Policy and Evaluation in the Age of Big Data

Jasjeet Sekhon

UC Berkeley

July 7, 2017

Causal Inference and Big Data

- Measuring human activity has generated massive datasets with granular population data: e.g.,
  - Electronic medical records
  - Genetic markers, Microbiome
  - Browsing, search, and purchase data from online platforms
  - Administrative data: schools, criminal justice, IRS
- Big in size and breadth: wide datasets
- Data can be used for personalization of treatments, creating markets, modeling behavior
- Many inferential issues: e.g., unknown sampling frames, heterogeneity, targeting optimal treatments, compound loss functions

Prediction versus Causal Inference

- Causal Inference is like a prediction problem: but predicting something we don't directly observe and possibly cannot estimate well in a given sample

- ML algorithms are good at prediction, but have issues with causal inference:
  - Interventions imply counterfactuals: response schedule versus model prediction
  - Validation requires estimation in the case of causal inference
  - Identification problems not solved by large data
  - Predicting the outcome mistaken for predicting the causal effect
    - targeting based on the lagged outcome

Classical Justifications Versus ML Pipelines

Two different justifications for statistical procedures:

1. (classical) statistical theory:
   it works because we have **relevant** theory that tells us it should
   Hopefully, this is not simply: "Assume that the data are generated by the following model ..." (Brieman 2001)

2. Training/test loop:
   it works because we have validated against ground truth and it works

Classical Justifications Versus ML Pipelines

Two different justifications for statistical procedures:

1. (classical) statistical theory:
   it works because we have **relevant** theory that tells us it should
   Hopefully, this is not simply: "Assume that the data are generated by the following model ..." (Brieman 2001)

2. Training/test loop:
   it works because we have validated against ground truth and it works

On the **normal distribution**:

> *"Everyone believes in it: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact." — Henri Poincaré (quoted by de Finetti 1975)*

Even Classical Justifications Should be Validated

- Question: coverage for the population mean. Is $n = 1000$ enough?
- Sometimes, no. Not for many metrics, even when they are bounded
- For some metrics, asking for 95% CI results in only 60% coverage
- Data is very irregular
  Many zeros, IQR: 0

$$\frac{p100 - p99}{p99 - p50} > 10,000$$

Conditional Average Treatment Effect (CATE)

Individual Treatment Effect (ITE): $D_i := Y_i(t) - Y_i(c)$

Let $\hat{\tau}_i$ be an estimator for D_i

$\tau(x_i)$ is the **CATE** for all units whose covariate vector is equal to x_i:

$$CATE := \tau(x_i) := \mathbb{E}\Big[D\Big|X = x_i\Big] = \mathbb{E}\Big[Y(t) - Y(c)\Big|X_i = x_i\Big]$$

Variance of Conditional Average Treatment Effect

$$CATE := \tau(x_i) := \mathbb{E}\Big[D\Big|X = x_i\Big] = \mathbb{E}\Big[Y(t) - Y(c)\Big|X_i = x_i\Big]$$

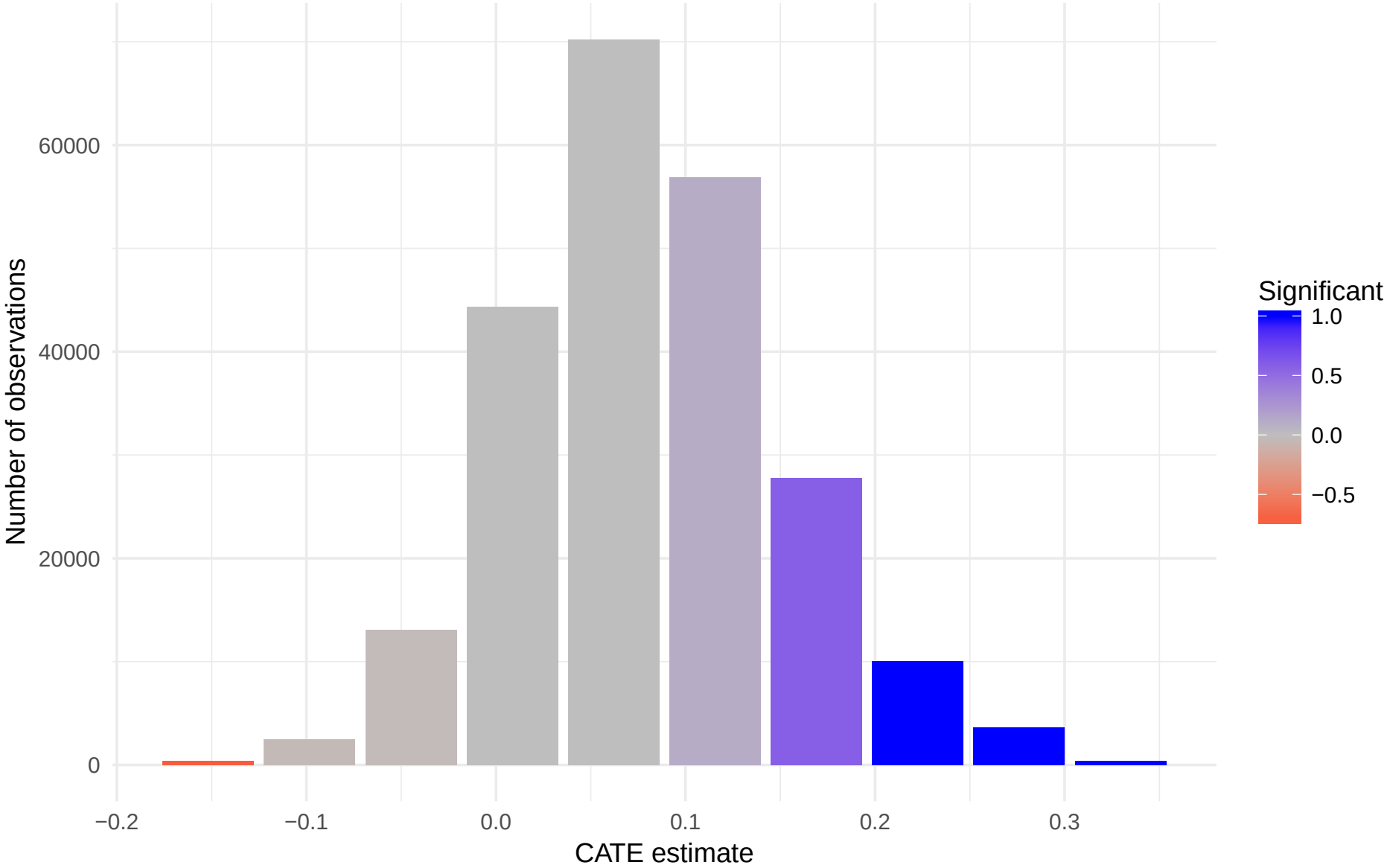
Decompose the MSE at x_i:

$$\mathbb{E}\left[(D_i - \hat{\tau}_i)^2 | X_i = x_i\right] =$$
$$\underbrace{\mathbb{E}\left[(D_i - \tau(x_i))^2 | X_i = x_i\right]}_{\text{Approximation Error}} + \underbrace{\mathbb{E}\left[(\tau(x_i) - \hat{\tau}_i)^2 | X_i = x_i\right]}_{\text{Estimation Error}}$$

- Since we cannot estimate D_i, we estimate the CATE at x_i
- But the error for the CATE is not the same as the error for the ITE

Supplementary

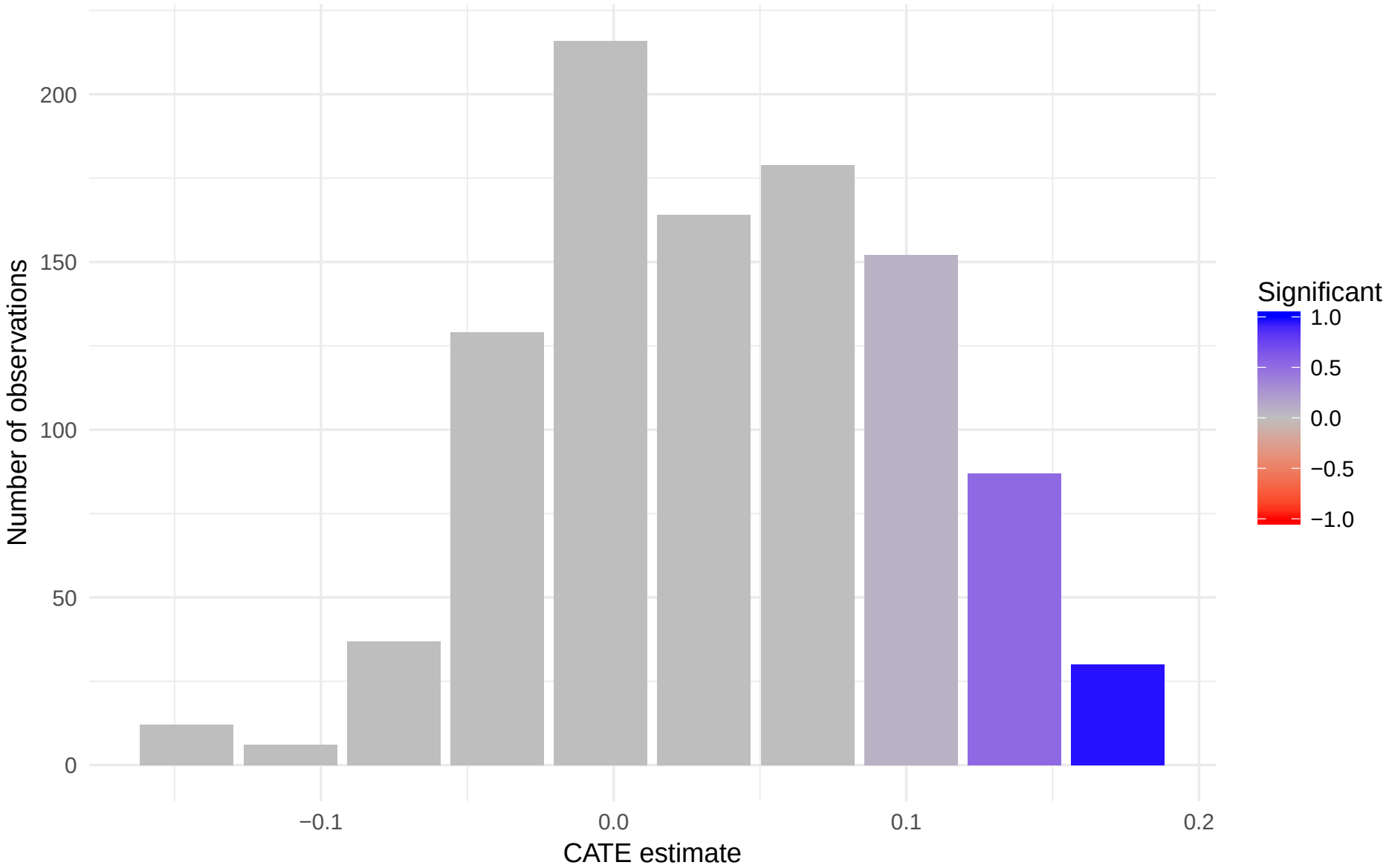
GOTV: Social pressure (Gerber, Green, Lairmer, 2008)



Pulmonary Artery Catheterization (PAC)

- Pulmonary Artery Catheterization (PAC): monitoring device commonly inserted into critically ill patients
- Detecting complications, but invasive to patients and significant expenditure
- Question: does PAC have effect on patient survival?
- Observational study (Connors et al, 1996): PAC had an adverse effect on patient survival and led to increased cost of care
- Subgroups and better methods for combining RCT and observational data: Hartman, Grieve, Ramsahai, Sekhon 2015

Pulmonary Artery Catheterization (PAC) Experiment



How to estimate the CATE?

Meta–learners

A meta–learner decomposes the problem of estimating the CATE into several sub–regression problems. The estimator which solve those sub–problems are called **base–learners**

- Flexibility to choose base–learners which work well in a particular setting
- Deep Learning, (honest) Random Forests, BART, or other machine learning algorithms

How to estimate the CATE?

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\ &= \mu_1(x) - \mu_0(x)\end{aligned}$$

How to estimate the CATE?

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\ &= \mu_1(x) - \mu_0(x)\end{aligned}$$

T–learner

1.) Split the data into control and treatment group,

2.) Estimate the response functions separately,

$$\hat{\mu}_1(x) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = 1]$$
$$\hat{\mu}_0(x) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = 0],$$

3.) $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$

How to estimate the CATE?

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\ &= \mu_1(x) - \mu_0(x)\end{aligned}$$

T–learner

1.) Split the data into control and treatment group,

2.) Estimate the response functions separately,

$$\hat{\mu}_1(x) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = 1]$$
$$\hat{\mu}_0(x) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = 0],$$

3.) $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$

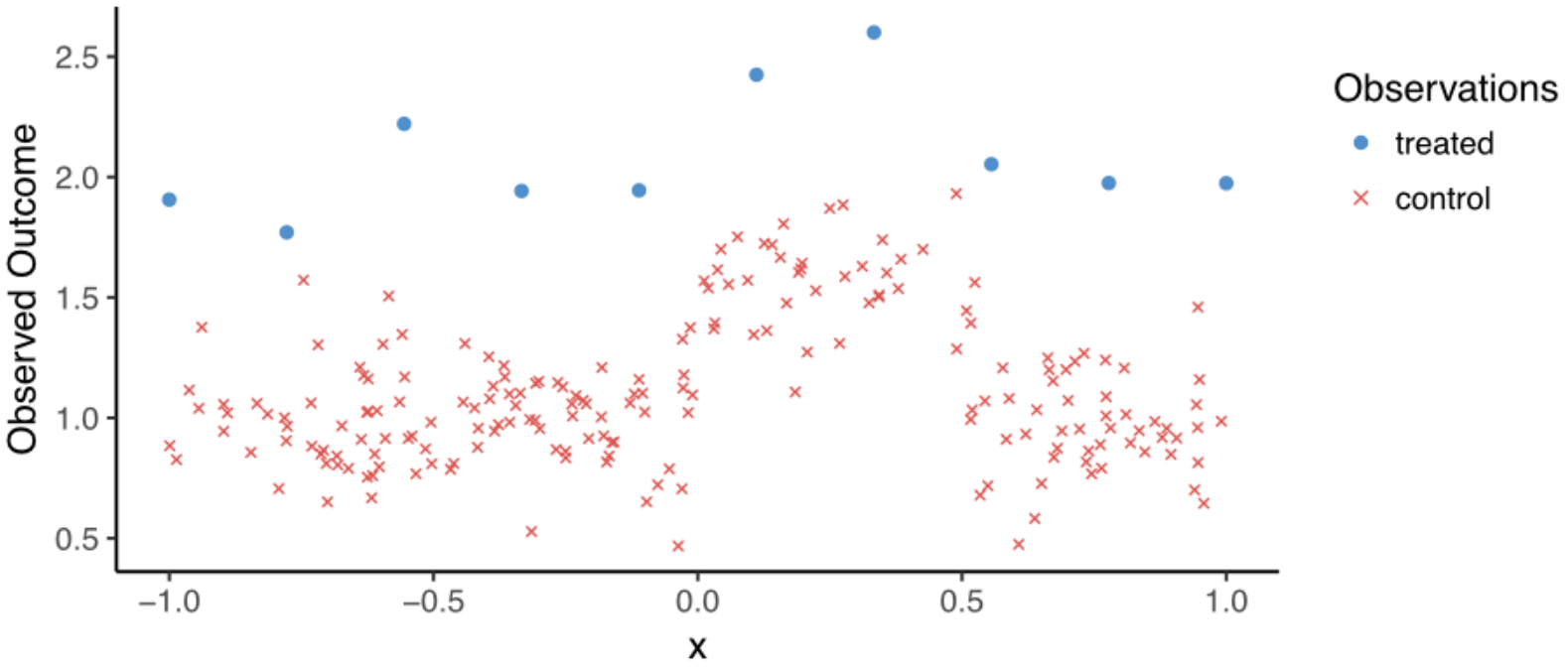
S–learner

1.) Use the treatment assignment as a usual variable without giving it any special role and estimate

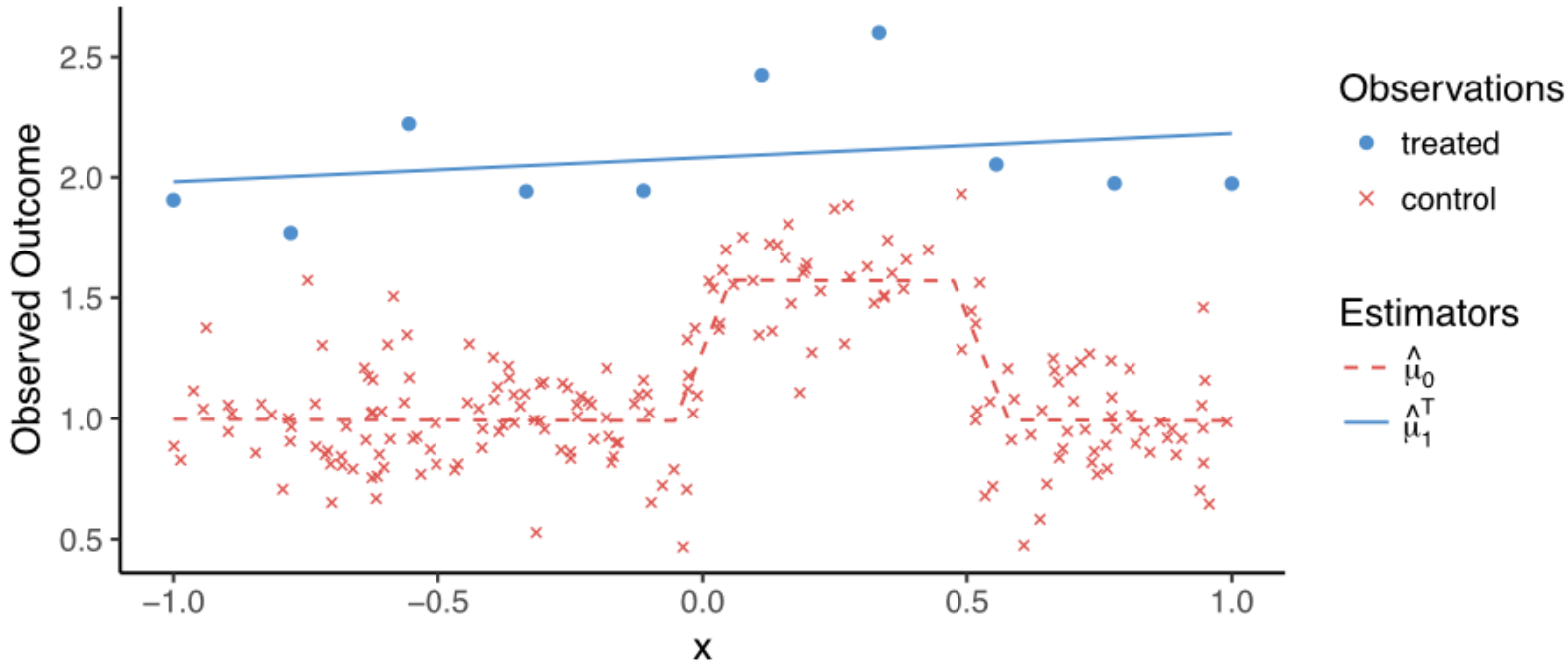
$$\hat{\mu}(x, w) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = w]$$

2.) $\hat{\tau}(x) := \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$

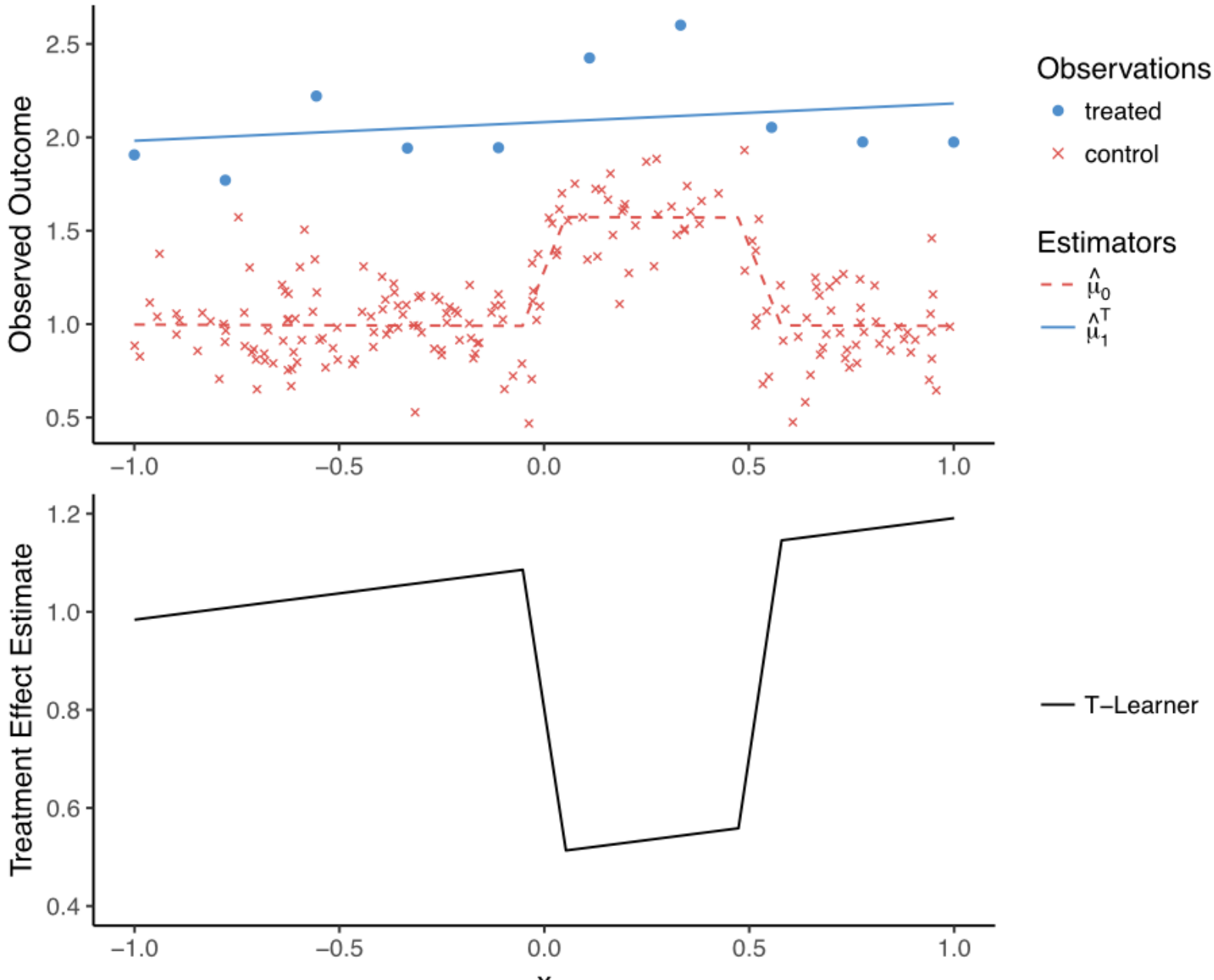
Motivating X



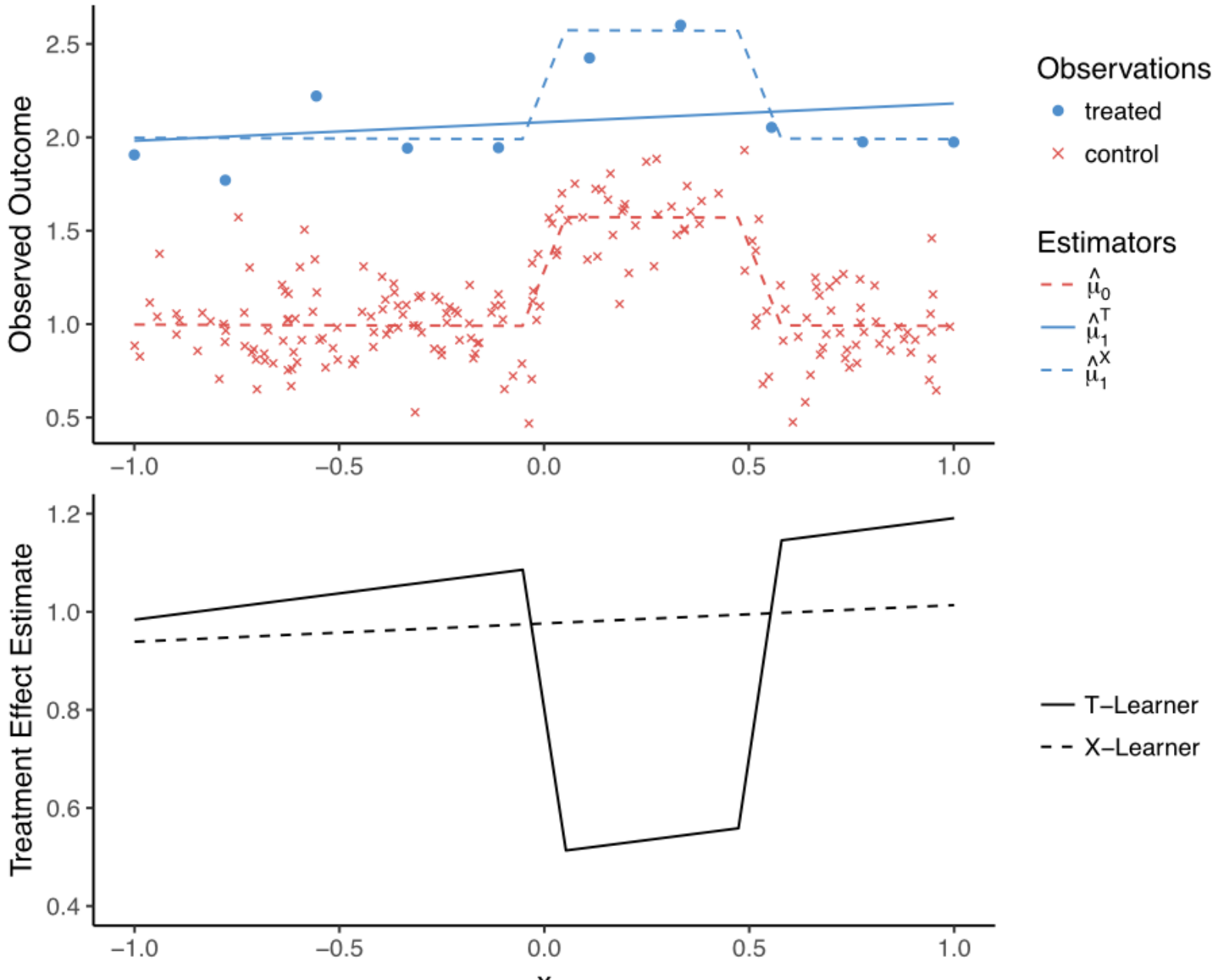
Motivating X



Motivating X



Motivating X



Formal defintion of the X–learner

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1) - \mu_0(x)|X = x]\end{aligned}$$

with $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$.

X–learner

1.) Estimate the control response function,

$$\hat{\mu}_0(x) = \hat{\mathbb{E}}[Y(0)|X = x],$$

2.) Define the **pseudo residuals**,

$$\tilde{D}_i^1 := Y_i(1) - \hat{\mu}_0(X_i(1)),$$

3.) Estimate the CATE,

$$\hat{\tau}(x) = \hat{\mathbb{E}}[\tilde{D}^1|X = x].$$

X in algorithmic form

procedure X–LEARNER(X, Y^{obs}, W)

$\hat{\mu}_0 = M_1(Y^0 \sim X^0)$ ▷ Estimate response function

$\hat{\mu}_1 = M_2(Y^1 \sim X^1)$

$\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_0(X_i^1)$ ▷ Compute pseudo residuals

$\tilde{D}_i^0 := \hat{\mu}_1(X_i^0) - Y_i^0$

$\hat{\tau}_1 = M_3(\tilde{D}^1 \sim X^1)$ ▷ Estimate CATE

$\hat{\tau}_0 = M_4(\tilde{D}^0 \sim X^0)$

$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$ ▷ Average

Algorithm 1: X–learner

Theorem 1

Theorem covers the case when estimating the base functions is not beneficial

Künzel, Sekhon, Bickel, Yu 2017

Assume we observe m control and n treatment units,

1.) Strong Ignorability holds: $(Y(0), Y(1)) \perp W|X \quad 0 < e(X) < 1$

2.) The treatment effect is linear, $\tau(x) = x^T \beta$

3.) There exists an estimator $\hat{\mu}_0$ with $\mathbb{E}[(\mu_0(x) - \hat{\mu}_0(x))^2] \leq C_x^0 m^{-a}$

Then the X-learner with $\hat{\mu}_0$ in the first stage, OLS in the second stage, achieves the parametric rate in n,

$$\mathbb{E}\left[\|\tau(x) - \hat{\tau}_X(x)\|^2\right] \leq C_x^1 m^{-a} + C_x^2 n^{-1}$$

If there are a many control units, such that $m \asymp n^{1/a}$, then

$$\mathbb{E}\left[\|\tau(x) - \hat{\tau}_X(x)\|^2\right] \leq 2C_x^1 n^{-1}$$

Theorem 2

Theorem covers the case when estimating the CATE function is not beneficial

Künzel, Sekhon, Bickel, Yu 2017

X–learner is minimax optimal for a class of estimators using KNN as the base leaner. Assume:

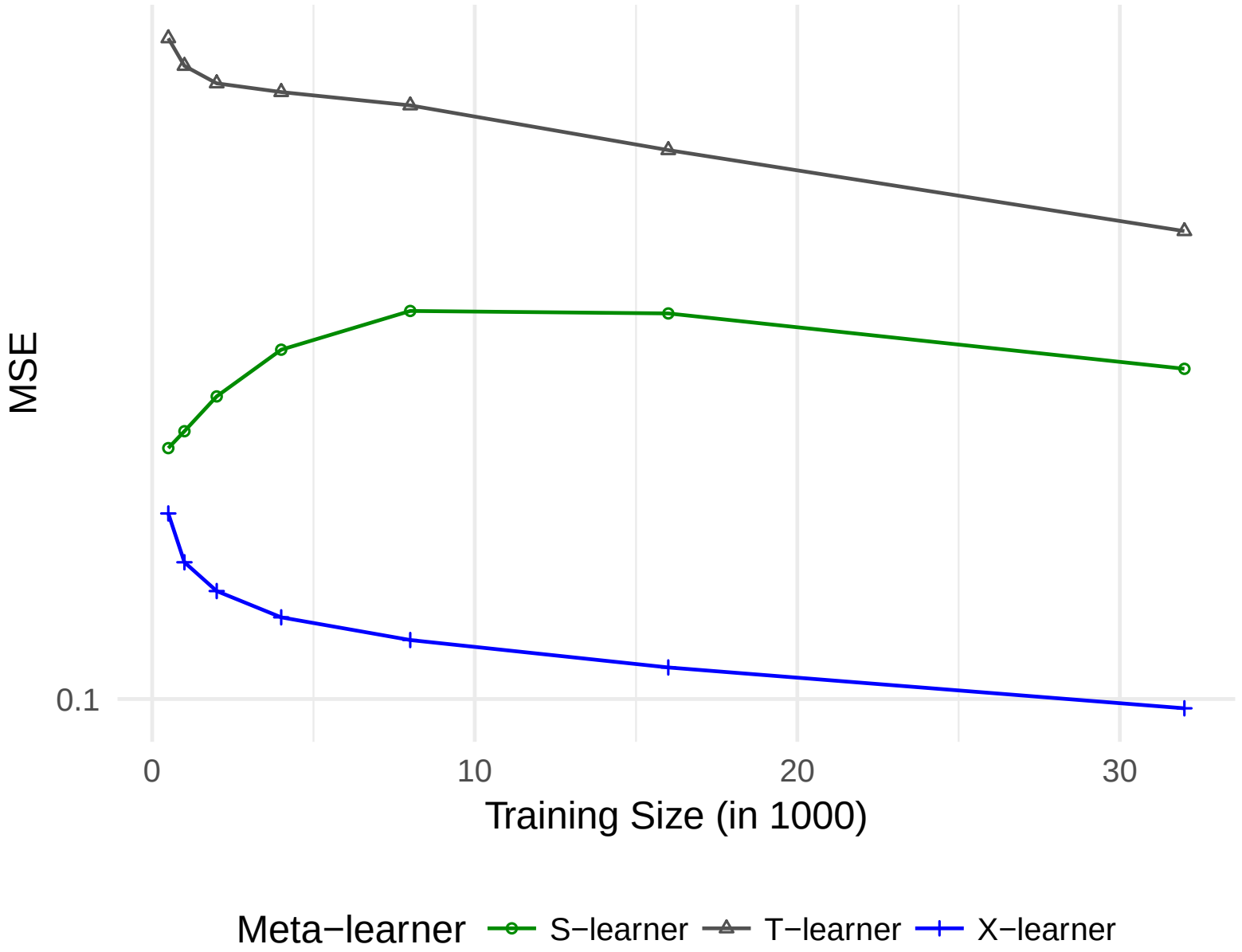
- Outcome functions are Lipschitz continuous
- CATE function has no simplification
- Features are uniformly distributed $[0, 1]^d$

The fastest possible rate of convergence for this class of problems is:

$$\mathcal{O}\left(\min(n_0, n_1)^{-\frac{1}{2+d}}\right)$$

- The speed of convergence is dominated by the size of the smaller assignment group
- In the worst case, there is nothing to learn from the other assignment group

Data Simulation: Social pressure and Voter Turnout



Conclusion

- We expect more from our experiments than ever before
- We should protect the Type I error rate—e.g., honest Random Forests, cross-fitting
- Lots of observational data, massive push to use it: could be used to help estimate control outcomes

Thanks

- Peter Bickel
- Richard Grieve
- Erin Hartman
- Sören Künzel
- Roland Ramsahai
- Bin Yu

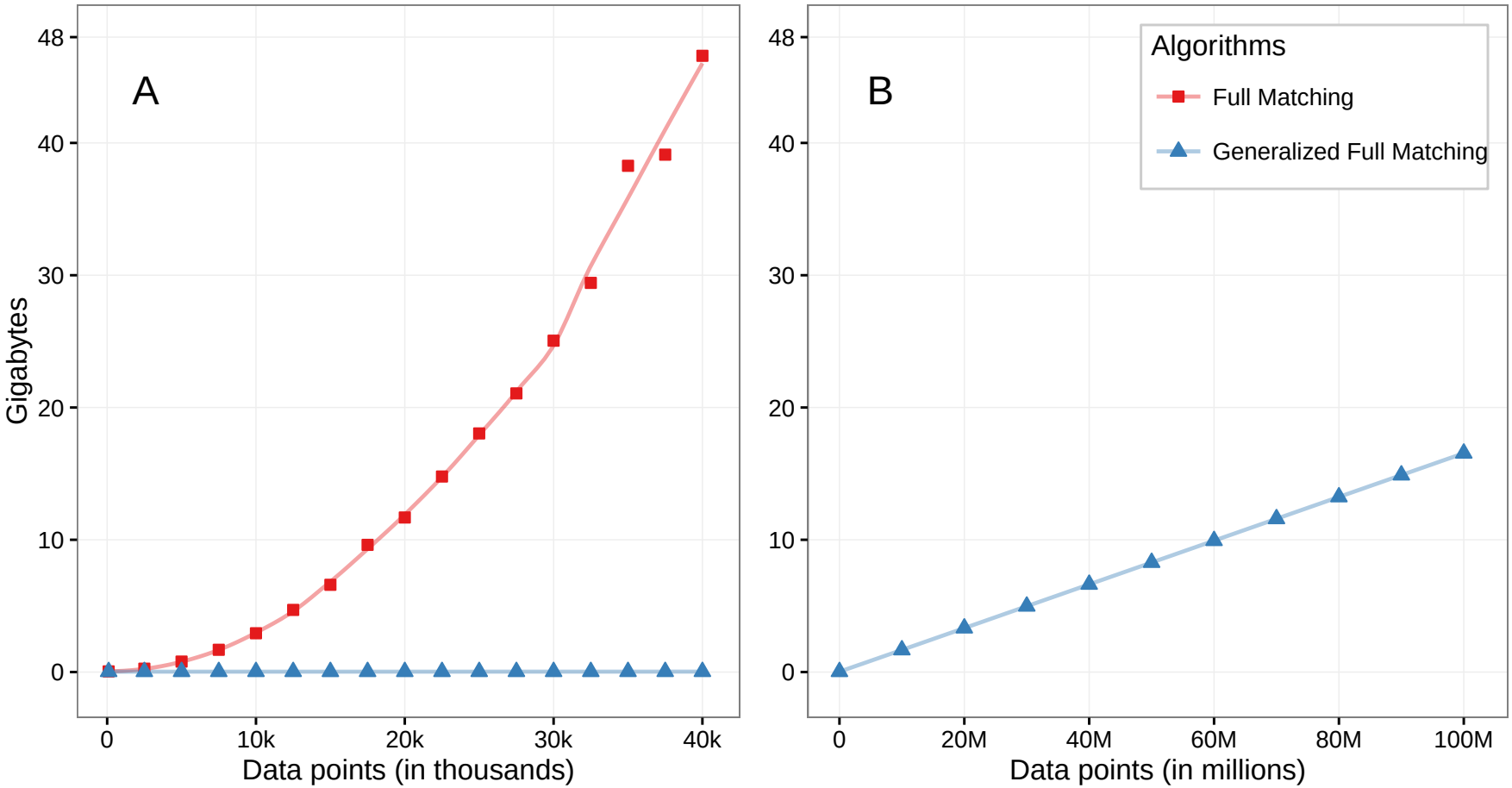
`http://sekhon.berkeley.edu`

Blocking/Post-Stratification

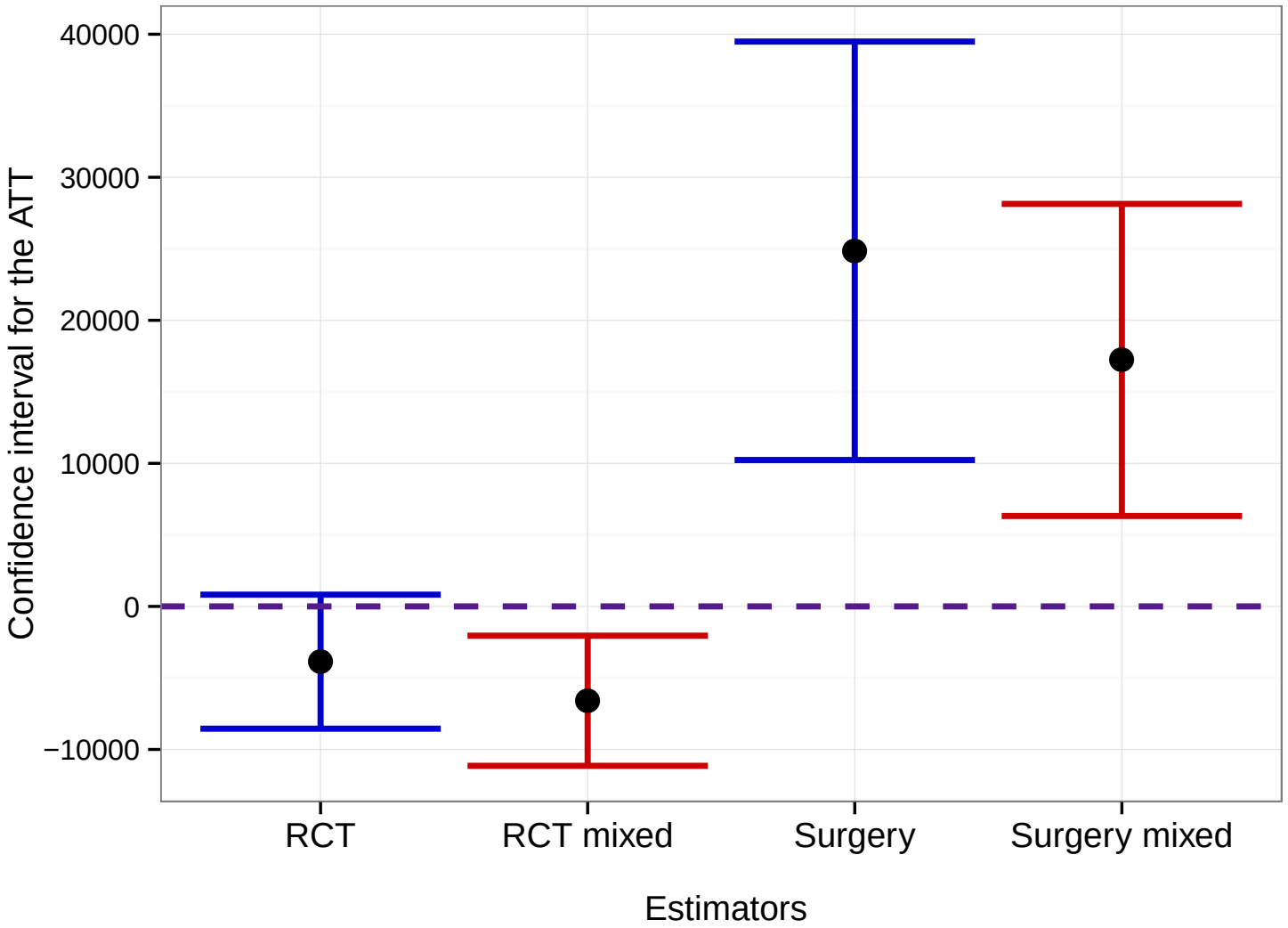
Minimizes the pair-wise Maximum Within-Block Distance: λ
(Higgins, Sävje, Sekhon 2016; Sävje, Higgins, Sekhon 2017)

- Any valid distance metric; triangle inequality
- We prove this is a NP-hard problem
- Ensures good covariate balance by design: approximately optimal: $\leq 4 \times \lambda$
- Works for any number of treatments and any minimum number of observations per block
- It is fast: $O(n \log n)$ expected time
- It is memory efficient: $O(n)$ storage
- Special cases
  1. with one covariate: λ
  2. with two covariates: $\leq 2 \times \lambda$

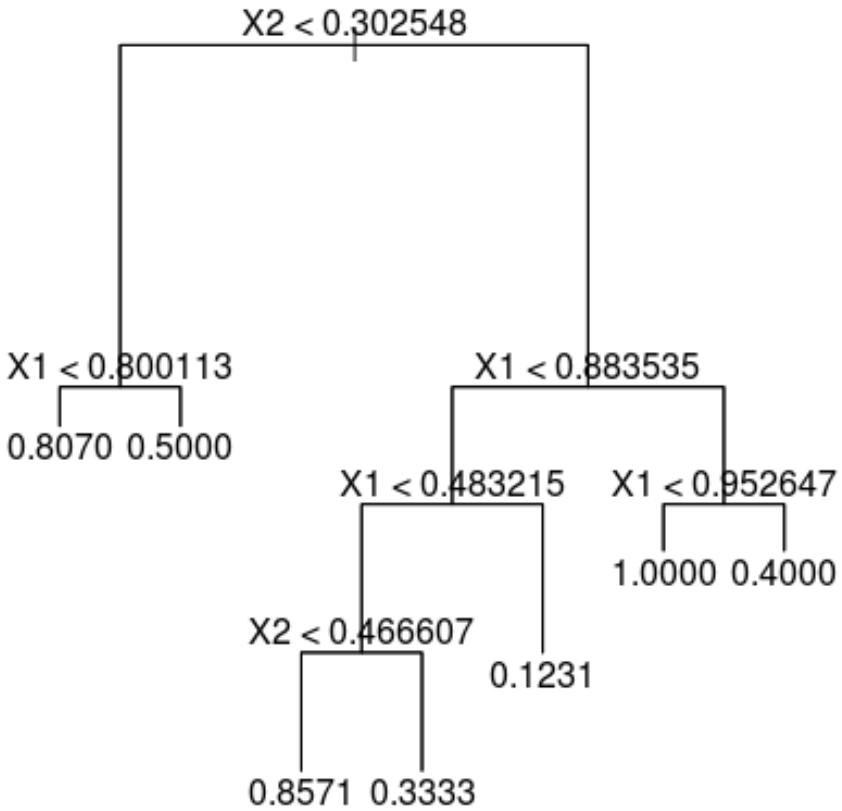
Space Complexity

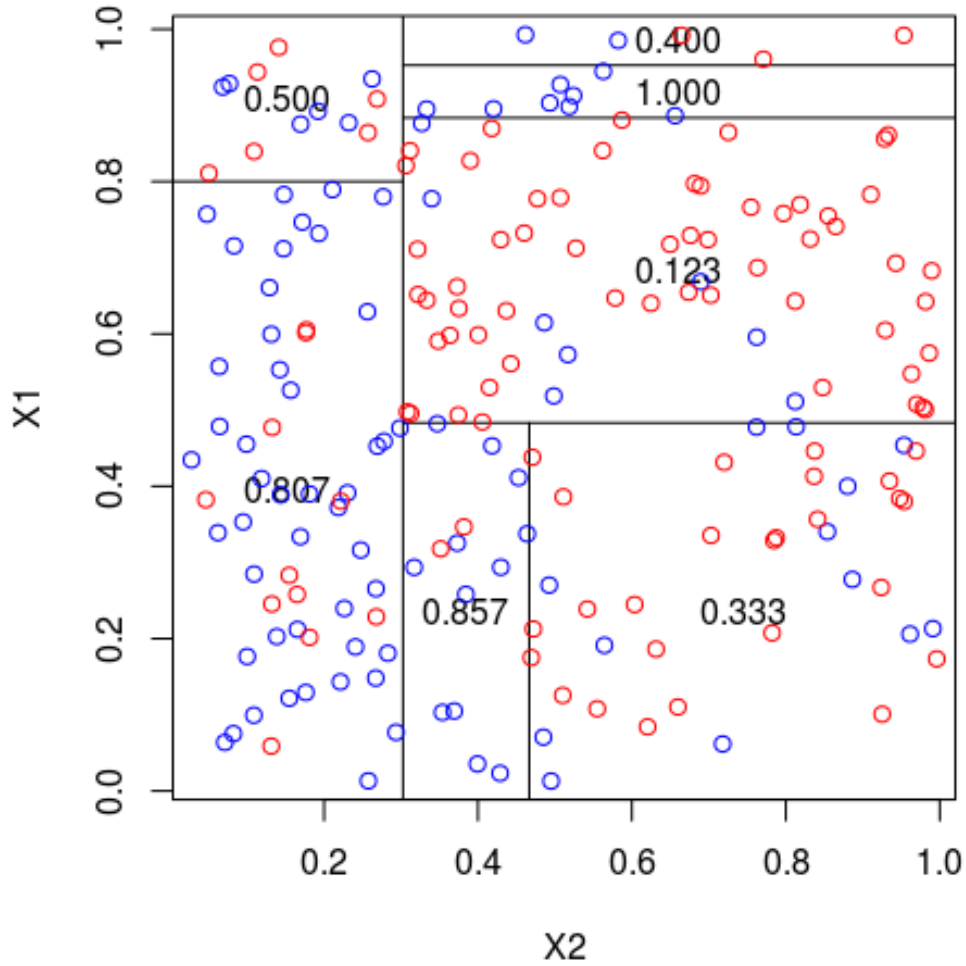


Pulmonary Artery Catheterization (PAC). Elective Surgery: Subgroups, effects on net incremental benefit



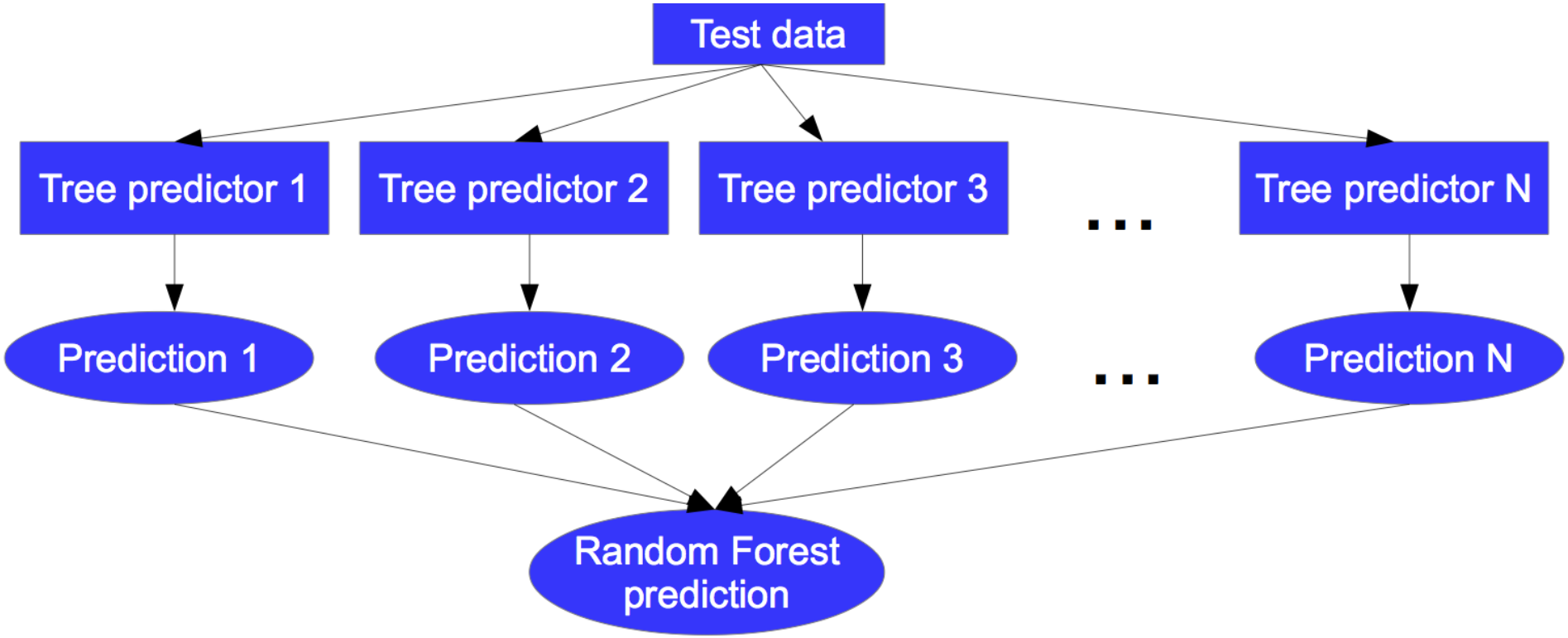
Regression Trees





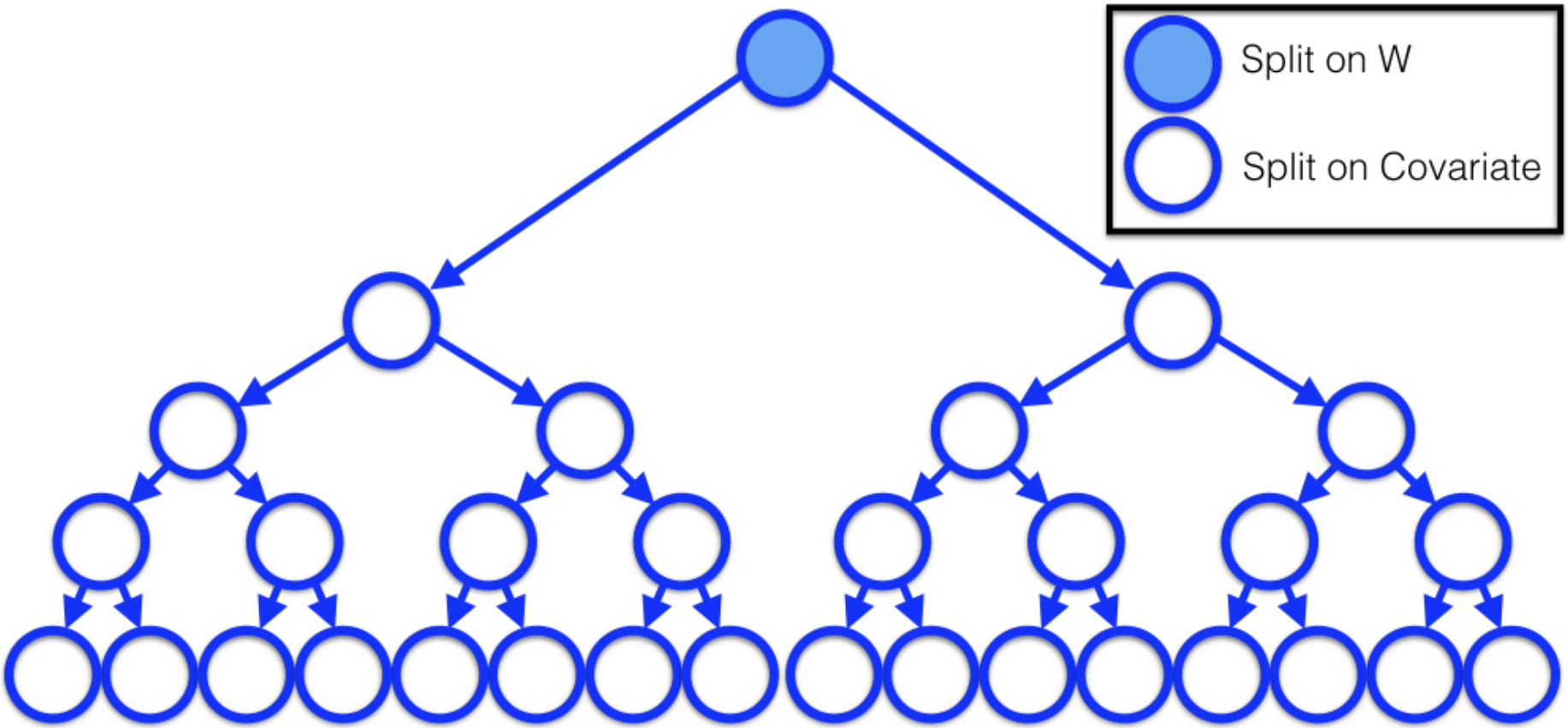
http://freakonometrics.hypotheses.org/1279

Random Forest = Many "random" Trees



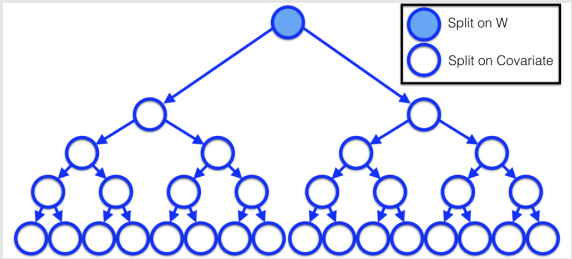
Supplementary

$\text{CATE} := \hat{\tau}(x) = f(x, w = 1) - f(x, w = 0)$



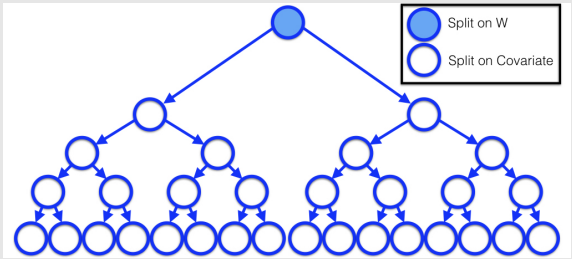
$\text{CATE} := \hat{\tau}(x) = f(x, w = 1) - f(x, w = 0)$

T-Learner



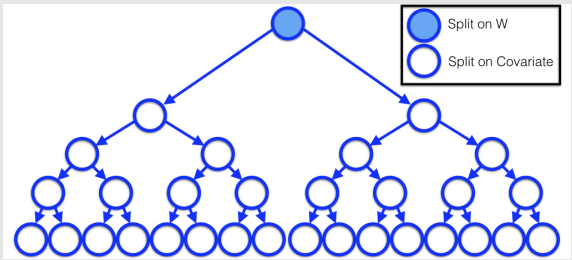
$\text{CATE} := \hat{\tau}(x) = f(x, w = 1) - f(x, w = 0)$

T-Learner

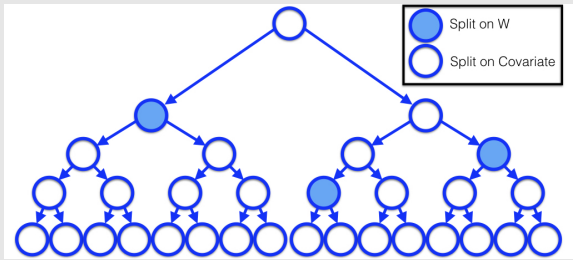


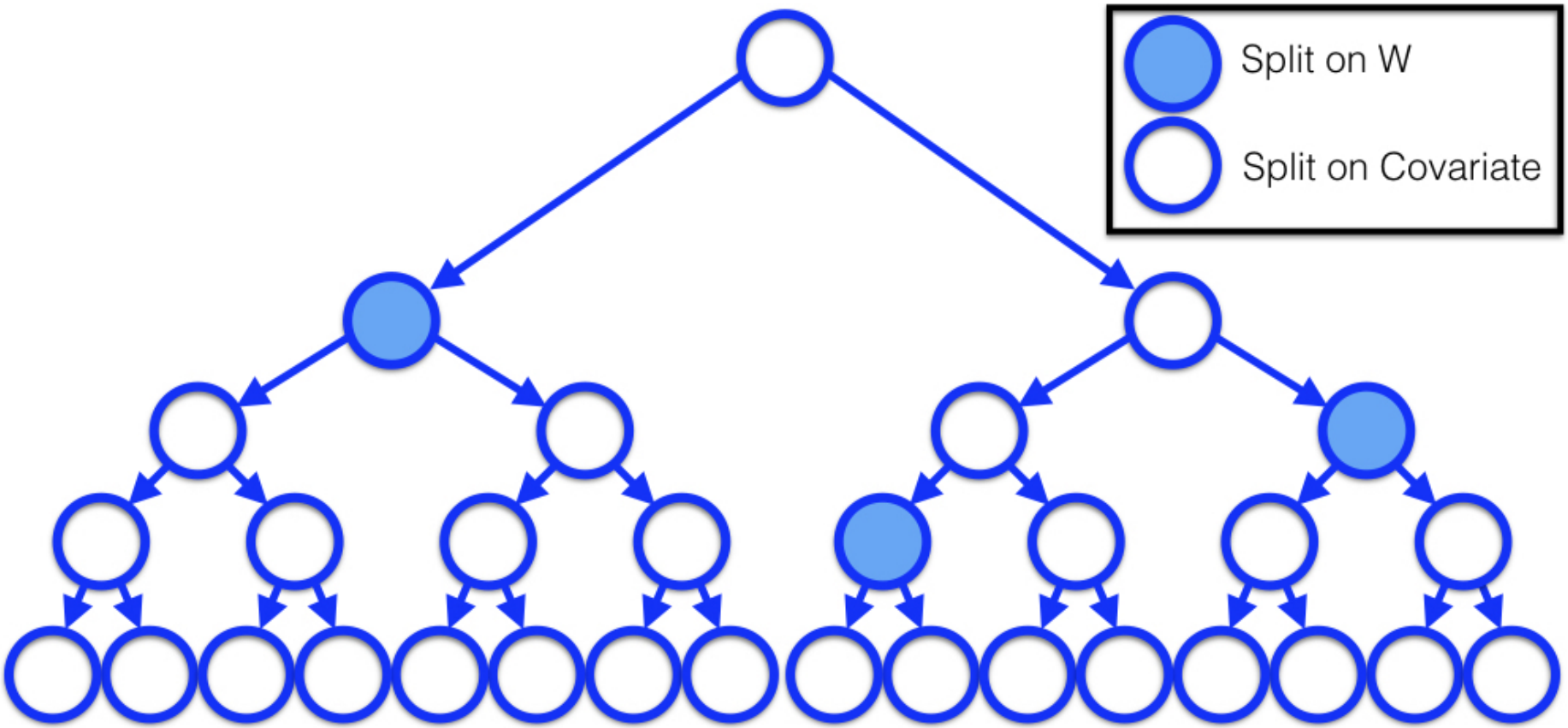
$\text{CATE} := \hat{\tau}(x) = f(x, w = 1) - f(x, w = 0)$

T-Learner



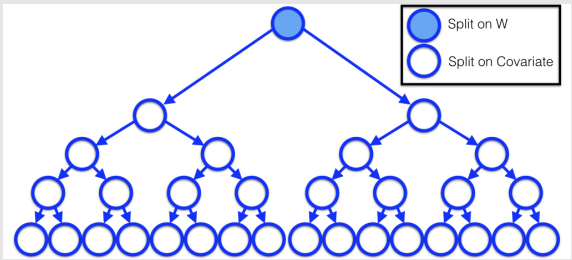
S-Learner



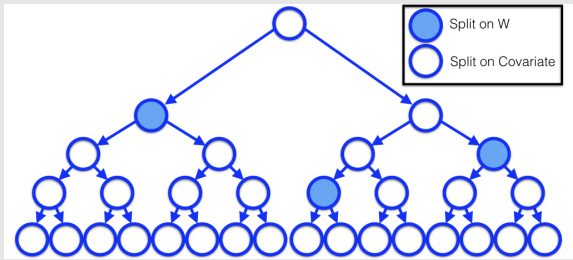


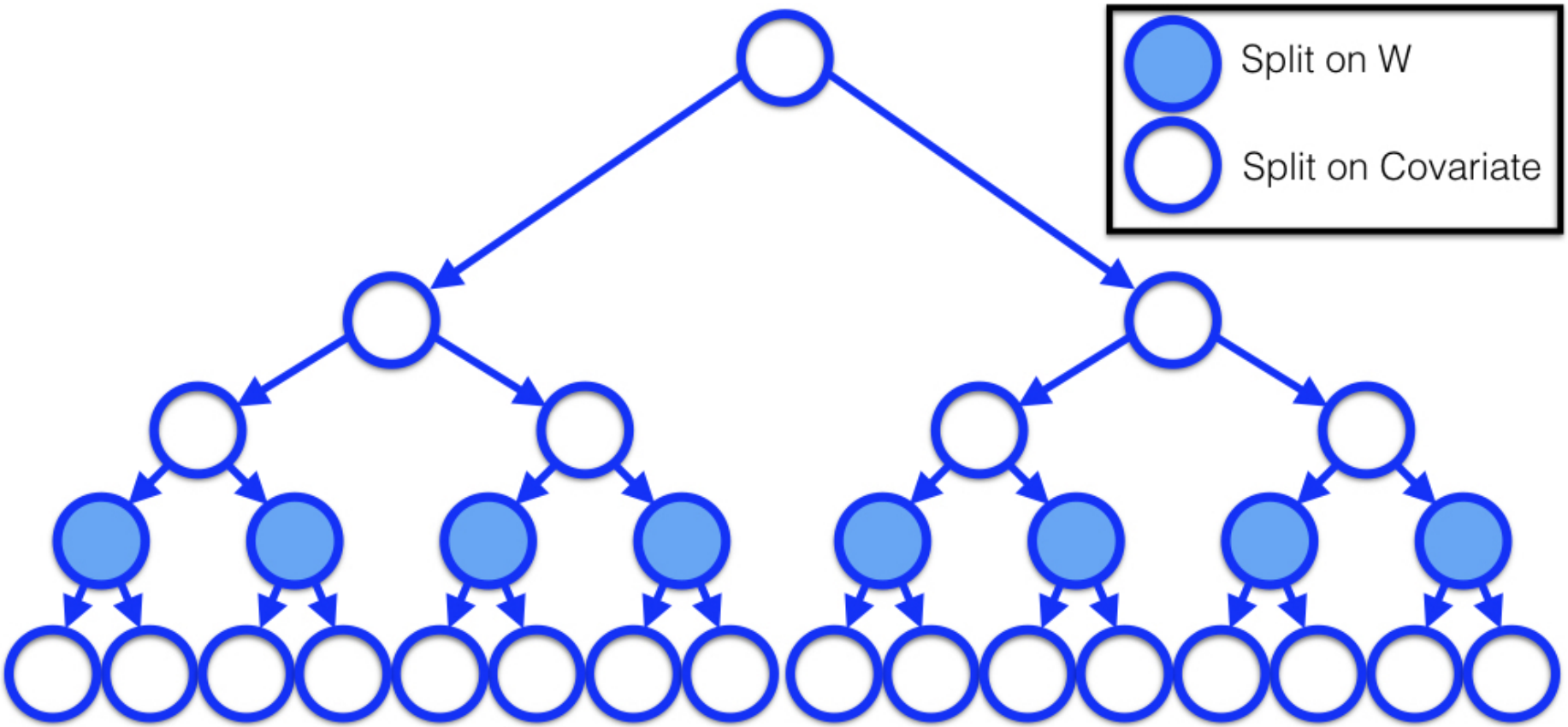
$\text{CATE} := \hat{\tau}(x) = f(x, w = 1) - f(x, w = 0)$

T-Learner



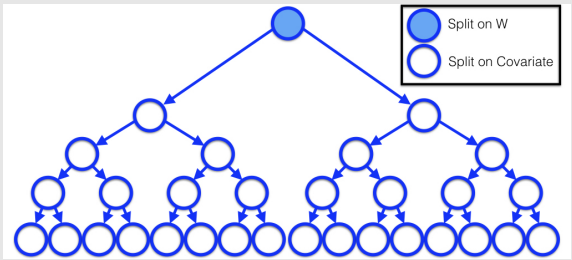
S-Learner



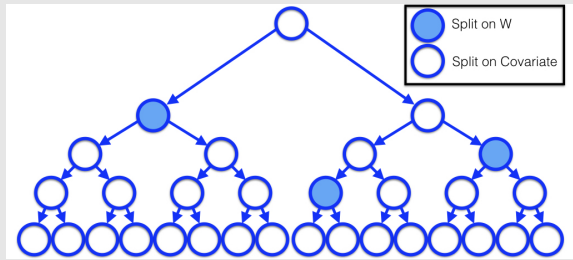


$\text{CATE} := \hat{\tau}(x) = f(x, w = 1) - f(x, w = 0)$

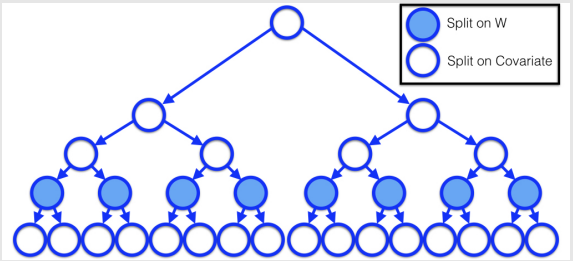
T-Learner

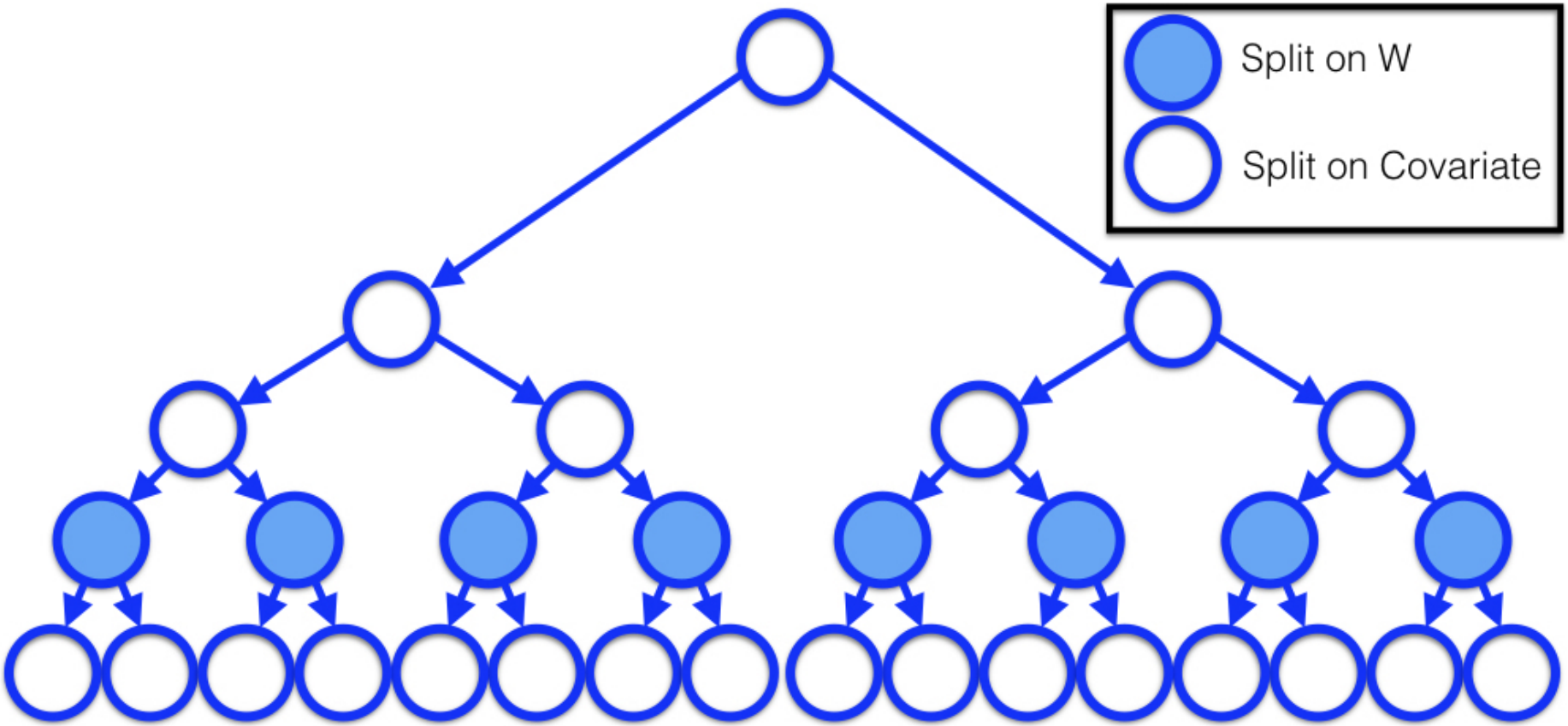


S-Learner



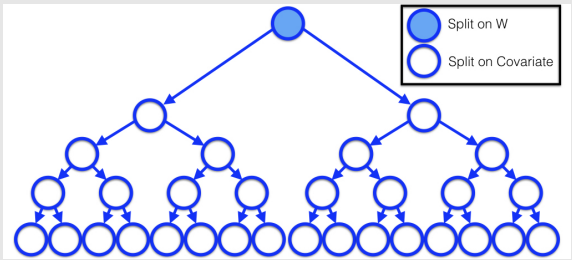
Causal Forest



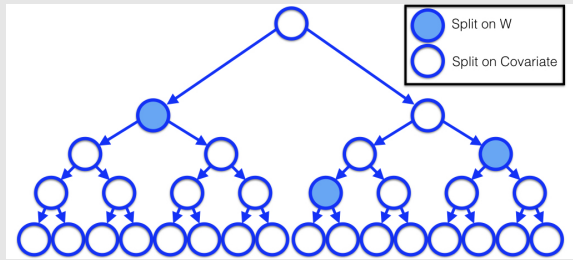


$\text{CATE} := \hat{\tau}(x) = f(x, w = 1) - f(x, w = 0)$

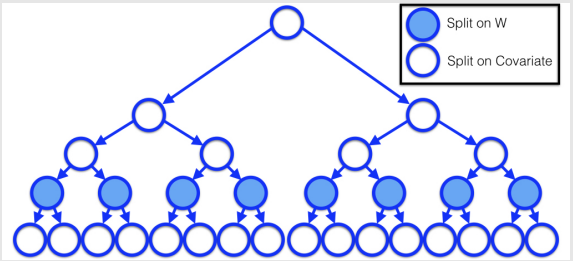
T-Learner



S-Learner



Causal Forest

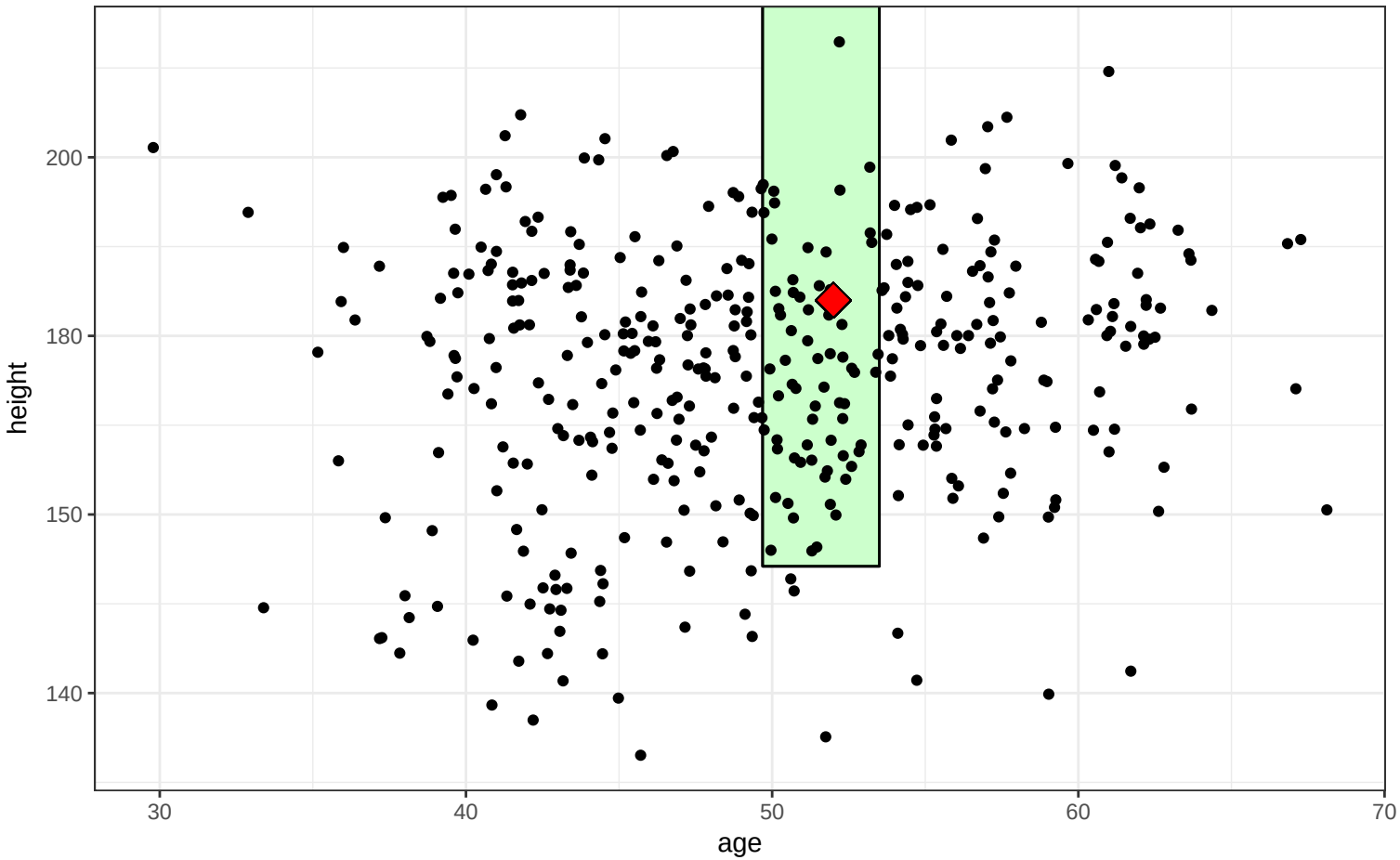


Honesty (Biau and Scornet, 2015; Scornet, 2015)

A tree estimator is **honest** iff the tree structure does not depend on the Y values used for leaf predictions:

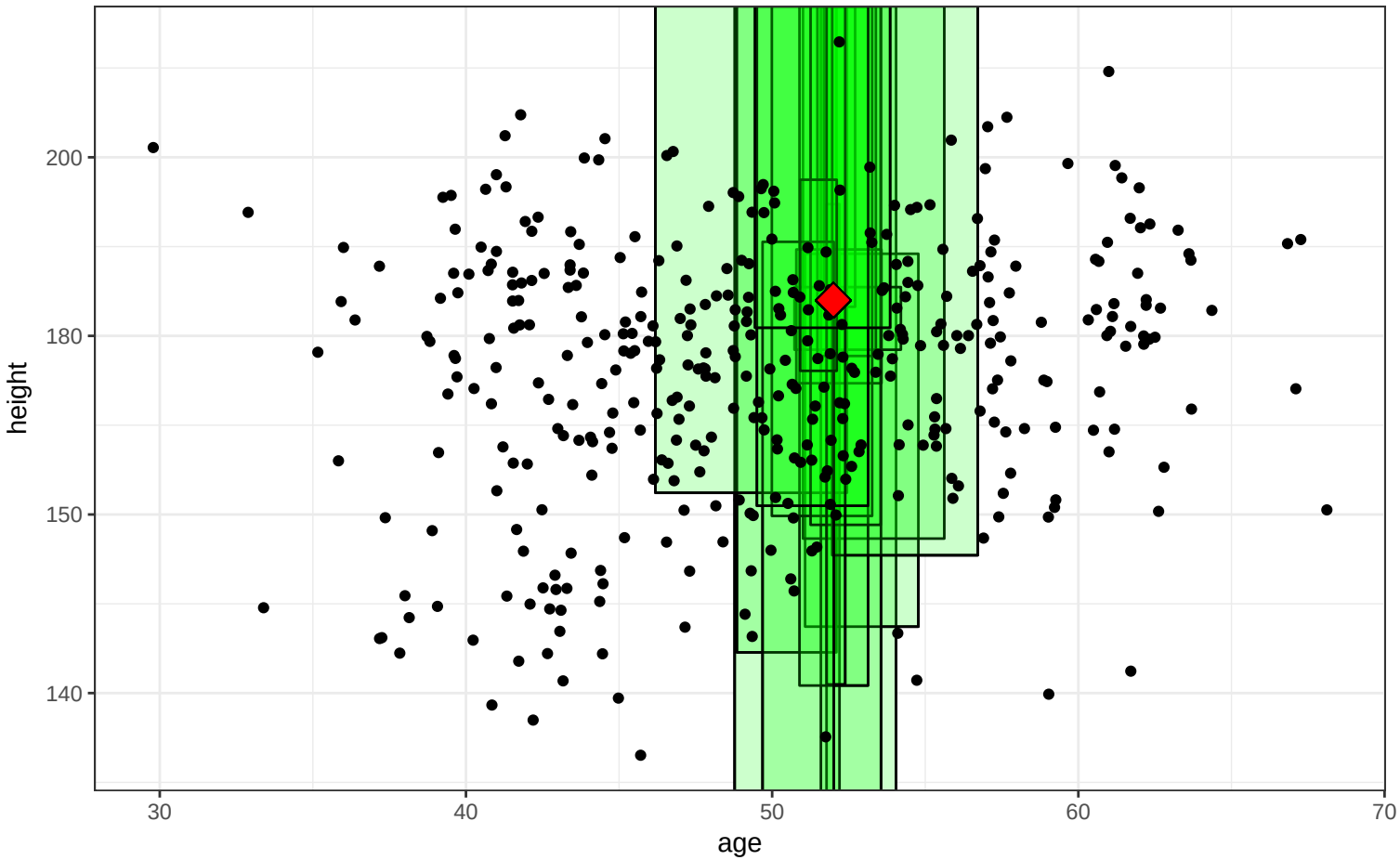
- Purely random tree
- Wager and Athey (2017) definition of Causal Forest: Split the data and use half of it to span the tree

The averaging effect of Random Forest



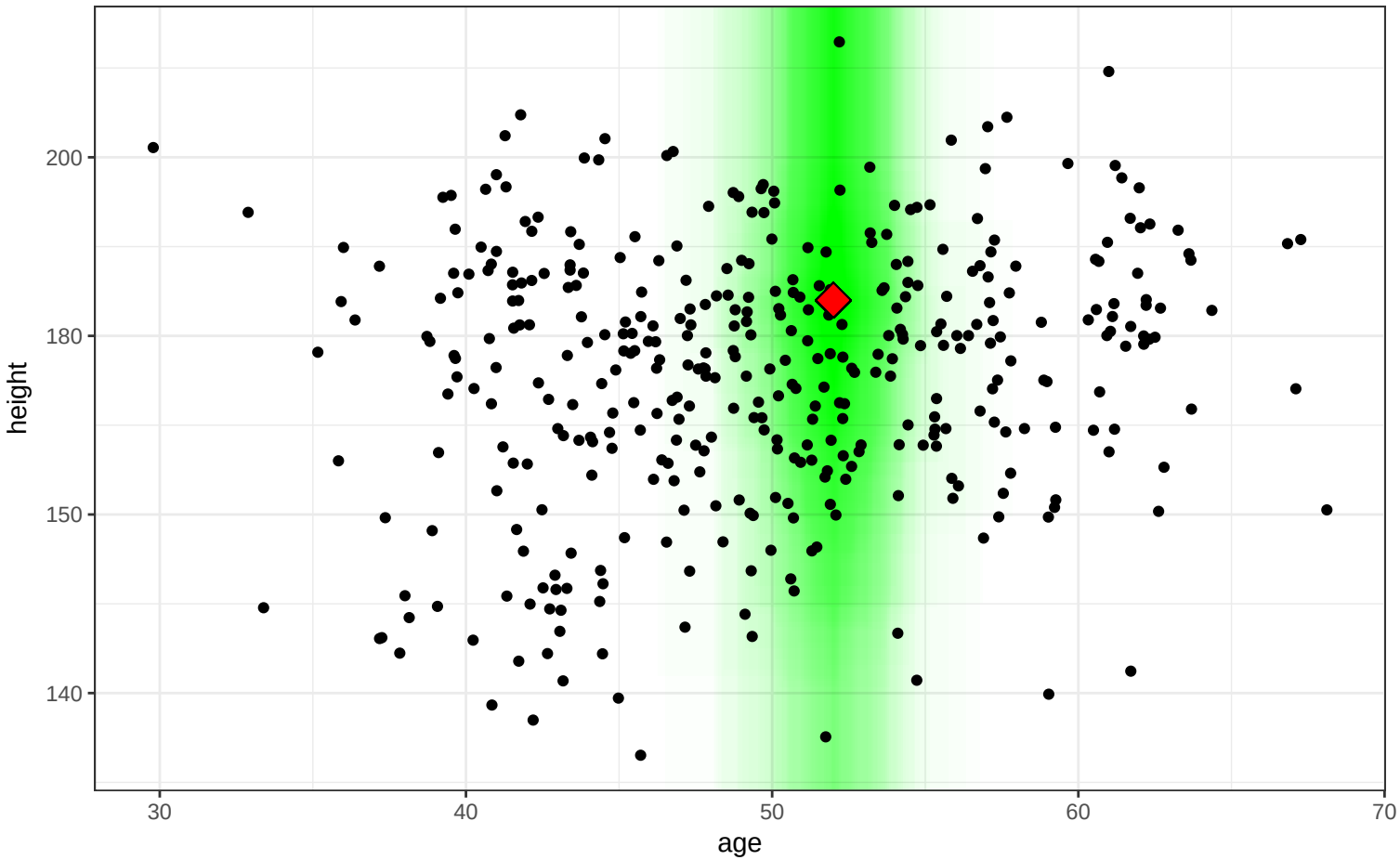
Back to RF

The averaging effect of Random Forest



Back to RF

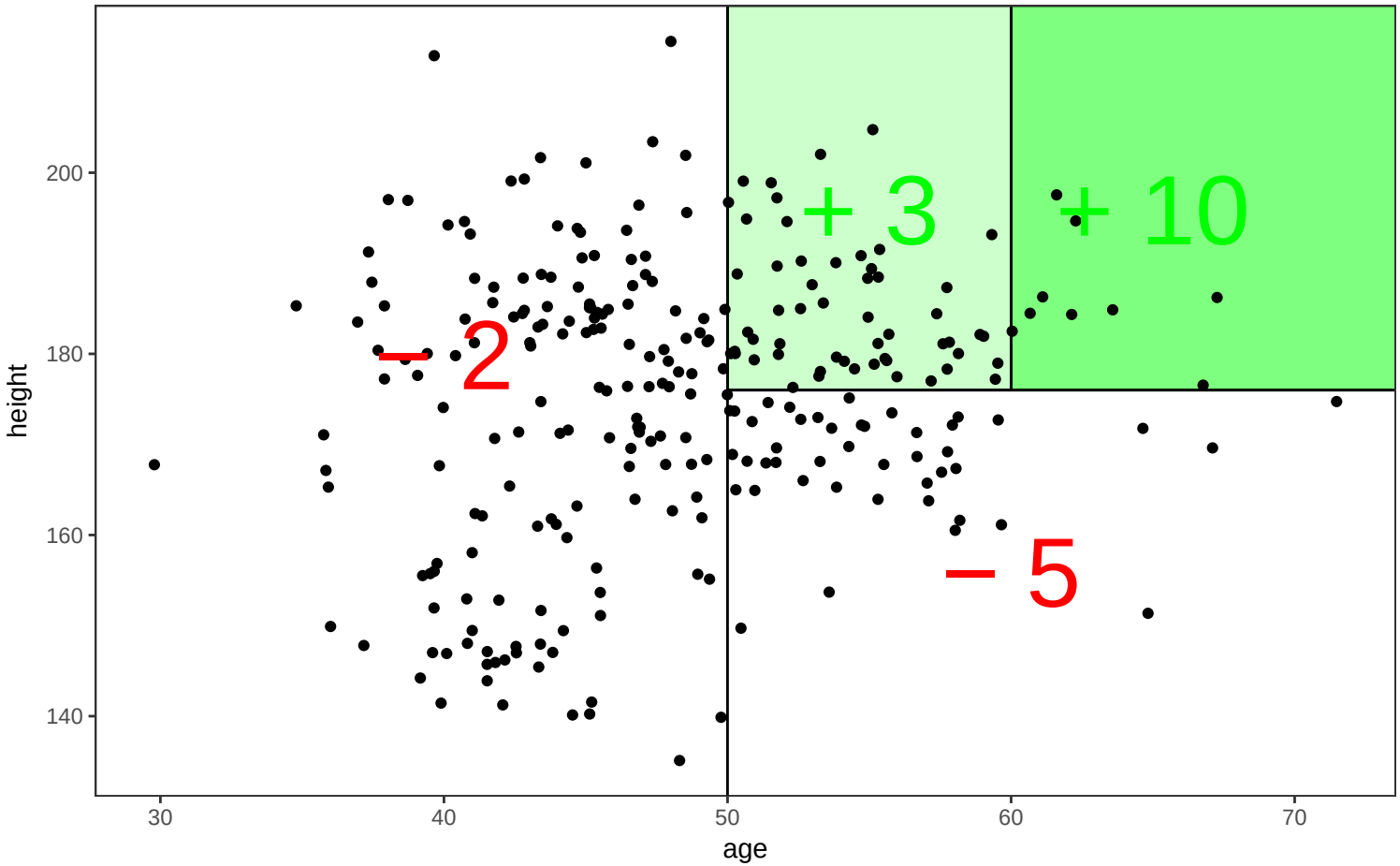
The averaging effect of Random Forest



Averaging leaves makes the weighing function of random forest smooth

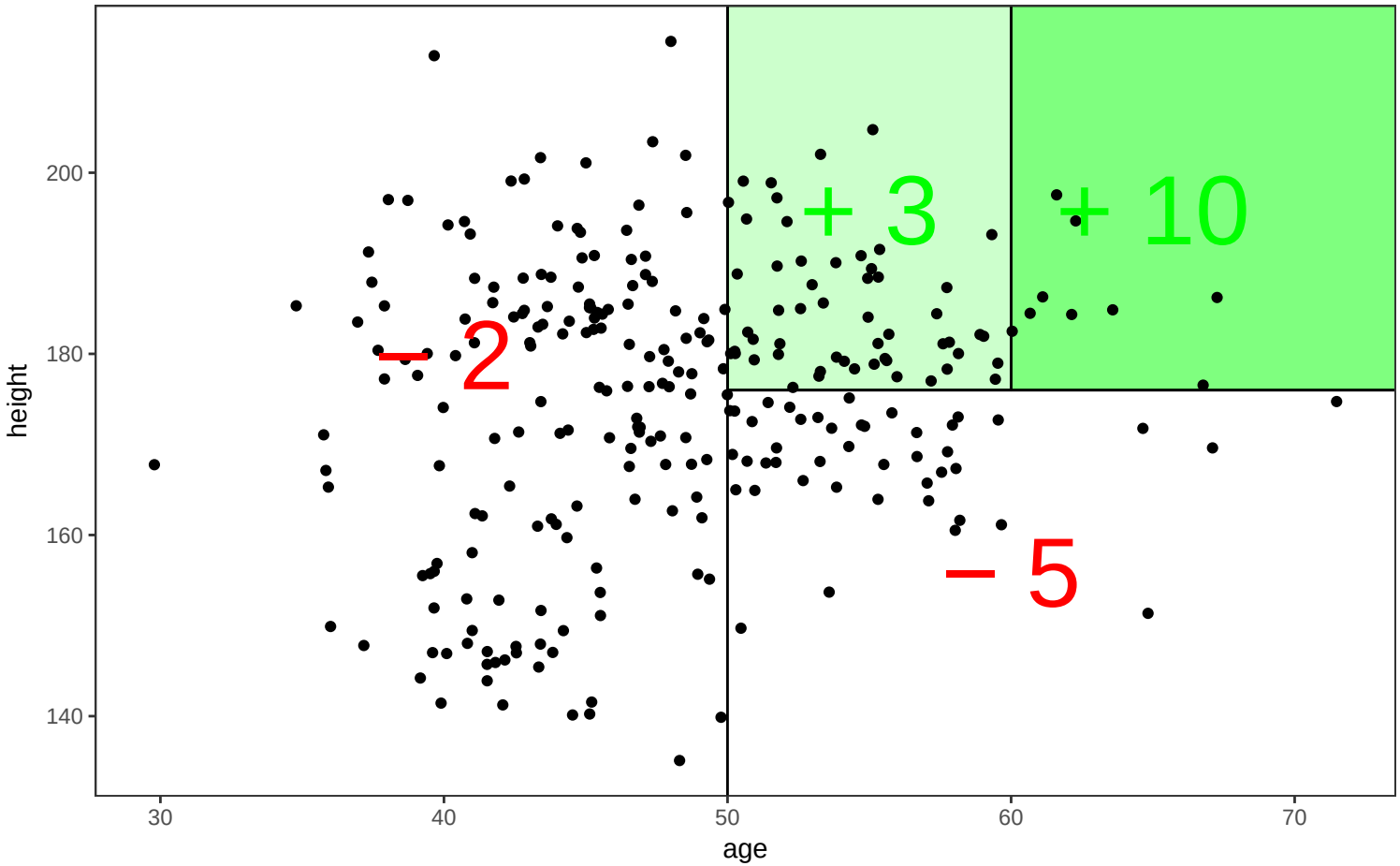
Back to RF

Honest versus adaptive fitting



Back to RF

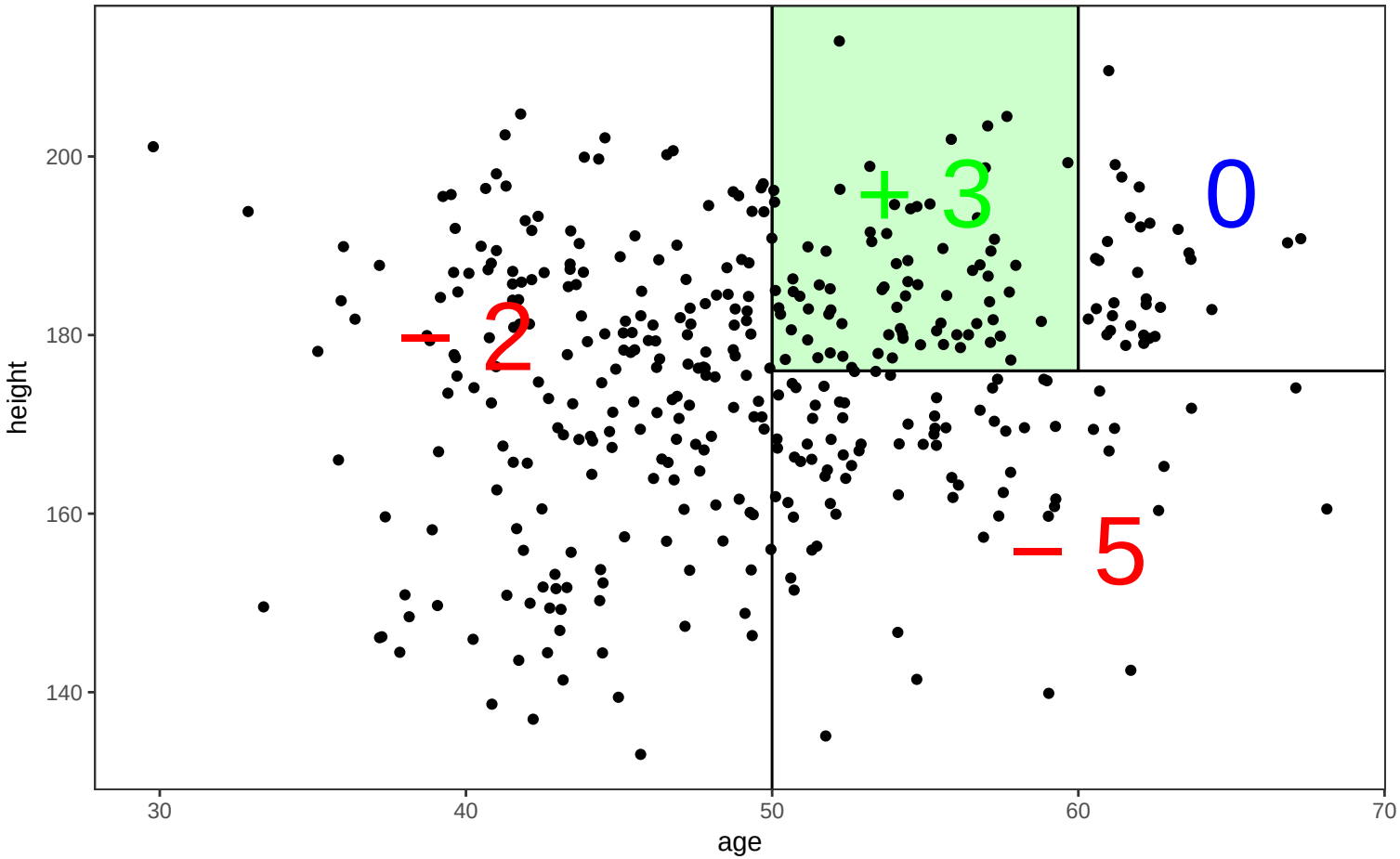
Honest versus adaptive fitting



Using the same data for the partitioning and the leaf estimates can lead to over-fitting

Back to RF

Honest versus adaptive fitting



Using the same data for the partitioning and the leaf estimates can lead to over-fitting

Back to RF

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 =$$

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 = \sigma^2 = \alpha$$

With one data point?

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 = \sigma^2 = \alpha$$

With one data point?

$$\begin{aligned} E(Y_i - Y_u)^2 &= E(Y_i - \mu + Y_u - \mu)^2 \\ &= E(Y_i - \mu)^2 + E(Y_u - \mu)^2 \\ &= 2\sigma^2 \\ &= 2\alpha \end{aligned}$$

General results for Cover-Hart class, which is a convex cone (Gneiting, 2012)
Back to CATE