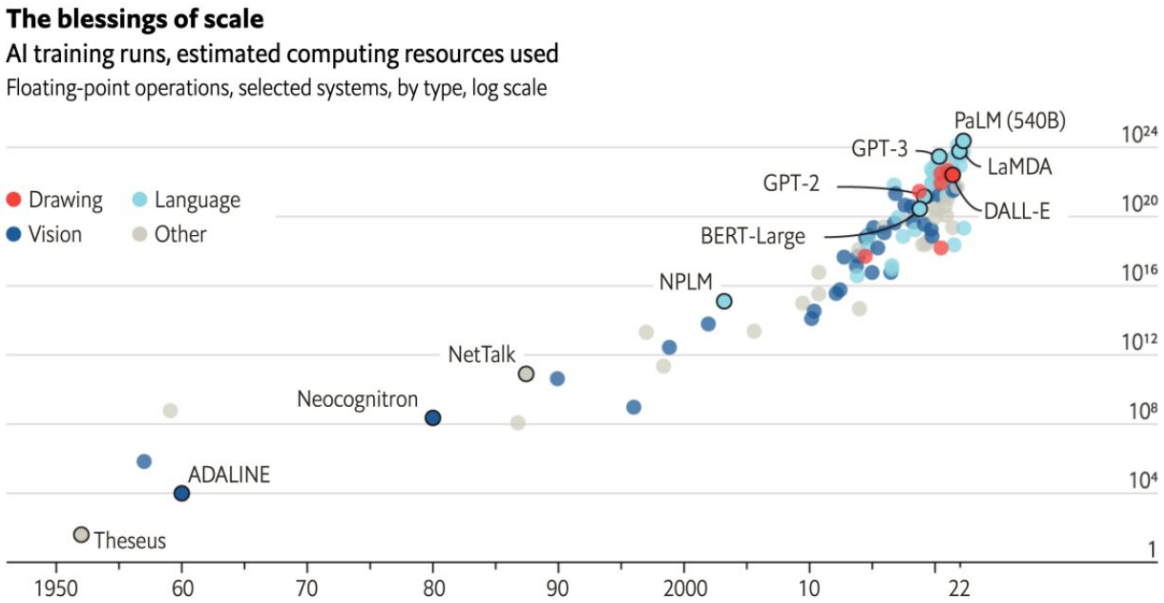
Penn Engineering
University of Pennsylvania

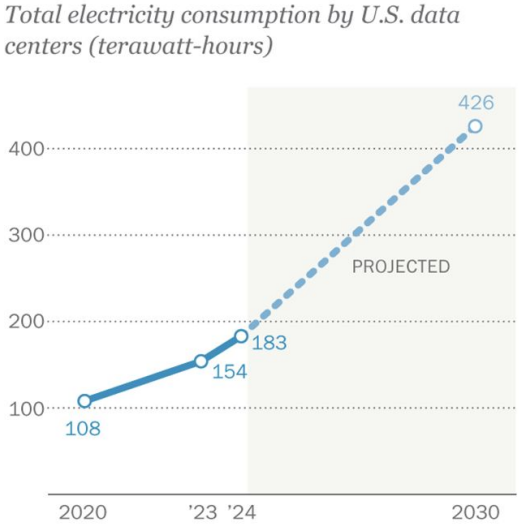
CEO: A Causal Evaluation and Optimization Framework for Datacenter Management Policies

William Meng, Tianyi Wu, Yi Ding, Benjamin Lee

Exploding Power Consumption in Datacenters

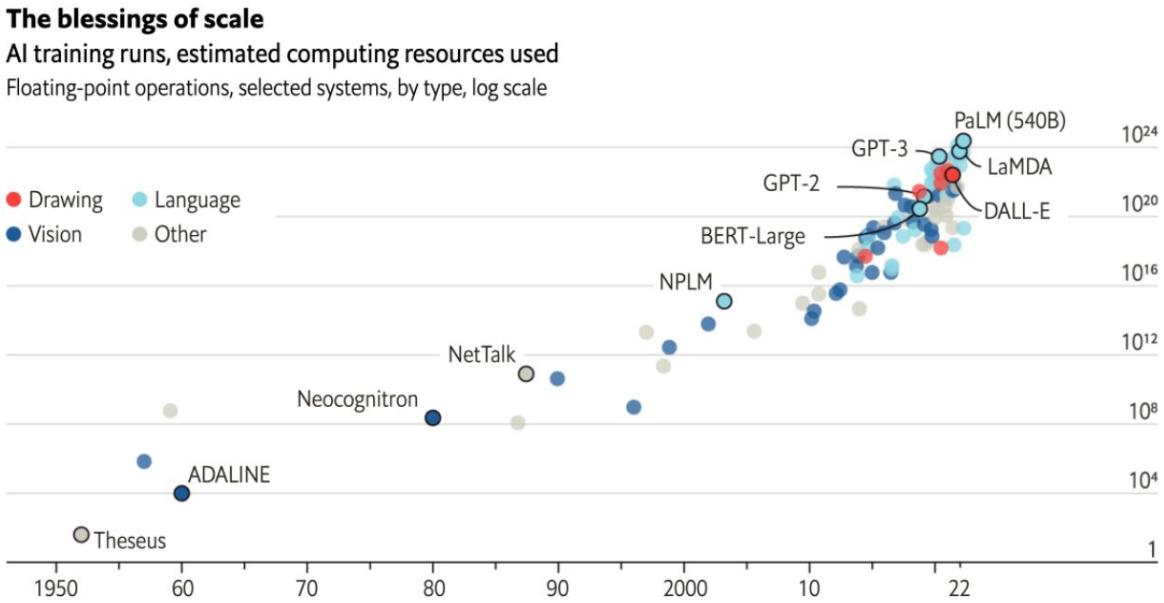


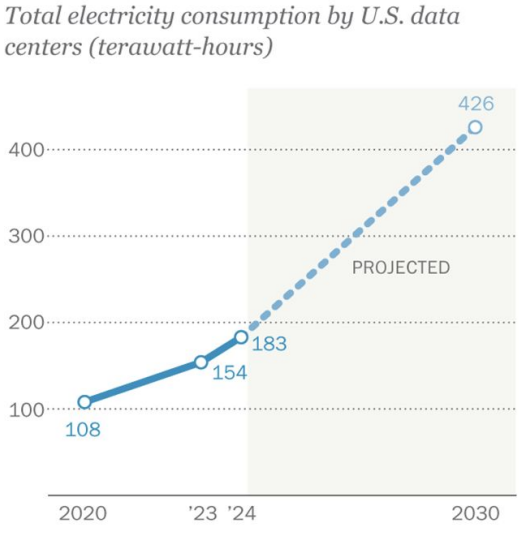
Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data



International Energy Agency, "Energy and AI" April 2025

Exploding Power Consumption in Datacenters





Large scale computation is ***expanding***
Demanding novel policies to improve ***efficiency***

Finding the Best Policy

Penn Engineering
University of Pennsylvania

GPU Optimization	LLM Optimization	Resource Optimization
Dynamic Frequency Scaling	KV Cache Offloading	Dynamic Resource Allocators
Power Capping	Prefix Caching	

Finding the Best Policy

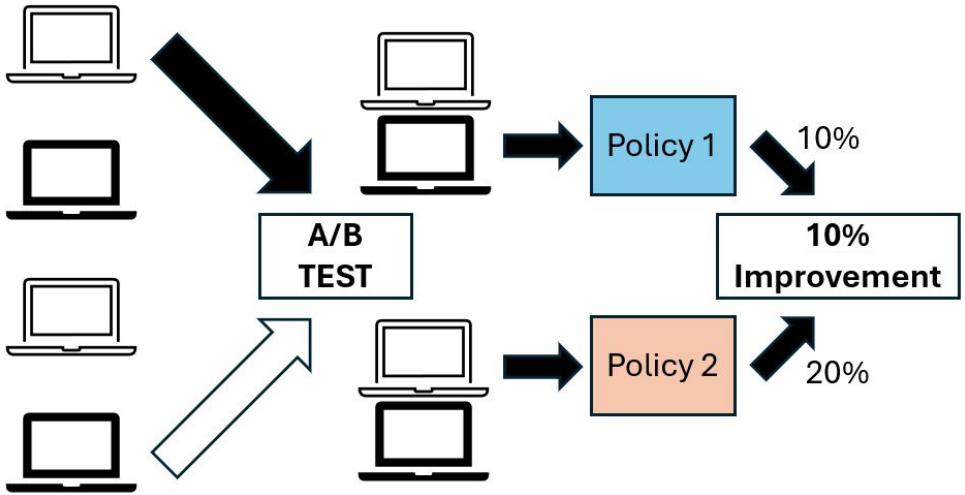
GPU Optimization	LLM Optimization	Resource Optimization
Dynamic Frequency Scaling	KV Cache Offloading	Dynamic Resource Allocators
Power Capping	Prefix Caching	

With such a large variety of solutions effective ***evaluation*** is imperative to establish their efficacy

A/B Tests

Run various policies under ***equivalent circumstances***

- Workload duplication
- Randomized Controlled Trial

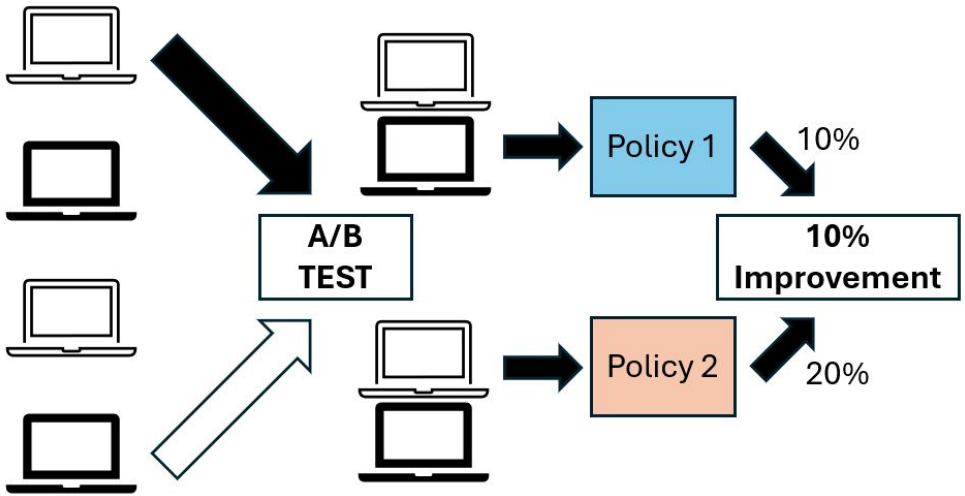


A/B Tests

Run various policies under ***equivalent circumstances***

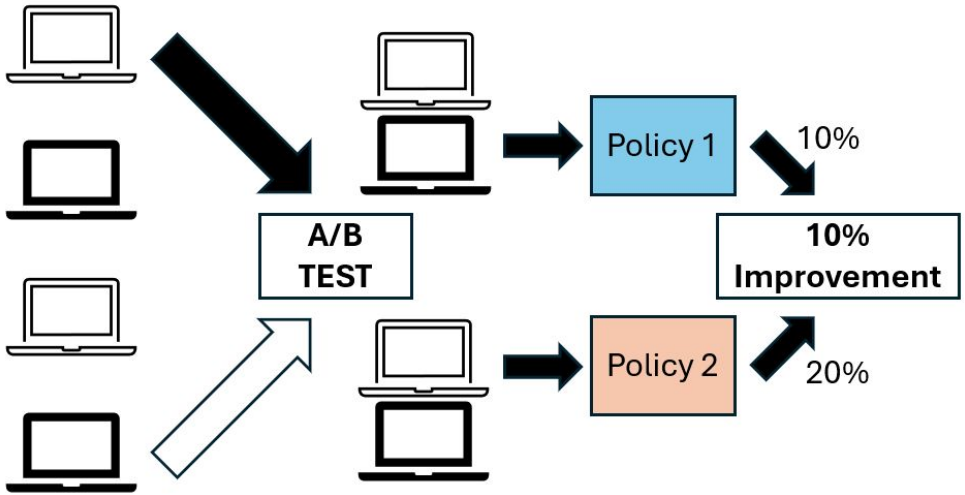
- Workload duplication
- Randomized Controlled Trial

Problematic at Scale



Challenges with A/B Testing

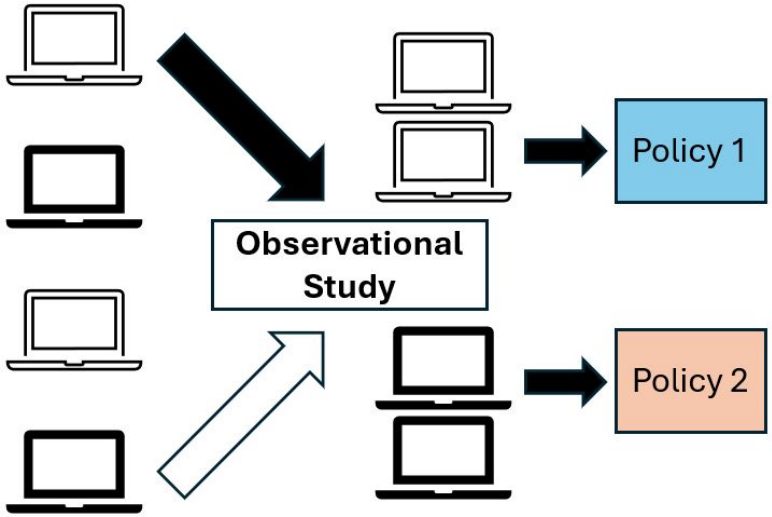
1. **Infrastructure**
   a. Equivalent test clusters
   b. Routing production traffic is expensive and impractical
2. **Consent**
   a. Randomization requires consent
   b. User adverse
3. **Data**
   a. Large costly datasets
   b. Large overheads



Observational Studies

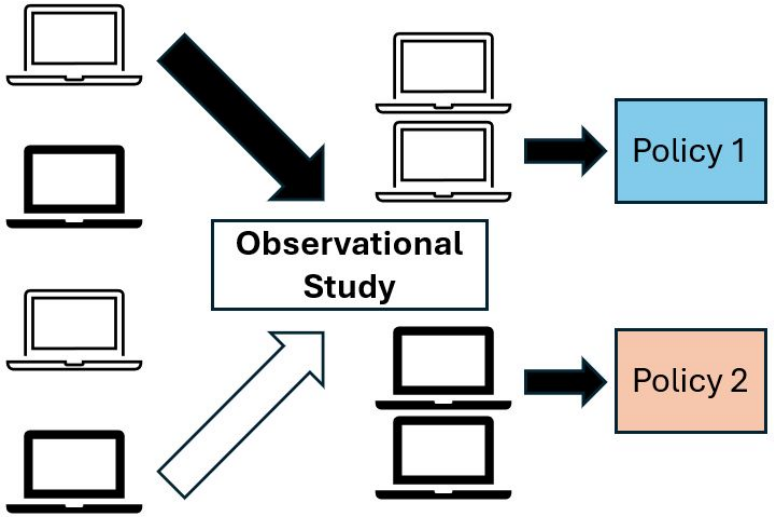
Assess performance in a ***natural*** setting

Allow production work to continue without ***disruption***



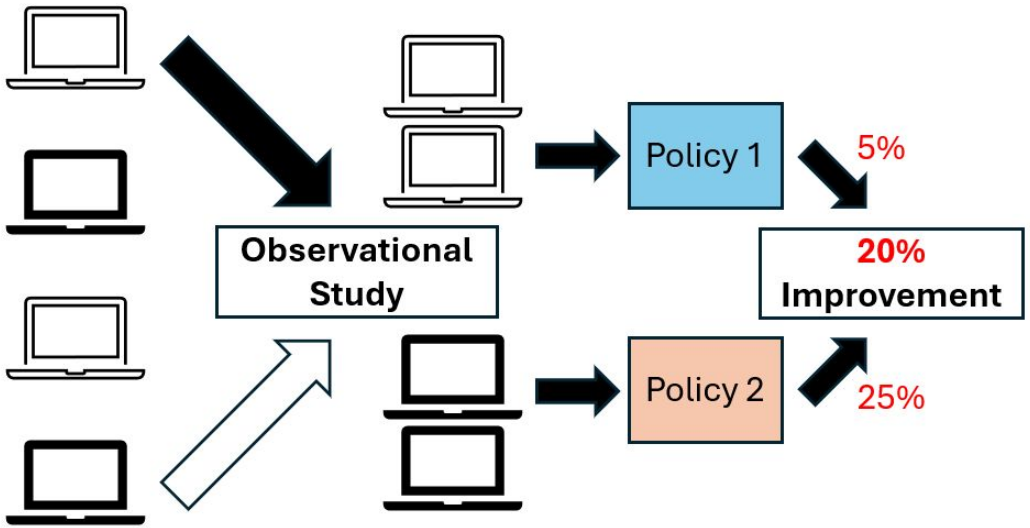
Benefits of Observational Studies

1. Cheap
2. Low Overhead
3. Doesn't Disrupt Production Flow



Challenges With Observational Studies

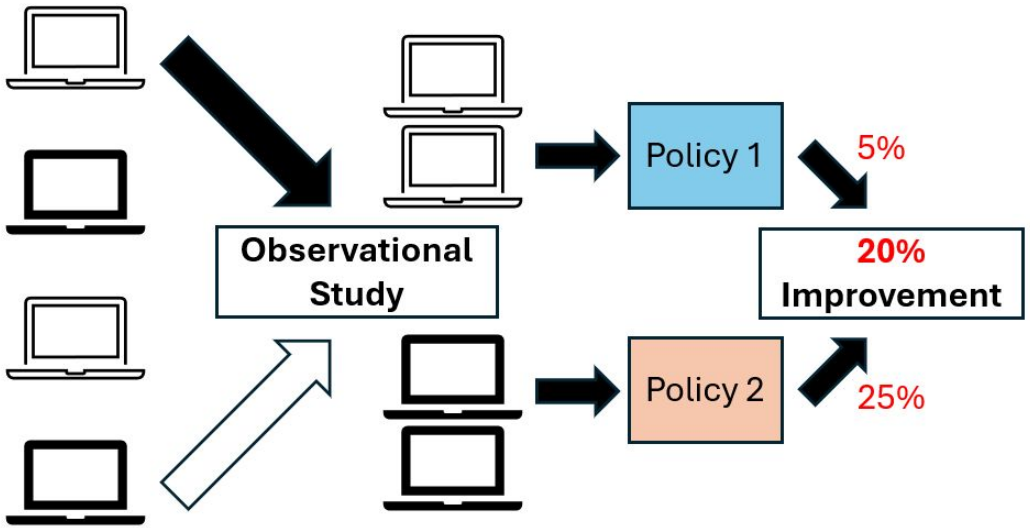
- Observational studies are ***uncontrolled***
- Policies see ***different*** types of workloads
- Imbalanced workload distributions induce ***bias***



Challenges With Observational Studies

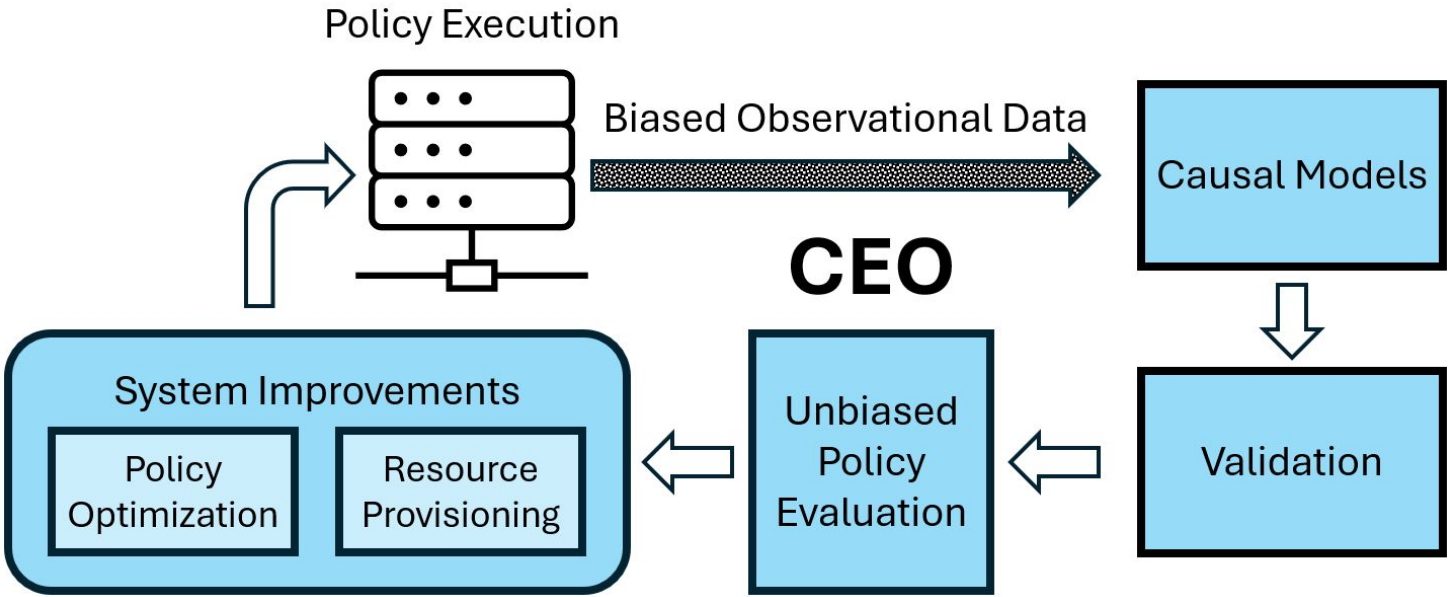
Penn Engineering
University of Pennsylvania

- Observational studies are ***uncontrolled***
- Policies see ***different*** types of workloads
- Imbalanced workload distributions induce ***bias***



How can we robustly evaluate policies using biased observational data?

Our Solution: CEO



CEO: Average Causal Effect

$$ACE = E[Y_1 - Y_0] = \frac{1}{N}\sum_{i=1}^{N}(Y_1(x_i) - Y_0(x_i))$$

Estimate the ***effect*** of a new policy on the outcome Y for a job with characteristics x_i

CEO: Average Causal Effect

$$ACE = E[Y_1 - Y_0] = \frac{1}{N}\sum_{i=1}^{N}(Y_1(x_i) - Y_0(x_i))$$

Problem

We can only measure ***either*** Y_1 or Y_0

CEO: Average Causal Effect

$$ACE = E[Y_1 - Y_0] = \frac{1}{N}\sum_{i=1}^{N}(Y_1(x_i) - Y_0(x_i))$$

Problem

We can only measure ***either*** Y_1 or Y_0

Use causal ML to estimate instead!

CEO: Outcome Regression

$$\mu(t, x) = E(Y|T = t, X = x)$$

Use ML to estimate the unmeasured outcome!

Allows us to fill in the missing outcome in our ACE equation enabling ***unbiased*** comparison

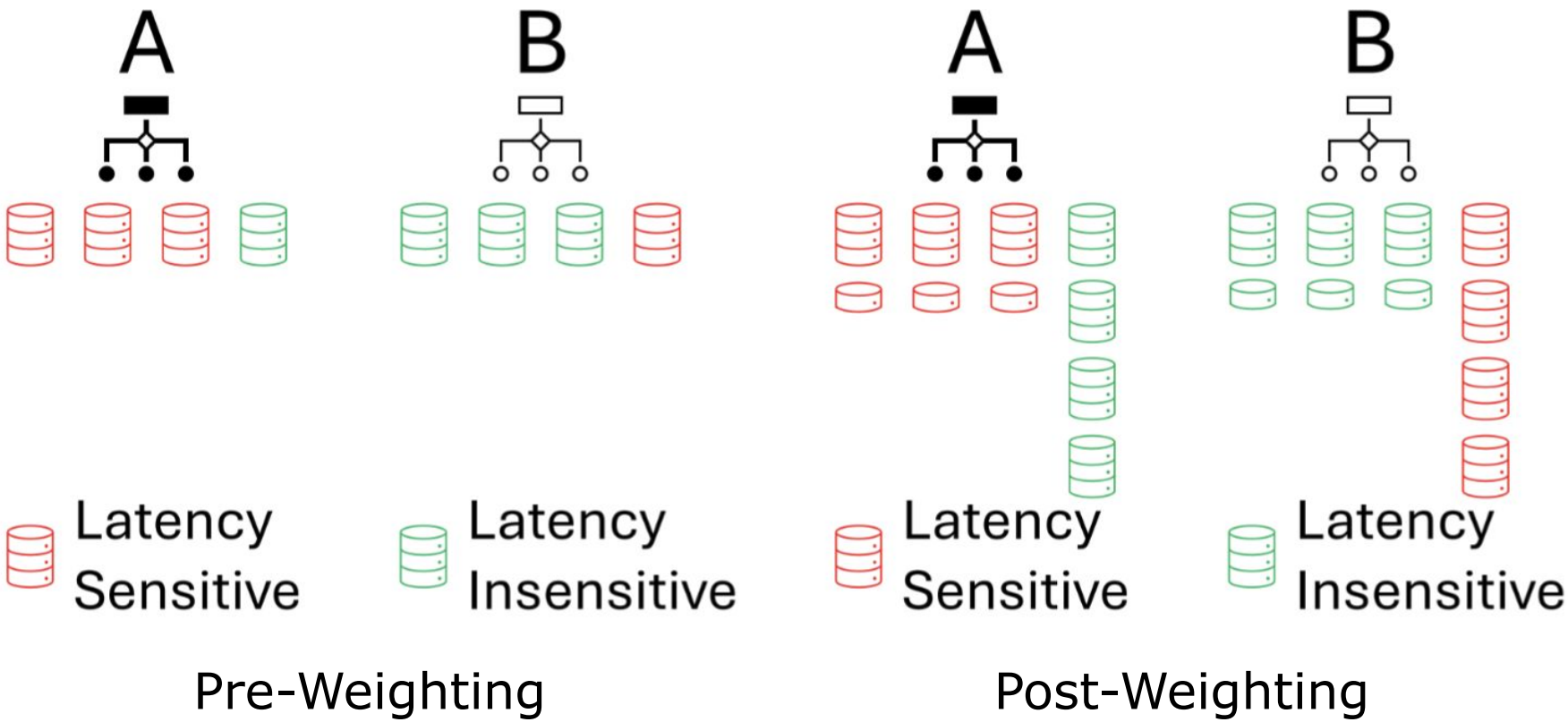
CEO: Propensity Score Weighting

$$P(x_i) = Pr(T = 1 | X = x_i)$$

Use ML to estimate the ***probability*** a job was treated given its characteristics

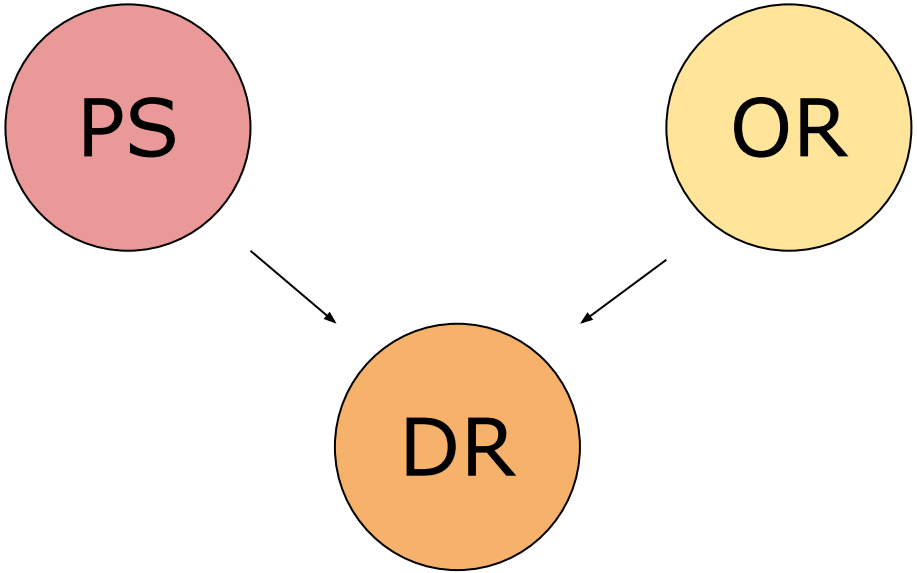
Use this probability to ***rebalance*** the initial dataset

CEO: Propensity Score Weighting



CEO: Doubly Robust

Merge PS and OR estimator into a ***single*** model



Allows for unbiased estimation as long as ***one*** model is correct

CEO: Autopilot Case Study

Penn Engineering
University of Pennsylvania

Autopilot: Google's Dynamic Resource Scheduler

- Dynamically adjust memory/CPU assigned to a job to reduce waste
- Evaluated over 10,000 jobs in an ***uncontrolled observational study***

CEO: Autopilot Case Study

Penn Engineering
University of Pennsylvania

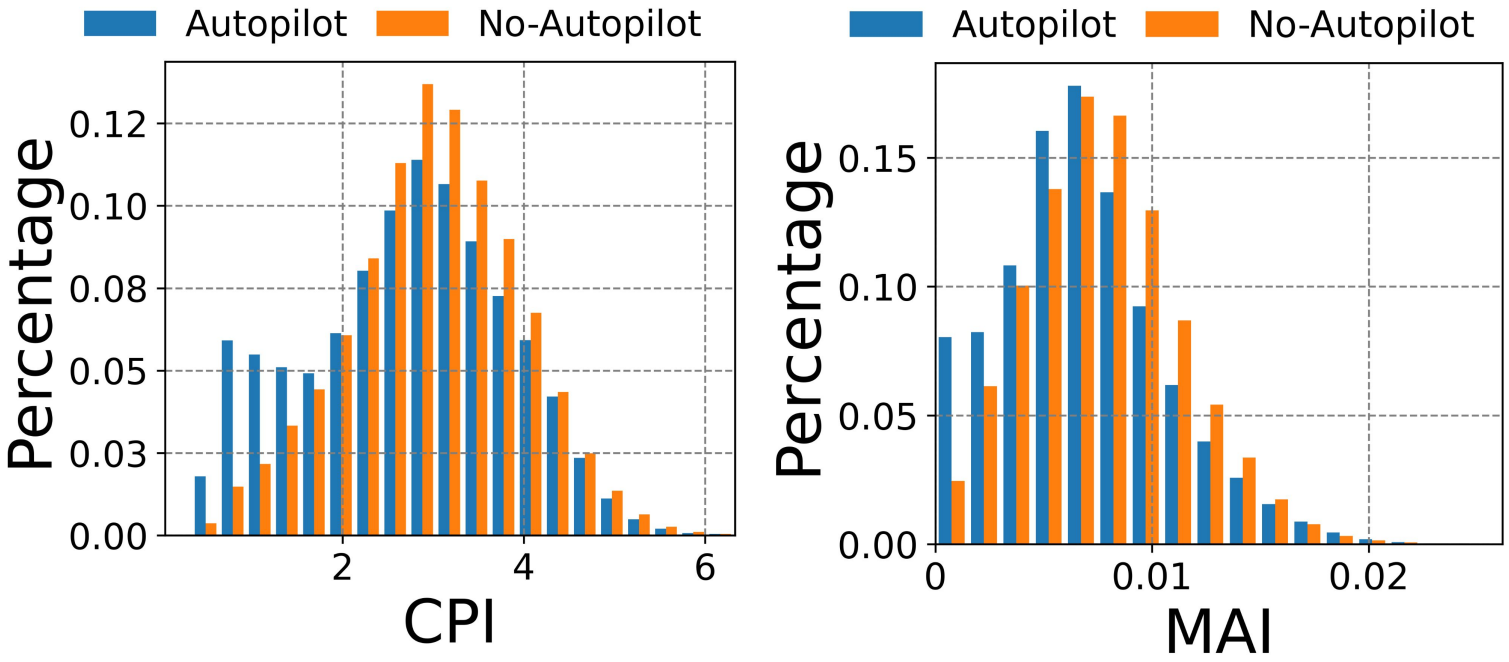
Autopilot: Google's Dynamic Resource Scheduler

- Dynamically adjust memory/CPU assigned to a job to reduce waste
- Evaluated over 10,000 jobs in an ***uncontrolled observational study***

Our Work

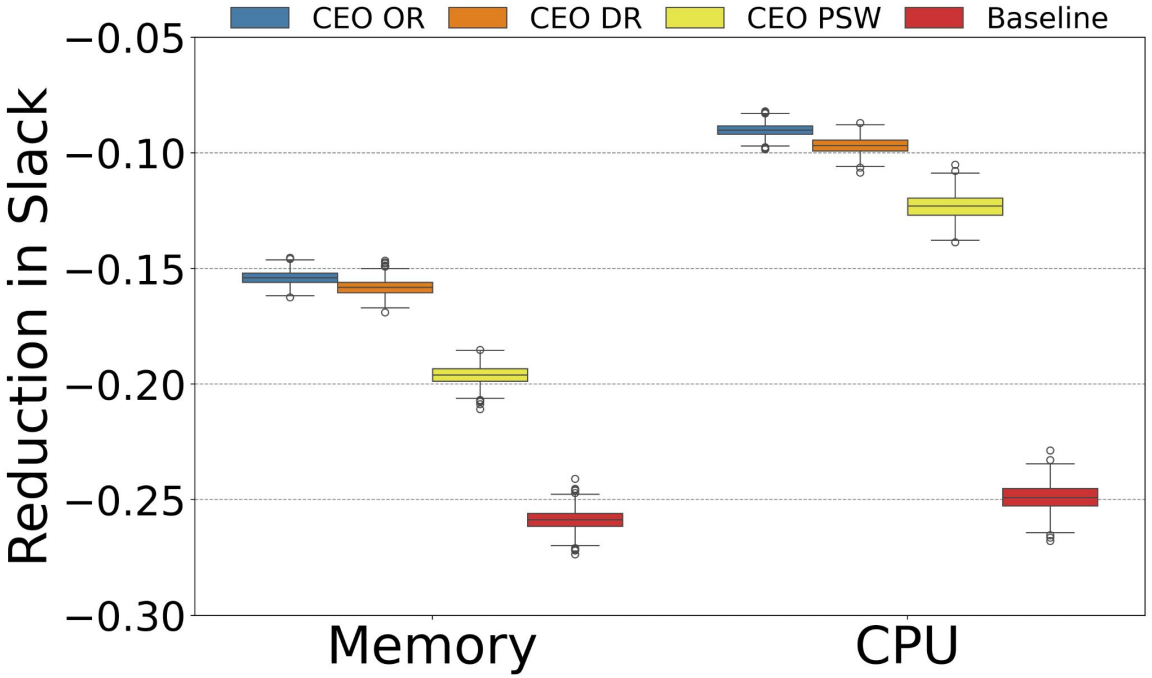
- ***Replicate*** Google's study using Google's 2019 Cluster Dataset (800,000 jobs)
- Use **CEO** to ***identify*** the existence and ***quantify*** the effect of bias
- ***Understand*** system ramifications of bias

CEO: Bias in Autopilot



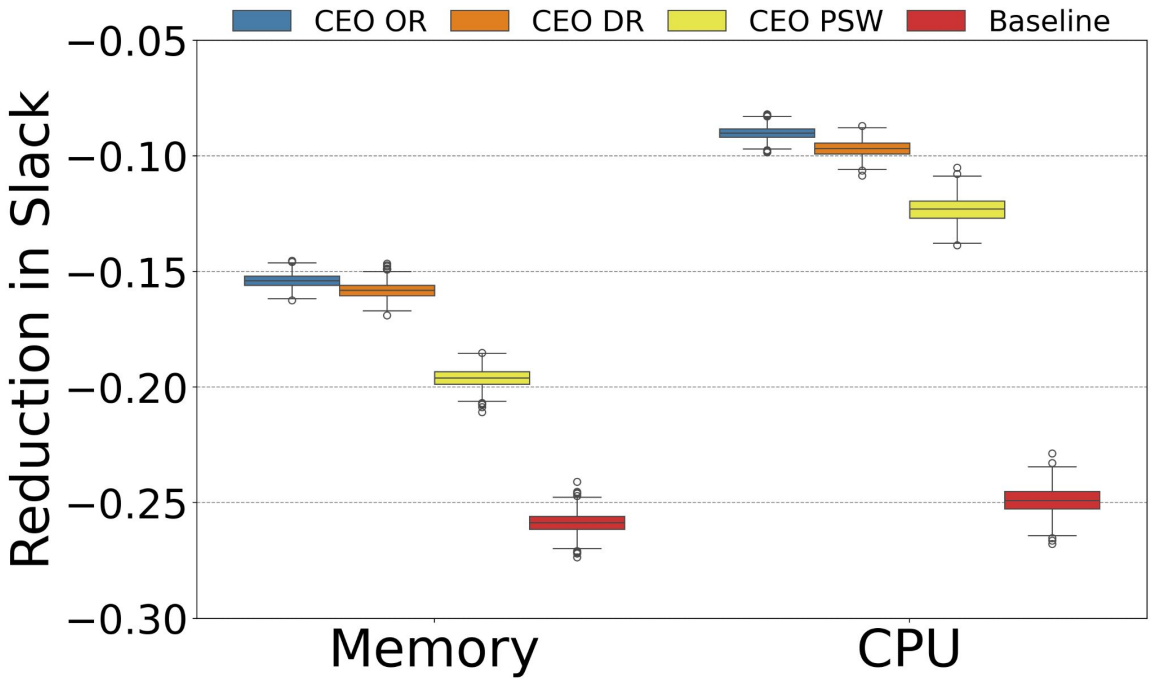
Autopiloted jobs are ***statistically different***
from Non-Autopiloted jobs

Estimating ACE Using CEO



Lower slack means more resources are saved

Estimating ACE Using CEO



Identifying ***possible*** magnitude of bias is important for production system decisions

CEO: System Implications

1. **Memory Provisioning**
   a. Naive memory savings would lead to ***overprovisioning*** memory bandwidth
   b. True savings are likely ***lower*** leading to out of memory errors
2. **CPU Provisioning**
   a. Proposed CPU savings may reduce power provisioning
   b. Due to savings being ***lower*** power consumption would be too ***high*** bypassing the power limit

CEO: System Implications

1. **Memory Provisioning**
   a. Naive memory savings would lead to ***overprovisioning*** memory bandwidth
   b. True savings are likely ***lower*** leading to out of memory errors
2. **CPU Provisioning**
   a. Proposed CPU savings may reduce power provisioning
   b. Due to savings being ***lower*** power consumption would be too ***high*** bypassing the power limit

Accurate policy estimation is imperative to safely deploying system resources at scale

Future Work

Unlike traditional ML, causal models can ***not*** be ***directly*** validated as ***ground truth*** does not exist

Future Work

Unlike traditional ML, causal models can ***not*** be ***directly*** validated as ***ground truth*** does not exist

How do we build trust in our model estimates?

Future Work: Validation

Penn Engineering
University of Pennsylvania

More Case Studies

- Use case studies where we measure ground truth then ***simulate*** observational data
  - Can CEO ***reproduce*** a ground truth from biased data?

Alternative Validation Techniques

- Standardized Mean Difference
  - How well does the Propensity Score Model rebalance the population?
- Synthetic Data

Conclusion

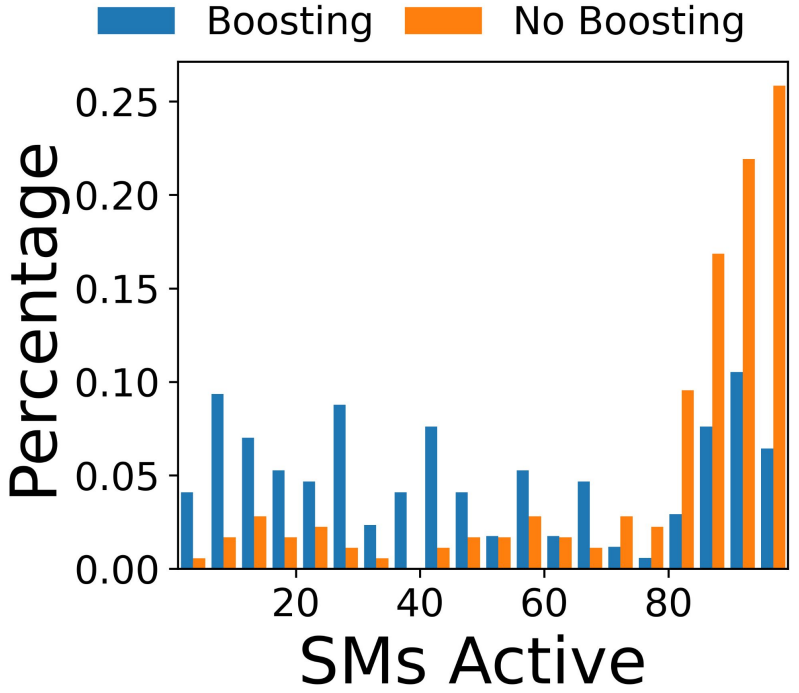
CEO exploits ***observational*** data to bypass inherent ***limitations*** of A/B testing at datacenter scale

CEO utilizes sophisticated ***Causal Models*** to overcome ***bias*** within observational data

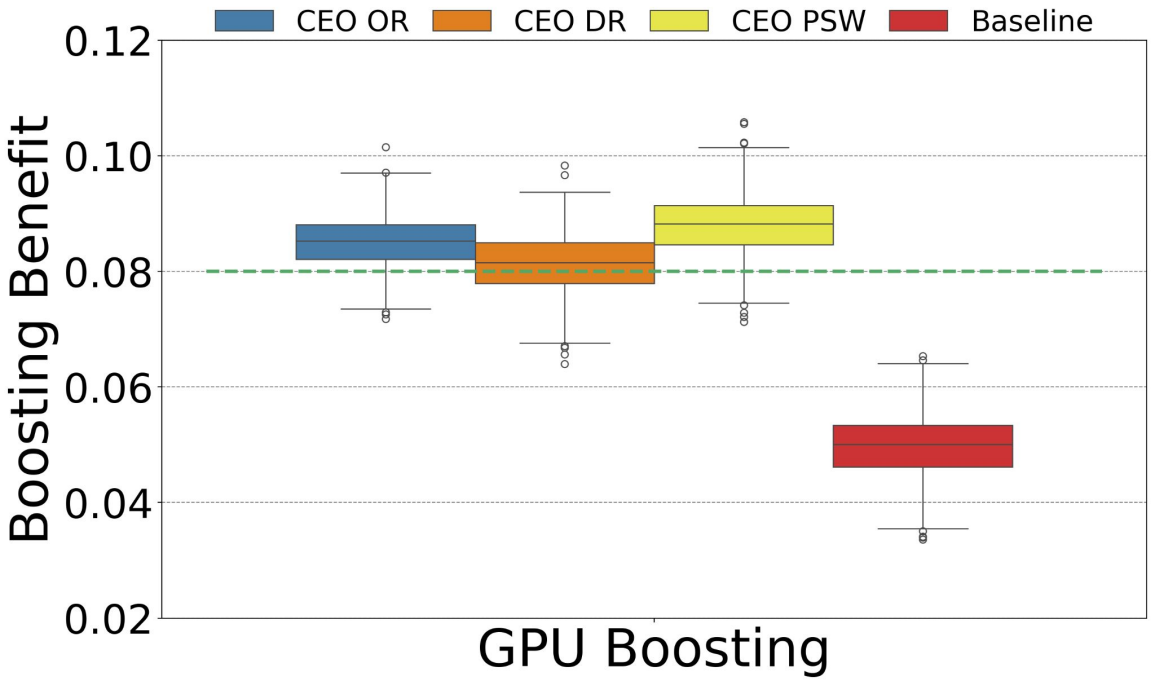
CEO's estimations have ***implications*** on system settings

CEO: GPU Boosting

- 349 ML models with boosting enabled/disabled
  - Computer Vision, NLP, Recommendation and Speech
- Generated a biased observational dataset
  - Boosting applied disproportionately to workloads with low GPU utilization

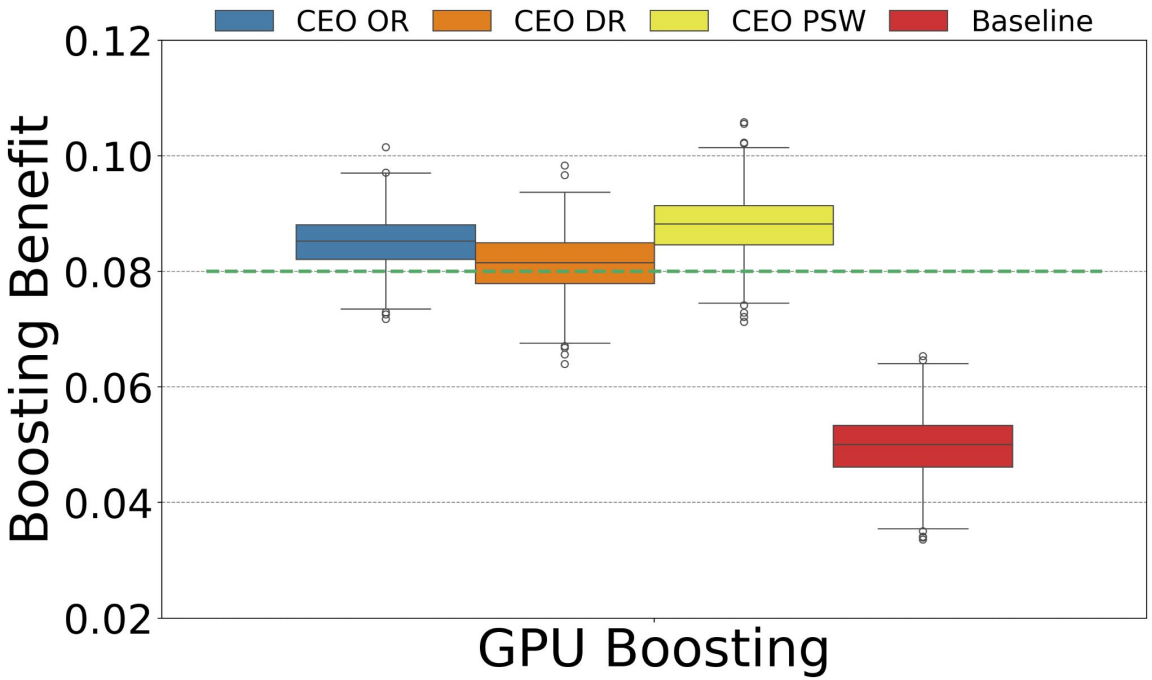


Reducing Bias Using CEO for GPU Boosting

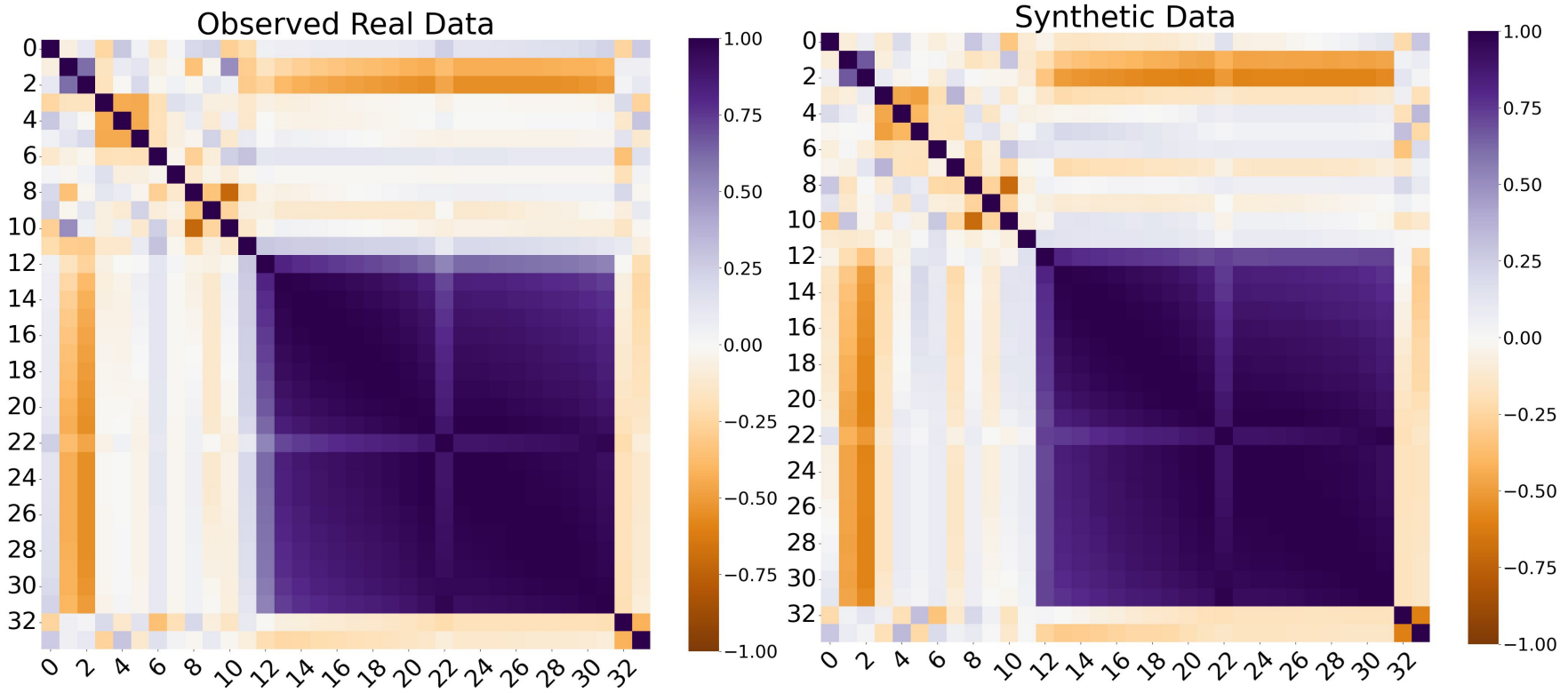


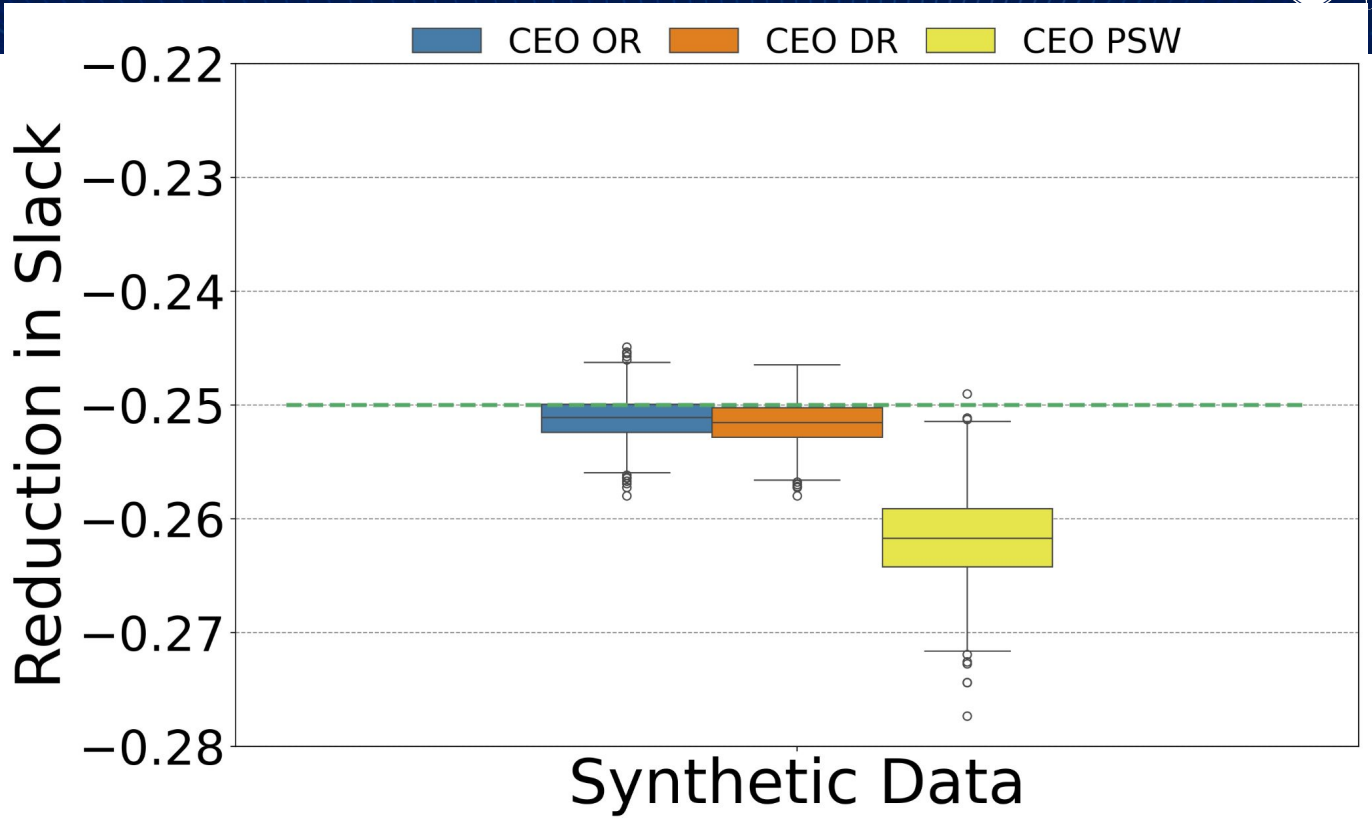
Green line represents the ***measured*** policy effect

Reducing Bias Using CEO for GPU Boosting



CEO can accurately estimate a policies effect from biased observational data.


Observed Real Data
Synthetic Data
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


CEO OR
CEO DR
CEO PSW
−0.22
−0.23
−0.24
−0.25
−0.26
−0.27
−0.28
Reduction in Slack
Synthetic Data