Machine Learning for Applied Causal Inference at Facebook

Dominic Coey

Facebook, Core Data Science

A common meme

machine learning is about black-box prediction

economics is about understanding causal mechanisms

Themes in machine learning

Learning relation between inputs and outputs from labeled examples

Themes in machine learning

Supervised learning

Learning relation between inputs and outputs from labeled examples

Themes in machine learning

Supervised learning

Learning relation between inputs and outputs from labeled examples

Finding simple, latent structure in complex data

Themes in machine learning

Supervised learning

Learning relation between inputs and outputs from labeled examples

Unsupervised learning

Finding simple, latent structure in complex data

Themes in machine learning

Supervised learning

Learning relation between inputs and outputs from labeled examples

Unsupervised learning

Finding simple, latent structure in complex data

Generating good behavioral policies in stochastic, dynamic environments

Themes in machine learning

Supervised learning

Learning relation between inputs and outputs from labeled examples

Unsupervised learning

Finding simple, latent structure in complex data

Reinforcement learning

Generating good behavioral policies in stochastic, dynamic environments

Some applications to causal inference

Supervised learning

”Compliance-weighting” for variance reduction

Unsupervised learning

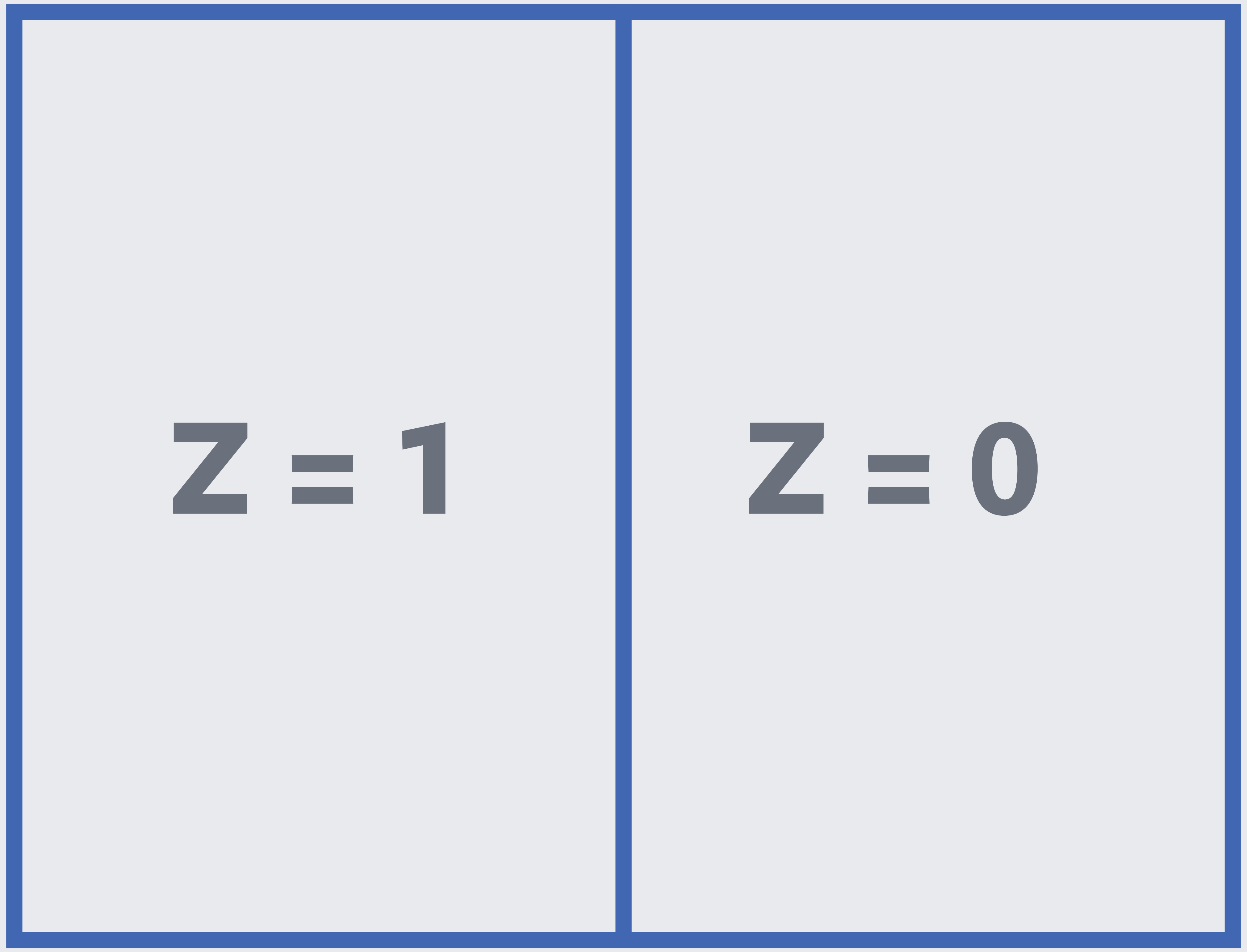
Graph partitioning for cluster randomization in experiments with spillovers

Reinforcement learning

Importance sampling for counterfactual evaluation of ML models

Compliance-weighted IV

Consider an experiment with one-sided non-compliance.

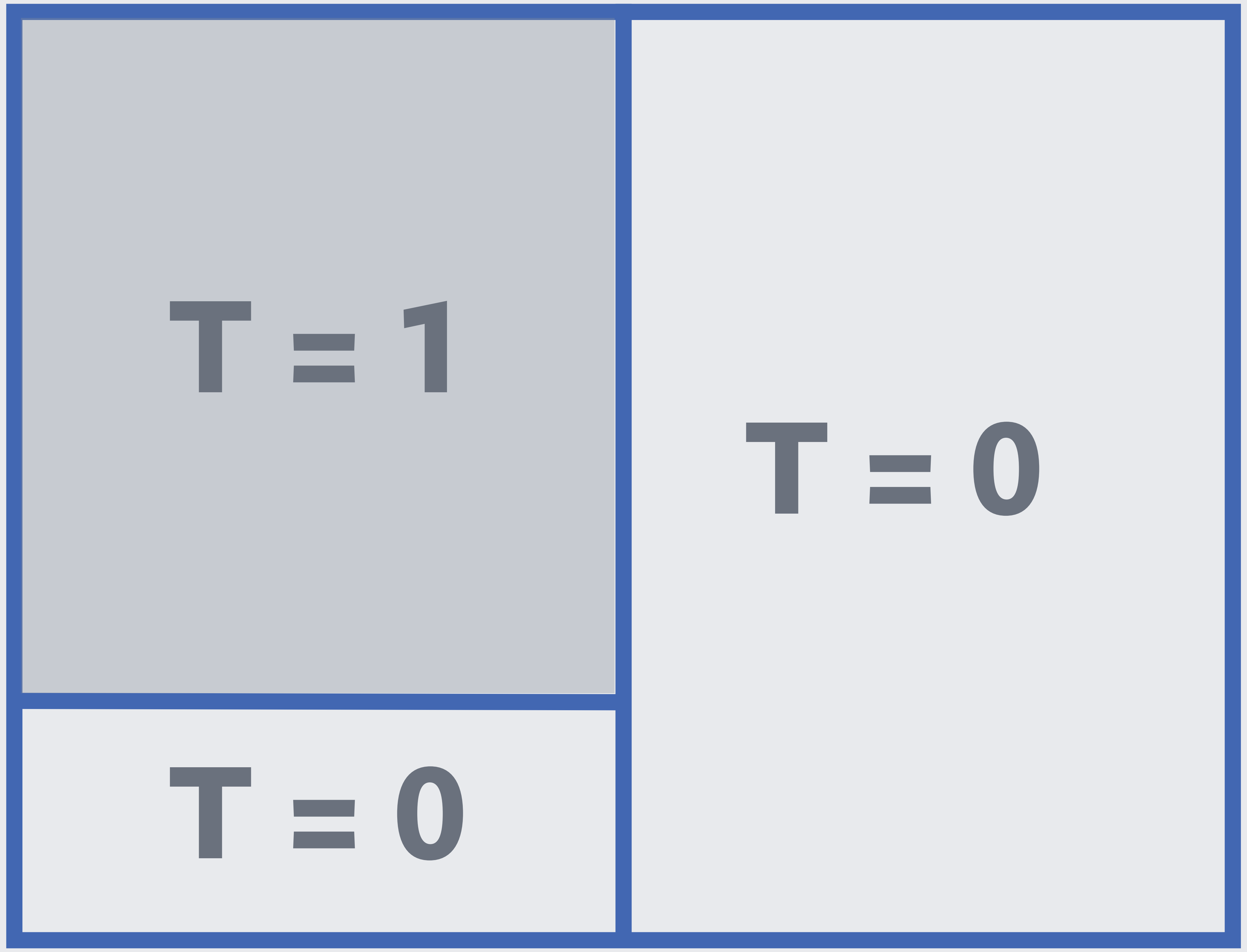


$Z = 1$

$Z = 0$

Compliance-weighted IV

Consider an experiment with one-sided non-compliance.



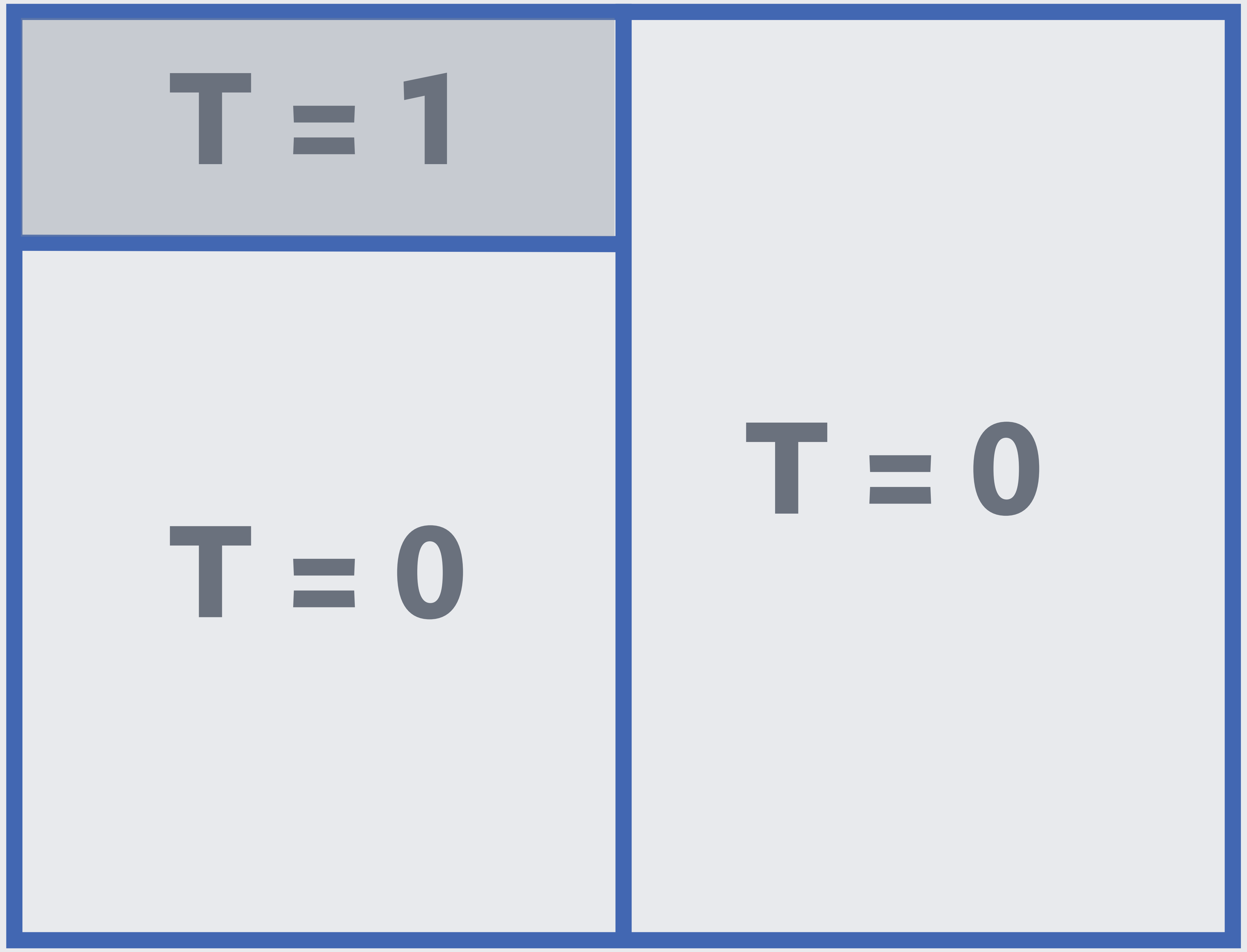
T = 1

T = 0

T = 0

Compliance-weighted IV

If few people in the treatment group receive treatment, IV estimates may be very noisy.



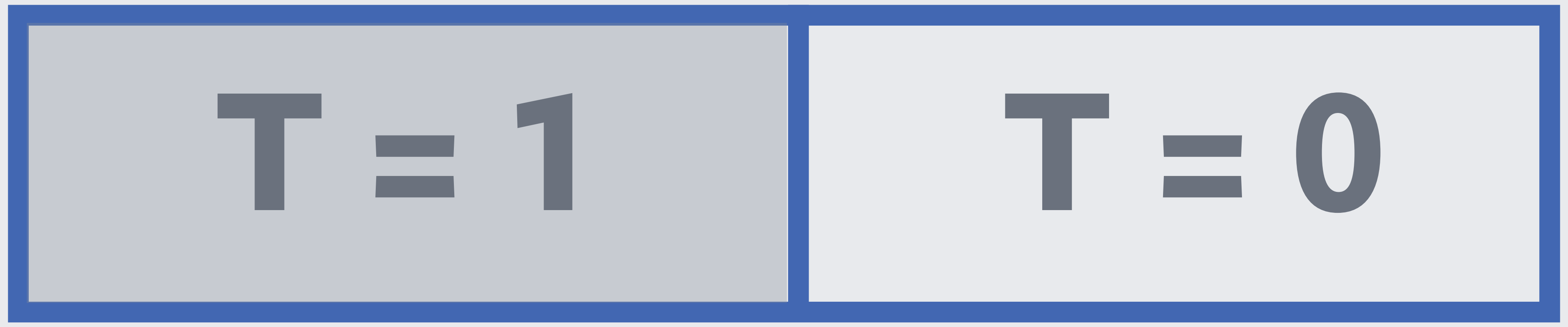
T = 1

T = 0

T = 0

Compliance-weighted IV

If compliance is observed, can simply restrict attention to compliers.



T = 1

T = 0

Non-compliers only contributing noise to the Z = 1 vs Z = 0 contrast. This reduces IV variance by a factor of ~1/Pr(complier).

Compliance status is typically not observed, but may have good predictors. Standard supervised learning problem.

Compliance-weighted IV: Optimality

This suggests downweighting those unlikely to comply.

Consider a class of weighted Wald estimators, where the weights are functions of individual covariates X_i.

Proposition. Assume i) zero intent-to-treat effect, ii) conditional mean and variance independence of potential outcomes and X_i. Then the variance-minimizing weighting function is the compliance probability, $Pr(complier_i \,|\, X_i)$.

Compliance-weighted IV: Inference

Estimating weighted LATE instead of LATE. With heterogeneous treatment effects, these will generally differ.

Precise definition of estimand rather subtle. Could consider the population weighted LATE evaluated at:

- the estimated weighting function, $\beta_{\widehat{w}}$
- the ″true″ weighting function, β_w

We study the former: $\sqrt{N}(\widehat{\beta}_{\widehat{w}} - \beta_{\widehat{w}})$, not $\sqrt{N}(\widehat{\beta}_{\widehat{w}} - \beta_w)$.

Compliance-weighted IV: Inference

What about inference? *Cross-fitting* (Athey & Wager, 2017; Chernozhukov et al. 2018) helps here:

- split experiment into two subexperiments
- estimate compliance probabilities on first half, use to generate weights for the second half and vice versa
- run weighted IV on both samples separately, and average results

Get valid standard errors from standard IV software. No special adjustments required.

Inference with Cross-fitting

Let $V(w)$ denote the usual sandwich estimator variance for weighted IV:

$$V(w) = (E[Z_i T_i' w_i])^{-1} E[Z_i Z_i' e_i^2 w_i^2] (E[T_i Z_i' w_i])^{-1} .$$

Under some relatively mild assumptions (chiefly existence of a first stage, consistency of $\widehat{w}$ for w), we have

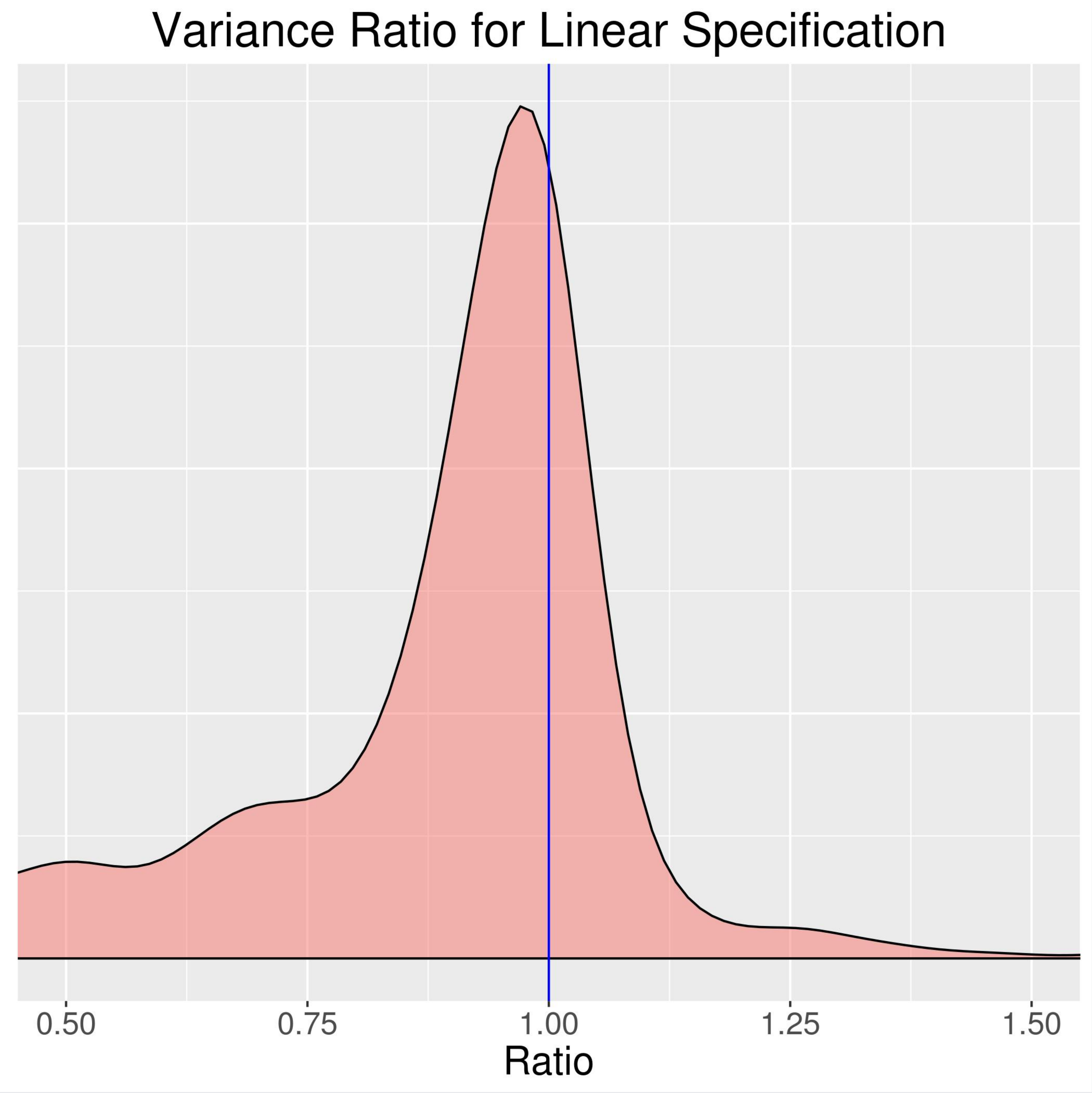
$$\sqrt{N}(\widehat{\beta}_{\widehat{w}} - \beta_{\widehat{w}}) \rightarrow_d \mathcal{N}(0, V(w)) .$$

Compliance-weighted IV: Empirical Results

Apply methodology to a sample of ~700 advertising effectiveness experiments, each with between 40k - 2m users.

Variance drops on average by 16%.

Compliance-weighted IV also has lower MSE for LATE than usual unweighted estimator.



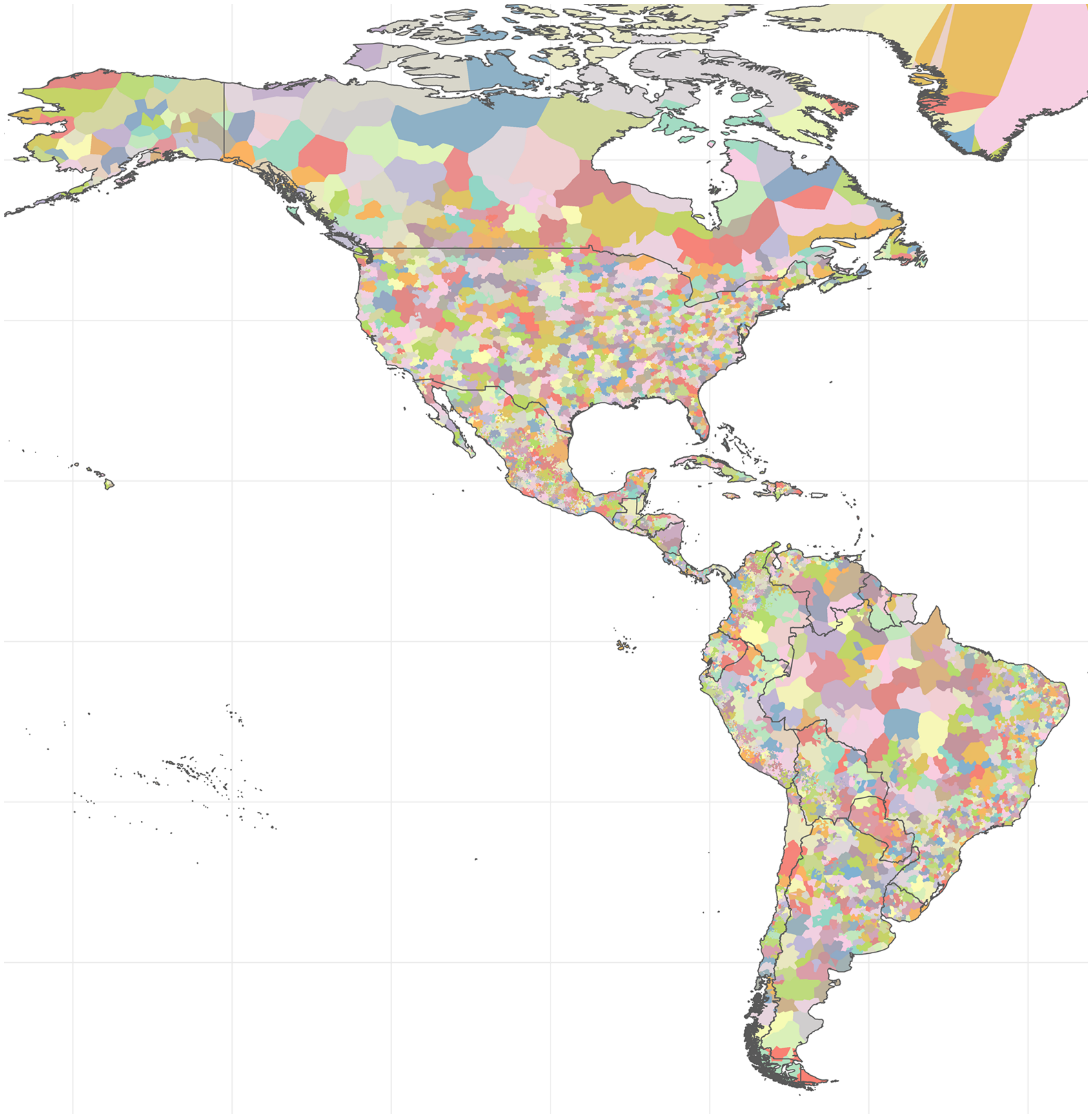
Graph partitioning for experimentation

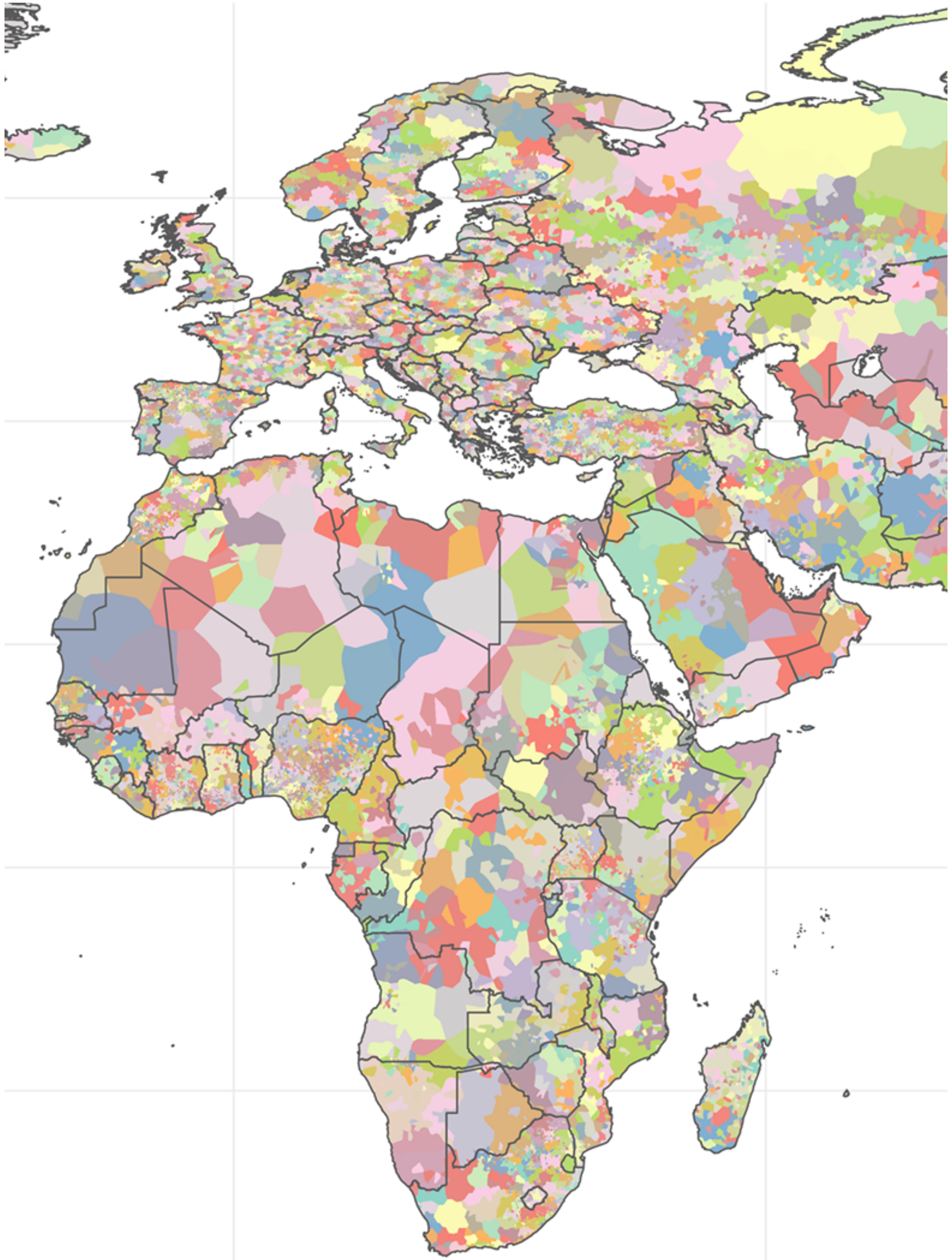
Classic assumption (SUTVA)

Outcomes for person i don't depend on treatment assignments of person $j \neq i$.

This is a *really* bad assumption in some common experimental scenarios, where one person can directly affect others' outcomes, e.g. hiring outcomes in job markets.

With the right data, can divide geographies into commuting zones, using graph partitioning algorithms.





Counterfactual evaluation of ranking models

Reinforcement learning *is* causal inference. Moreover, many supervised learning problems are RL problems in disguise.

Want to show relevant ads to users. Need to predict click-through rates.

Can have a bad model with *perfectly* calibrated predictions on production data. The model endogenously determines the data observed.

Counterfactual evaluation of ranking models

	Ad A	Ad B
True CTR	1	0.1
Estimated CTR, Model 1	0	0.1
Estimated CTR, Model 2	0.5	0

Counterfactual evaluation of ranking models

	Ad A	Ad B
True CTR	1	0.1
Estimated CTR, Model 1	0	**0.1**
Estimated CTR, Model 2	**0.5**	0

Model 1 shows ad B, has perfectly calibrated predictions.

Model 2 shows ad A. Poor predictive accuracy, but much better outcomes than model 1.

Counterfactual evaluation of ranking models

Why not just run the experiment of model 1 vs 2? In practice not so easy: way more ideas to test than experimental bandwidth.

Instead, use an idea common in the contextual bandits literature: let's randomize between showing ad A and ad B, then estimate counterfactuals by reweighting (importance sampling).

Natural next step: from counterfactual policy *evaluation* to counterfactual policy *learning*.

Conclusion

Doing causal inference well may involve solving pure prediction problems.

With complex data, simplified data representations are typically necessary for doing data science of *any* kind, let alone causal inference.

Economic policy questions *are* prediction questions, where the prediction target is counterfactual outcomes.