Doubly Robust Machine Learning & Causal Forest for Political Science

Sam Fuller* & Jack T. Rametta**

*Harvard University
**University of California, Davis

January 2025
Binghamton University
Political Science Research Workshop

Overview

Outline

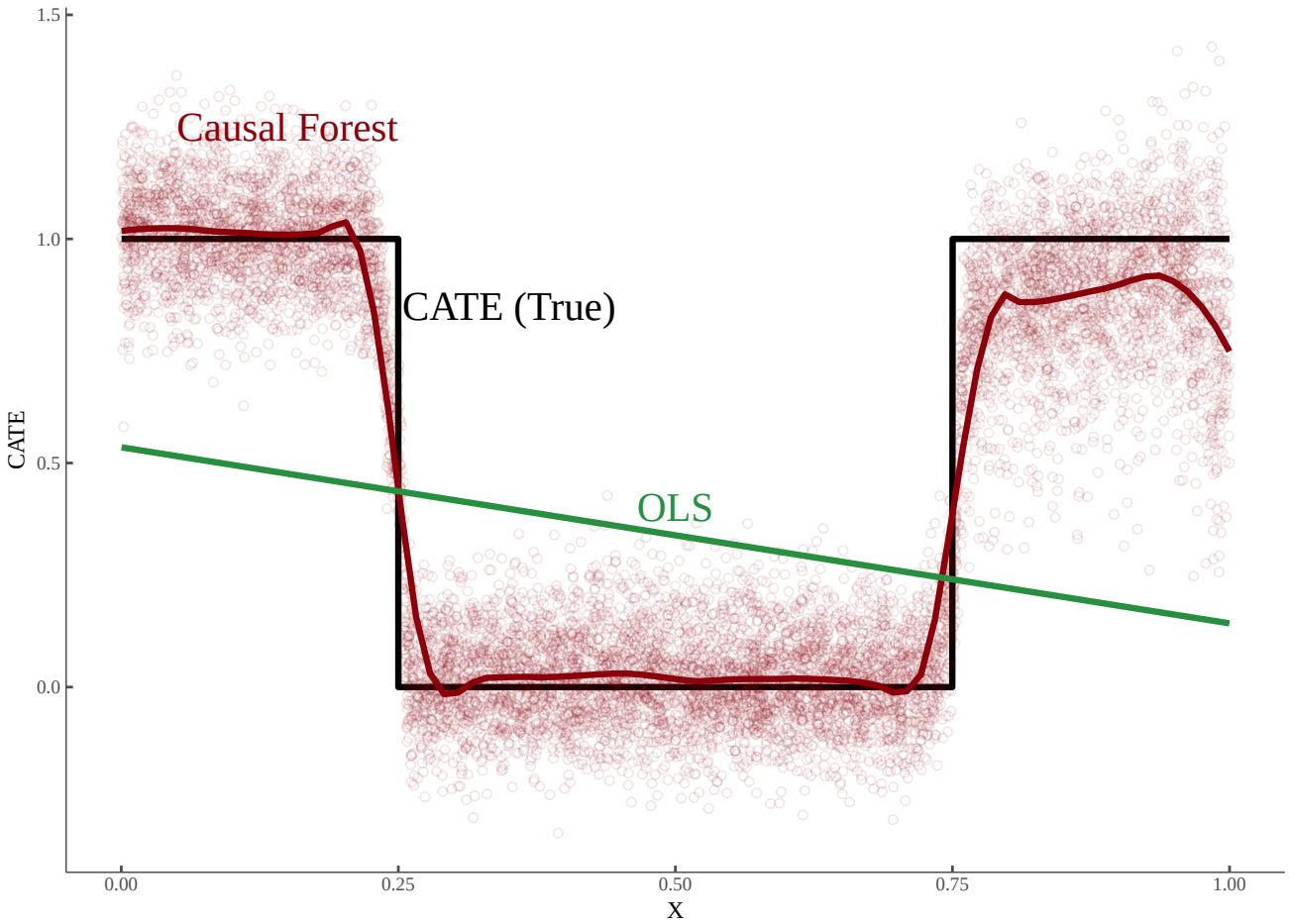
Today's presentation proceeds as follows:

1. Where are these methods applicable? Setting up a framework
2. Establishing the Doubly-Robust Machine Learning framework
3. A Machine learning refresher, trees and (causal) forests
4. Implementing DRML
5. Simulation evidence
6. An application to Voelkel et al. (2024)

Questions During the Presentation

If you have any questions during the course of the presentation, do not hesitate to ask! I'm happy to clarify during or after the presentation.

Do Social Scientists Really Need ML? Can't We Just Use OLS?



↪ Nonlinearities, interactions, heterogeneity lurking beneath the surface. Researchers are fallible!

Background on Causal Machine Learning

Rapid expansion of ML methods for causal inference (Grimmer, Roberts, and Stewart 2021):

- Bayesian Additive Regression Trees (Chipman, George, and McCulloch 2010)
- Causal Forest (Wager and Athey 2018)
- Double Machine Learning (Chernozhukov et al. 2023)
- Ratkovic's PLCE and MDEI estimators (Ratkovic and Tingley 2023)

New forest-based ML methods (Montgomery and Olivella 2018) promise theoretical + practical advantages for main effects, HTEs, and beyond.

But old-school, parametric methods still dominate in political science. Why?

A healthy skepticism of these new approaches.
↪ But much of this strong skepticism is missing the point (e.g, Morucci and Spirling 2024)

Our Goal & Contributions

Goal: Provide a conceptual framework & empirical foundation for political scientists to confidently apply ML methods for both experimental and observational analyses.

Our contributions...

1. Introduce new framework: "Doubly Robust Machine Learning" (DRML)
2. Outline benefits and requirements
3. Provide expansive simulation evidence for the robust benefits of DRML

$\hookrightarrow$ Hope to address concerns, provide basis for adoption of this approach.

Research Contexts

A Motivating *Experimental* Question

Let's assume we have some information treatment and we want to see if it affects vote choice:

information (W) → vote-choice (Y)

But we know from a whole bunch of political science and psychological research that information is processed differently by different people. So we might want to see reactions differ by covariates (PID, gender, education, etc.).

So really we're looking at:
information (W) × covariates (X) → vote-choice (Y)

Conditional treatment effects are our bread and butter in *a lot* of political science research. (More on this later.)

A Motivating *Observational* Question

But let's we instead want to know how the effect of knowing information about an event, say a speech or a policy-implementation, affects vote-choice. The kinds of people who are exposed to information are much different than those who aren't (e.g., more politically sophisticated, stronger partisanship, etc.).

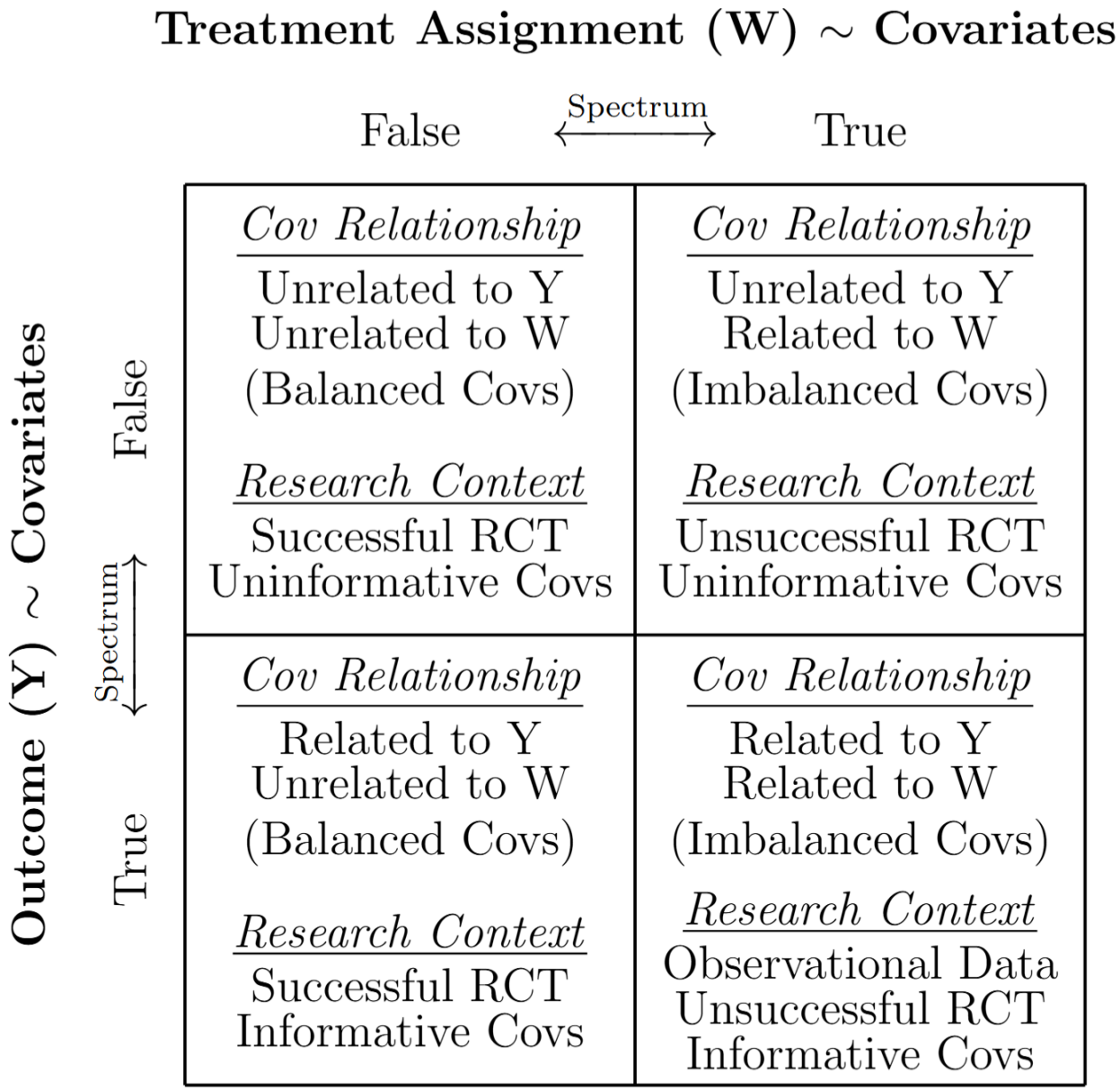
Then really really we're looking at:
$(\textbf{covariates}(\mathbf{X}) \rightarrow \textbf{information}(\mathbf{W})) \times \textbf{covariates}(\mathbf{X}) \rightarrow \textbf{vote-choice}(\mathbf{Y})$

Oh, and those covariates also influence the vote-choice themselves!
$X \rightarrow Y$.

This is all to say, generally, we can think of research contexts on two spectra:

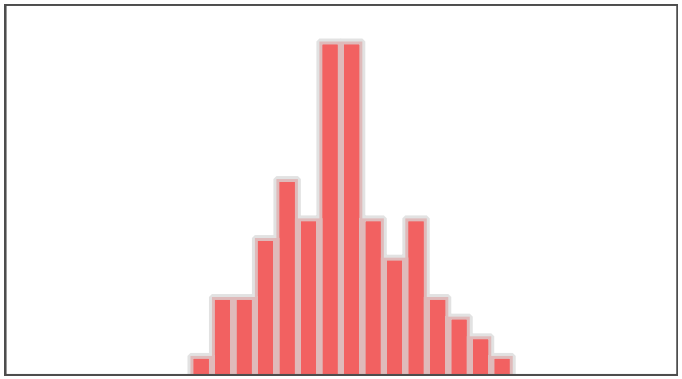
1. Are covariates (X) related to the outcome?
2. Are covariates (X) related to treatment assignment?

Research Contexts for DRML

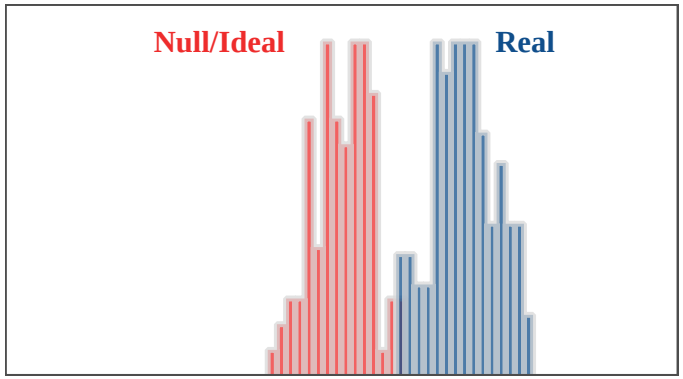


Exogenous Treatment or Selection?

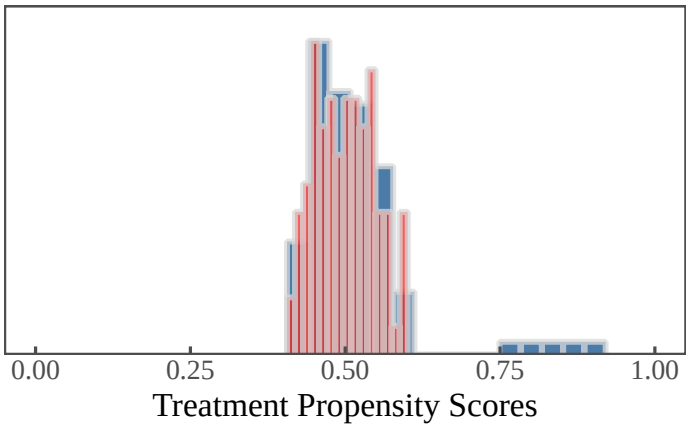
Random Assignment



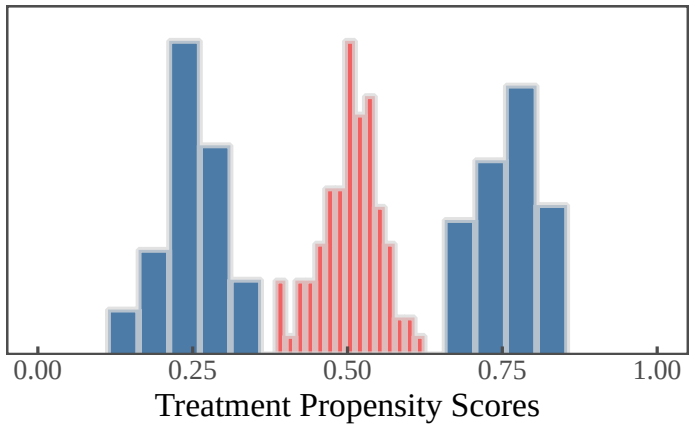
Mean Shift



Clusters of Extreme Propensities



Bimodality/Separation



Doubly Robust Machine Learning (DRML)

The DRML Framework

DRML treatment effect estimation works within the standard linear model framework. Specifically, it uses a partially linear model:

DRML

$$Y = \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon|W, X) = 0 \ \} \text{Outcome Model}$$
$$W = \gamma(X) + V, \quad \mathbb{E}(V|X) = 0 \ \} \text{Treatment Model}$$

- Y is our outcome variable.
- $W = \{0, 1\}$ denotes treatment, and θ is the treatment effect.
- The vector $X = (X_1, ..., X_p)$ is composed of all pre-treatment covariates, broadly construed.
- ϵ and V are random errors for each equation.

$\hookrightarrow$ We model both $\omega(X)$ (outcome $\sim$ covs) and $\gamma(X)$ (treatment $\sim$ covs) using some flavor of ML.

DRML Cont.

In this setup, $\omega(X)$ and $\gamma(X)$ are nuisance parameters. Only need good predictions!

In experimental contexts, DRML with good randomization turns into a fancy regression adjustment (with some distinct benefits).

Depending on the implementation, DRML employs cross-fitting or sample-splitting to guard against overfitting.

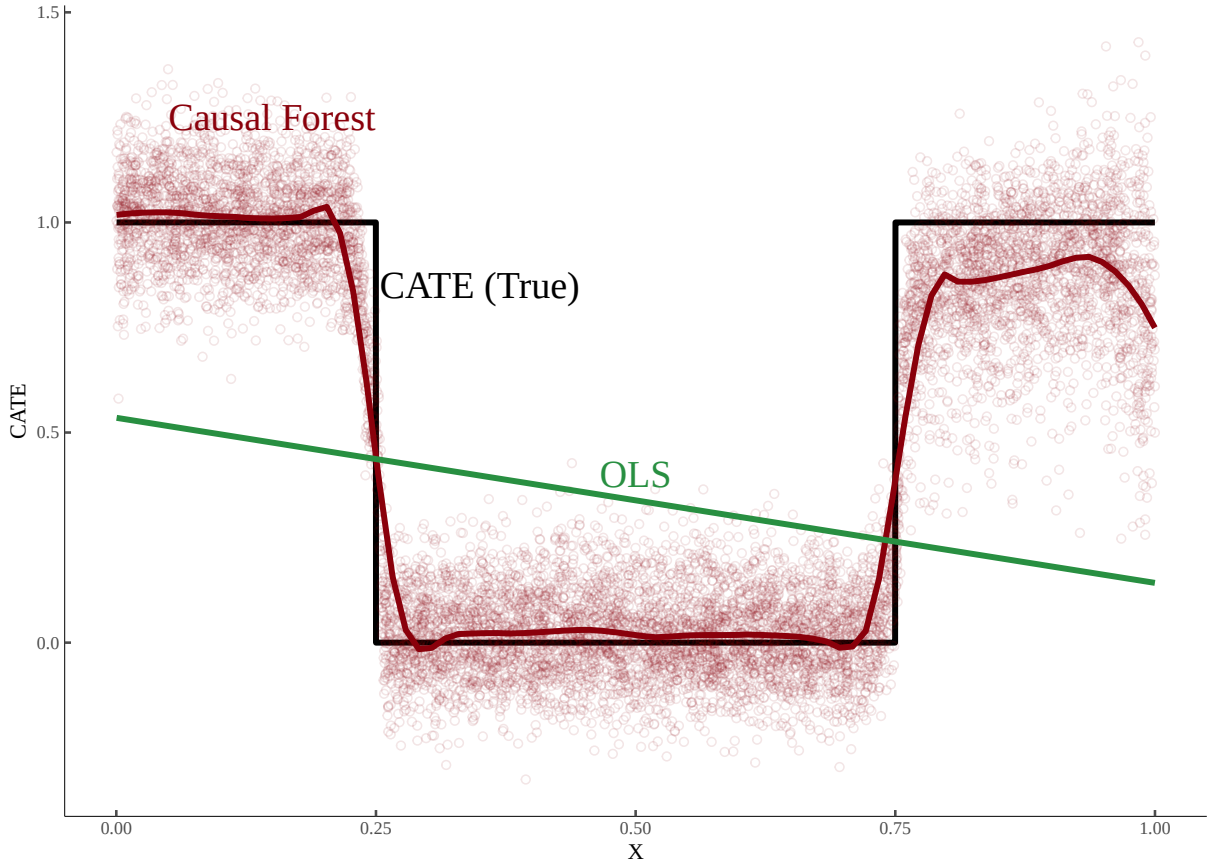
$\hookrightarrow$ We employ a variant of DRML where the models are honest random forests, also known as causal forests (Athey, Tibshirani, and Wager 2019).

Note: Some gains possible from other options but this model is fast, highly predictive, and has nice asymptotic guarantees.

What is Machine Learning and Why Should I Care?

Why Machine Learning?

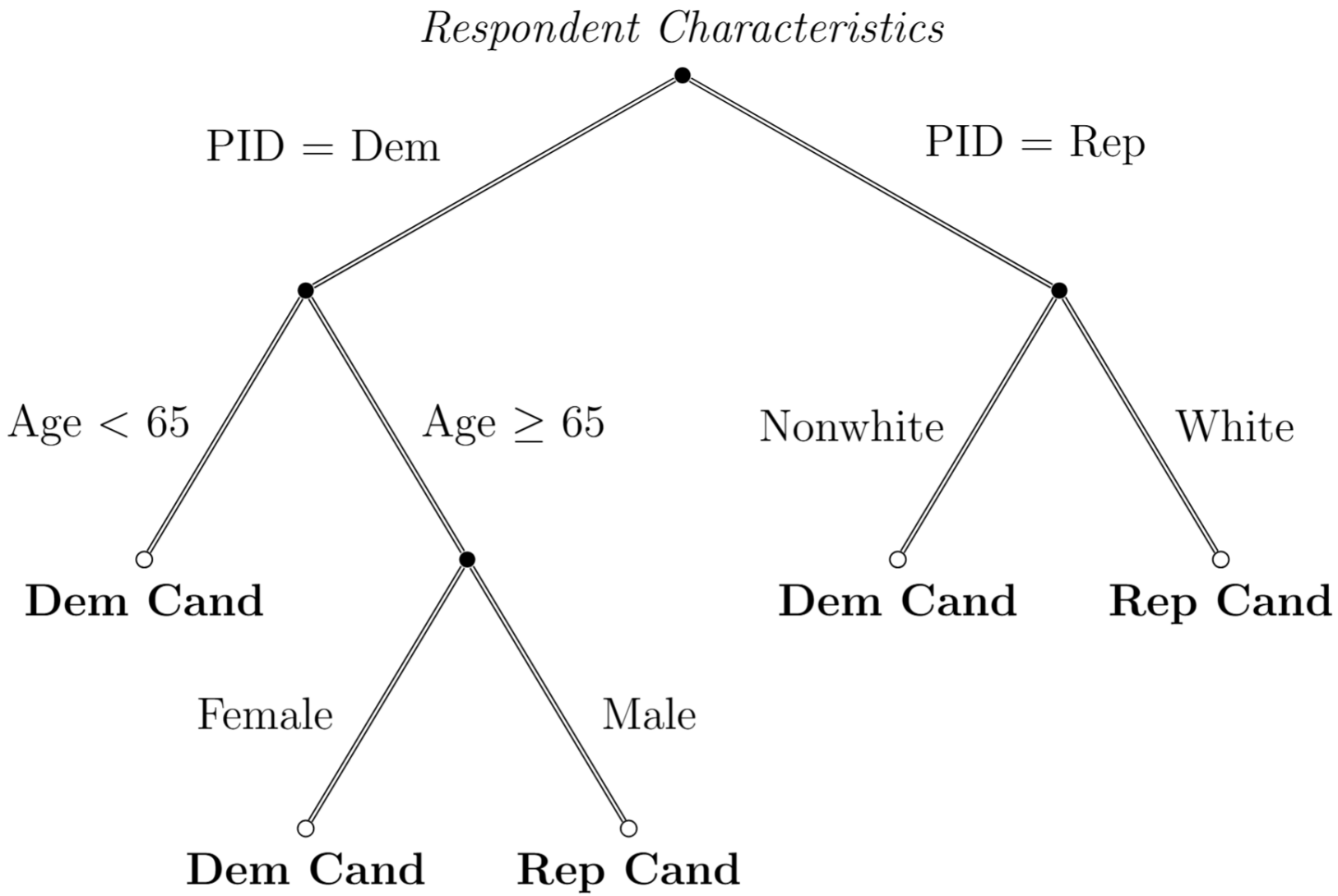
Looking back at our first figure, ML methods are **very** good at prediction.



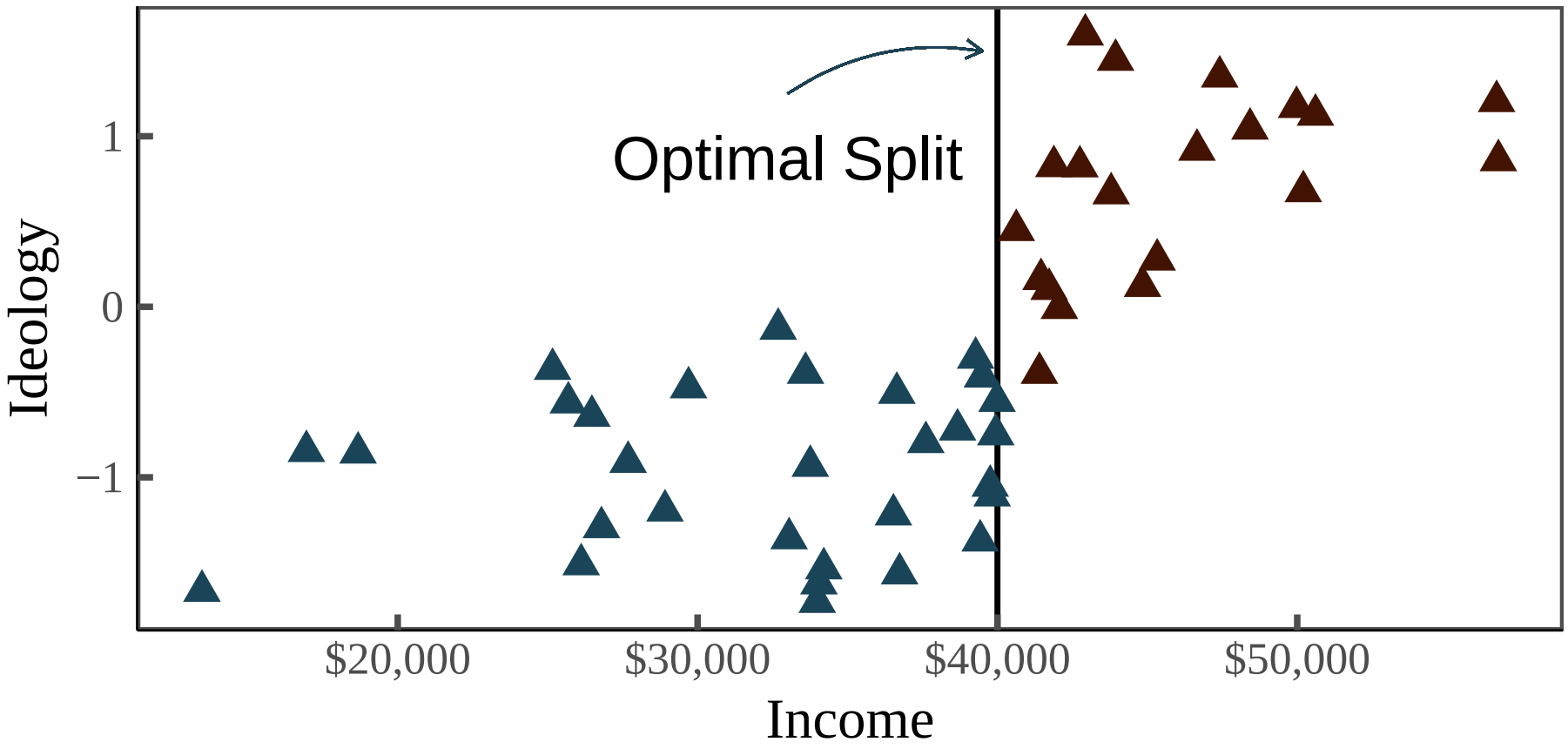
A Motivating Research Example

Predict vote-choice using standard public opinion data (PID, age, race, income, etc.)

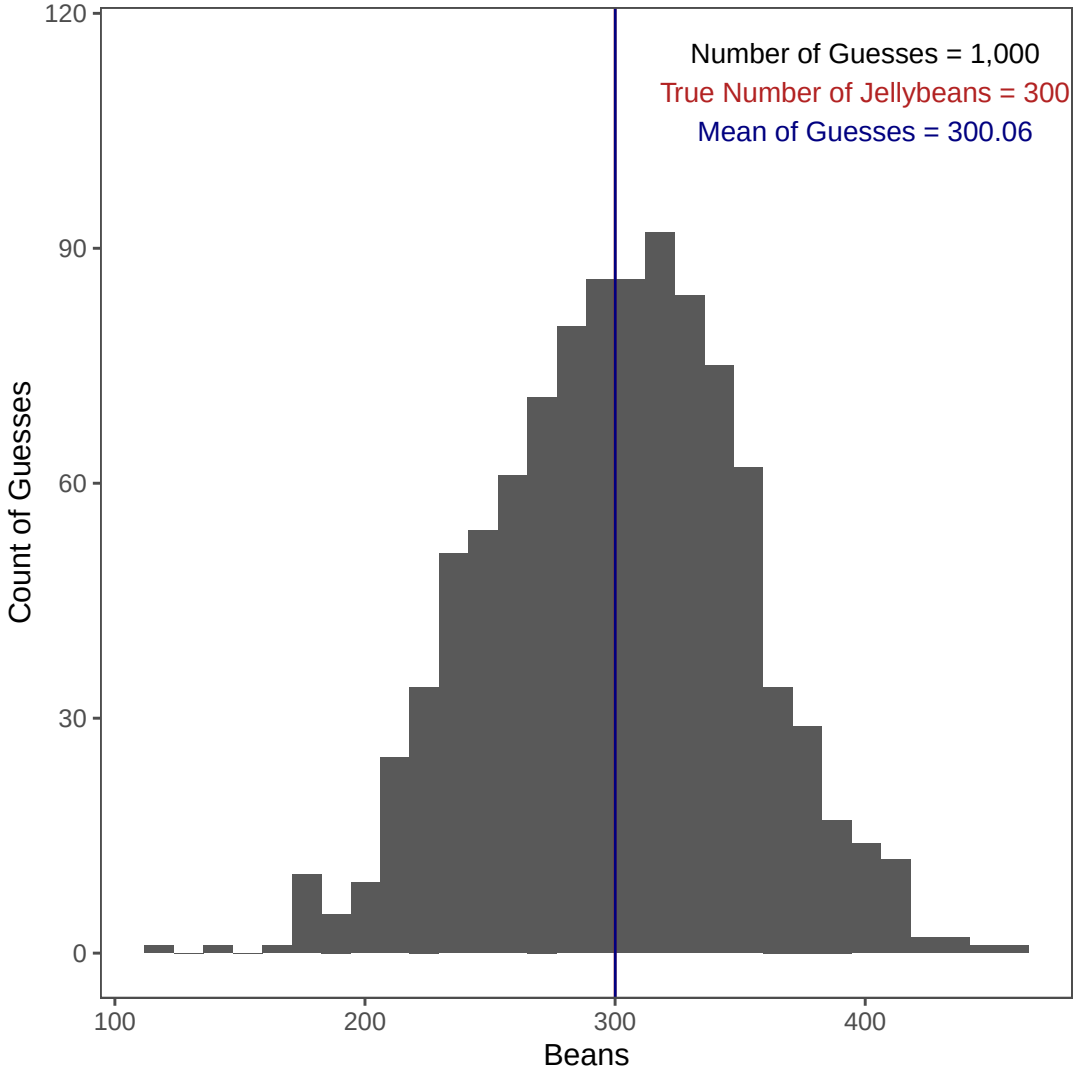
A Simple Decision Tree



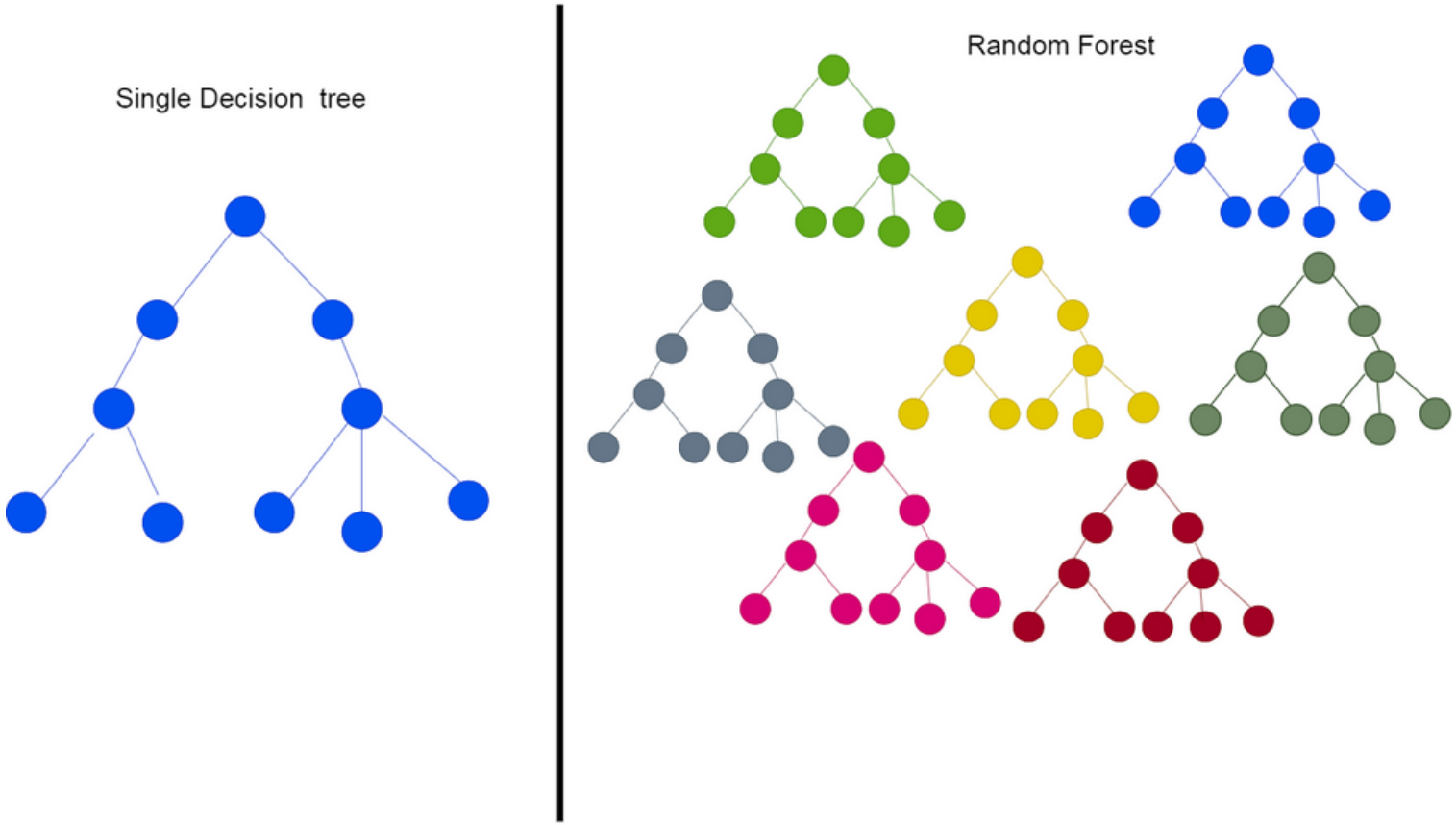
A Graphical Rendition of a Decision Tree



The Wisdom of the Crowds

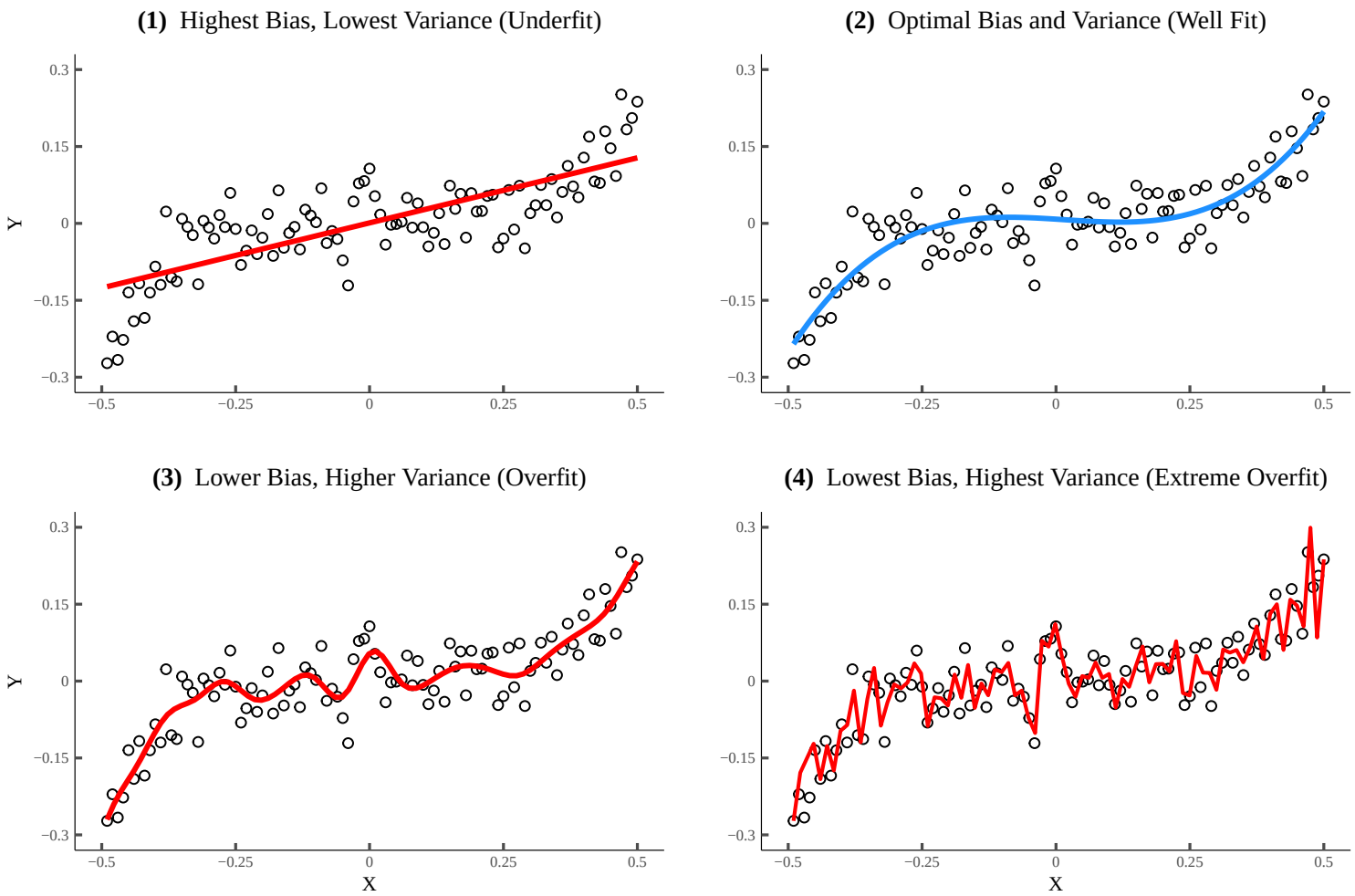


Random Forests



- Here we use a random sample of predictors chosen from the full set
  - This is the random component of random forests
- A bunch of "dumb" predictions are better than one "smart" prediction

Bias-Variance Tradeoff: We Don't Want to Overfit



- So ML often uses sample-splitting (testing and training sets)

Model Tuning and Cross-Validation

The model is trained on the training set and the test set is used to check the model's performance.

So we estimate a bunch of models using some number of "tuning" parameter sets and then choose the model that performs the best on the test set.

Common Random Forest tuning parameters include:

1. The number of variables any given tree can use
2. The minimum number of observations to be in the bottom of any given prediction
3. The maximum number of splits for any given tree

The Goal is Always the Same

Minimize error on the test set to create the most accurate, but generalizable model.

Causal Forests

Causal Forests adapt Random Forests to get known statistical properties (consistent estimates, Gaussian asymptotic sampling distribution, and estimable variance to build CIs).

To do this they:

1. Alter the tree-growing algorithm to ensure that trees have bins/leaves L that are small enough such that:
   1. Y_is where $i \in L(x)$ are independently distributed *and*
   2. Combinations of (Y_i, W_i)—where W represents treatment—are as if they come from a randomized experiment.
2. Make sure each tree is trained on a *subsample*, each split contains a minimum number of treatment and control units, and **honesty** is employed in each leaf.
   1. *Within* a tree there is a training/testing split: some portion (`honesty.fraction`) of the data is used to *estimate effects* (test; $\mathcal{I}$) and the other portion is used to determine splits (train; $\mathcal{J}$).
   2. Splits are made by maximizing the variance of the estimated treatment effect conditional on covariates $\hat{\tau}(X_i)$ for $i \in \mathcal{J}$ between splits.

So... What Does This Mean?

The long and short of it is this:

1. Causal forests have statistically known properties (like OLS) but are far more flexible and powerful at prediction.
   - The overall "box" may still be black, but we now know the properties of CF's ATEs and CATEs!
2. Because they are built to maximize the variance of the treatment effect as predicted by covariates (think demographics) they are phenomenal at:
   1. **Detecting** treatment effect heterogeneity.
   2. **Estimating** conditional average treatment effects.

- There are more assumptions made by CF to recover the statistical guarantees, but many of these are assumed by other models (unconfoundedness) or are directly testable (overlap in treatment propensities).
- Also/however, in cases of complete randomization these assumptions are essentially guaranteed. Hence the beauty of RCTs.

How to (and Why You Should) Implement DRML

DRML Benefits

1. Model specification doesn't fall from the sky!
2. High dimensional relations and interactions? No problem.
3. Nonlinear functional form? No problem.
4. Kills temptation for the Table 2 fallacy.
5. Sample splitting/cross-fitting good for prediction & inference.
   - Athey and Imbens (2016), Ratkovic and Tingley (2023), and Blackwell and Michael P Olson (2022a), etc.

$\hookrightarrow$ In our sims, DRML significantly improves precision relative to unadjusted, reg. adjusted, and the popular Lin estimator.

Understanding Your Data and Treatment Assignment

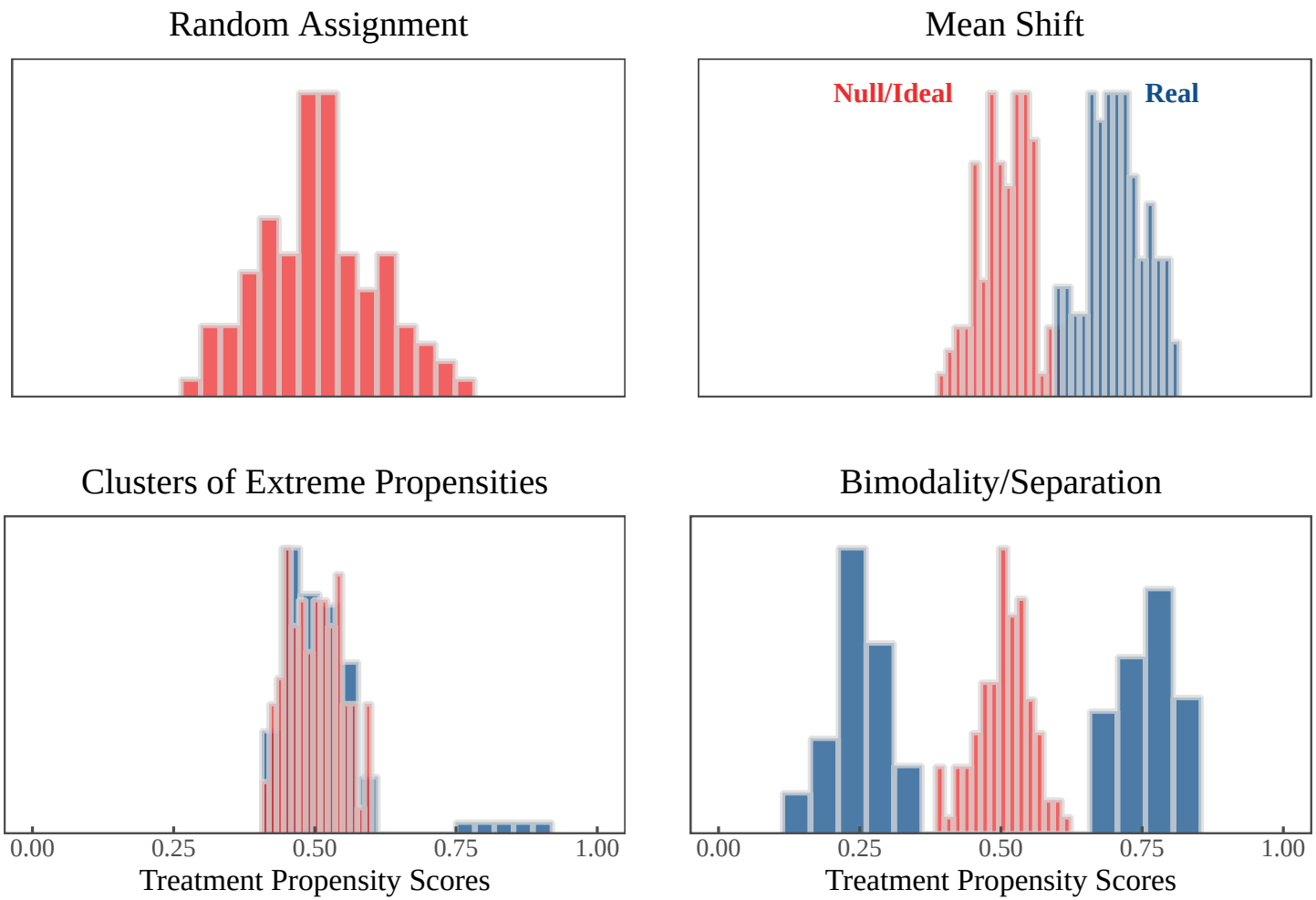
First, regardless of if your experiment is **observational** or **experimental**, you should estimate a treatment propensity model.

- We are very biased, but there's a nice little package called `MLBalance` (Rametta and Fuller 2024) that implements this.

In **observational contexts**, both a combination of theory (especially to argue that there are no unobserved confounders) and estimating a treatment propensity model are needed to argue for causality.

In **experimental contexts**, an explanation of the treatment assignment mechanism and a treatment propensity model is plenty sufficient.

Examples of Treatment Propensity Models



- You can also break these down by **treatment** and **control** groups. Especially useful for observational analyses.

DRML ATE Estimation: Reducing Noise & Bias

DRML

$$Y = \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon|W, X) = 0 \ \} \text{Outcome Model}$$
$$W = \gamma(X) + V, \quad \mathbb{E}(V|X) = 0 \ \} \text{Treatment Model}$$

Assume we *only* care about estimating ATEs and not CATEs.

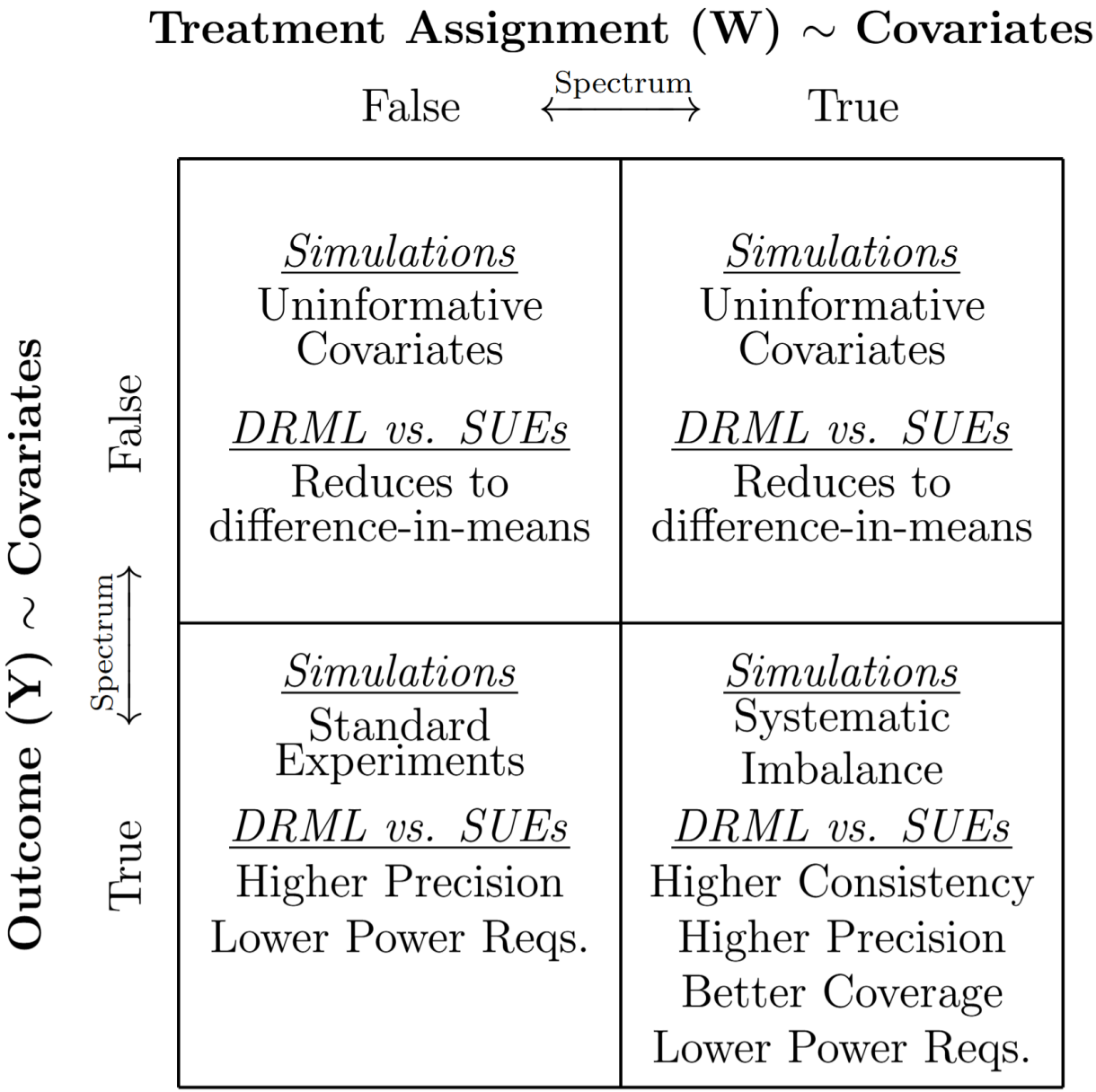
In **experimental contexts**:

- Then we still benefit significantly from reducing the noise introduced by covariates that inform baseline rates of our outcome.
- Increased precision and lower power requirements for smaller effects.

In **observational contexts**:

- We can significantly reduce bias from nonrandom assignment.
- This is similar to matching and other propensity score methods.
- The overlap estimator (Li, Morgan, and Zaslavsky 2018) can provide credible ATT estimates with significant nonrandom assignment.

ATE Estimation: DRML vs. SUEs



Conditional Effects

How should we estimate conditional effects (CATEs)?

---

1. Multiple interactions in a simple regression?
   - Could overfit, likely misspecified (Beiser-McGrath and Beiser-McGrath 2020). Need lots of power.
2. Separate single interaction regression models?
   - Could introduce omitted interaction bias! (Blackwell and Michael P. Olson 2022b)
3. Off-the-shelf machine learning methods (RF, boosting, etc.)?
   - Not designed for causal inference, likely overfit (Ratkovic and Tingley 2023).

DRML vs. Traditional: Testing for HTEs

Standard approaches tend to rely on preregistration and interactions in regressions (treatment $\times$ conditioning variable).

This approach, as we illustrate in our simulations, can lead to significant bias *and* missing important, unexpected heterogeneity.

Many tests for detecting heterogeneous treatment effects:

- Best linear fit test (Chernozhukov et al. 2022)
- The bound test (Athey and Wager 2019)
- Rank-weighted average treatment effect (RATE) tests (Wager 2024)

Check omnibus test, then proceed to CATEs IFF you pass HTE tests.
$\hookrightarrow$ You can use variable importance measures to identify conditioning variables.

Simulation Evidence

Simulation Overview

We run a lot of simulations...

For ATEs we run four sets of sims:

- Standard experimental context. RA + simple informative covariates
- High dimensional context. RA + complex informative covariates
- Systematic imbalance/confounding. No RA + Informative covariates
- Uninformative covariates. RA + Uninformative covariates

For CATEs we compare CF to saturated OLS in standard experimental context.

- By dimension (1 vs. 2)
- By linearity (linear, nonlinear)

$\hookrightarrow$ DRML can increase power and precision, shrink confidence intervals. Significant performance gains in common sample sizes.

Simulation Setup & Details

Competing estimators...

- CF-AIPW
- CF-Overlap Weighting
- Lin Estimator
- Unadjusted Difference in Means

Standard Experimental Context Results

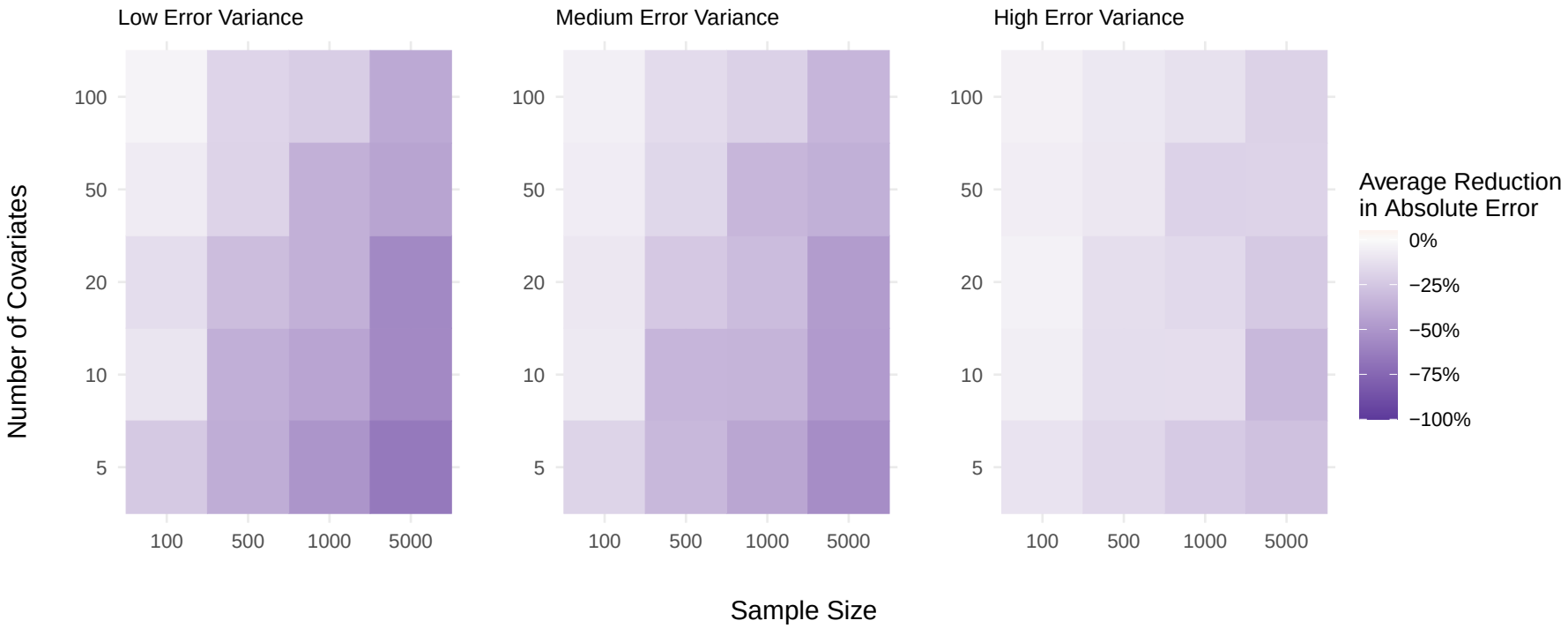
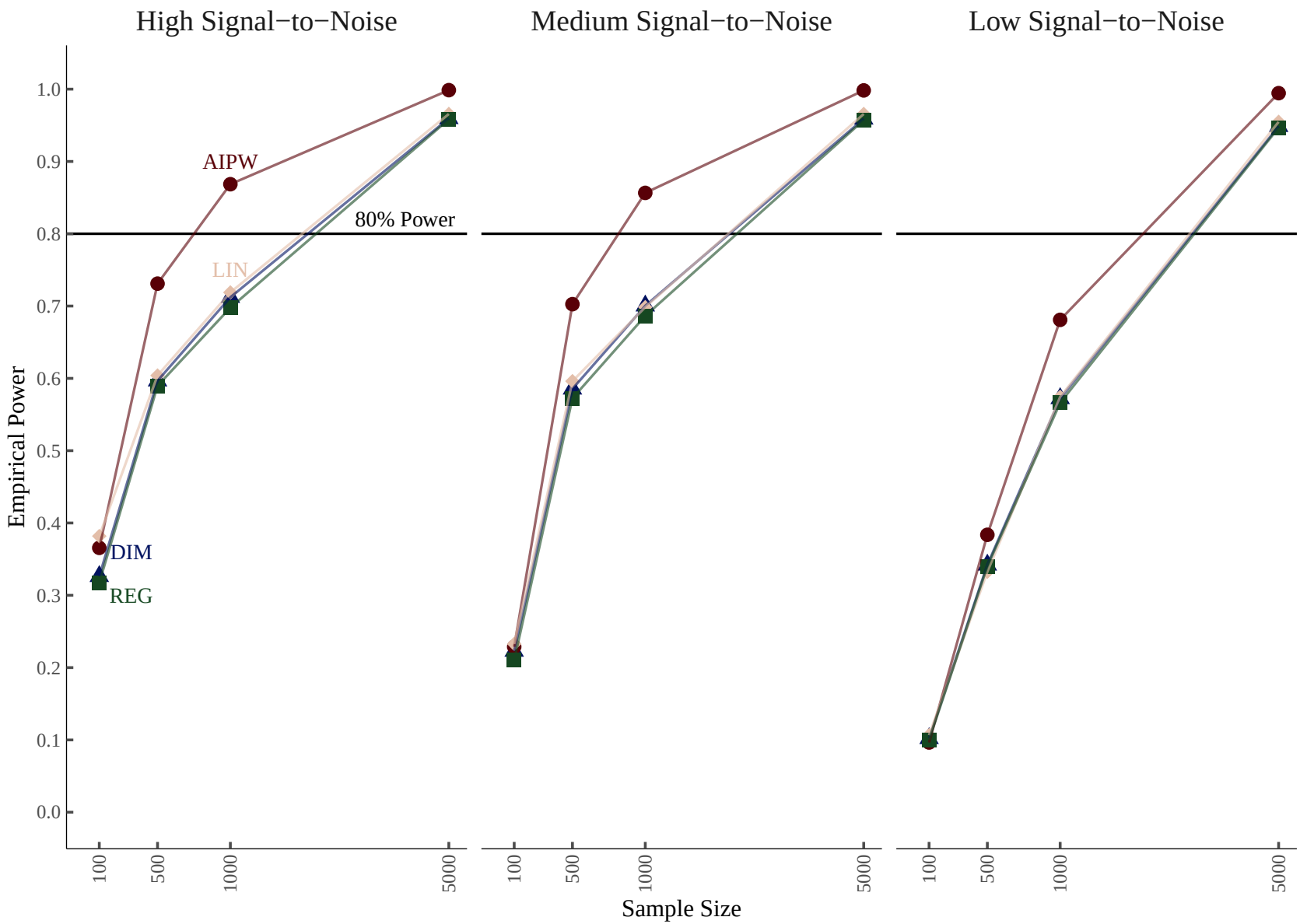


Figure: Average reduction in absolute error of the DRML treatment effect estimates versus unadjusted difference in means, regression-adjusted, and Lin-adjusted effects.

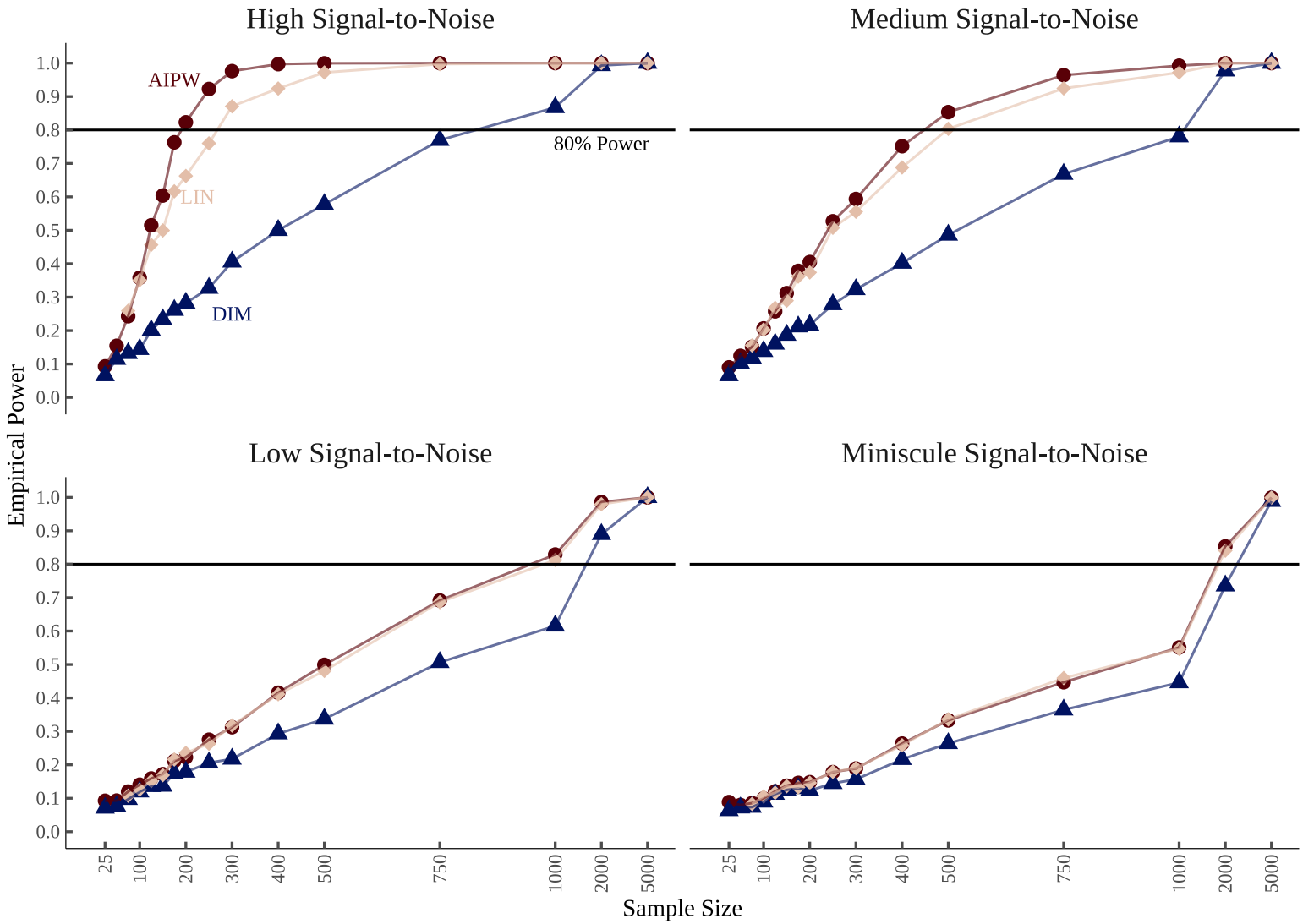
480 permutations x 2,000 seeds = 960,000 simulations covering many DGPs.

$\hookrightarrow$ (*untuned*) DRML wins in $\sim$ 99% **of cases.**

Power Results: Standard Experimental Context

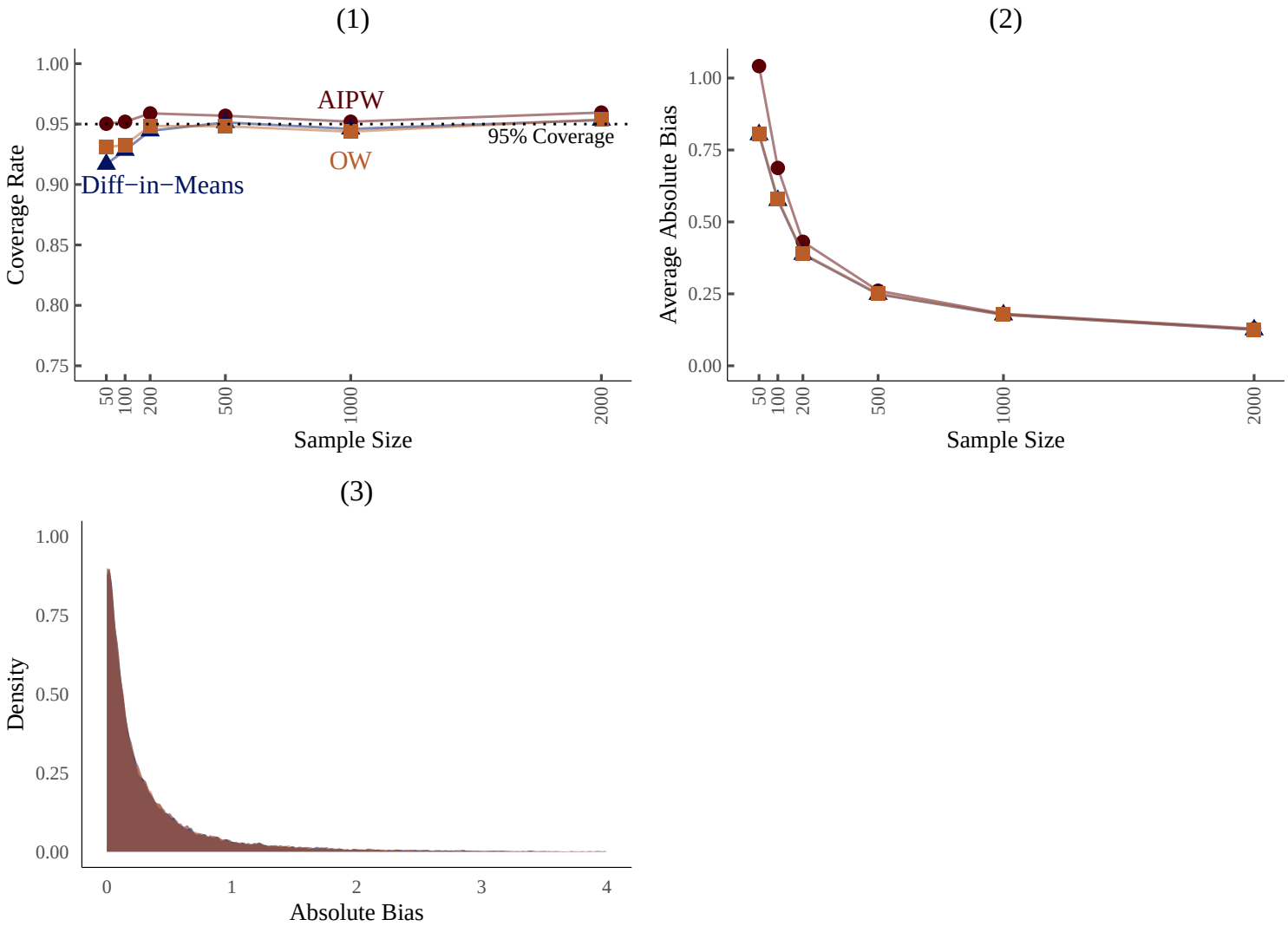


Power Results: High Dimensional Context



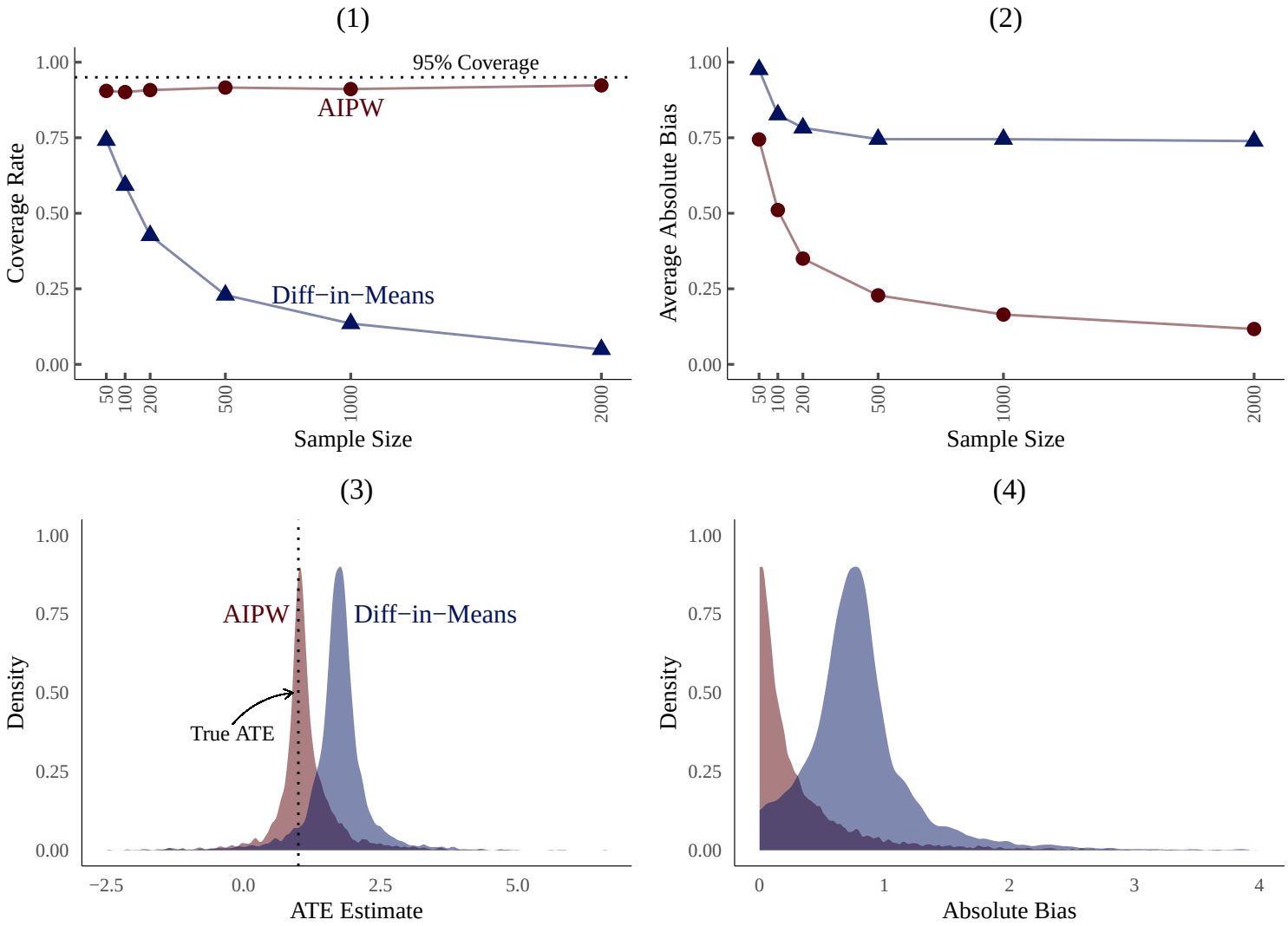
↪ DRML leaves no variance reduction on the table.

DRML with Uninformative Covariates



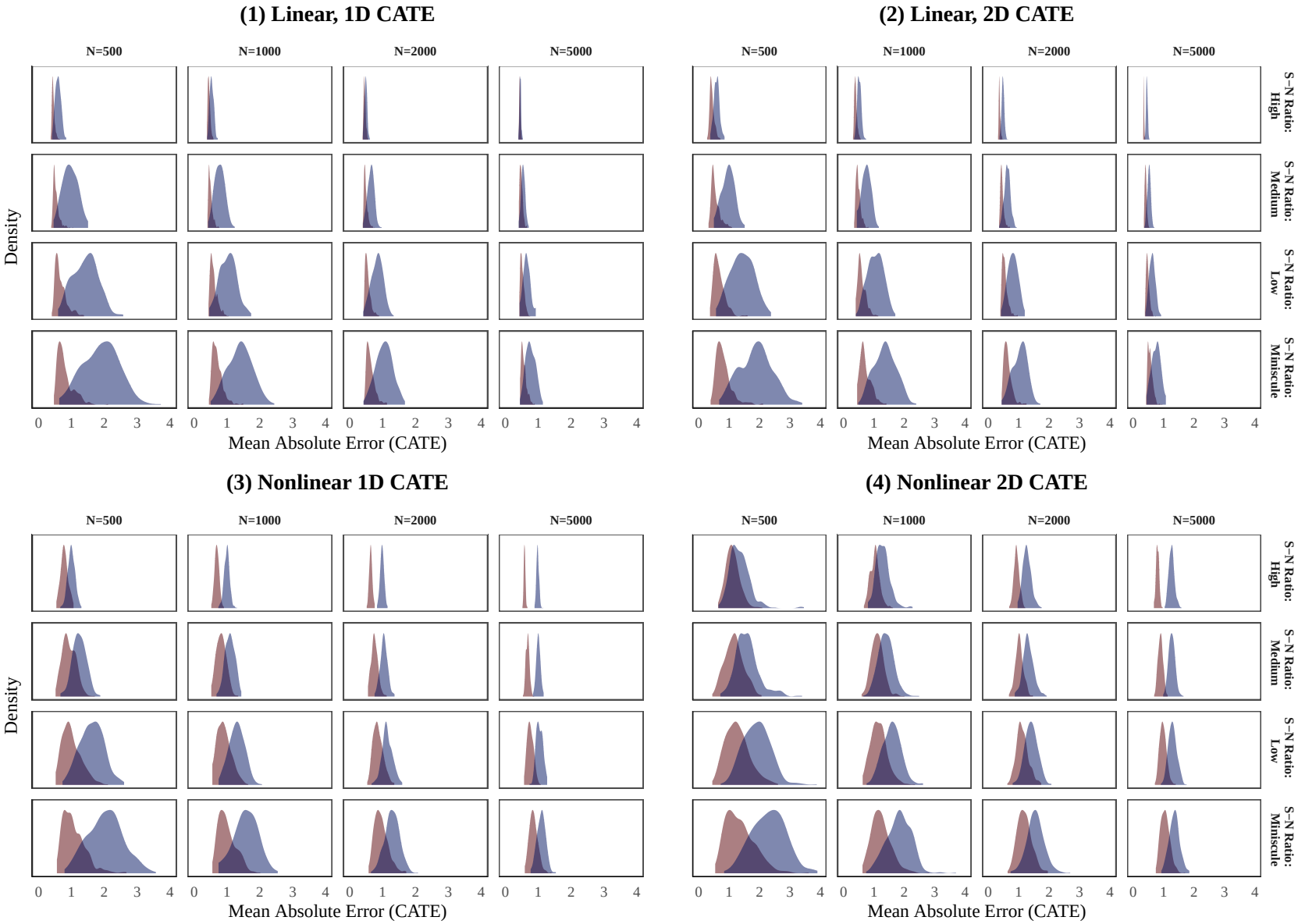
↪ Small increase in bias, DRML CIs improve small sample coverage.

DRML for Systematic Imbalance/Confounding



↪ Imbalance over confounder(s) biases diff-in-means, DRML corrects

CF vs. OLS for CATEs



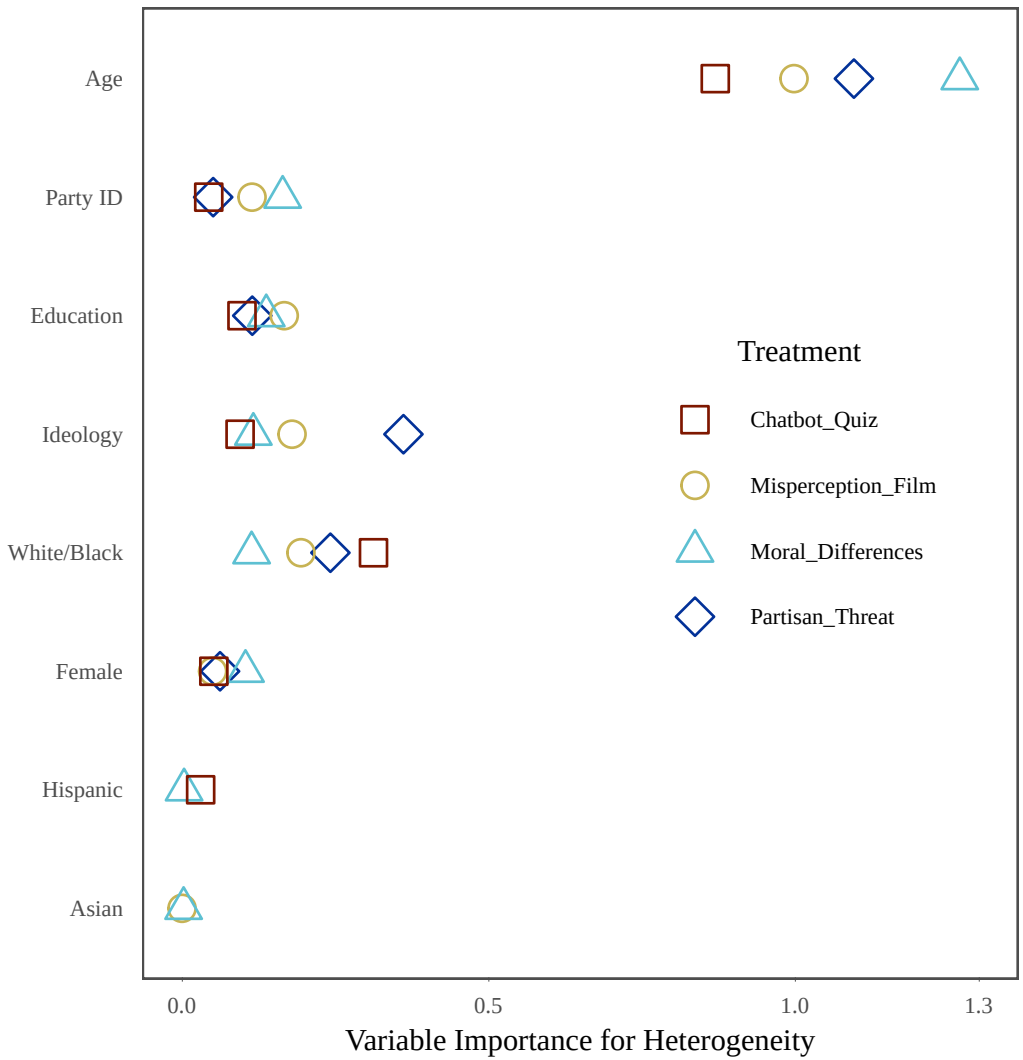
An Application to Voelkel et al. (2024)

Variable Importance: Detecting Heterogeneity

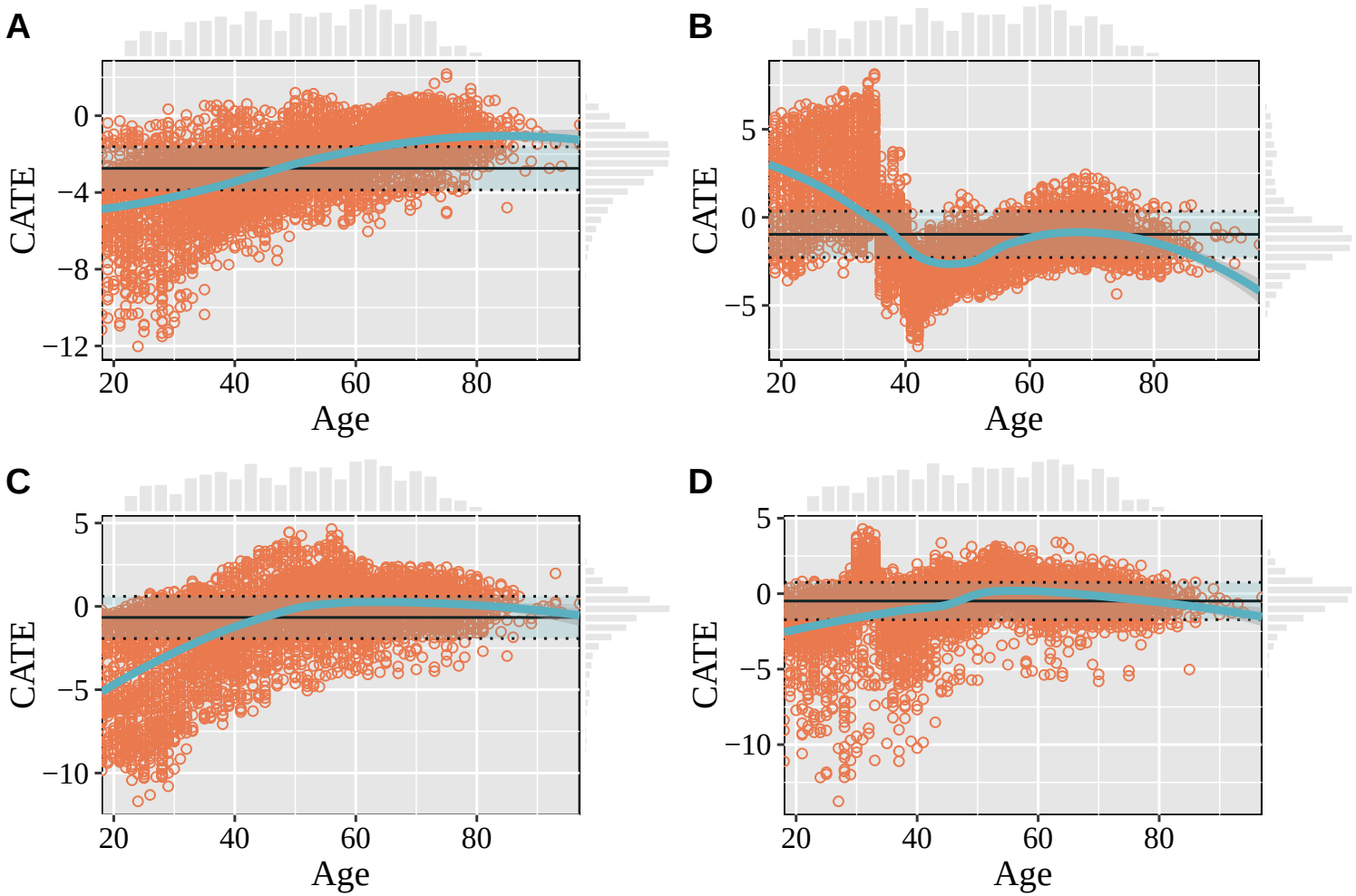
Voelkel et al. (2024) Details

- Megastudy of 30k respondents across 25 treatments
- Aimed at reducing partisan affect and related phenomena (e.g., support for political violence)
- Only 4 interventions were statistically significant for reducing SPV
- We identified 4 with heterogeneity (3 from null ATE estimates)

Our VI Results →



The Conditional Effect of Age on Treatments



A) Misperception Film, B) Befriending Meditation,
C) Chatbot Quiz, and D) Partisan Threat.

Conclusion

What We (Hopefully) Covered

What Causal Machine Learning is and how it works in the Doubly Robust Machine Learning Framework.

- Also a *decent* understanding of trees and forests, especially causal forests.

Why CML is useful.

- For both experimental and observational research.
- A bunch of simulations and an application to Voelkel et al. (2024).

A basic understanding of the process for incorporating CML/DRML into your own research.

Obviously we didn't have time to go into detail on all of these things, but please read our working paper for more information!

Future Work (to Keep an Eye Out For)

1. ML for Systematic Attrition Detection
2. Unifying Unsupervised and Supervised ML
3. Implications of ML for Experimental Design
4. A Meta-Reanalysis of CATEs in Experimental Political Science

Shameless Plug: ***Causal Machine Learning for Observational and Experimental Research***

If any of this is interesting to you and you would like to learn more, my coauthor Jack T. Rametta (UC Davis) and I are teaching an **online, 1-week ICPSR Summer Program Workshop** on this topic.

Thank You! I Look Forward to Your Feedback.

Appendix

Unifying Unsupervised and Supervised ML

While supervised ML (e.g., random forests, neural networks) has grown in popularity in recent years, the same is not exactly true for unsupervised machine learning (AKA scaling or dimensional analysis).

Although ML-powered estimators are certainly the most powerful and flexible option available to researchers to estimate robust treatment effects, it still has its limitations: namely multicolinearity.

Measurement and dimensional reduction, informed by theory and supported by exploration, can help transcend these issues and lead to the best research possible.

Implications of ML for Experimental Design: Transparency Over Simplicity

Potential Question #1

How do all of these methods relate to existing experimental standards and practices, such as pre-registration?

Potential (Skeptical) Question #2

Doesn't ML increase the likelihood of p-hacking and specification searching? Aren't these models black boxes?

We aim to address these concerns and discuss in detail how machine learning should be incorporated into existing ideas of experimental design, including pre-registration.

A Meta-Reanalysis of CATEs in Political Science

We've also begun a reanalysis of CATEs estimated using standard methods, instead using causal forest models.

We're not alone in our skepticism of subgroup analyses using simple models (Hainmueller, Mummolo, and Xu 2019; Ratkovic 2021; Blackwell and Michael P Olson 2022a).

Our initial results suggest that a significant portion of simple CATEs are not identified when using CF and DoubleML...

For example, we reanalyzed a very popular cueing experiment from recent years: Barber and Pope (2019).

$\hookrightarrow$ While we don't have any pretty figures to illustrate our results, suffice it to say that we do not find any evidence of treatment effect heterogeneity and the estimated CATEs are nowhere near the same as those reported in the paper.