Case studies from a methodologist on an experimentation platform team

Laura Cosgrove
Senior Data Scientist

Microsoft ExP

Who is Microsoft ExP?

- We operate the ExP A/B testing platform, founded in 2006.
- Our mission: "*Empowering product teams to innovate and make data-driven decisions through trustworthy experimentation at scale.*"
- **Experimentation process:** Randomly divide a population into groups, assign variants to the random groups, and measure differences with causal attribution and quantifiable statistical uncertainty

Our A/B Testing Partners

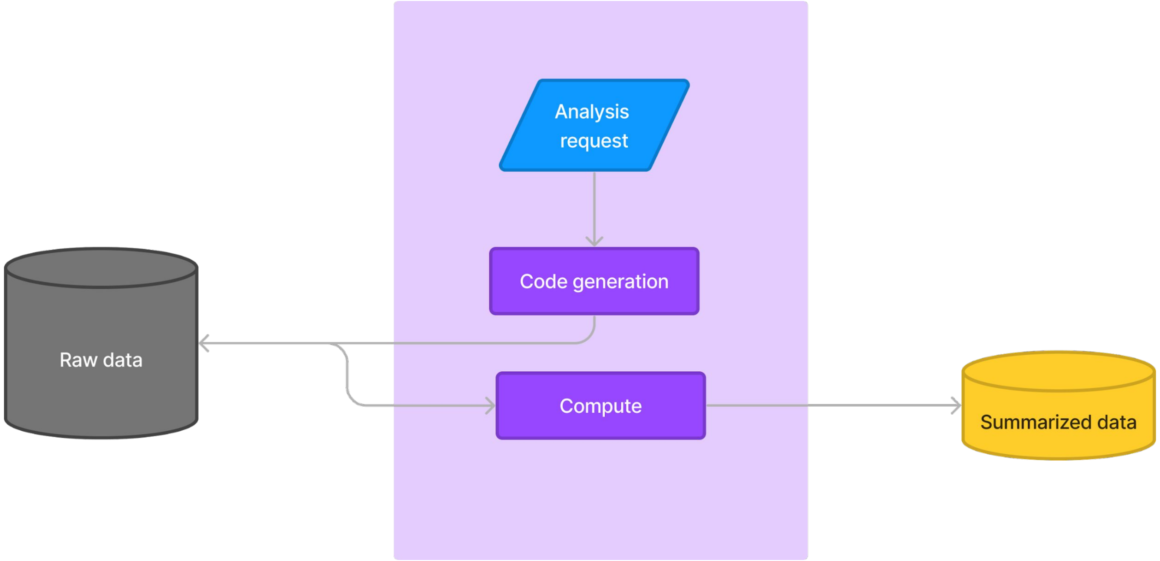
Role of a methodologist on an experimentation platform team

- Our methodology work aims to:
  - Identify opportunities to improve trustworthy evidence-based decision making
  - Update methodology in response to new scenarios
  - Develop methodology-informed user experiences

- Our methodology work does not aim to:
  - Optimize analysis for one or few experiments
  - Own policies or decisions

Platform

- Microsoft ExP supports multiple compute backends, along with the experiment needs of 40+ organizations
- >100k experiments/year



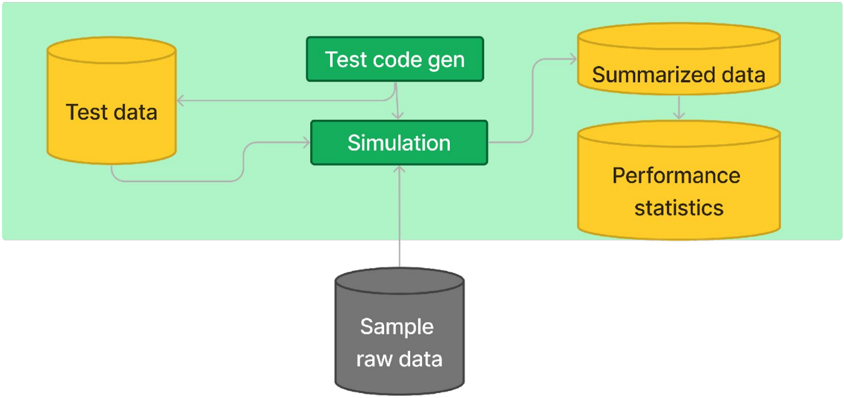
[Metric computation for multiple backends - Microsoft Research]

Navigating the data landscape

Investigation and verification through test code generation is essential to methodology improvements

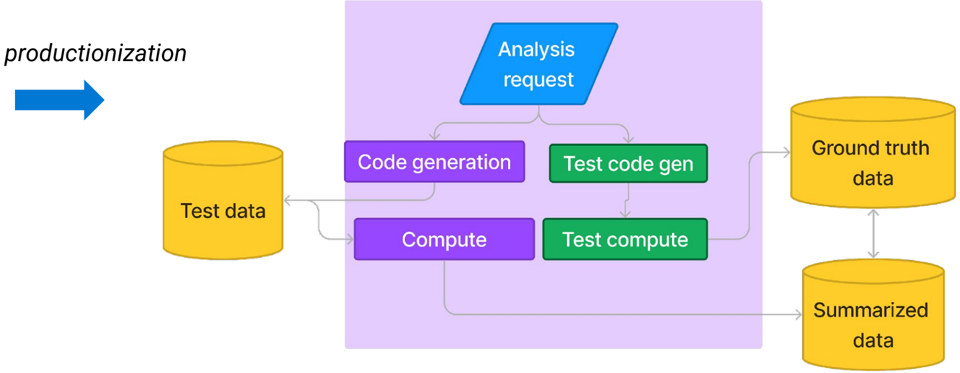
Simulation-based analysis

+ case study data

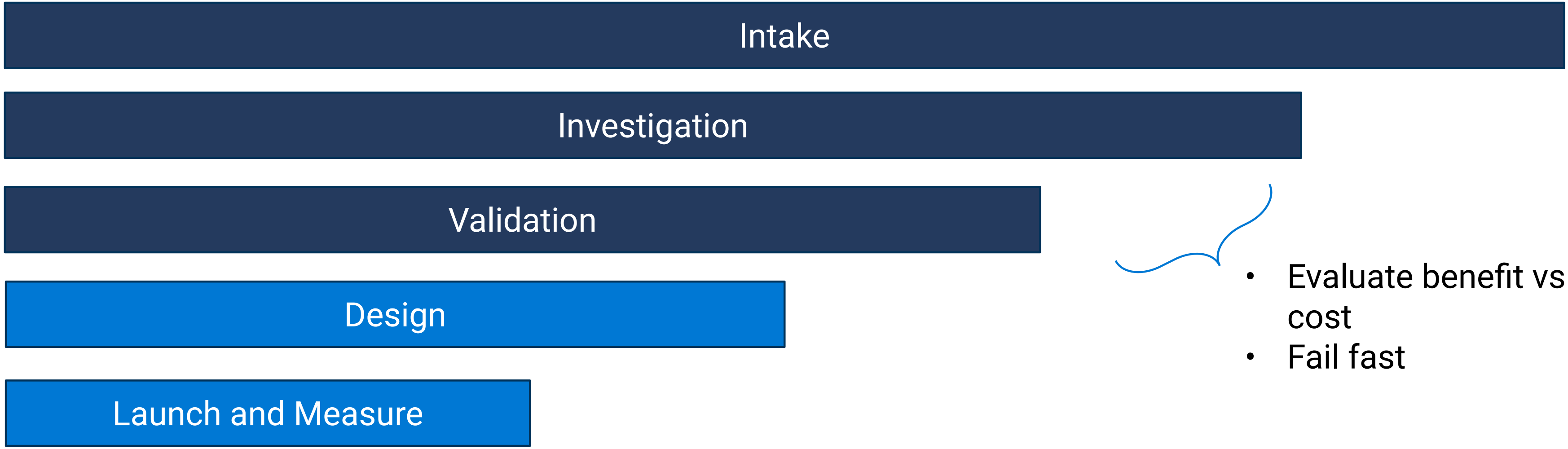


productionization

Integration testing framework



Methodology research process



Intake

Investigation

Validation

Design

Launch and Measure

- Evaluate benefit vs. cost
- Fail fast

Case studies from the investigation-validation feedback loop

- **Grounding** our investigation with simulated data
  - VR heterogeneity
  - Safe deployment

- **Sizing our impact** with summarized data
  - Safe deployment partner application

- Shuffling sample raw data to **simulate realistic performance**
  - ML assisted VR

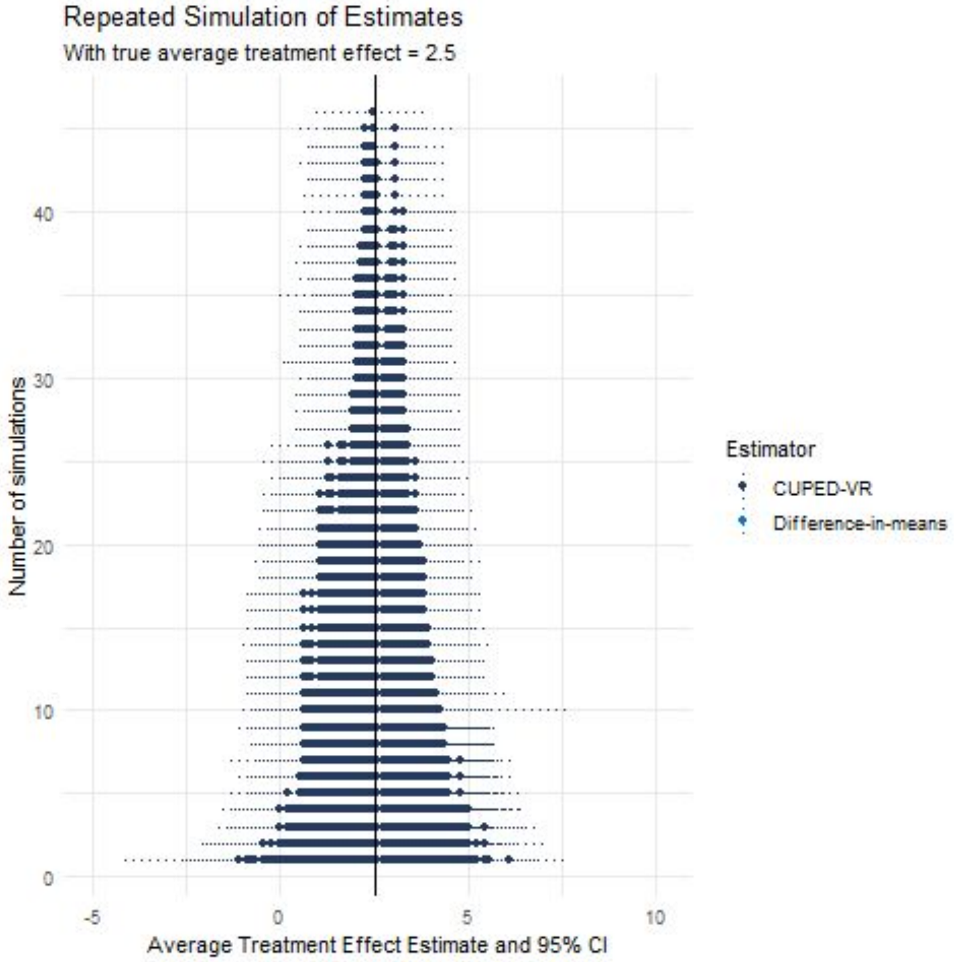
Case study 1

Variance Reduction heterogeneity

What is
Variance Reduction?

- A user's past behavior is a predictor of their future behavior
  - The users who were most active last week are likely to also be relatively active this week
  - The users who were least active last week are likely to also be relatively inactive this week
- How can we leverage that information?
  - We can adjust for the explained variance. And reduce the sample standard deviation of the metric while keeping the metric an unbiased estimator

- **Effective traffic multiplier:** For a given test, the ratio of unadjusted estimated variance to raw variance estimates the amount of traffic that would need to be added to the simple difference estimator to provide the same level of variance reduction as VR.

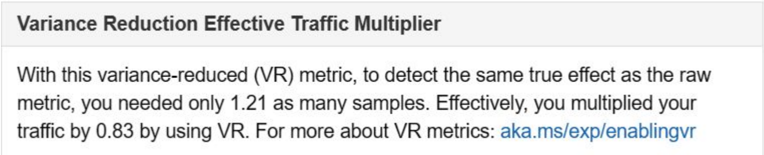


https://aka.ms/exp/vr-deep-dive

Can we improve VR performance under heterogeneous treatment effects?

Problem

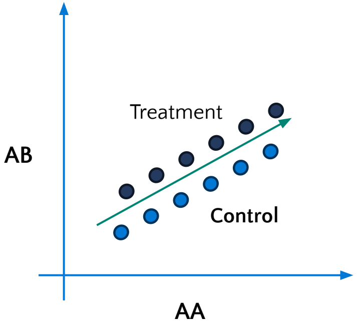
In some partners, we find ~2% of observed effective traffic multipliers < 1 within segments (Conditional treatment effects).



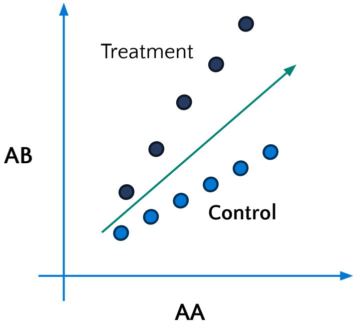
Variance Reduction Effective Traffic Multiplier

With this variance-reduced (VR) metric, to detect the same true effect as the raw metric, you needed only 1.21 as many samples. Effectively, you multiplied your traffic by 0.83 by using VR. For more about VR metrics: aka.ms/exp/enablingvr

Heterogeneous treatment effects and VR



Estimate of AA <> AB relationship is not modified by treatment. Residuals are small in treatment and control.



Estimate of AA <> AB relationship is modified by treatment – or AA value modifies treatment effect.

Level-setting: Identify good candidate estimators

Because the goal is to **ground** theoretical performance relative to difference in means, benchmark with fully simulated data: high heterogeneity and low sample size

Estimator	Characteristics	True efficiency gain and average bias	Estimated Frequency of ETM < 1
CUPED		-1.5% efficiency gain (Average simulation-based variance estimate rel. to simulation-based DiM variance estimate) 0% average bias (Mean of point estimates rel. to mean of DiM point estimates)	28%
ANCOVA1	Similar point estimate to CUPED		74%
ANCOVA2	Best under unequal ratios and heterogeneity (1)		74%
"Better" CUPED	Similar point estimate to ANCOVA2 (2)		0%

1. Negi and Wooldridge, 2021

2. Lin, 2013

Lesson learned: ETM > 0 reflects real performance

Each of the regression-adjusted estimators have nearly identical simulation-based variance – this is closest to the true variance of the estimator, and is what we should report.

Estimator	Characteristics	True efficiency gain and average bias	Estimated Frequency of ETM < 1	Relative under-coverage based on estimated variance
CUPED		-1.5% efficiency gain (Average simulation-based variance estimate rel. to simulation-based DiM variance estimate) 0 average bias (Mean of point estimates rel. to mean of DiM point estimates)	5%	1.2%
ANCOVA1	Similar point estimate to CUPED		10%	-0.2%
ANCOVA2	Best under unequal ratios and heterogeneity (1)		5%	-0.3%
“Better” CUPED	Similar point estimate to ANCOVA2 (2)		0%	2.3%

Implications

1. An “effective traffic multiplier” in an individual study is just an estimate of the variance reduction of the estimator.
    - ETM < 1 is **more likely observed in the null state** where there is truly no gain from Regression Adjusted estimators
        - Small sample sizes
        - Balanced design *and* strongly heterogeneous treatment effects conditional on AA metric value

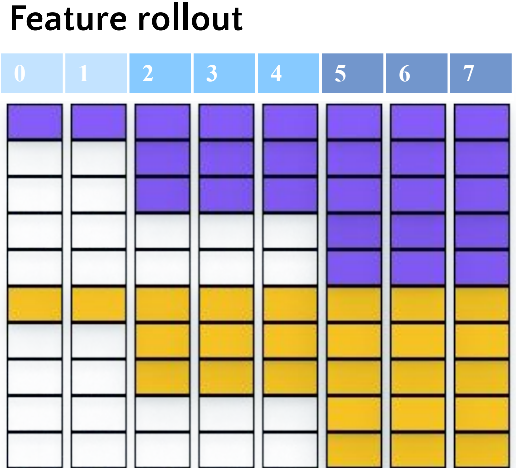
2. CUPED has asymptotic but not numerical equivalency with RA estimators, and we can get there by changing CUPED to “Better CUPED”
    - Transitioning to an equivalent delta estimator to ANCOVA2 theoretically helps under 0.5 << p << 0.5 and strong treatment heterogeneity, but differences in performance between the approaches are small in simulations
    - If we align to ANCOVA2, move to a corresponding variance estimator *(consider variance in theta estimates)*

It will not resolve the “scary effective traffic multiplier” observed estimate, and estimators are very close.

Case study 2

Safe deployment

Can we get 7-day insights faster?



- As a feature owner, I follow safe deployment practices even when A/B testing

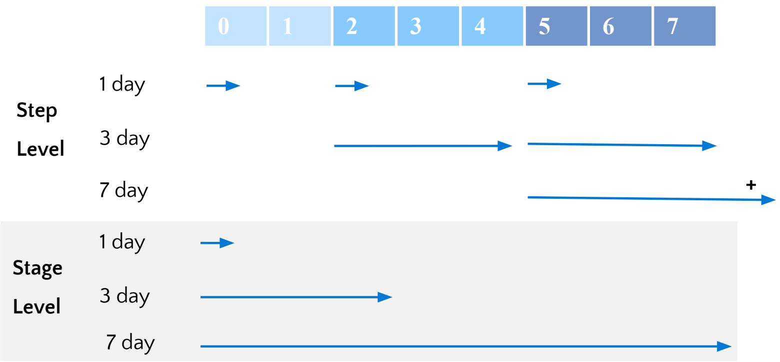
- I start my new feature A with the rollout steps:
  - 10% -> 30% -> 50%

- There is consistent treatment assignment between steps

- While ramping up, I monitor A/B metrics, and I want to monitor >= 7-day metrics before shipping to 100%

Can we get 7-day insights faster?

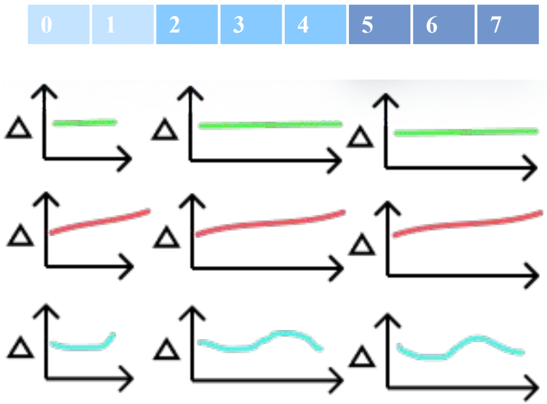
Scorecard schedules



- Default for ExP is step-level analysis
- For this feature owner, minimum date to get 7-day insights is day 12
- A stage-level analysis might give us a 7-day insight on day 7

Can we get 7-day insights faster?

Potential treatment effects



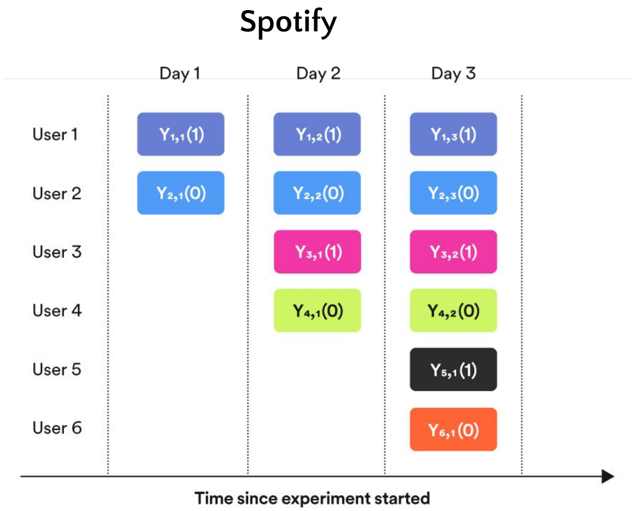
- Versus a 1-day analysis, a 7-day analysis might detect a constant treatment effect with more power

- Or it might detect a time-dependent treatment effect, such as increasing with exposure length

- Or a heterogenous effect, such as a light-or-heavy user effect

Level-setting: Open-ended metrics lead to a changing estimand

- Cohort-based metrics use time since exposure (TSE) to measure outcomes (1) (2)
    - Example: average minutes played during the first seven days of exposure
    - Only users with enough exposure are included in the analysis
- Cohort-based metrics have a fixed estimand, unlike open-ended metrics
- Open-ended metrics have a changing estimand that depends on the “intake distribution”
    - Implication: This means the information accrued at each sequential analysis is not fully new, disrupting the alpha-spending approach
    - Solution: use a longitudinal model to *estimate the correlation* between successive deltas and adjust the group sequential test accordingly



Spotify

	Day 1	Day 2	Day 3
User 1	$Y_{1,1}(1)$	$Y_{1,2}(1)$	$Y_{1,3}(1)$
User 2	$Y_{2,1}(0)$	$Y_{2,2}(0)$	$Y_{2,3}(0)$
User 3		$Y_{3,1}(1)$	$Y_{3,2}(1)$
User 4		$Y_{4,1}(0)$	$Y_{4,2}(0)$
User 5			$Y_{5,1}(1)$
User 6			$Y_{6,1}(0)$

Time since experiment started

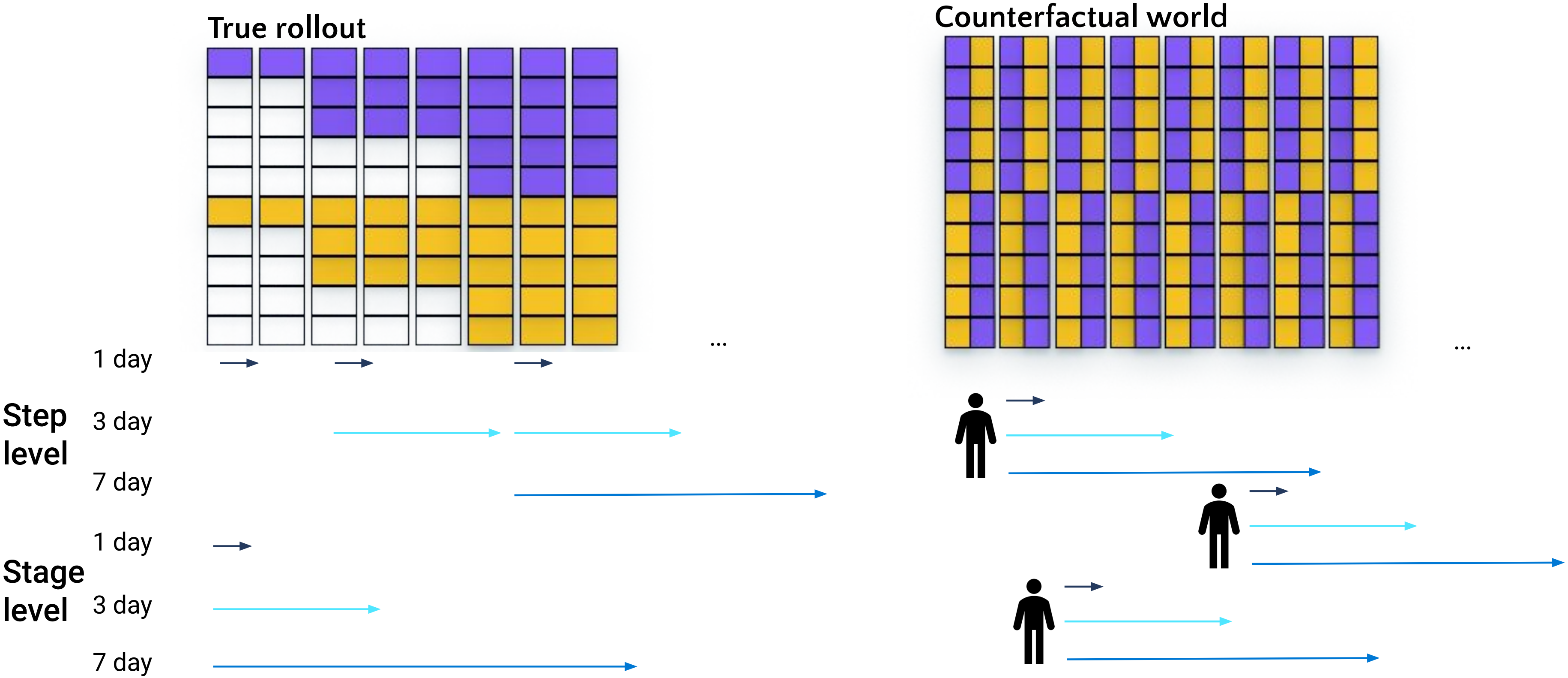
$$E[\hat{\delta}_{1,\mathrm{ROB-OLS}}] = E\left[\frac{Y_{1,1}(1)}{1} - \frac{Y_{2,1}(0)}{1}\right] = \delta_1$$

$$E[\hat{\delta}_{2,\mathrm{ROB-OLS}}] = E\left[\frac{Y_{1,1}(1) + Y_{1,2}(1) + Y_{3,1}(1)}{3} - \frac{Y_{2,1}(0) + Y_{2,2}(0) + Y_{4,1}(0)}{3}\right] = \frac{2}{3}\delta_1 + \frac{1}{3}\delta_2$$

1. Bringing Sequential Testing to Experiments with Longitudinal Data (Part 1): The Peeking Problem 2.0 - Spotify Engineering : Spotify Engineering (atspotify.com)
2. Novelty and Primacy: A Long-Term Estimator for Online Experiments (Gupta et al., Microsoft)

Choosing an estimand

Because the goal is to ground theoretical performance, benchmark with fully simulated data, but *mimic realistic rollout and analysis policy*



True rollout

Counterfactual world

...

...

Step level

1 day

3 day

7 day

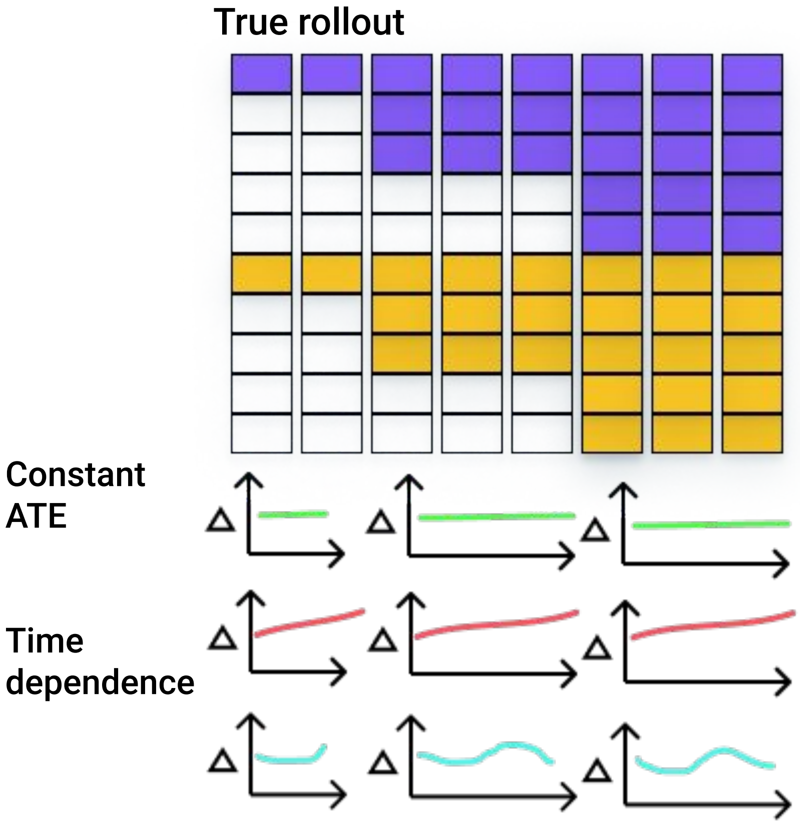
Stage level

1 day

3 day

7 day

Lesson learned: Method performance is meaningless without the relevant estimand



ATE

- Y1 – Y0 | experiment period, if all users were immediately eligible for treatment

CATE1

- Y1- Y0 | 1 days exposed

CATE3

- Y1- Y0 | 3 days exposed

CATE7

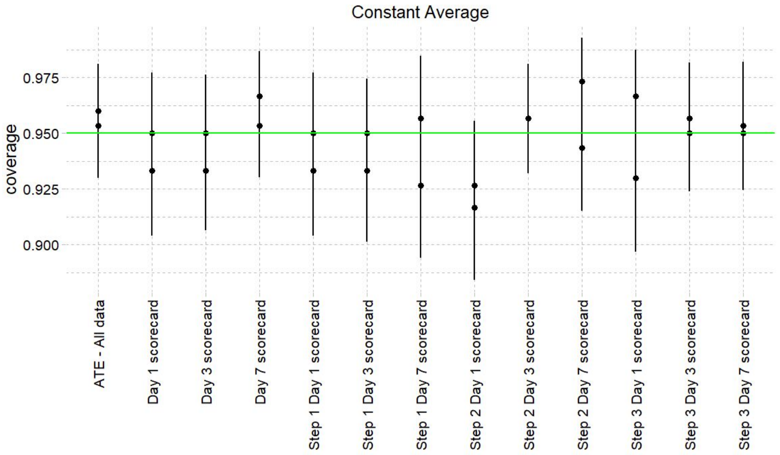
- Y1- Y0 | 7 days exposed

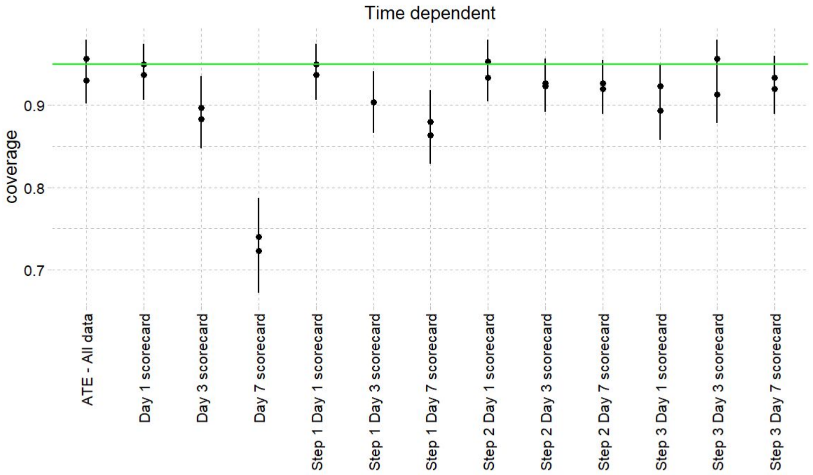
Performance of estimators

With day effects, trend in metric values over days, historical usage affecting metrics and day in analysis:

Constant average treatment effects are well-estimated regardless of whether it’s scoped to one step, or across steps.

Time-dependent treatment effects are poorly estimated due to dilution of existing users with new users





Implications

- Cohort-based metrics are safer metrics for:
  - Cross-step analysis in a rollout scenario
  - Sequential testing

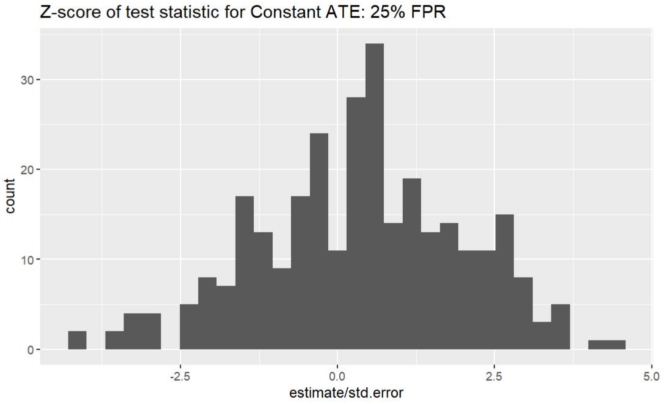
- Other trustworthy option for user-based sequential testing: model the time structure

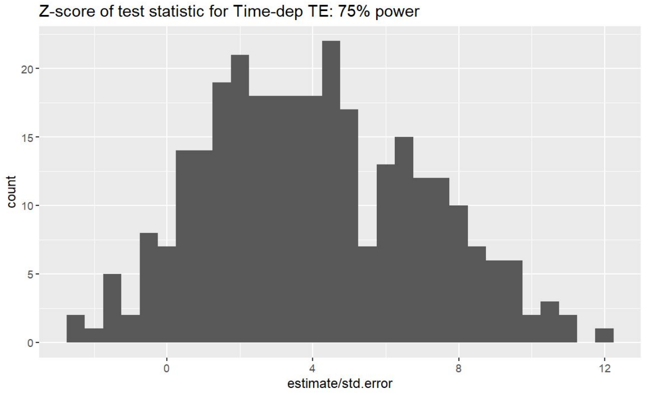
- Cross-step analysis is valid without time-dependent effects
  - **Next: Can we estimate the potential application with real, aggregated data?**

Under constraints, how do we estimate if we are in a world of constant average treatment effects?

- Test: Difference in CATEs = 0
  - Point estimate is CATE_7 – CATE_1
  - For the test statistic, we ignore assumed-positive correlation between:
    - CATE_1 = Avg(Y1-Y0 | 1 days exposed) and
    - CATE_7 = Avg(Y1- Y0 | 7 days exposed)

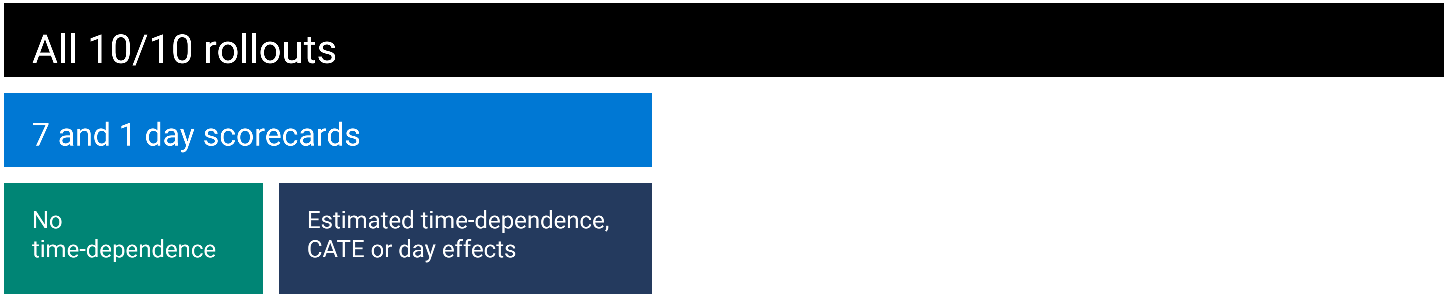
- We detect more time dependence than we should in constant average treatment effects, and have 75% power for time-dependent treatment effects
  - Proceed with a SWAG estimate


Z-score of test statistic for Constant ATE: 25% FPR


Z-score of test statistic for Time-dep TE: 75% power

Partner application

18% of rollouts can be recommended cross-step analysis



All 10/10 rollouts

7 and 1 day scorecards

No time-dependence

Estimated time-dependence, CATE or day effects

- Median rollout schedule: 10 days in a 10/10 step and 14 days in a 50/50 step = **28 days**
  - Impact: 28 -> 21 days for 18% of rollouts

Implications

- Cohort-based metrics are safer metrics for:
  - Cross-step analysis in a rollout scenario
  - Sequential testing

- Other trustworthy option for user-based sequential testing: model the time structure

- Cross-step analysis is valid without time-dependent effects, but naïve approach of cross-step analysis is not sufficient for the majority of feature rollouts

- **Overall:** Motivated deeper investment for improving time-to-decision
  - Design input in SPRT implementation

Case study 3

ML assisted multivariate VR

What is the expected gain from extending VR implementation to a multivariate option?

Because the goal is to **simulate realistic performance**, use real sample A vs. A test as input to simulations

Model / Approach		Efficacy	Cost
Multivariate VR	CUPAC (Machine Learning, nonlinear, LGBM)	⚡⚡⚡	$$$
	CUPAC (Machine Learning, linear, Ridge)		
	MLR* with all covariates		
	MLR with fixed list of covariates (optimally segmented outcomes by date)		
Current Production: CUPED (univariate VR)			
No VR (simple t-test)		⚡	$

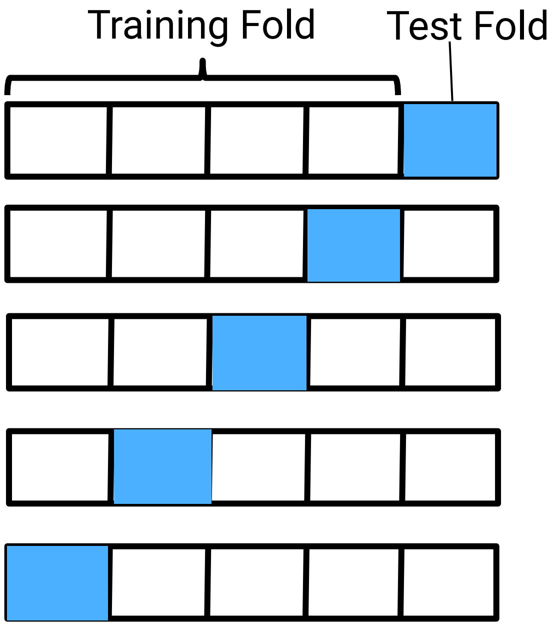
Higher efficacy is better

Lower cost is better

*MLR: Multi-linear regression

Model Evaluation: 5-Fold Cross-Validation for CUPAC

(Single) cross-fitting: avoid estimation bias induced by regularization and overfitting*



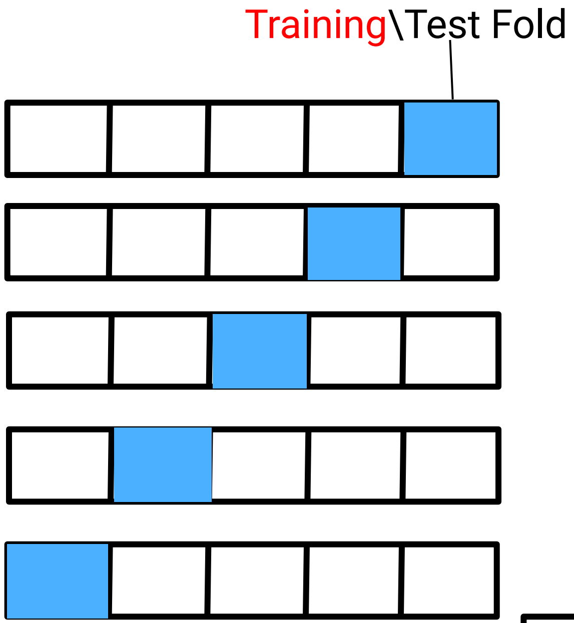
	Point Estimate	Variance Estimate
Iteration 1	P1	V1
Iteration 2	P2	V2
Iteration 3	P3	V3
Iteration 4	P4	V4
Iteration 5	P5	V5
	Median(P1,...P5) Or Mean(P1, ...)	Between-fold variance [variance(V1, ..., V5)] + within-fold variance [average(V1, ..., V5)]

Model Training Frequency
Note: If we recommend a ML (CUPAC)/variable selection method, we won't ask the team to run this model for each metric in each experiment.
Goal to train the model for each metric per semester to reduce computation cost

* Ref: Double Machine Learning for causal inference | by Borja Velasco | Towards Data Science

Model Evaluation: 5-Fold Validation for Non-ML Models

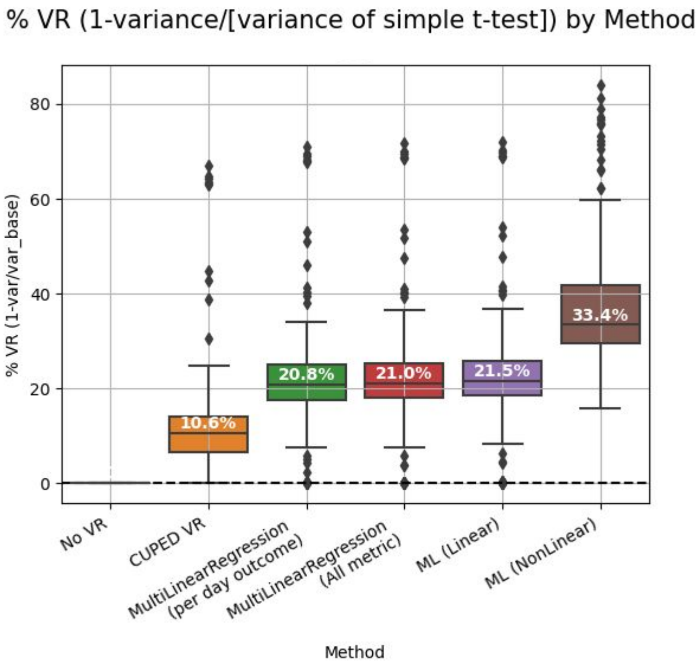
5-fold evaluation: obtain comparable model performance criteria

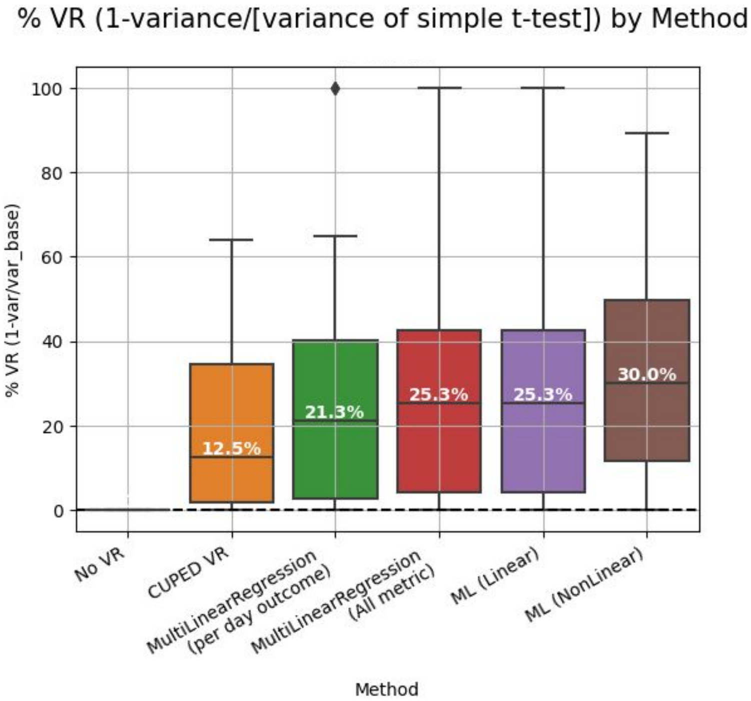


	Point Estimate	Variance Estimate
Iteration 1	P1	V1
Iteration 2	P2	V2
Iteration 3	P3	V3
Iteration 4	P4	V4
Iteration 5	P5	V5
	Median(P1,...P5) Or Mean(P1, ...)	Between-fold variance [variance(V1, …, V5)] + within-fold variance [average(V1, …, V5)]

All Methods Comparison – VR*

Pattern: CUPED VR < Cost-effective MLR [per day outcome] < Machine-learning Based





Implications

Model / Approach		Efficacy	Cost
Multivariate VR	CUPAC (Machine Learning, nonlinear, LGBM)	⚡⚡⚡	$$$
	CUPAC (Machine Learning, linear, Ridge)		
	MLR with all covariates		
🏆	MLR with fixed list of covariates (optimally segmented outcomes by date)		
Current Production: CUPED (univariate VR)			
No VR (simple t-test)			
		⚡	$

Higher efficacy is better

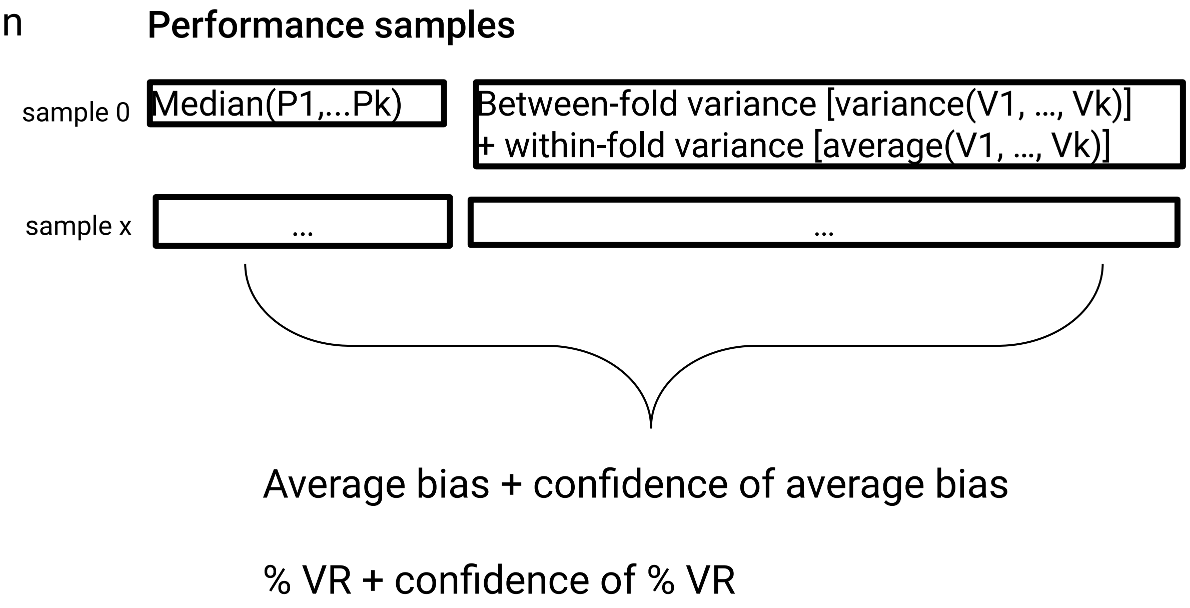
Lower cost is better

Lesson learned: simulate levels of variation to characterize the estimator

- *What we want:* An estimate of how well these methods will perform in the population using case study data (A vs. A test)

- *What we can leverage:*
  - Subsampling (changing the index of folds)
  - Shuffling of assignment column
  - Bootstrap resample

- Subsampling is not sufficient for average bias interval including 0 using case study data



Performance samples

sample 0	Median(P1,...Pk)	Between-fold variance [variance(V1, ..., Vk)] + within-fold variance [average(V1, ..., Vk)]
sample x	...	...

Average bias + confidence of average bias

% VR + confidence of % VR

Questions, implications, and lessons

1. Can we improve VR performance with heterogeneous treatment effects?
   - No clear performance gain: known method limitation
   - An estimate of performance like effective traffic multiplier in an individual study will not always meet theoretical performance guarantees across studies

2. Can we get 7-day insights faster in a rollout scenario?
   - Shift to cohort-based metrics, or model the time structure
   - Clarity about the reference estimand from the start will help analysis extensions in the future

3. How do we estimate if we are in a world of constant average treatment effects?
   - Can estimate for ~1/2 of experiments for one partner, and can recommend cross-step analysis in 2/5 of these cases
   - Investment in deeper solution is justified to improve time-to-decision

4. What is the expected gain from extending VR implementation to a multivariate option?
   - For our partners, we found the best tradeoff from date-segmented multivariate VR
   - Comparing performance criteria across methods needs certainty estimates, and these need resampling

Acknowledgements

- Ada Wang
- Jen Townsend
- and the whole Microsoft ExP team

Slide references

1. Microsoft Research. (n.d.). Metric computation for multiple backends.
2. Microsoft Research. (n.d.). Deep Dive Into Variance Reduction.
3. (2020). Econometric Reviews, 39(10), 1071-1094. https://doi.org/10.1080/07474938.2020.1824732
4. Winston, L. (n.d.). Agnostic Learning vs. Prior Knowledge. Department of Statistics, University of California, Berkeley. https://www.stat.berkeley.edu/~winston/agnostic.pdf
5. Spotify Engineering. (n.d.). Bringing Sequential Testing to Experiments with Longitudinal Data (Part 1): The Peeking Problem 2.0. https://engineering.atspotify.com/2021/06/08/bringing-sequential-testing-to-experiments-with-longitudinal-data-part-1-the-peeking-problem-2-0/
6. Gupta, M., et al. (n.d.). Novelty and Primacy: A Long-Term Estimator for Online Experiments. Microsoft.
7. Velasco, B. (n.d.). Double Machine Learning for causal inference. Towards Data Science. https://towardsdatascience.com/double-machine-learning-for-causal-inference-832f4b3aa1e1

Candidates

- Multi-linear regression with all covariates (in the same set of metrics)

$$Y \sim I_T + \boldsymbol{X}$$

  - $\boldsymbol{X}$: a list of pre-experiment covariates (may include pre-experiment outcome Y_{pre}, independent of I_T)
- Multi-linear regression with variable selection (ways to select $\boldsymbol{X}$)
  - Option 1: from Lasso, keep variables with coefficients != 0
  - Option 2: from Ridge, keep variables with coefficients != 0
  - Option 3: forward variable selection
- CUPAC
  - Replacing a large number of variables with a single predicted value: using machine learning models

Benchmark methods

- Simple t-test
- CUPED method (Univariate VR metrics)