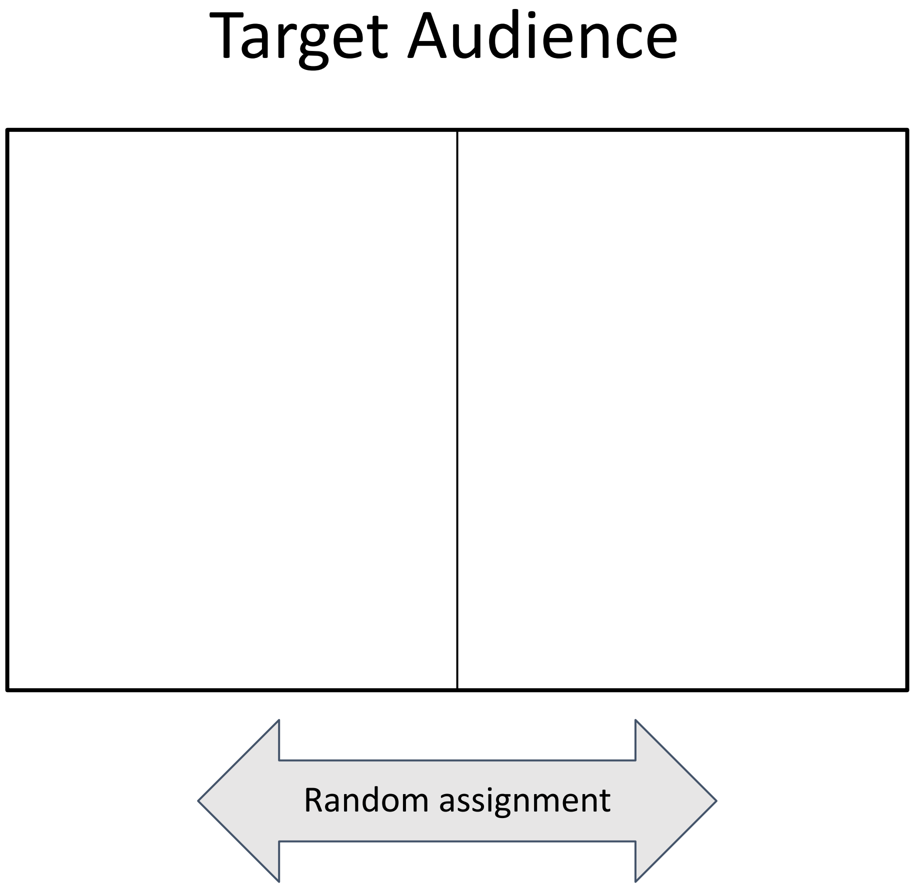
Predicted Incrementality by Experimentation (PIE)

Florian Zettelmeyer
Kellogg School of Management, Northwestern University

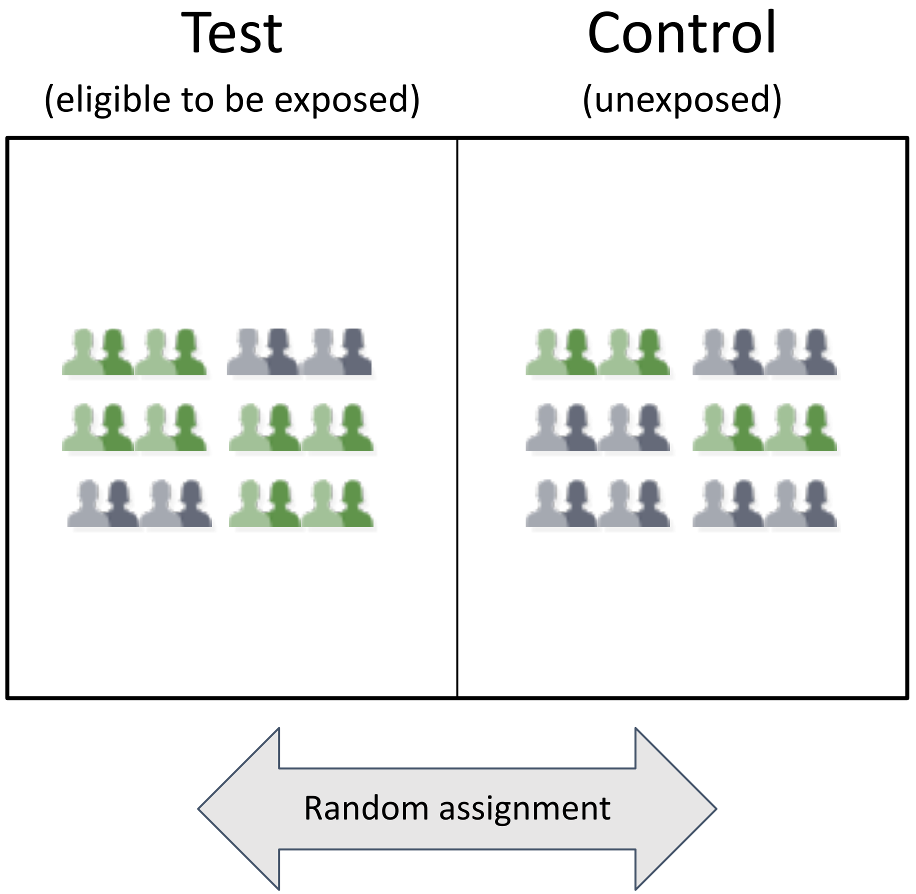
Joint work with Brett Gordon (Northwestern) and Robert Moakler (Meta)

MSI 2024 Summit

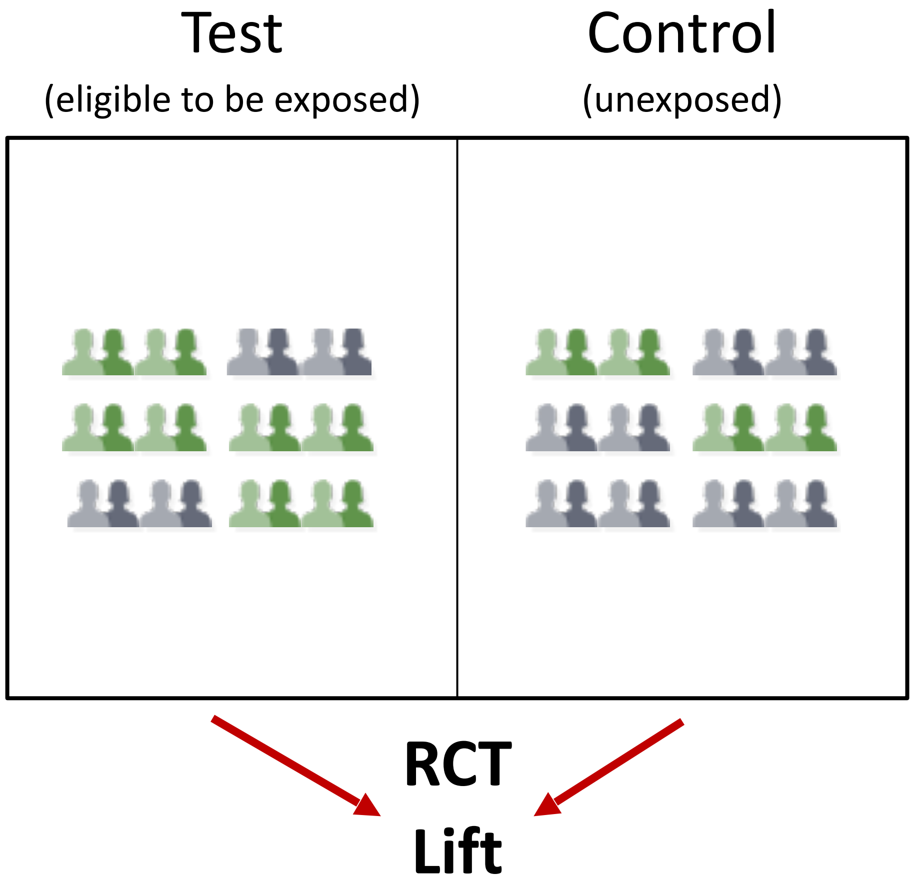
Randomized controlled trials (RCTs) are recognized as the “gold standard” to measure incremental (causal) effects



Randomized controlled trials (RCTs) are recognized as the "gold standard" to measure incremental (causal) effects



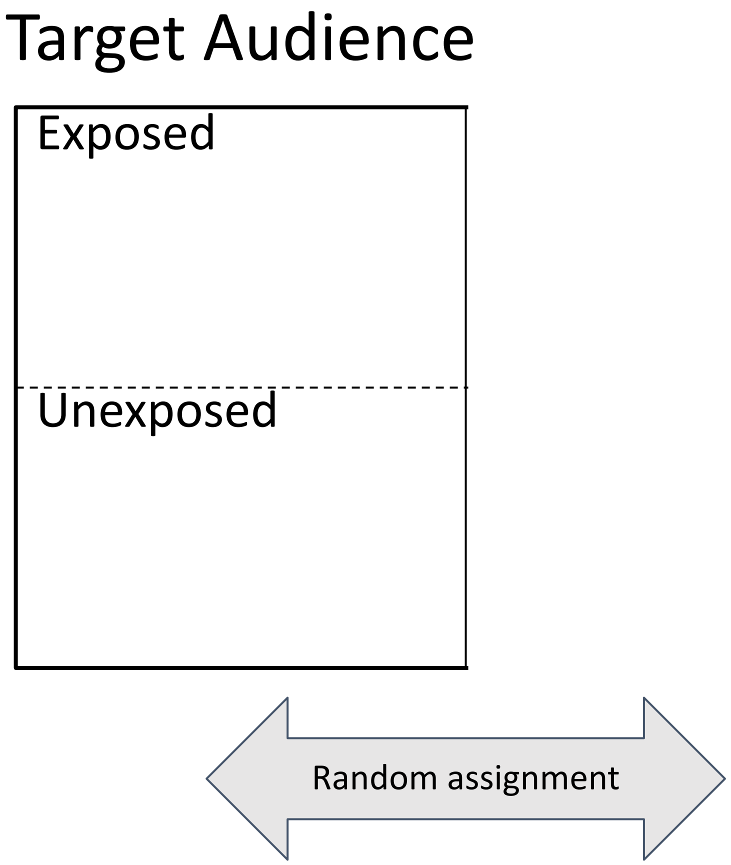
Randomized controlled trials (RCTs) are recognized as the "gold standard" to measure incremental (causal) effects



Randomized controlled trials (RCTs) are recognized as the "gold standard" to measure incremental effects

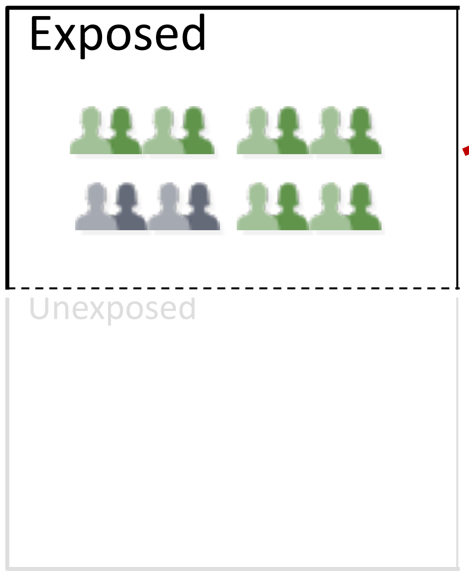
But what if you can't run an RCT?

Imagine we don't have an RCT…



Proxy metrics are a common approach

Target Audience

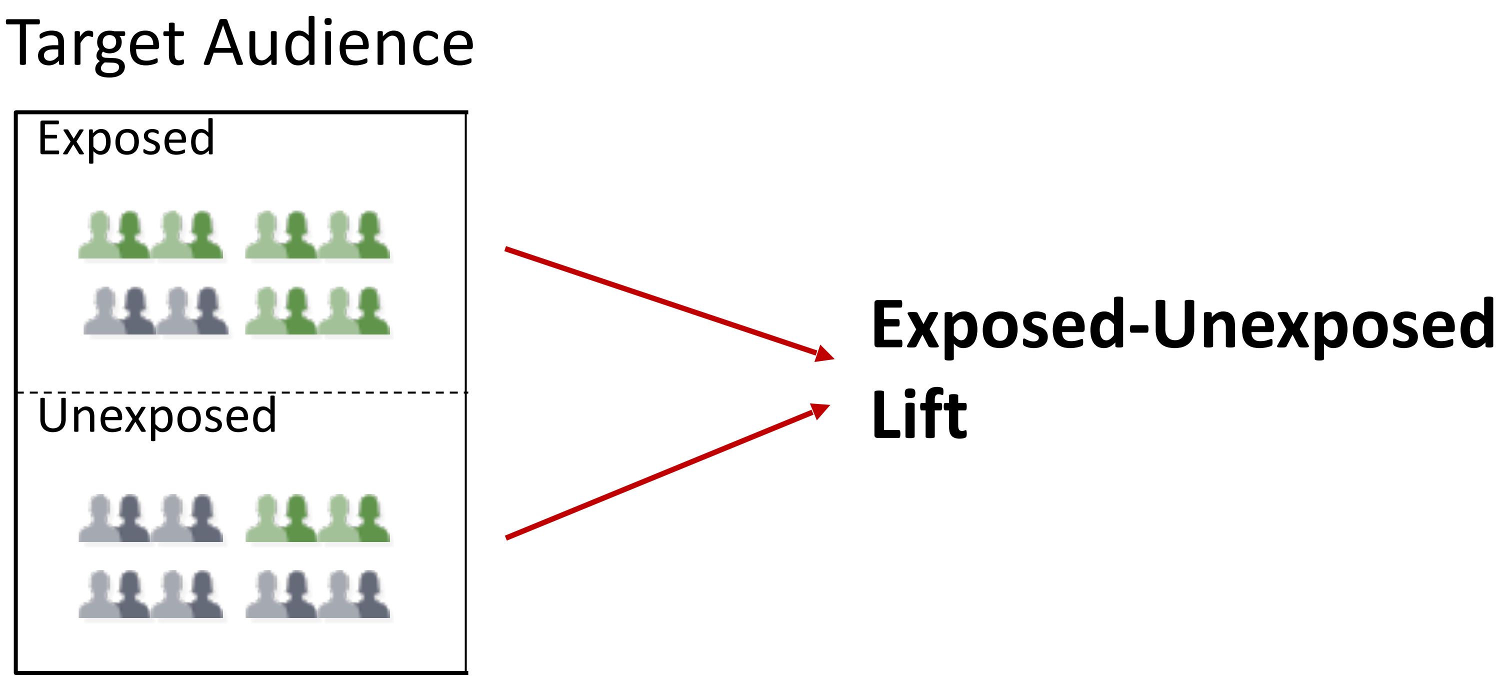


Last Click (LC):

- Start with an outcome (e.g., purchase)
- Attribution window (e.g., 7 days)
- "Attribute" purchase to ad that was clicked last in attribution window
- No counterfactual (control group)

Problem: Can over or underestimate incrementality

Alternatively, we could compare outcomes between people who saw versus did not see the ad campaign



Problem: Suffers from “selection bias” into ad exposure

"Undoing" the selection induced by ad-targeting algorithms using causal inference approaches has been unsuccessful

EVIDENCE

- *Gordon, Zettelmeyer, Chapsky, Bhargava (2019), Marketing Science*
  - Compare RCTs with (observational) program evaluation approaches
  - 15 studies, hand-selected -> cannot come close replicating RCT results
- *Gordon, Moakler, Zettelmeyer (2023), Marketing Science*
  - 1673 RCTs, representative
  - SPSM, Double/Debiased ML + Deep Learning for propensity score
  - From 30 to (nearly all ~ 5000) logged features at FB
  - Equally depressing …

RCT Lift vs. Lift from SPSM & DML

Funnel Level of Outcome	Median Lift		
	RCT	SPSM	DML
Upper	29%		
Mid	18%		
Lower	5%		

The **RCT lift** estimates and …
- **SPSM** are statistically different in 1482 / 1673 = 89% of the RCTs
- **DML** are statistically different in 1258 / 1673 = 75% of the RCTs

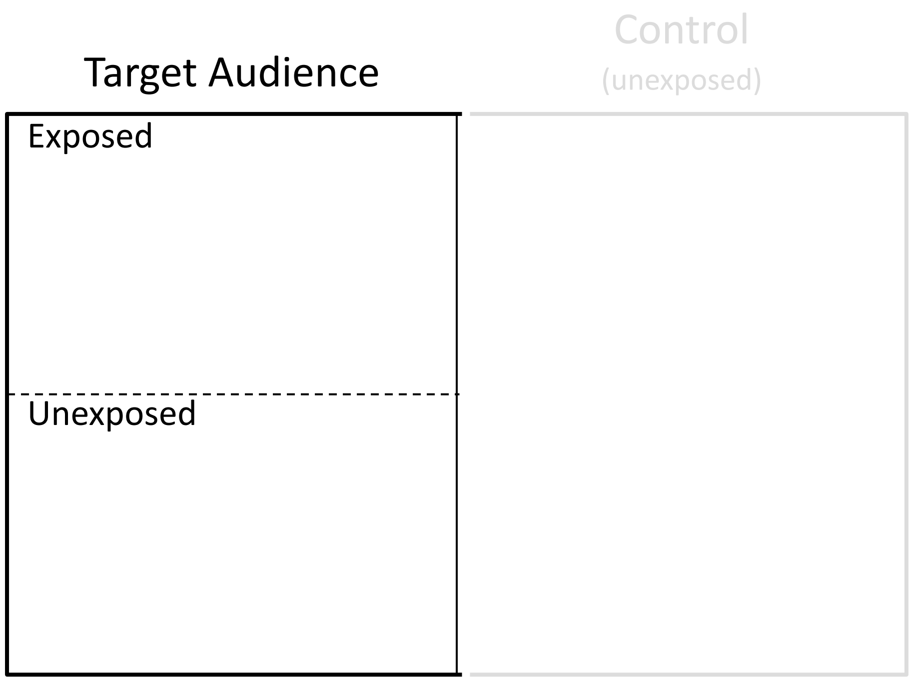
So, what should advertisers do?

<u>Don't have the data</u> for observational methods

And <u>can't run RCTs</u> all the time

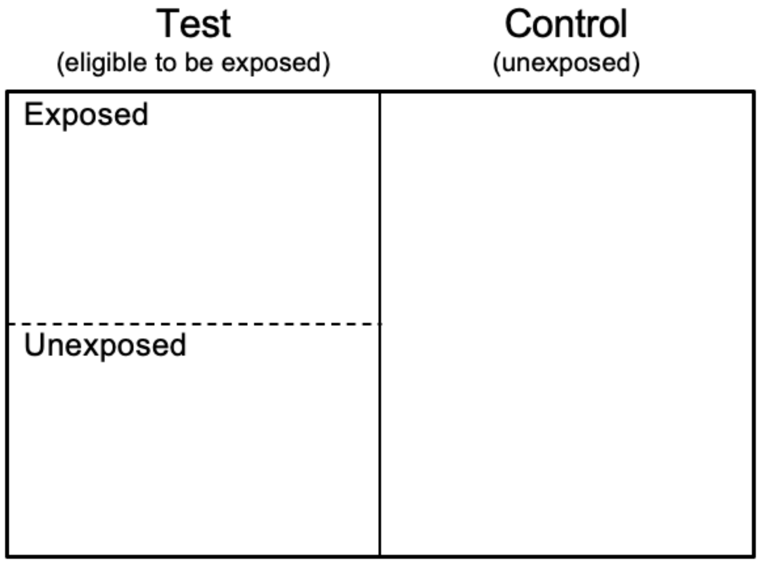
We have tried to estimate the causal effect of advertising without RCTs by controlling for user-level selection bias

“Traditional” Causal Inference methods



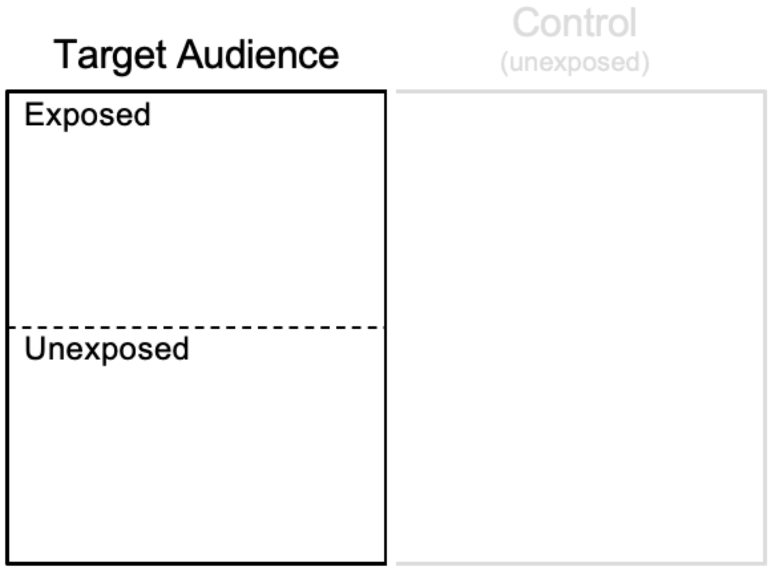
We often have RCTs for a subset of advertising campaigns…

Ad campaigns as RCTs



RCT Lift

Ad campaigns not as RCTs

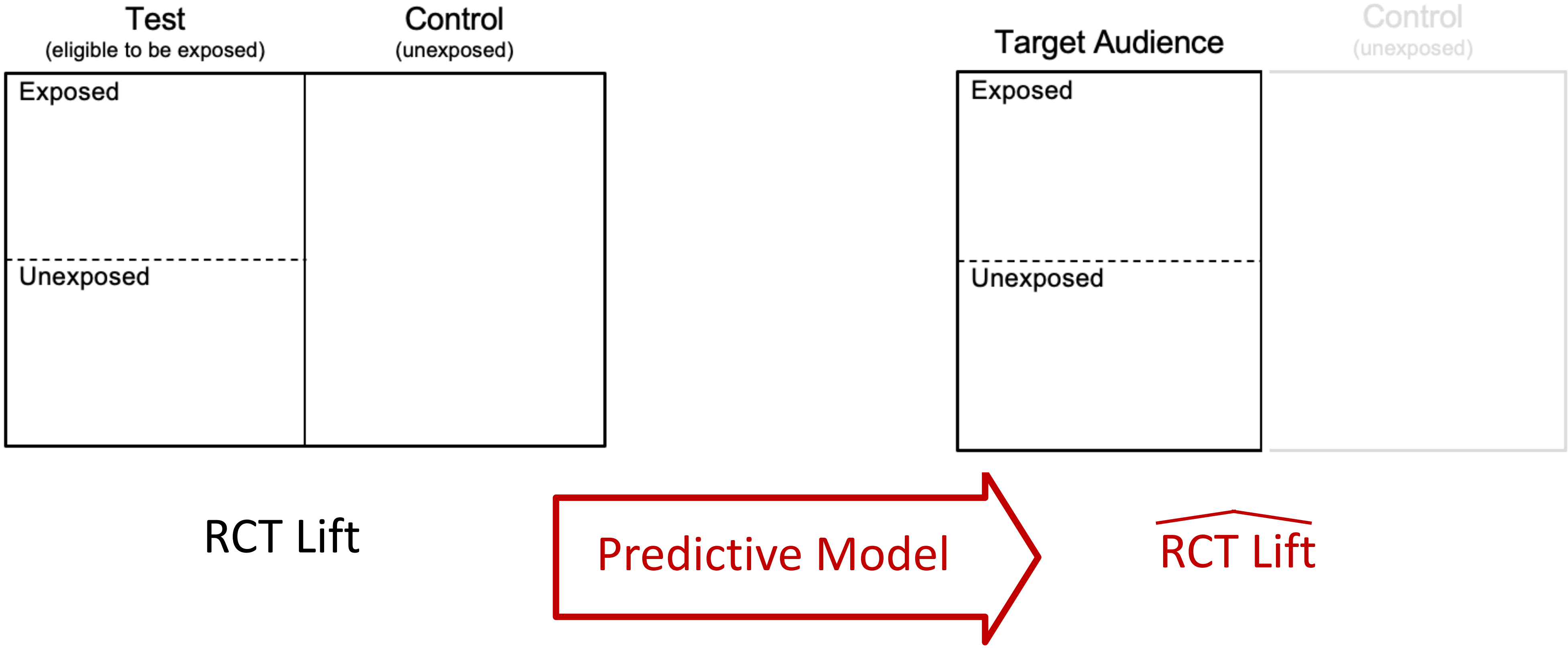


RCT Lift?

Predictive Incrementality by Experimentation (PIE)

Ad campaigns as RCTs

Ad campaigns not as RCTs



Using a database of RCTs, how well could we predict a new campaign's RCT Lift if it was <u>not</u> run as an RCT?

- Unit of observation is an RCT campaign, not a user
- Shift to predictive models, instead of those from causal inference
- Use our RCT dataset to assess the performance of PIE models

PIE – a first cut

$$\# \text{ Incremental Conversion in } \mathrm{RCT}_r = \theta(\# \text{ of } \mathrm{LC}^w \text{Conversions in } \mathrm{RCT}_r) + \mathrm{Error}_r$$

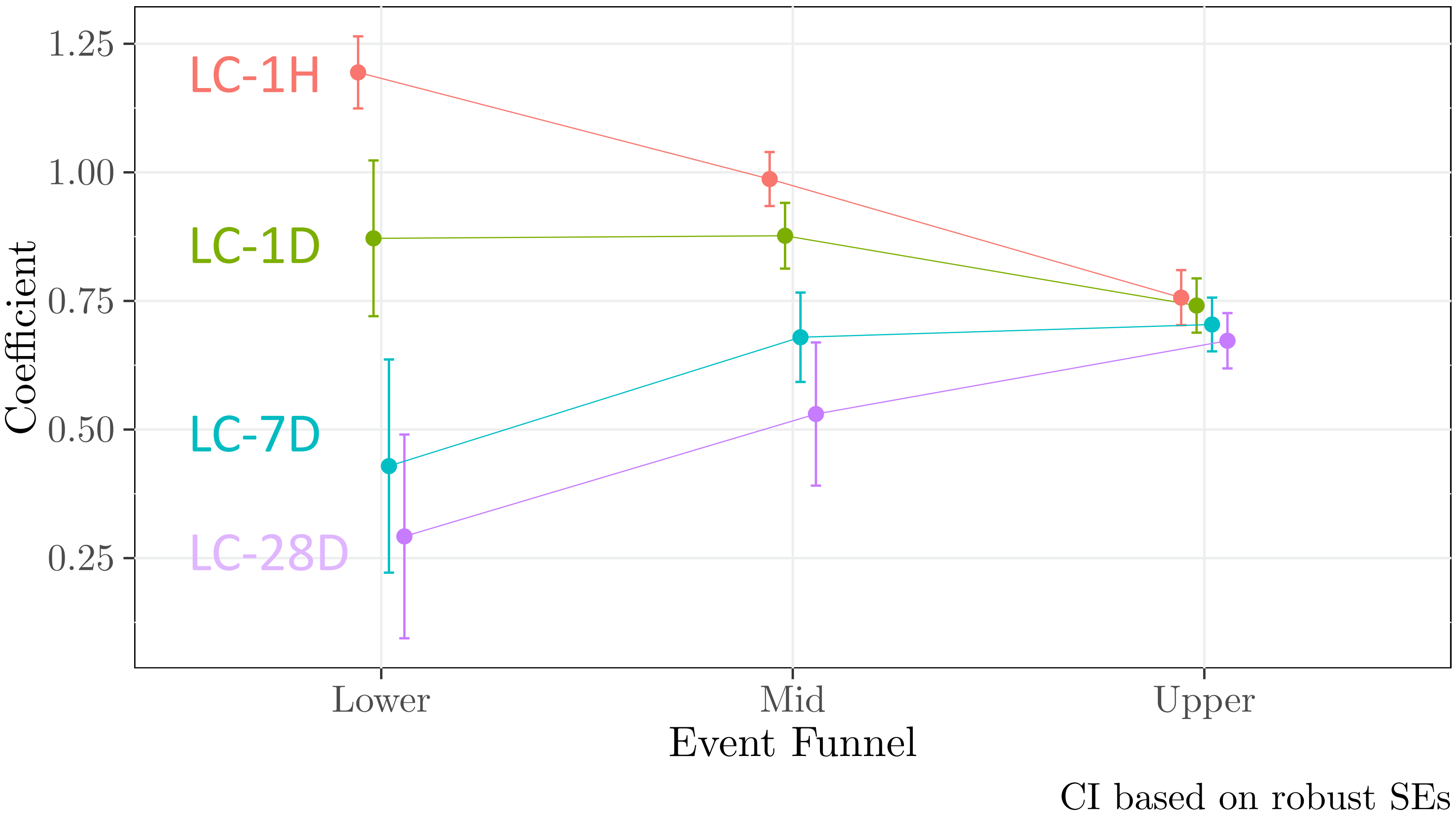
"Calibration Factor"

We estimate this model separately by:

- Funnel levels (Lower, Mid, Upper)
- Last click attribution windows w ∊ {1 hour, 1 day, 7 days, 28 days}

Calibration factor model by funnel and attribution window

Calibration Factor



To generalize this approach, we normalize incremental conversions by ad spend and add more features

RCT Incremental Conversions per Dollar (ICPD)

$$\mathrm{ICPD}_r = f\left(X_r^{\mathrm{pre}}, X_r^{\mathrm{post}}; \theta\right) + \varepsilon_r \ ,$$

Advertiser-campaign characteristics known before campaign was run

Campaign: targeting criteria, bidding params, optimization goal, budget, etc.
Advertiser: vertical, experimentation experience, etc.

Proxy metrics known after campaign was run

Last click conversion counts by {1H, 1D, 7D, 28D}

If available, other post-campaign metrics could be used (e.g., view-through conversion counts)

Key: None of the features rely on the RCT control group

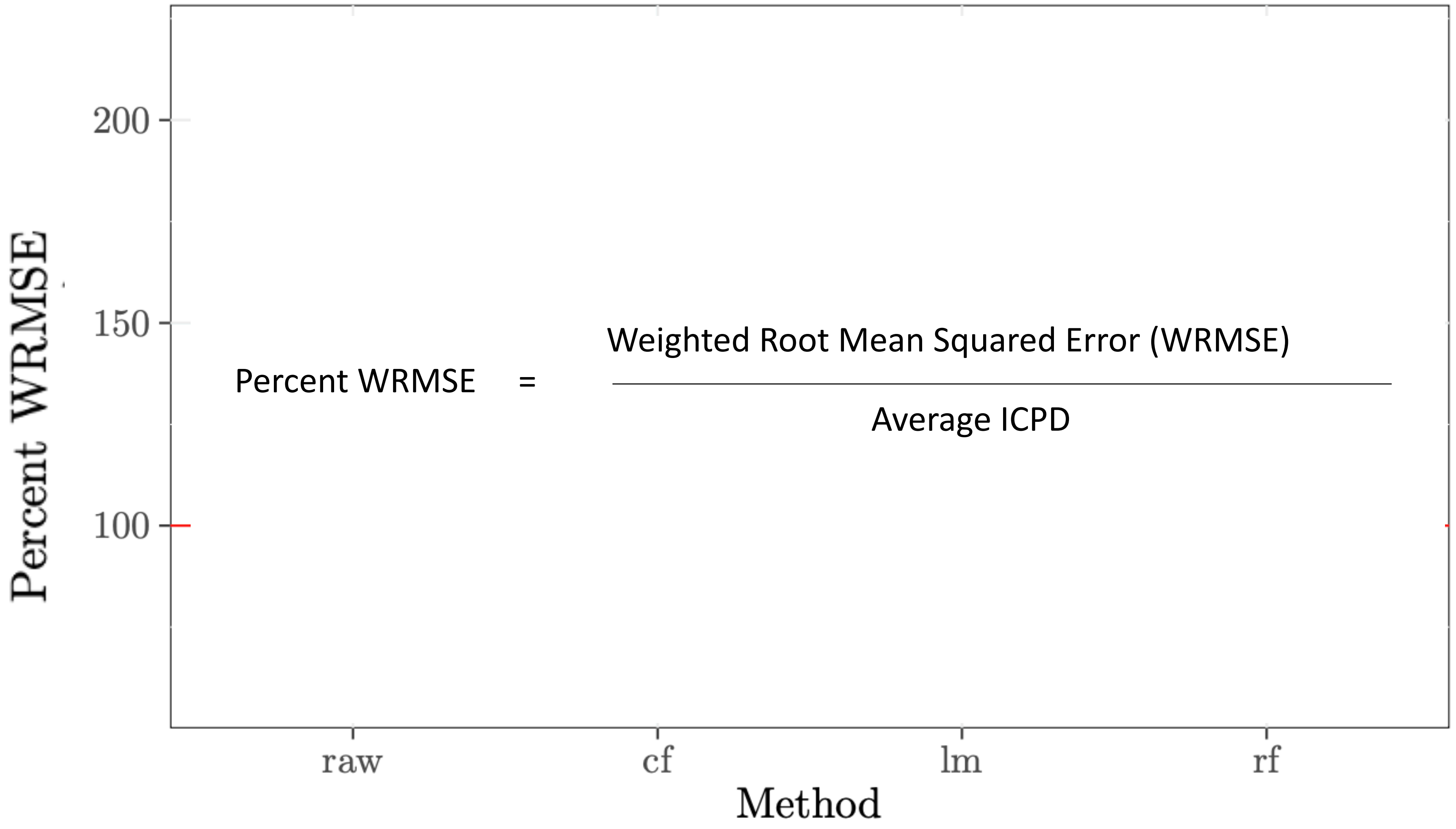
We try a variety of models and two feature sets

$$f\left(X_r^{\text{pre}}, X_r^{\text{post}}; \theta\right)$$

- Models for
  - **raw:** how well does each last click metric perform by itself?
  - **cf:** calibration factor model (the "first cut")
  - **lm:** linear regression
  - **rf:** random forest

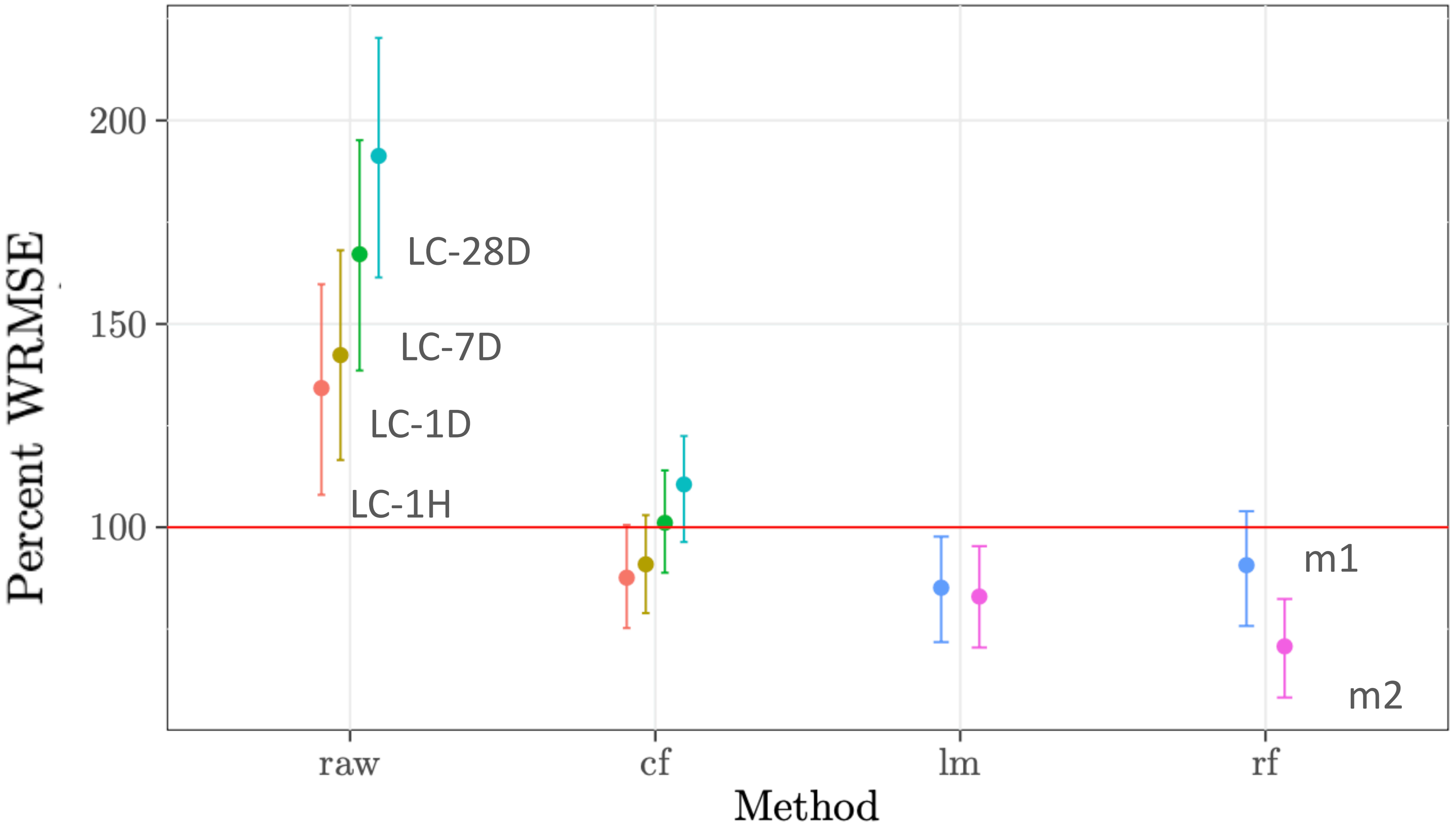
- Features used in **lm** and **rf**
  - **m1:** $X_r^{\text{post}} = \{\text{LC-1h}, \text{LC-1d}, \text{LC-7d}, \text{LC-28d}\}$ $\quad X_r^{\text{pre}} = \{\ \}$
  - **m2:** $X_r^{\text{post}} = \{\text{LC-1h}, \text{LC-1d}, \text{LC-7d}, \text{LC-28d}\}$ $\quad X_r^{\text{pre}} = \{\text{everything}\}$

Assess the models using Percent WRMSE based on Leave-One-Out Predictions

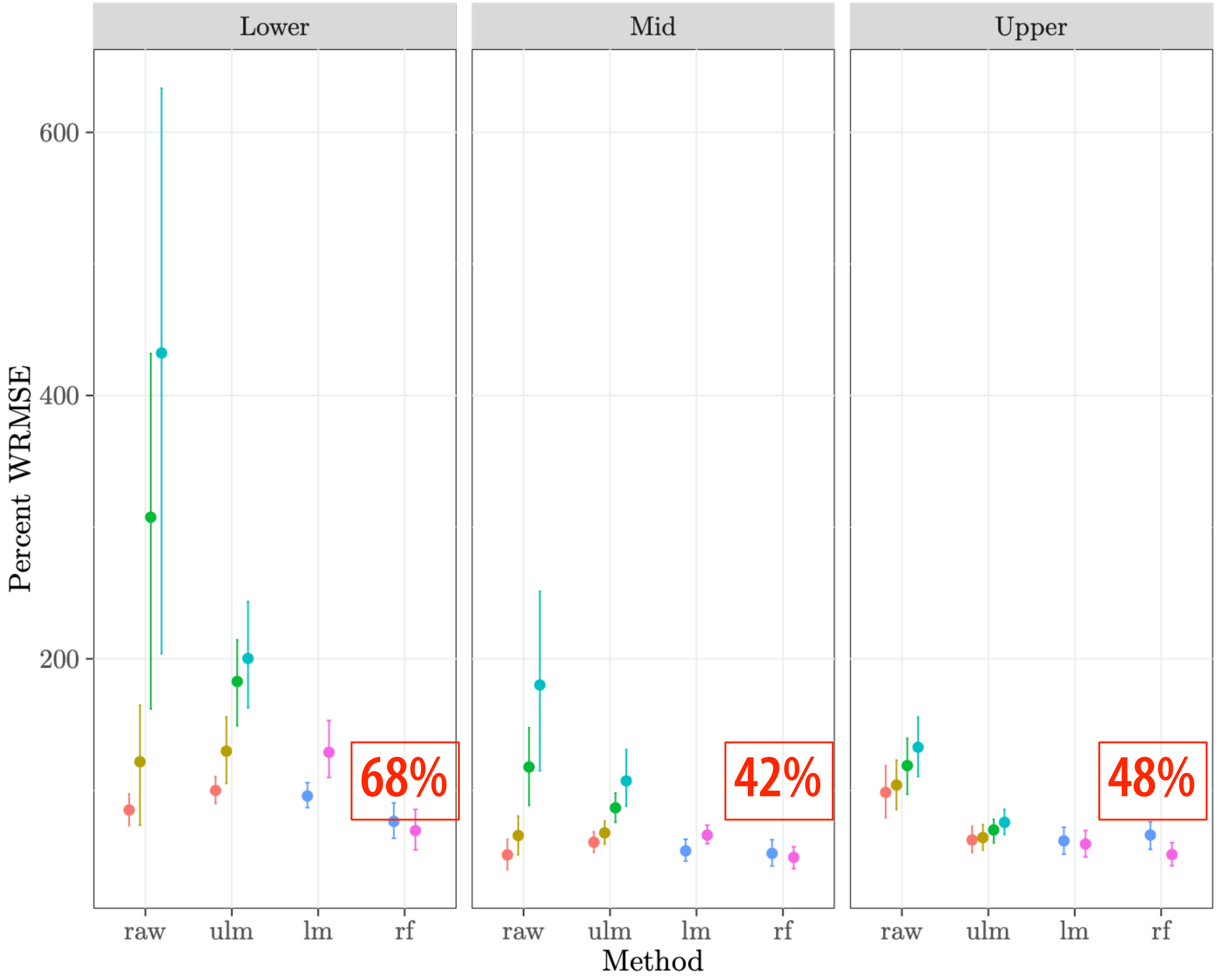


$$\text{Percent WRMSE} = \frac{\text{Weighted Root Mean Squared Error (WRMSE)}}{\text{Average ICPD}}$$

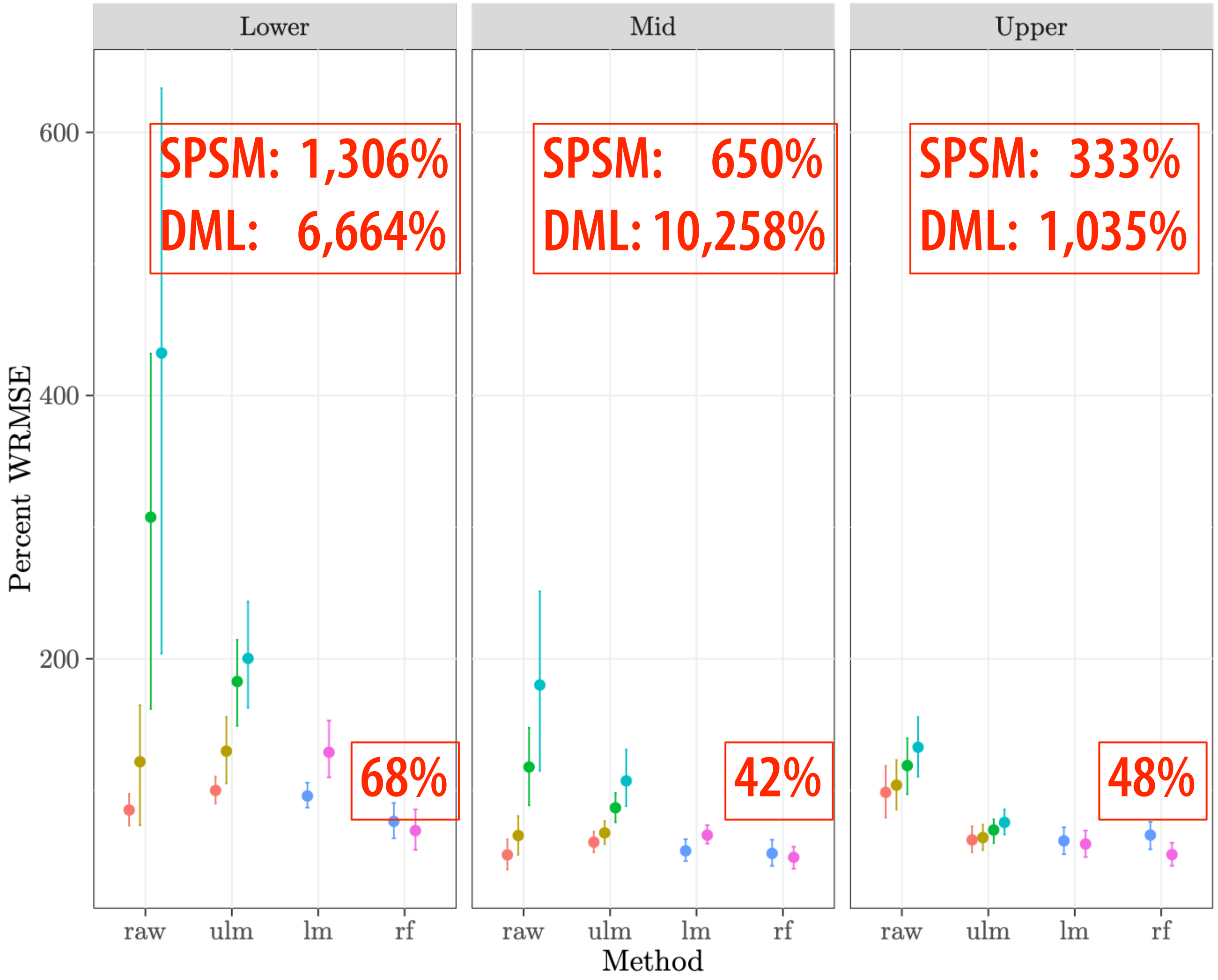
A random forest (rf) with all features performs best



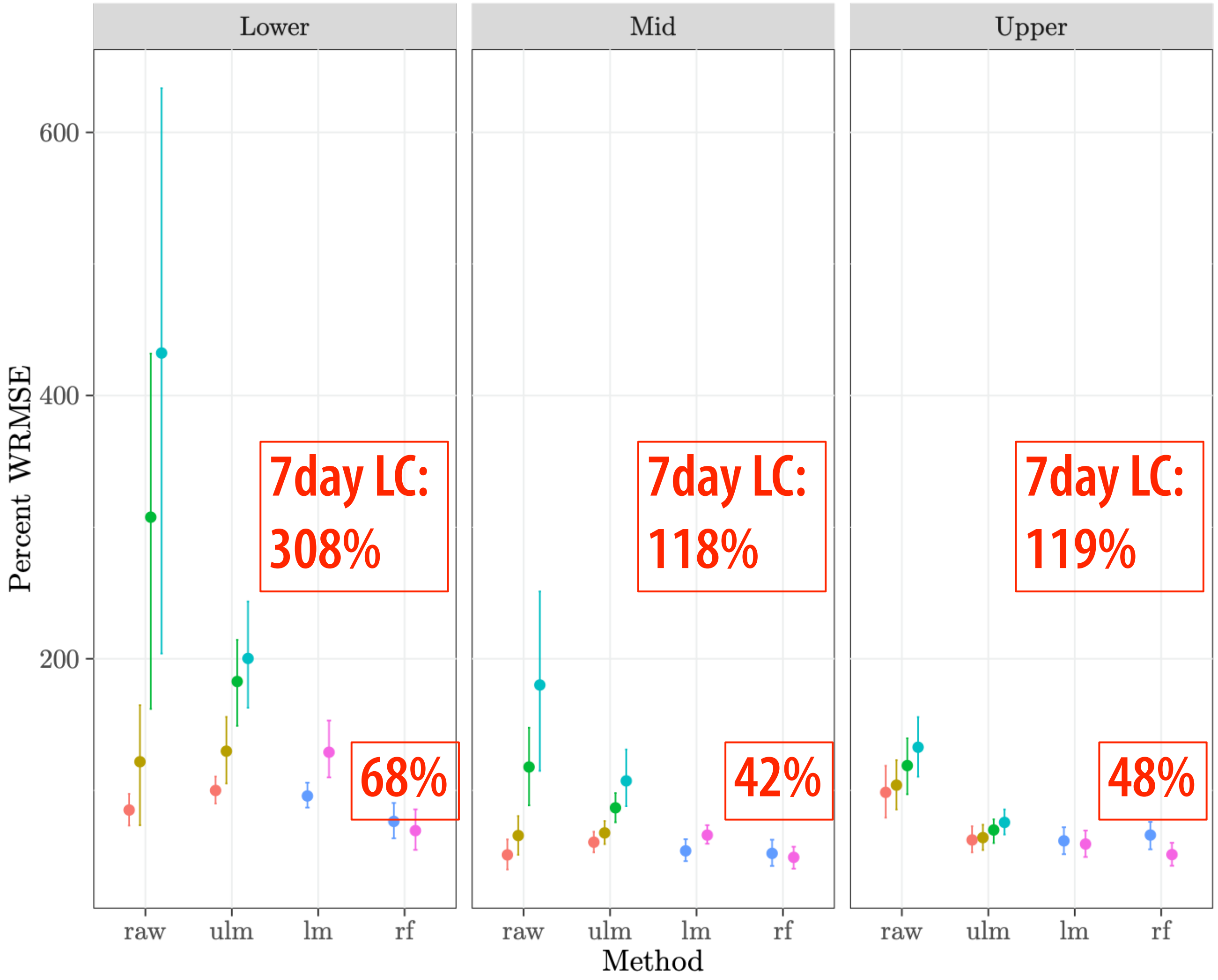
The percent WRMSE for the best specification is around 50%



PIE does much better than SPSM or DML



PIE also does substantially better than the industry-standard 7-day Last-Click Attributed Conversions



Recap: What exactly is PIE?

PROBLEM - Advertisers can't rely on observational data, nor can they always run RCTs …
…but they still need to measure advertising effects

Recap: What exactly is PIE?

PROBLEM
- Advertisers can't rely on observational data, nor can they always run RCTs …
  …but they still need to measure advertising effects

KEY IDEA
- Advertisers can still run RCTs for a subset of campaigns …
- Use RCTs to predict ad effects for new campaign's that were not run as an RCT

Recap: What exactly is PIE?

PROBLEM - Advertisers <u>can't rely</u> on observational data, nor can they <u>always</u> run RCTs …
…but they still need to measure advertising effects

KEY IDEA - Advertisers can still run RCTs for a <u>subset</u> of campaigns …
- Use RCTs to predict ad effects for new campaign's that were <u>not run as an RCT</u>

STANDARD PREDICTION
- Estimate ad effects using campaign and user characteristics <u>before campaign ran</u>
- Uses "<u>pre-determined</u>" features: e.g. FinTech ads have higher lift than CPG ads

Recap: What exactly is PIE?

PROBLEM
- Advertisers <u>can't rely</u> on observational data, nor can they <u>always</u> run RCTs …
  …but they still need to measure advertising effects

KEY IDEA
- Advertisers can still run RCTs for a <u>subset</u> of campaigns …
- Use RCTs to predict ad effects for new campaign's that were <u>not run as an RCT</u>

STANDARD PREDICTION
- Estimate ad effects using campaign and user characteristics <u>before campaign ran</u>
- Uses "<u>pre-determined</u>" features: e.g. FinTech ads have higher lift than CPG ads

PIE
- Predict ad effects using performance features <u>after the campaign starts</u>
- Uses "<u>post-determined</u>" features: e.g. clicks, last-click conversion, page views, …
- Anything in the <u>treatment group</u> that is correlated with causal ad effects
- Move from causal inference to a <u>prediction problem</u>

PIE will work when post-determined features are <u>predictive</u> and the relationship is <u>stable</u>

WHEN PIE IS LIKELY TO WORK

- **PREDICTIVE**: RCTs need to measure causal effect (not too noisy)
  Post-determine features need to contain some causal signal (empirical question)

- **STABLE:** We need the nature of this relationship to be stable over time (i.e., no concept shift)

So, is PIE useful for practice?

HOW SHOULD WE MEASURE CAUSAL ADVERTISING EFFECTS?

- Attribution models are biased
- Causal inference models don't work
- RCTs are only viable option … but are infeasible at scale
- PIE makes RCTs scalable
- In our testing PIE has smaller confidence intervals than raw RCTs

Thank you!