds of clients.

Key Limits

Cost Speed Scalability

PSA = Public-Service Announcement (a no-op ad shown to a control group)

The pipeline: a counterfactual from addressable users

Novelty is the middle: addressable-audience controls plus symmetric per-pair attribution windows.

01

Exposure logs

Target campaign impressions

02

Addressable controls

Users seeing other advertisers' impressions

03

CEM + windows

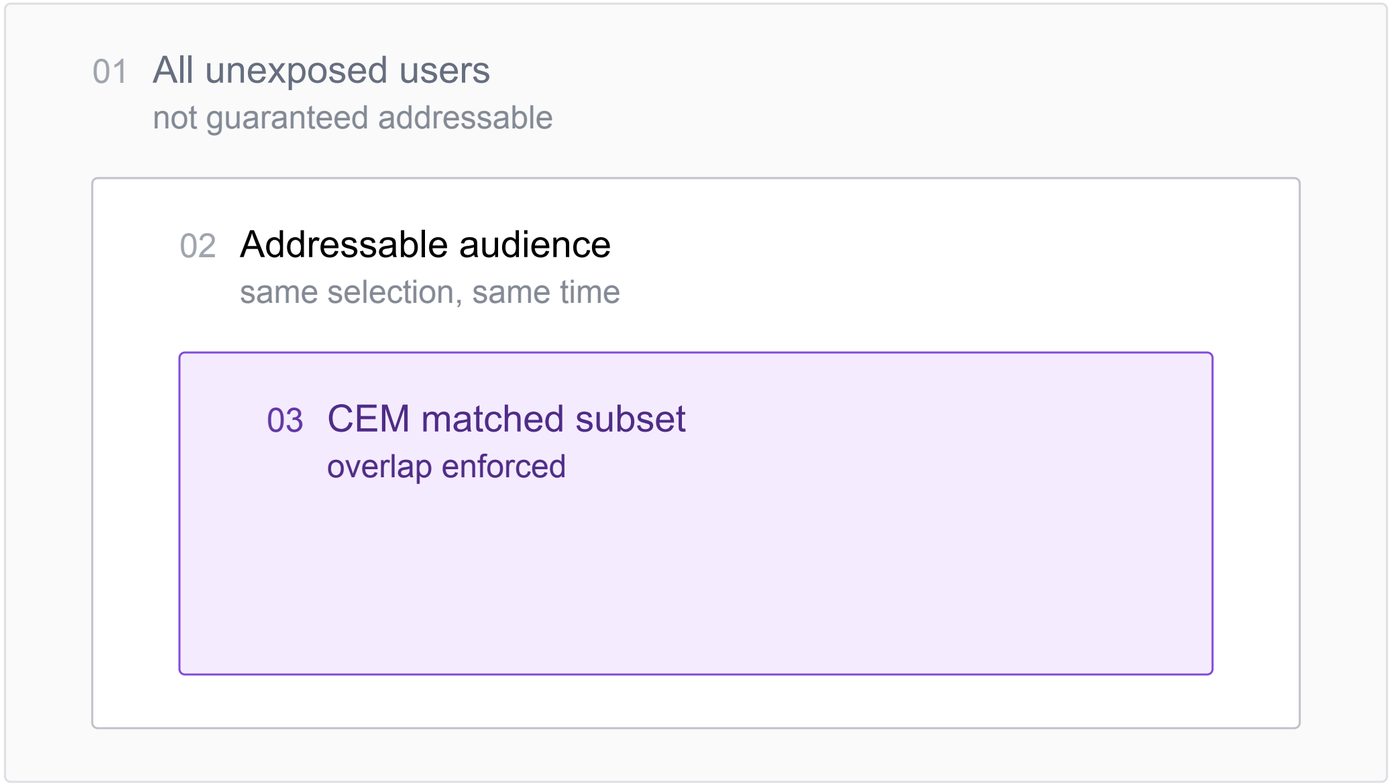
1:k controls; per-pair copied attribution windows

04

DML / routing

Causal lift and confidence intervals

Control construction reduces first-order selection bias



~0.2% of the treated group was not matched

CEM handles overlap; DML handles residualization

CEM

- Groupby-bucket-join matching
- Bounded covariate imbalance
- Drops no-overlap cells

R-learner DML

- Cross-fit propensity
- Residualized outcome
- AIPW confidence intervals

Dense outcome	Sparse outcome	Very sparse
Direct DML	Surrogate DML	IPW

Assumption: the sparse route relies on a site-visit surrogate; thresholds are deployment-fixed.

IPW = Inverse Propensity Weighting · AIPW = Augmented IPW

Validation triangulates, but does not prove, causality

No observational study proves causality alone, so we ask three independent checks to agree.

3/4

geo lift tests agree with DML — the only high-precision test rejects the naive baseline, not DML

8/8

A/A placebos cover zero

1.4%

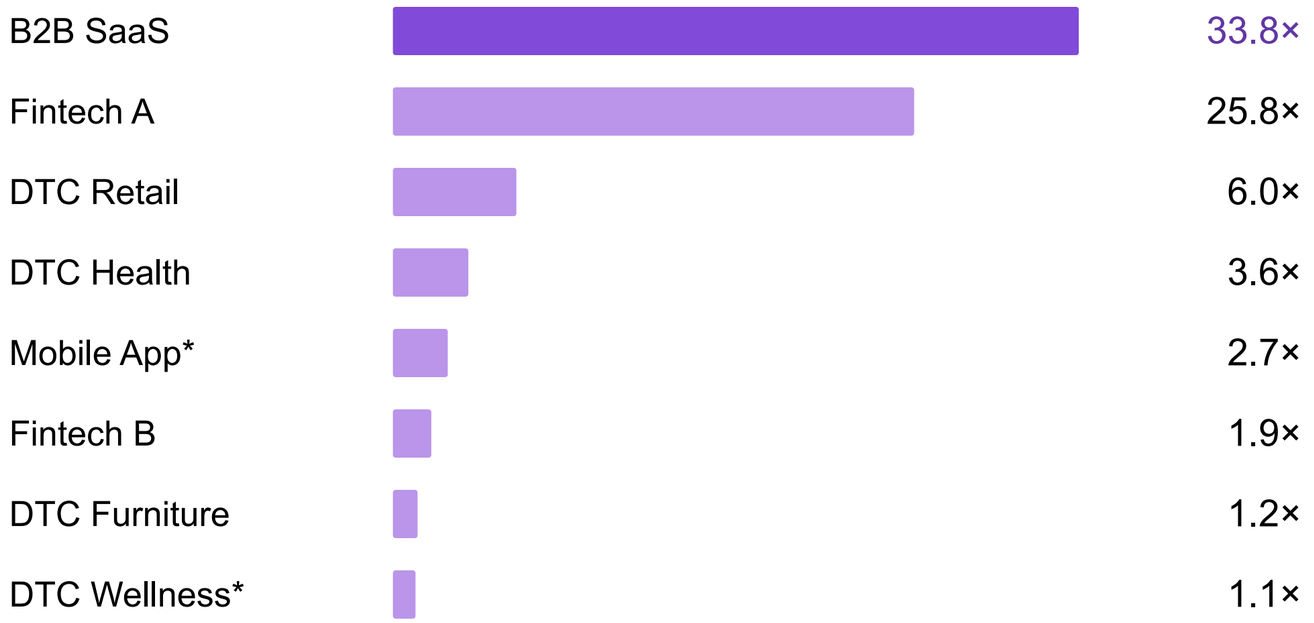
LL-DW PSID-1 error under CEM + R

Interpretation: encouraging triangulation, not proof; geo calibration is still limited — detailed sweeps and misses are in backup for Q&A.

LL-DW = LaLonde-Dehejia-Wahba, a public causal-inference benchmark · PSID = Panel Study of Income Dynamics (its control pool) · A/A = a placebo test comparing two untreated groups

DML reports higher visit lift than SBI

Interpretation: evidence of systematic estimator disagreement across eight advertisers; the mechanism hypothesis is SBI's pre-impression baseline over-sampling the advertiser's existing-user pool.



DML ÷ SBI, 28-day incremental visits · * post-filter · DTC = Direct-to-Consumer

Public RCT sanity check: CEM does most of the correction

LaLonde-Dehejia-Wahba PSID-1 · ATT error vs experimental truth of $1,794

Raw DiM error

$16,971

CEM-only DiM error

$2,787

CEM + R point error

$25

84% of DiM error removed by CEM alone

78% of NSW-treated retained by CEM

1.4% CEM + R PSID-1 point error

ATT = Average Treatment Effect on the Treated · DiM = Difference in Means · NSW = National Supported Work (the randomized treated group)
RCT = Randomized Controlled Trial

Takeaway

A practical observational path to continuous incrementality measurement for platform operators.

01

Where it fits

Platforms with impression logs, addressable controls, covariates, and privacy-compliant linking of ad exposures to conversions

02

What supports it

Geo triangulation, A/A placebos, public RCT recovery

03

What remains

Residual confounding, sparse-surrogate diagnostics, identity-graph features, and more geo tests/calibration

tatari

Thank you — questions?

jaeyeong@tatari.tv

Geo validation estimator sweep

Backup

Full sweep; the operational pipeline is row D. ✓ = inside the geo SCM 95% CI.

Method	DTC Pet Food	Fashion all	Fashion new	Vitamins
Geo SCM (95% CI)	−66 [−240, 123]	987 [572, 1413]	479 [308, 603]	61 [9, 116]
A. Naive + DiM	419 ✗	2,909 ✗	1,279 ✗	99 ✓
B. CEM + DiM	246 ✗	1,043 ✓	409 ✓	77 ✓
C. Naive + R	−85 ✓	781 ✓	28 ✗	69 ✓
D. CEM + R (operational)	−88 ✓	853 ✓	285 ✗	26 ✓
E. Naive + IPW	−52 ✓	912 ✓	14 ✗	74 ✓
F. CEM + IPW	−7 ✓	931 ✓	342 ✓	37 ✓

SCM = Synthetic Control Method · CI = Confidence Interval

Production A/A and SBI table

Backup

Upper-funnel 28-day visit estimand.

Advertiser	SBI 28d Incr.	DML 28d Incr.	DML/SBI	A/A 28d Lift
Fintech A	78.9K	2.04M	25.8×	+0.92%
Fintech B	116.2K	226K	1.9×	−4.70%
DTC Health	534.3K	1.90M	3.6×	+2.17%
DTC Furniture	123.1K	147K	1.2×	+7.84%
DTC Wellness*	126.6K	137K	1.1×	+1.12%
B2B SaaS	6.5K	220K	33.8×	+2.87%
Mobile App*	3.3K	8.9K	2.7×	−0.39%†
DTC Retail	47.4K	286K	6.0×	+1.74%

* post-filter · † unfiltered A/A due to small-N cascade.

LaLonde estimator table

Backup

ATT mean ($); experimental truth = $1,794.

Estimator	PSID Raw	PSID CEM-c	PSID CEM-w	CPS Raw	CPS CEM-c	CPS CEM-w
DiM	−15,177	−986	−8,397	−8,486	−166	−5,542
R (operational)	939	1,819	1,645	897	1,998	1,872
S	−808	420	−169	−115	669	124
T	−1,513	1,372	543	−481	1,708	1,035
X	−1,137	1,336	529	501	1,852	1,230
DR	−69	1,726	1,352	1,057	1,895	1,502
NonParamDML	2,156	1,359	1,908	1,589	2,384	2,246
DRLearner	611	886	612	1,106	1,870	970

CPS = Current Population Survey (a second external control pool)

Runtime table

Backup

Nie-Wager Setup B DGP; fixed 10× i3.4xlarge cluster.

n per group	EconML NonParamDML	EconML DRLearner	PySpark R-learner
10^5	91 s	92 s	59 s
10^6	996 s	1,013 s	127 s
10^7	timeout	timeout	1,098 s

PySpark is 8× faster at 10^6 and the only method completing at 10^7.

CEM cell to per-pair window mechanics

Backup

CEM creates cells; the implementation then samples 1:k matched controls per treated anchor.

01

CEM cell

Same coarsened pre-treatment activity / demographics

02

1:k matching

k=5 default; modular stride; replacement if a cell is small

03

Copy window

Treated anchor and endpoint copied to each matched control

04

Variance

Cluster-robust IF SEs; IP-clustered bootstrap for replacement dependence

Open paper gap: direct ablation of the per-pair window remains future work.

Data access, privacy, and generalizability

Backup

Data needed	Impression logs, addressable-audience control pool, pre-treatment covariates, outcome logs, privacy-compliant linking of exposures to conversions
Best fit	CTV measurement platforms, DSPs, walled gardens, or platform-user-pool variants with equivalent eligibility data
Harder fit	Ordinary advertisers, MMP-only reporting, or SRN channels without impression-level eligible/control data
Privacy position	Avoids new randomization or holdouts; relies on existing delivery/outcome logs and in-platform matching; model outputs are user-level uplift, aggregated for reporting today — user-level targeting/recommendation is planned
Identifier risk	Geo-IP vendor disagreement and the coarse post-hoc ID-resolution filter remain operational risks; continuous identity-graph features are future work

MMP = Mobile Measurement Partner

Sparse routing assumptions

Backup

The router is practical for production, but the surrogate route has a causal assumption.

Dense

Direct DML

Use when outcome counts support residualized outcome modeling.

Sparse

Surrogate DML

Use site visits as a dense mediator/proxy; may attenuate effects when brand paths bypass visits.

Very sparse

IPW

Use Hajek-normalized weights with exact/bootstrap inference when DML is unstable.

Known gap: threshold calibration and surrogate-vs-direct diagnostics on dense outcomes should be expanded.

Limitations

Backup

Unmeasured confounding	Cannot rule out latent intent, targeting, or auction dynamics correlated with addressability. Four facts bound it: DML sits at/below the geo midpoint on every test, 8/8 A/A placebos cover zero, the active-user mechanism (not our matching) drives the SBI gap, and LL-DW recovers ATT within 1.4% — encouraging bounds, not proof. Formal Rosenbaum sensitivity is future work.
Estimand	Matched-subset ATT after CEM overlap enforcement, not full-population ATE (Average Treatment Effect).
Geo validation	Encouraging triangulation, but only one geo test is tight enough to strongly discriminate among estimators; more geo tests are planned to sharpen the comparison.
Post-hoc filter	Two advertisers use a coarse ID-resolution flag filter; identity-graph features should replace this.
Surrogate routing	Useful for rare conversions; threshold calibration and surrogacy diagnostics should be expanded.

Scope: a platform-level measurement method

Requires

Impression logs

Addressable-audience pool

Pre-treatment covariates

Outcome linkage

Fits

CTV measurement platforms

DSPs

Walled gardens

Identity-safe environments

Does not assume

Ordinary advertiser data

SRN black-box reporting

Public cross-sharing

no new randomization or holdout; in-platform matching; user-level uplift, aggregated for reporting.

CTV = Connected TV · DSP = Demand-Side Platform · SRN = Self-Reporting Network

System scale: a production engineering benchmark

At 10^7 rows per group, single-node references time out; PySpark finishes in about 18 minutes.

n per group	EconML NonParamDML	EconML DRLearner	PySpark R-learner
10^5	91 s	92 s	59 s
10^6	996 s	1,013 s	127 s
10^7	timeout	timeout	1,098 s

Benchmark: Nie-Wager Setup B DGP, fixed 10× i3.4xlarge cluster; distributed Spark vs single-node EconML. Spark makes the estimator operational — an engineering benchmark, not algorithmic superiority.

DGP = Data-Generating Process