# Low-Dimensional Representations v. Fully Nonparametric Estimation

## Causal Trees

- Move the goalpost so that we are estimating a small number of parameters, but get guaranteed coverage with many observations per parameter and by using sample splitting
- Easy to interpret, easy to mis-interpret
- Can be many trees
- Leaves differ in many ways if covariates correlated; describe leaves by means in all covariates

## Causal Forests

- Attempt to estimate $\tau(x)$
- Can estimate partial effects
- In high dimensions, still can have omitted variable issues
- Confidence intervals lose coverage in high dimensions (bias)
- Can be used as an input into estimation of lower dimensional objects, such as average effects for high-CATE individuals or average benefits of a targeted treatment assignment

# Baseline method: $k$-NN matching

Consider the $k$-**NN matching** estimator for $\tau(x)$:

$$\hat{\tau}(x) = \frac{1}{k} \sum_{\mathcal{S}_1(x)} Y_i - \frac{1}{k} \sum_{\mathcal{S}_0(x)} Y_i$$

where $\mathcal{S}_{0/1}(x)$ is the set of $k$-nearest cases/controls to $x$. This is consistent given **unconfoundedness** and regularity conditions.

- **Pro**: Transparent asymptotics and good, robust performance when $p$ is small.
- **Con**: Acute curse of dimensionality, even when $p = 20$ and $n = 20k$.

**NB**: Kernels have similar qualitative issues as $k$-NN.

# Adaptive nearest neighbor matching

**Random forests** are a a popular heuristic for adaptive nearest neighbors estimation introduced by Breiman (2001).

- **Pro**: Excellent empirical track record
- **Con**: Often used as a black box, without statistical discussion

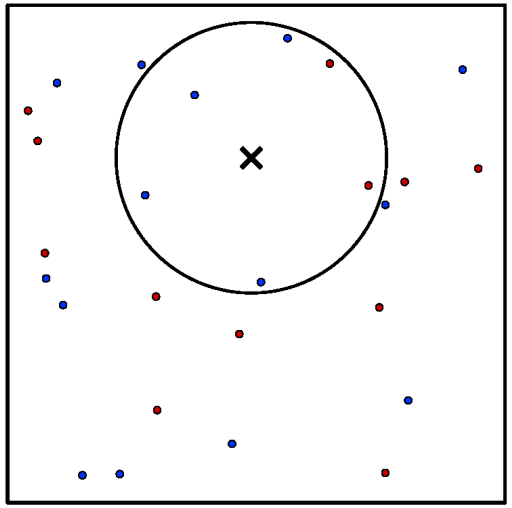
There has been considerable interest in using forest-like methods for treatment effect estimation, but without formal theory.

- Green and Kern (2012) and Hill (2011) have considered using **Bayesian forest algorithms** (BART, Chipman et al., 2010)
- Several authors have also studied related **tree-based methods**: Athey and Imbens (2016), Su et al. (2009), Taddy et al. (2014), Wang and Rudin (2015), Zeilis et al. (2008)

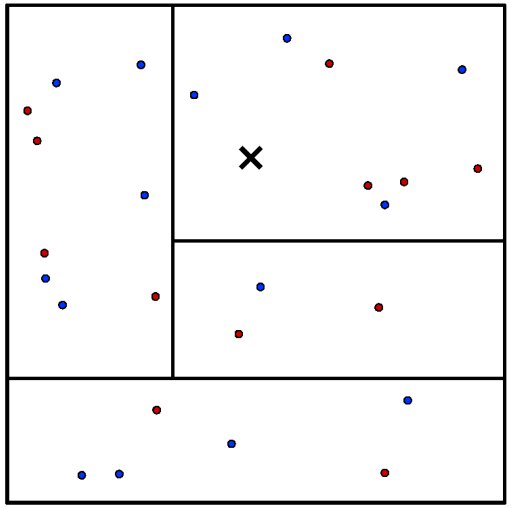
Wager and Athey (2018) provide the first formal results allowing random forest to be used for provably valid **asymptotic inference**

# Making $k$-NN matching adaptive

Athey and Imbens (2016) introduce the **causal tree**: defines neighborhoods for matching based on **recursive partitioning** (Breiman, Friedman, Olshen, and Stone, 1984), advocate sample splitting (w/ modified splitting rule) to get assumption-free confidence intervals for treatment effects in each leaf.



Tree-based neighborhood



Tree-based neighborhood

# From trees to random forests (Breiman, 2001)

Suppose we have a training set $\{(X_i,\ Y_i,\ W_i)\}_{i=1}^n$, a test point $x$, and a tree predictor

$$\hat{\tau}(x) = T(x;\ \{(X_i,\ Y_i,\ W_i)\}_{i=1}^n)$$

**Random forest idea**: build and average many different trees $T^*$:

$$\hat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b^*(x; \{(X_i, Y_i, W_i)\}_{i=1}^n)$$

# From trees to random forests (Breiman, 2001)

Suppose we have a training set $\{(X_i,\ Y_i,\ W_i)\}_{i=1}^n$, a test point $x$, and a tree predictor

$$\hat{\tau}(x) = T(x; \{X_i,\ Y_i,\ W_i\}_{i=1}^n)$$

**Random forest idea**: build and average many different trees $T^*$:

$$\hat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b^*(x; \{(X_i, Y_i, W_i)\}_{i=1}^n)$$

We turn $T$ into $T^*$ by: - Bagging / subsampling the training set (Breiman, 1996); this helps smooth over discontinuities (Bühlmann and Yu, 2002). - Selecting the splitting variable at each step from $m$ out of $p$ randomly drawn features (Amit and Geman, 1997).

# Statistical inference with regression forests

**Honest trees** do not use the same data to select partition (splits) and make predictions.
**Theorem.** (Wager and Athey, 2018) Regression forests are asymptotically **Gaussian and centered**,

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathbb{N}(0, 1), \quad \sigma_n^2(x) \to_p 0$$

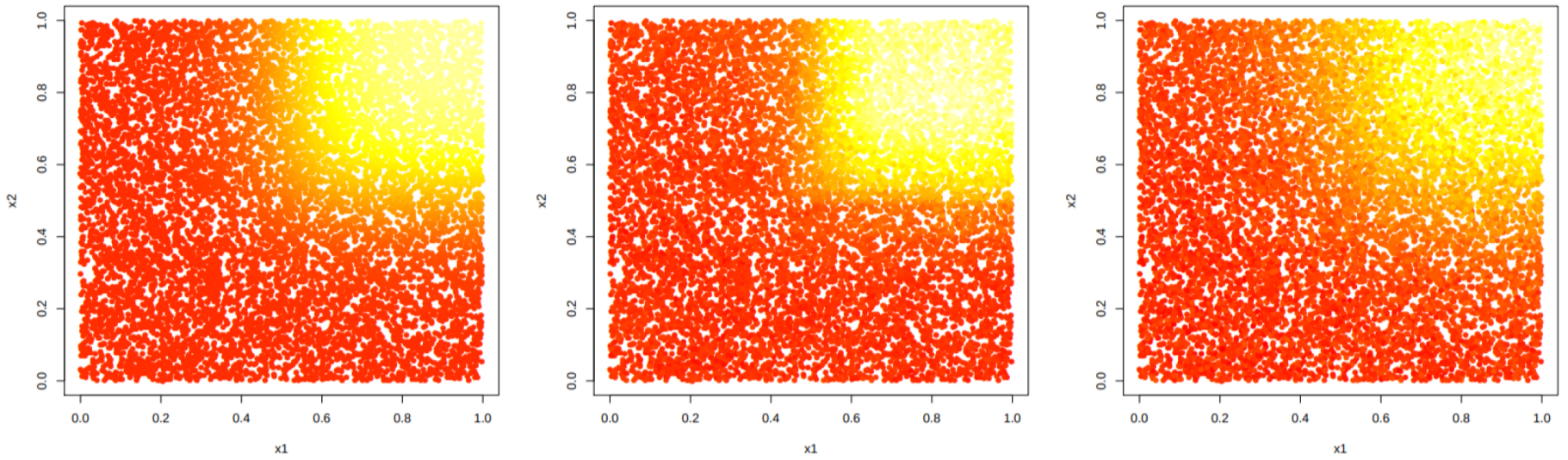
given the following assumptions (+ technical conditions):

1. **Honesty.** Individual trees are honest
2. **Subsampling.** Individual trees built on random subsamples of size $s \asymp n^\beta$, where $\beta_{\text{min}} < \beta < 1$
3. **Continuous features.** Features, $X_i$, have a density that is bounded away from 0 and $\infty$.
4. **Lipschitz response.** Conditional mean function $\mu(x) = \mathbb{E}[Y \mid X = x]$ is Lipschitz continuous.

# Causal forest example

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 6$; treatment assignment is random. All **the signal is concentrated along two features**. Plots depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.

**true effect $\tau(x)$** **causal forest** **$k$-NN estimate**

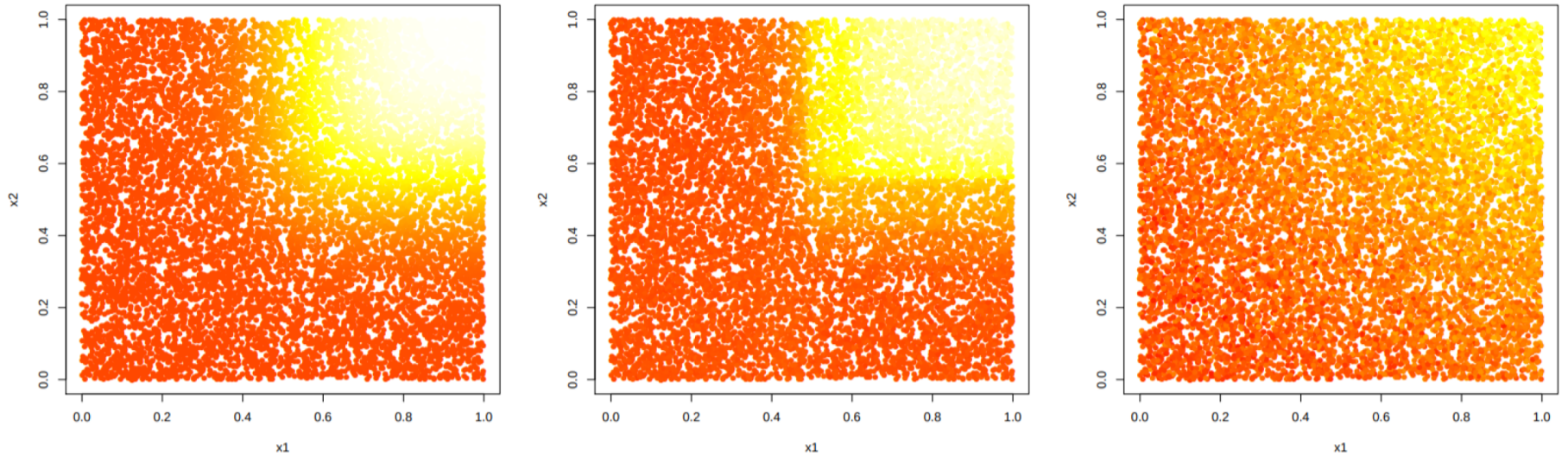



Software: `causalTree` for `R` (Athey, Kong, and Wager, 2015) available on `GitHub: susanathey/causalTree`

# Causal forest example

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 20$; treatment assignment is random. All **the signal is concentrated along two features**. Plots depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.

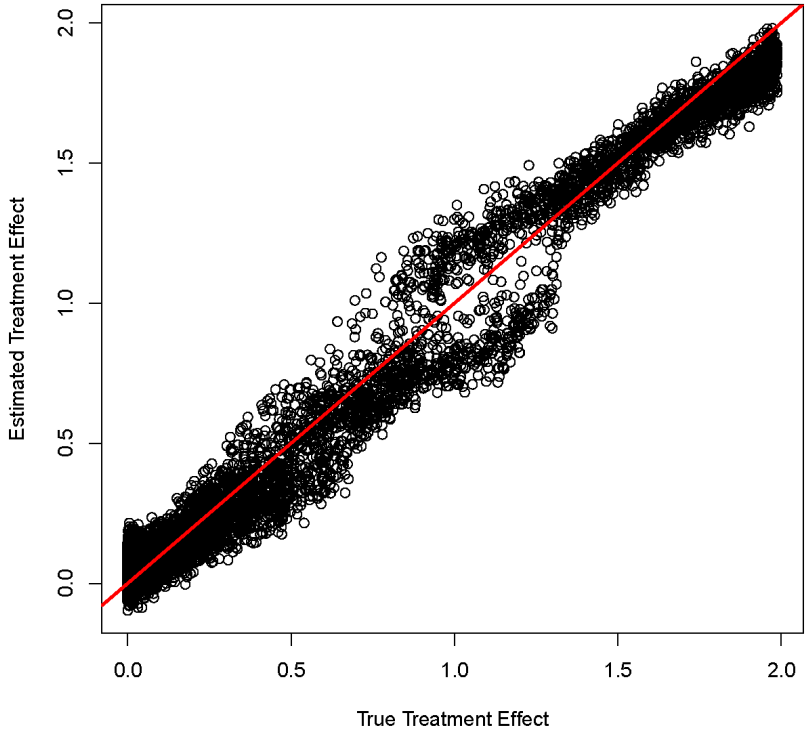
**true effect $\tau(x)$** **causal forest** **$k$-NN estimate**



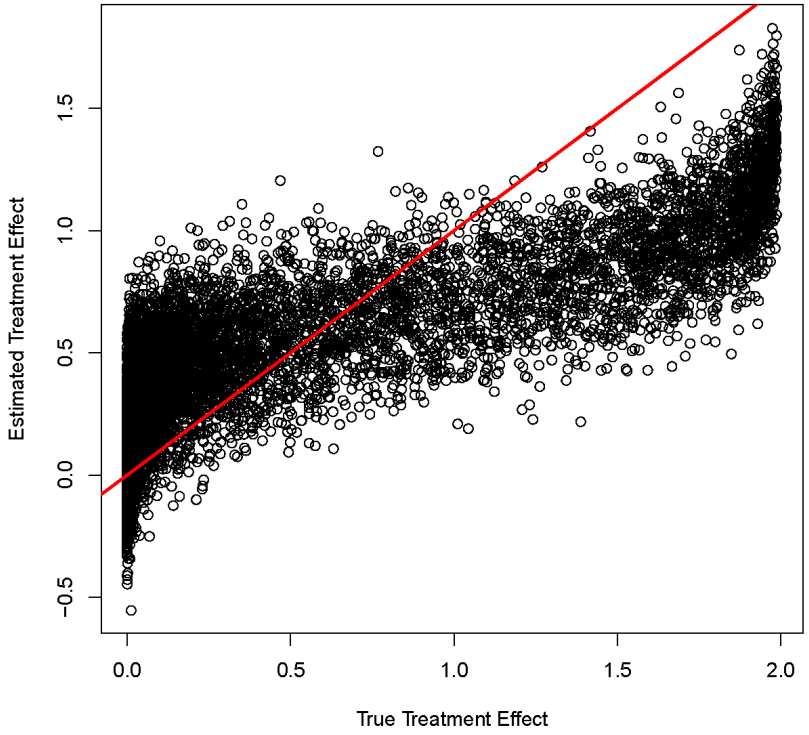

Software: `causalTree` for `R` (Athey, Kong, and Wager, 2015) available on `GitHub: susanathey/causalTree`

# Causal forest example

Causal forest dominates $k$-NN for both bias and variance. With $p = 20$, relative mean-squared error (MSE) for $\tau$ is $\frac{\text{MSE for } k\text{-NN (tuned on test set)}}{\text{MSE for forest (heuristically tuned)}} = 19.2$




causal forest




$k$-NN estimate

# Applications in Economics and Marketing

- Hitsch and Misra (2017): Use causal forests to target catalog mailings. Causal forest detects significant heterogeneity, performs better than alternatives including LASSO and off-the-shelf random forest
- Davis and Heller (2017): Analyze heterogeneous impacts of summer jobs using causal forest
- Athey, Kelleher, and Spiess (2021): Applications for financial aid

# Application: General Social Survey

The General Social Survey is an extensive survey, collected since 1972, that seeks to measure demographics, political views, social attitudes, etc. of the U.S. population.

Of particular interest to us is a **randomized experiment**, for which we have data between 1986 and 2010.

- **Question A**: Are we spending too much, too little, or about the right amount on **welfare**?
- **Question B**: Are we spending too much, too little, or about the right amount on **assistance to the poor**?

**Treatment effect:** how much less likely are people to answer **too much** to question B than to question A.

- We want to understand how the treatment effect depends on **covariates**: political views, income, age, hours worked, ...

**NB**: This dataset has also been analyzed by Green and Kern (2012) using Bayesian additive regression trees (Chipman, George, and McCulloch, 2010)

# Application: General Social Survey

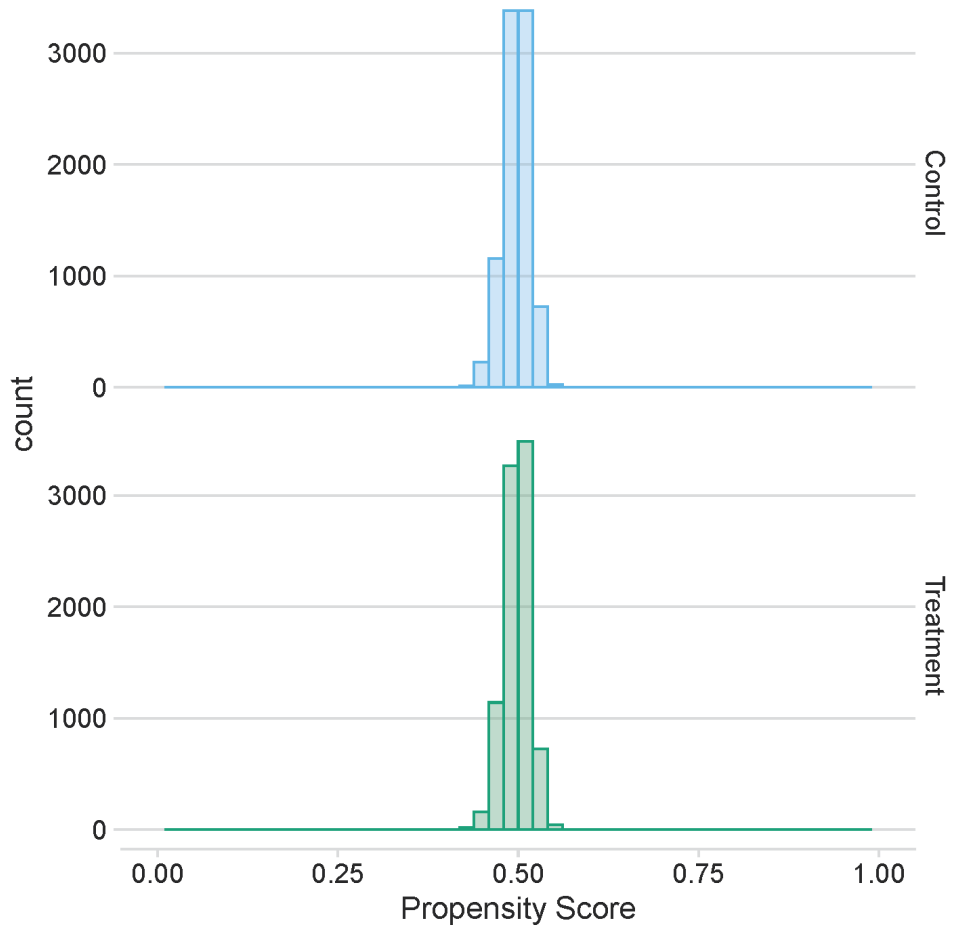
```
## Propensity model
W.hat.mod <- grf::regression_forest(X = as.matrix(select(train_df, -Y, -W))
                                    , Y = train_df$W
                                    , num.trees = 200
                                    , ci.group.size = 1
                                    , tune.parameters = "all")

W.hat.rf <- W.hat.mod$predictions
## Outcome model
Y.hat.mod <- grf::regression_forest(X = as.matrix(select(train_df, -Y, -W))
                                    , Y = train_df$Y
                                    , num.trees = 200
                                    , ci.group.size = 1
                                    , tune.parameters = "all")
Y.hat.rf <- Y.hat.mod$predictions
```

# Application: General Social Survey

Implement causal forest using `grf`.

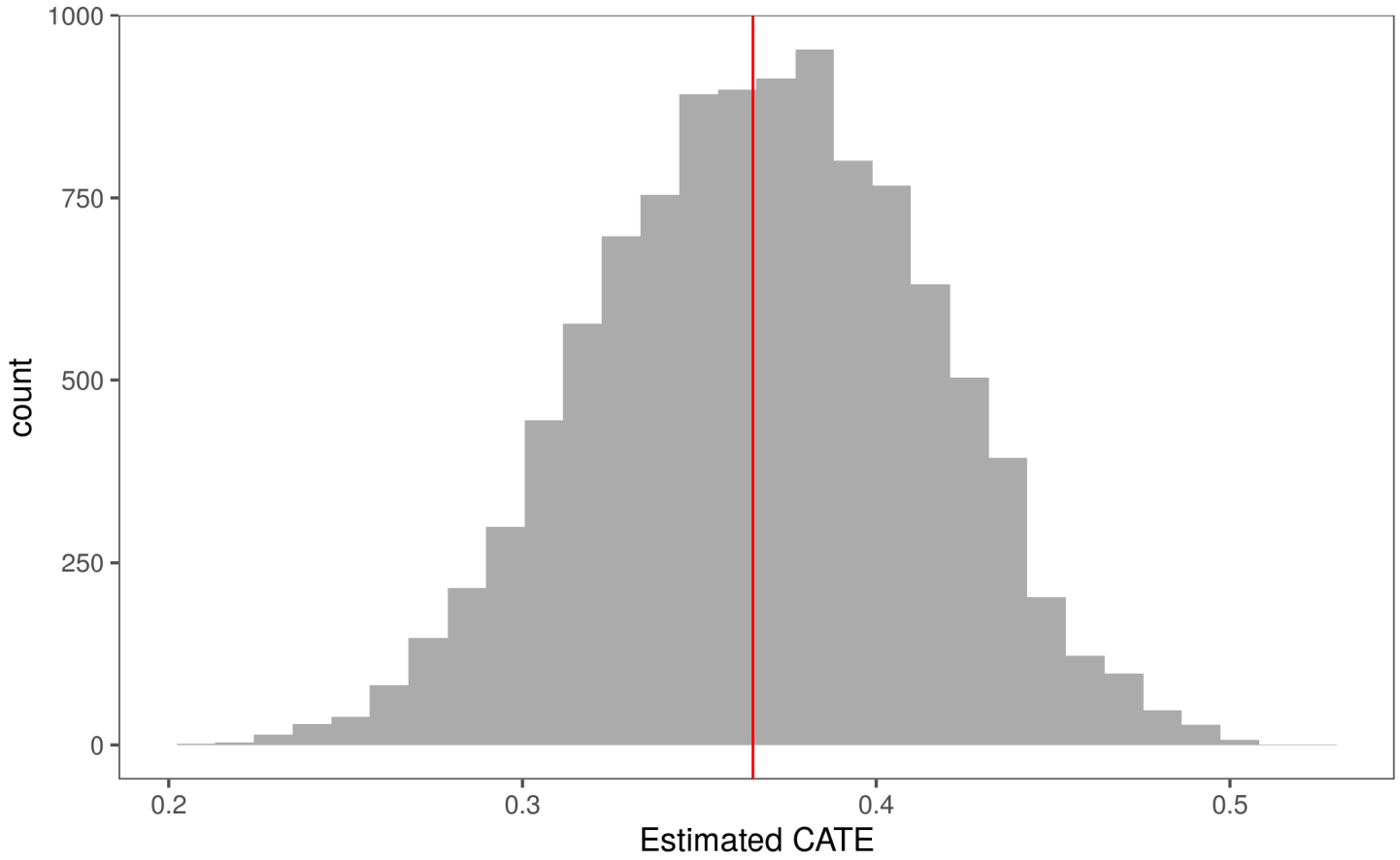
```
cf <- grf::causal_forest(
  X = as.matrix(select(train_df, -Y, -W)),
  Y = train_df$Y,
  W = train_df$W,
  Y.hat = Y.hat.rf,
  W.hat = W.hat.rf,
  num.trees=200)
```

# Verifying Randomization/Balance/Overlap



# Out-of-bag Conditional ATE

```
oob_cf_pred <- predict(cf, estimate.variance = TRUE)
train_df_predicted$cate_orthog <- oob_cf_pred$predictions
```



# Quantifying Heterogenity

Common in prediction models to assess calibration. One approach is to estimate the slope of the relationship between actual and predicted CATE. Slope of 1 is well calibrated; slope above zero is evidence of heterogeneity.

1. Best Linear Predictor (Chernozhukov, Demier, Duflo, and Fernandez-Val, 2018)

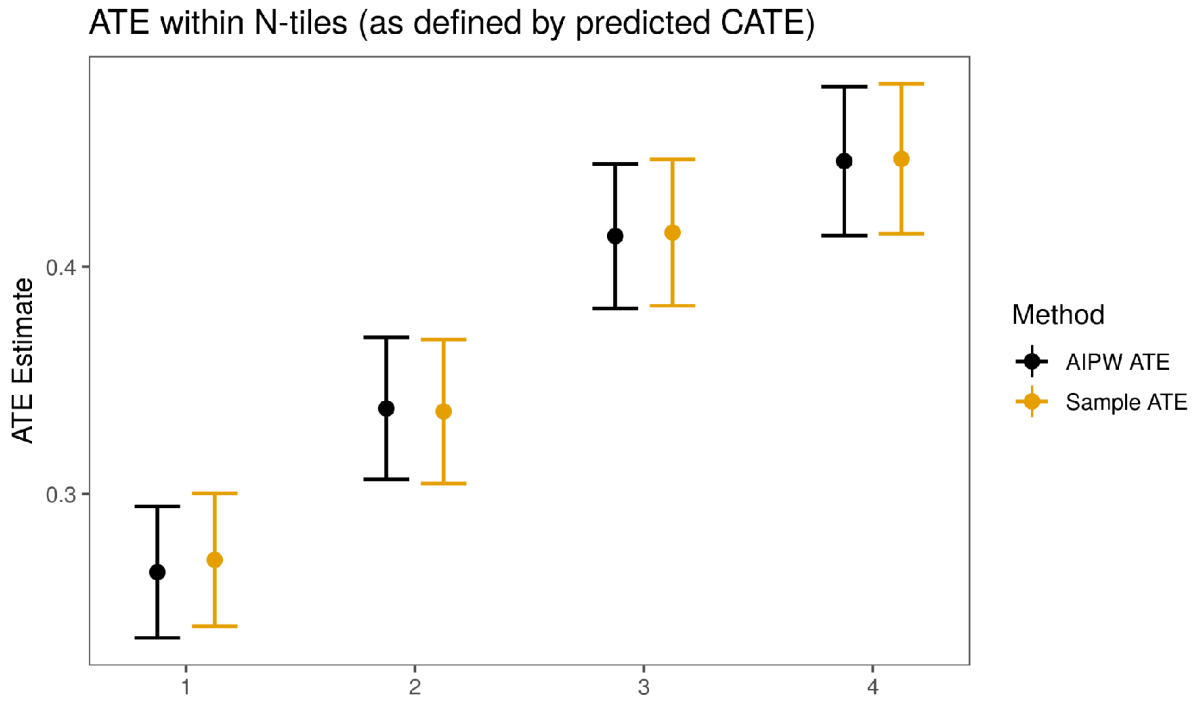
```
test_calib_orthog <- grf::test_calibration(cf)
```

```
Best linear fit using forest predictions (on held-out data)
as well as the mean forest prediction as regressors, along
with one-sided heteroskedasticity-robust (HC3) SEs:

                              Estimate Std. Error t value    Pr(>t)
mean.forest.prediction        0.995229   0.021511 46.2670 < 2.2e-16 ***
differential.forest.prediction 1.579928   0.164924  9.5797 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Quantifying Heterogenity

2. Quantile Plots - Compare ATE by quantile of sorted CATE

Hold out fold $k$ of the data. On the remaining data estimate CATE and define a mapping $\hat{Q}_{-k}$ from $x$ to, e.g., quartiles of the CATE distribution. Apply $Q_{-k}$ to the covariate values $x$ in fold k. Repeat this for all folds; at that point, each observation has been assigned to one of the four quartiles. Finally, estimate the ATE by quartile.



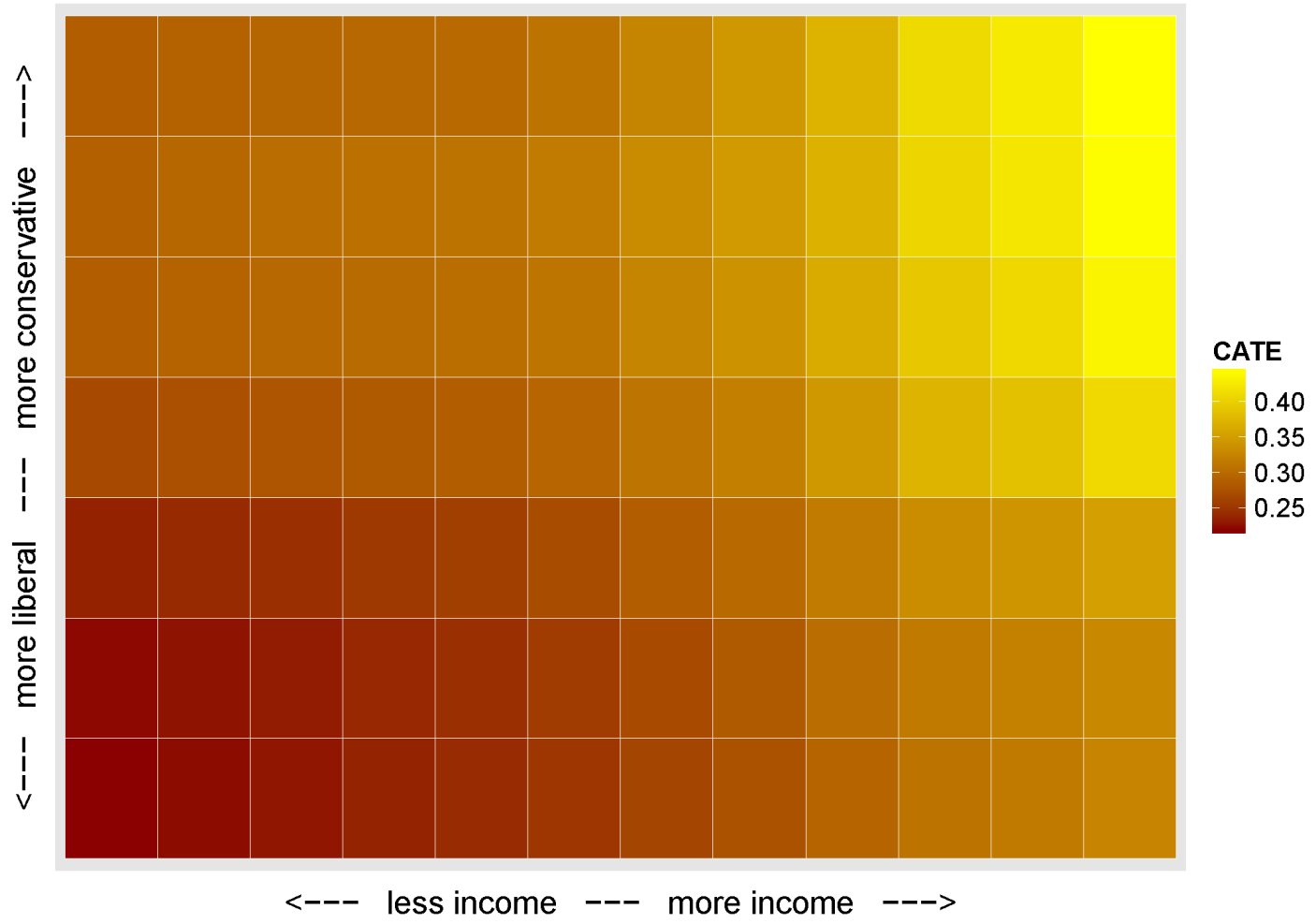
# Quantifying Heterogenity

3. Comparison of Means at end of the CATE distribution (e.g. bottom vs. top deciles)

| | 1 | 10 | p.overall |
|---|---|---|---|
| | *N=1057* | *N=1056* | |
| age | 3.05 (1.08) | 3.43 (0.82) | <0.001 |
| income | 6.35 (1.67) | 7.23 (0.34) | <0.001 |
| educ | 5.68 (1.15) | 4.76 (0.72) | <0.001 |
| polviews_X2 | 0.38 (0.49) | 0.02 (0.13) | <0.001 |
| polviews_X3 | 0.19 (0.39) | 0.05 (0.22) | <0.001 |
| polviews_X4 | 0.26 (0.44) | 0.40 (0.49) | <0.001 |
| polviews_X5 | 0.03 (0.17) | 0.23 (0.42) | <0.001 |
| polviews_X6 | 0.06 (0.23) | 0.22 (0.41) | <0.001 |
| polviews_other.values | 0.08 (0.27) | 0.09 (0.28) | 0.634 |
| sex_X1 | 0.49 (0.50) | 0.61 (0.49) | <0.001 |

# Application: General Social Survey

A causal forest analysis uncovers **strong treatment heterogeneity**



# FASFA Text Message Experiment

Collaboration with Ideas42 and CUNY: Athey, Kelleher, and Spiess (2021).

**Research Questions**

- For whom are nudges most effective
- How well do different targeting approaches do?

# FASFA Text Message Experiment

## Experiment

- Run in 2017 and 2018 by ideas42 and the City University of New York (CUNY).
- Control group: business-as-usual emails from the college.
- Treatment groups: supplementary behavioral emails and text messages.
- Matriculated students from CUNY community colleges who had not yet renewed FAFSA in February of the study year.
- 2017: 25,167 students from 3 colleges
- 2018: 40,638 students from 5 colleges
- ATE: on-time submission increases by (in percentage points) $6.4 \pm 0.6$ (2017) and $12.1 \pm 0.7$ (2018), increasing early filing rates from 37% to 43% and 38% to 50%, respectively

# FASFA Text Message Experiment

**Text Message Content (using BMCC texts as an example):**

| Msg. # | Send Date and Time | Content |
|---|---|---|
| 0 | Wed, March 1 @ 6pm | **Part 1:** Hi {First Name}! This is the CUNY Student Persistence Team. To help you finish the year strong we will send you a few helpful texts.<br><br>**Part 2:** Reply CANCEL if you don't want help setting yourself up for success.<br><br>**Response to "cancel":** Thanks for letting us know. You will no longer receive texts from us. |
| 1 | Tues, March 14 @ 6pm | {First Name}, you must renew your FAFSA each year. This year it's easier -- you can use the same tax info as last year! Go to http://bit.ly/FAFSABMCC today. |
| 2 | Tues, March 28 @ 6pm | Renew your FAFSA and do it right the first time! Stop by the Financial Aid Lab (S115-C) and get help renewing today. |
| 3 | Wed, April 12 @ 6pm | Renew your FAFSA today! Many people renew in 30min or less at http://bit.ly/FAFSABMCC. Tip: use the IRS data retrieval tool to renew quickly. |
| 4 | Tues, April 25 @ 6pm | Unsure how to renew FAFSA? That's OK! Many students go before/after class to FinAid Lab (S115-C) for free help. Hrs: M/Th 10-6, F 10-5. |
| 5 | Tues, May 2 @ 6pm | {First Name Last Name}: FAFSA Status—NOT RENEWED. Renew now at http://bit.ly/FAFSABMCC |

# HTE in the FASFA Experiment

Use causal forest to estimate HTE: $\hat{\tau}(x)$.

Calibration using transformed outcome approach: regress $Y_i \cdot \frac{2W_i - 1}{2}$ on $\hat{\tau}(x)$.

| Study year | Covariates used for estimation | Slope estimate | SE | $t$-stat | $p$-value |
|---|---|---|---|---|---|
| 2017 | Early (before semester) | 0.4733 | 0.2545 | 1.8601 | 0.0314 |
| | Late (mid-semester) | 0.4767 | 0.2437 | 1.9561 | 0.0252 |
| 2018 | Early (before semester) | 0.5555 | 0.3212 | 1.7296 | 0.0419 |
| | Late (mid-semester) | 0.7338 | 0.2932 | 2.5024 | 0.0062 |

Slope coefficient estimates for the calibration regression of actual outcomes on treatment-effect estimates interacted with normalized treatment following Chernozhukov et al (2019). We see significant evidence of heterogeneity, but the model is not well calibrated (slope coefficients far from 1).

# HTE in the FASFA Experiment

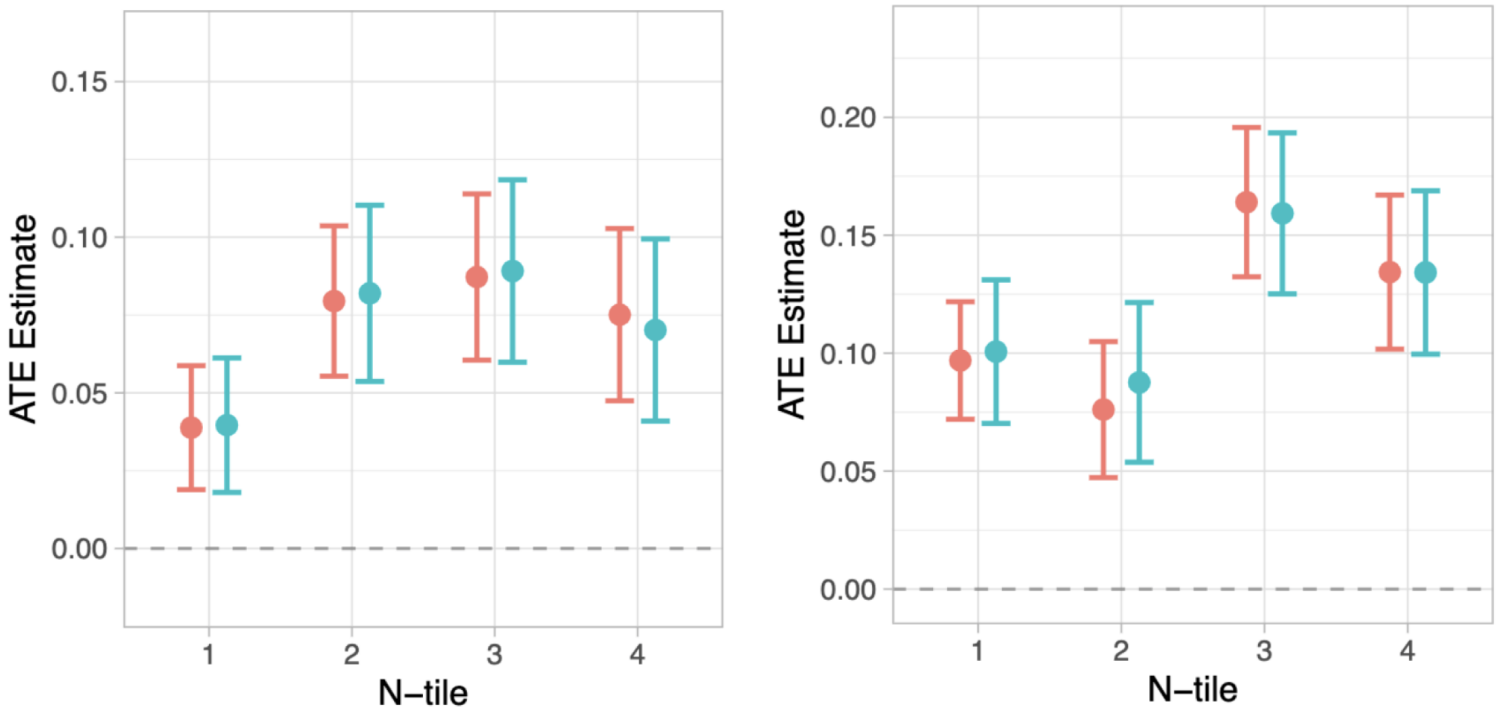
## Quartiles of CATE

- Hold out one fold of data $k$. Build model of CATE on remaining data, and use that model to build a derived model mapping $x$ to each of four quartiles based on $\hat{\tau}_k(x)$.
- Put observations in held-out fold $k$ into four quartiles based on their $x$.
- Estimate the ATE in each of the four quartiles for fold $k$ using the outcomes in fold $k$.
- Since outcomes in fold $k$ were not used to build model or assign quartiles, this yields unbiased estimate of ATE for the quartile.
- To make use of all of our data, we do this exercise across all folds of the data, a procedure referred to as "cross-fitting." We average the ATE's for a given quartile across folds.

# HTE in the FASFA Experiment

## Results

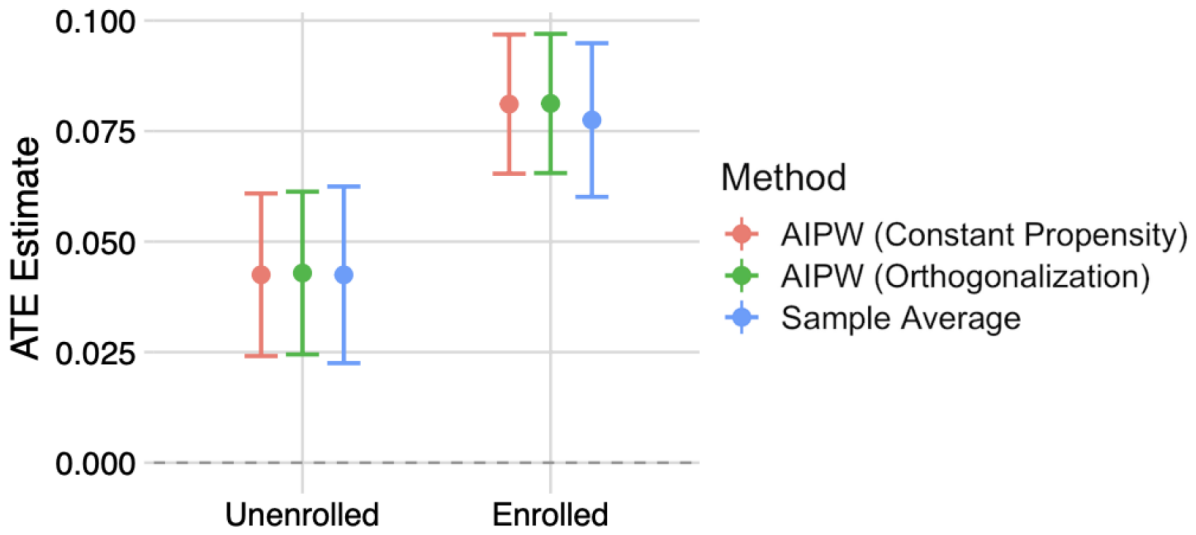
- Plotting ATE by CATE-quartile reveals estimates noisy (simple averages red, AIPW blue).
- With strong heterogeneity, higher quartiles should have higher ATE. We see we are in a low-signal environment.

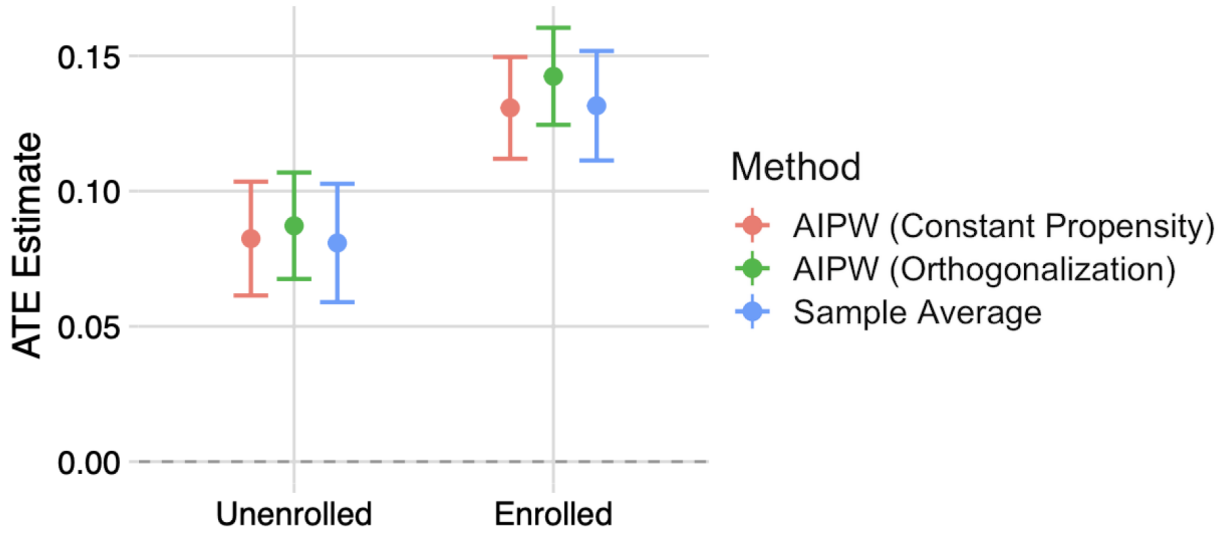



**(a)** 2017 **(b)** 2018

# ATE by Subgroup: Enrollment

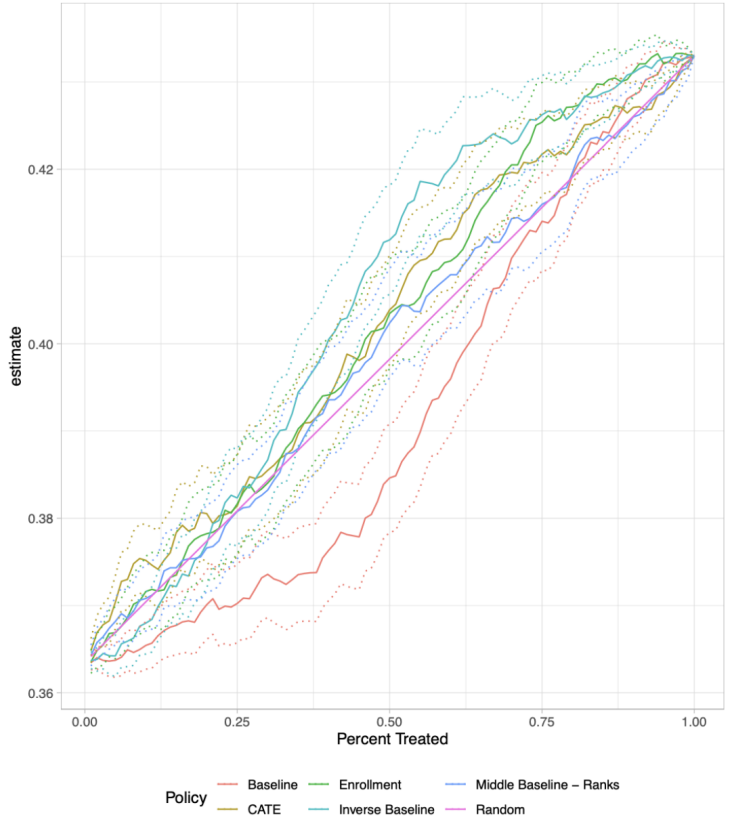
- Inspecting X's most associated with CATE heterogeneity, enrollment status is top
- Some students have dropped out, less likely to consider enrolling next year -> low CATE.
- If we had started with this hypothesis, would have identified strong heterogeneity.
- Subgroup analysis does BETTER than that based on our causal forest CATEs.
- ML comes at a cost: using the data to search for heterogeneity adds some noise, which can dominate in low signal environment.



# Evaluating benefits of targeted treatment assignment

- Imagine we can only target $Z\%$ of individuals. Which ones should get treatment?
- Use grf to estimate treatment effects, and target individuals to those with highest CATEs
- Alternatives: target by enrollment status; most or least likely to fill out form w/o treatment



# Forest Weaknesses

- Many economic datasets have smooth relationships
- Many relationships are monotonic or U-shaped
- Forests fit a line as a step function; very inefficient
- A variety of ML methods might improve but little theory
- Solution: Local Linear Forests + theory

# Step Functions (Forests) v. Locally Linear Forest

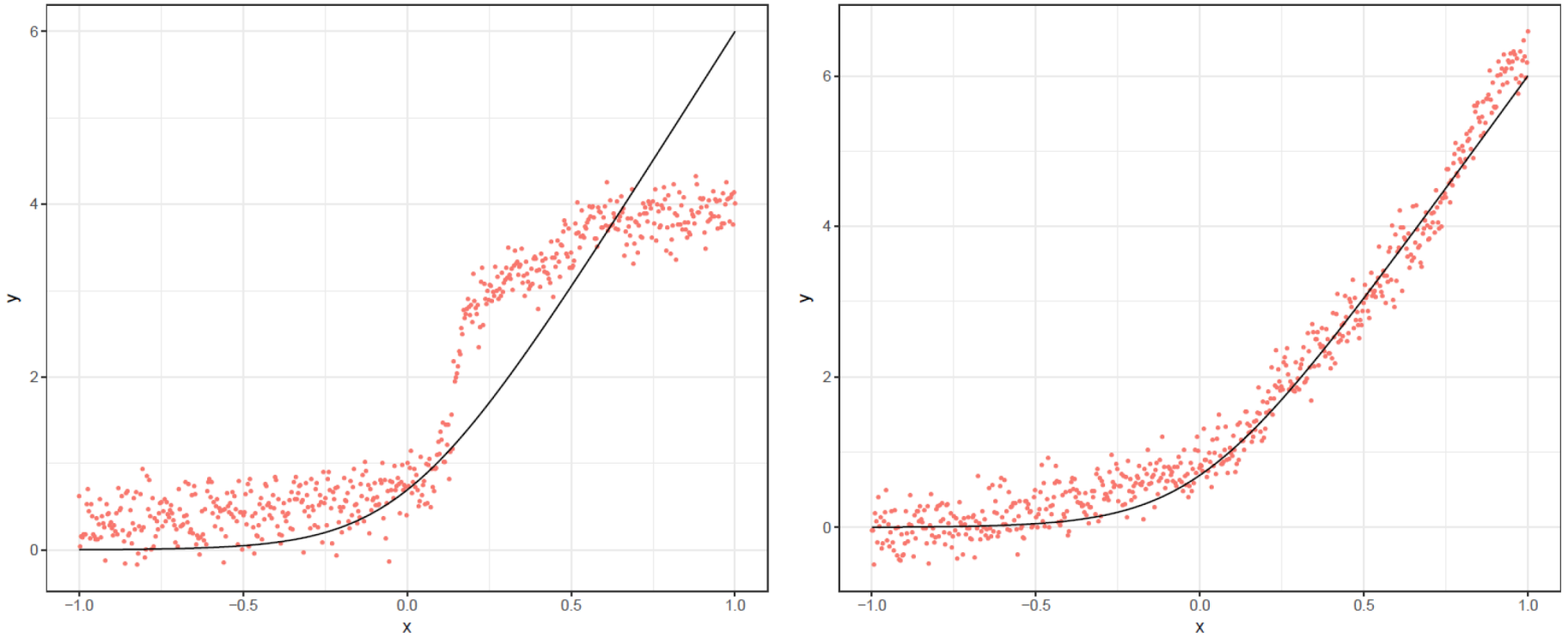
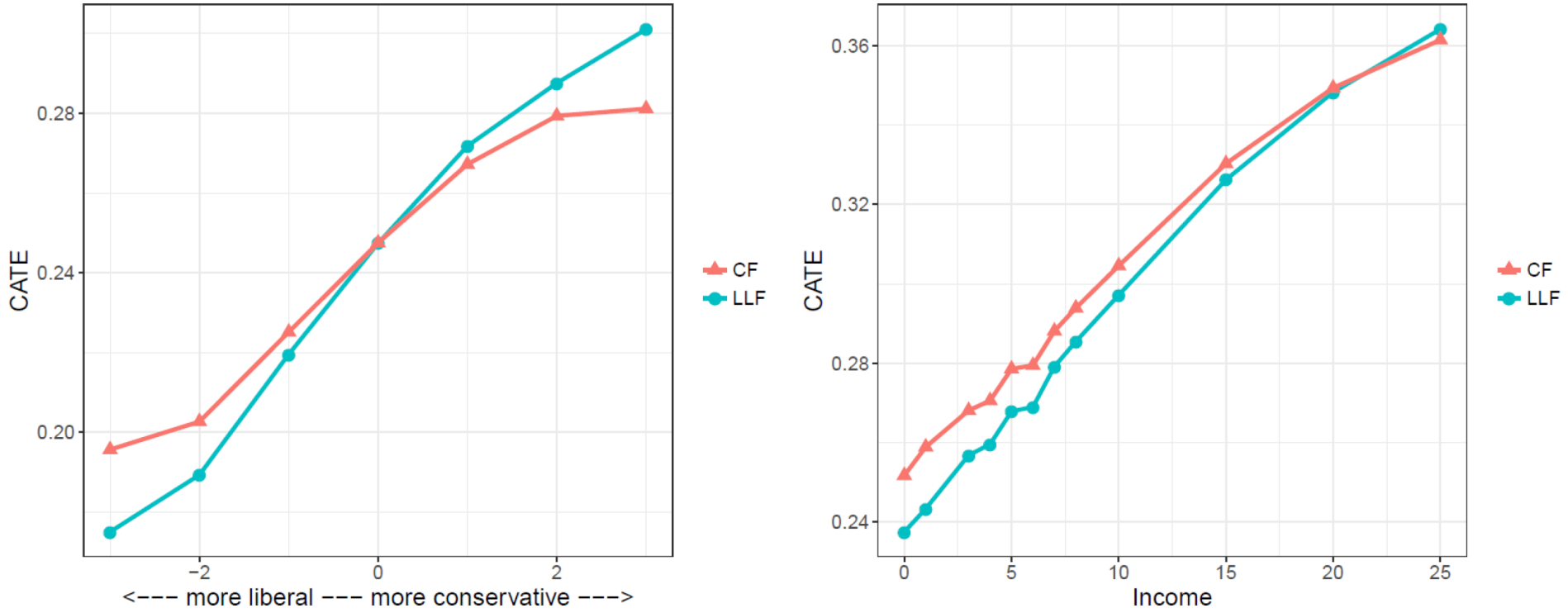



*Figure 1: Predictions from random forests (left) and locally linear forests (right) on 600 test points. Training and test data were simulated from equation (1), with dimension* $d = 20$ *and errors* $\epsilon \sim N(0, 20)$*. Forests were trained also on* $n = 600$ *training points and tuned via cross-validation. Here the true conditional mean signal* $\mu(x)$ *is in black, and predictions are shown in red.*

# Causal Forest v. Locally Linear Causal Forest



# Locally Linear Forest

Locally linear regression with ridge penalty:

$$\begin{pmatrix} \hat{\mu}(x) \\ \hat{\theta}(x) \end{pmatrix} = \arg\min_{\mu,\theta} \sum_{i=1}^{n} \alpha_i(x)(Y_i - \mu(x) - (X_i - x)\theta(x))^2 + \lambda||\theta(x)||_2^2$$

In matrix form:

$$\begin{pmatrix} \hat{\mu}(x) \\ \hat{\theta}(x) \end{pmatrix} = (X^T A X + \lambda J)^{-1} X^T A Y$$

Weights are determine from forest a la GRF, accounting for regression in splitting for efficiency.

**Theorem.** (Friedberg, Athey, Tibshirani, and Wager (2018): Assuming $\mu(x)$ is twice continuously differentiable, estimates are asymptotically normal. Faster rate of convergence than GRF; result exploits smoothness assumptions. Extends to causal LLF.