Generalized Random Forest

Seminar Data Science for Economics

Madina Kurmangaliyeva
m.kurmangaliyeva@uvt.nl

Spring 2020

Tilburg University

https://madina-k.github.io/dse_mk2021

GRF application to the National Study of Learning Mindsets (Athey and Wager, 2019)

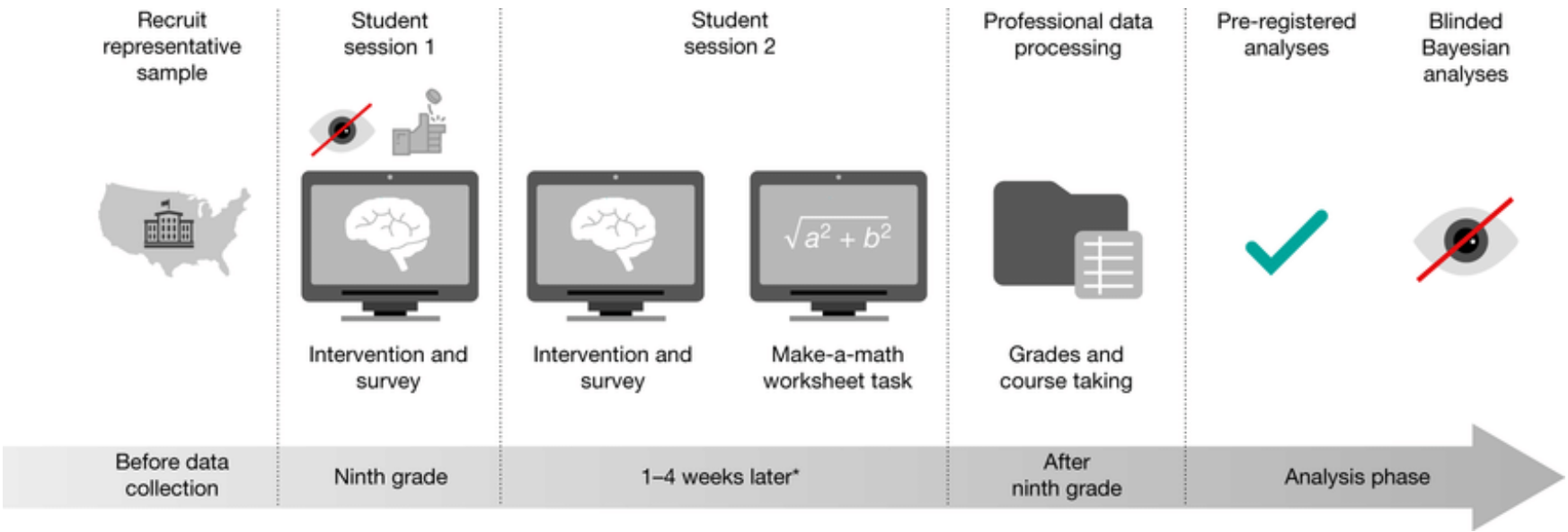
https://madina-k.github.io/dse_mk2021

National Study of Learning Mindsets

- Large RCT: a short online program designed to foster a growth mindset during the transition to high school
- a low-cost "nudge-like" intervention to change student behavior: "hard work can make you more intelligent" (**learning mindset**)
- as opposed to a **fixed mindset**: "abilities are fixed from birth"
- Primary purpose: seeks to understand heterogeneity in the intervention.

A national experiment reveals where a growth mindset improves achievement

David S. Yeager[1]*, Paul Hanselman[2]*, Gregory M. Walton[3], Jared S. Murray[1], Robert Crosnoe[1], Chandra Muller[1], Elizabeth Tipton[4], Barbara Schneider[5], Chris S. Hulleman[6], Cintia P. Hinojosa[7], David Paunesku[8], Carissa Romero[9], Kate Flint[10], Alice Roberts[10], Jill Trott[10], Ronaldo Iachan[10], Jenny Buontempo[1], Sophia Man Yang[1], Carlos M. Carvalho[1], P. Richard Hahn[11], Maithreyi Gopalan[12], Pratik Mhatre[1], Ronald Ferguson[13], Angela L. Duckworth[14] & Carol S. Dweck[3]



https://madina-k.github.io/dse_mk2021

Variables

Y	a continuous measure of achievement
W	a binary treatment
S3	Student: self-reported expectations for success in the future
C1	Student: race/ethnicity
C2	Student: gender
C3	Student: first-generation status
XC	School: urbanicity (i.e. rural, suburban, etc.)
X1	School: mean of students' fixed mindsets
X2	School: achievement level, as measured by test scores and college preparation for the previous 4 cohorts of students
X3	School: racial/ethnic minority composition
X4	School: poverty concentration
X5	School: size - Total # of students

Questions

The main research questions to be addressed include:

1. Was the mindset intervention effective in improving student achievement?
2. Test hypothesis: the effect is moderated by school-level pre-existing mindset norms (X1) and achievement (X2):
    - Effect largest in middle-achieving schools or is decreasing in school-level achievement
3. Exploring possible role of other variables in moderating treatment effects.

Challenges and the GRF

- Multiple hypothesis testing outside a pre-analysis plan (General)
- Observational study: Originally an RCT, but synthetic data led to corr(W, S3)≠ 0 (General)
- Clustering at school level (GRF-specific)

Generalized Random Forests:

- uses Random Forests machinery
- non-parametrically estimates treatment effects
- captures heterogeneity in treatment effects in a data-driven manner

Results using GRF

Using Generalized Random Forest (Athey, Tibshirani, and Wager, 2019):

- Treatment had a large positive effect on average
- **No** evidence of strong heterogeneity in treatment effects
- **Some** evidence that $X1$ – school-level mean of student's fixed mindsets
- **No** evidence that $X2$ – school achievement level – mediates the effect

All achieved with just a few lines of code in R using *grf* package.

Plan for today

1. Brief recap of random forests
2. random forests $\longrightarrow$ causal trees $\longrightarrow$ generalized random forests
3. `grf` package commands in R and the mindset program evaluation

https://madina-k.github.io/dse_mk2021

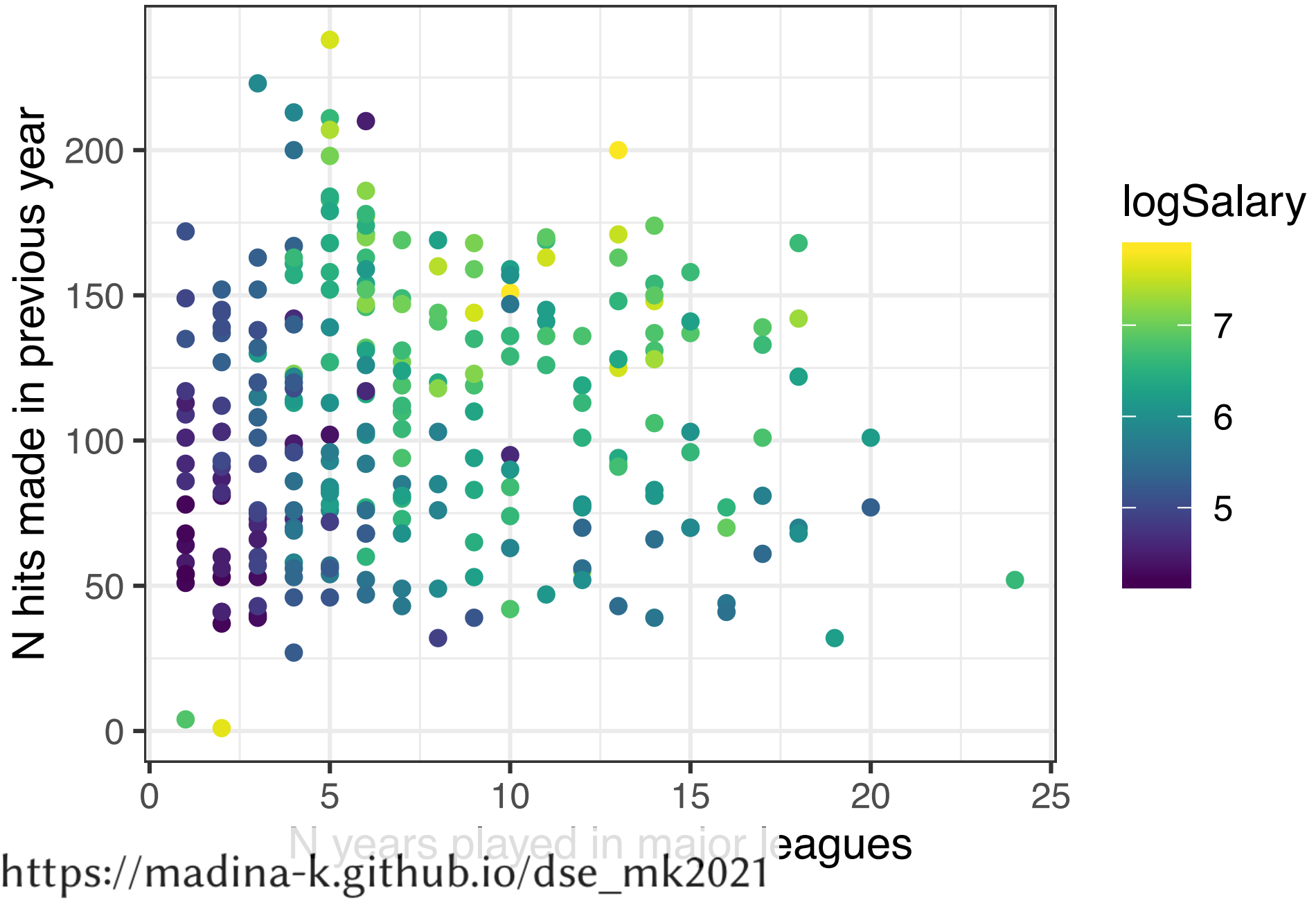
Decision trees (briefly)

Decision trees

Decision trees:

- are very popular for prediction tasks
- help to **partition**/**segment** the predictor space
- ... for both **regression** (i.e., continuous target variables) and **classification** (categorical target variables)

Baseball players' salary plot: Color indicates log(Salary)

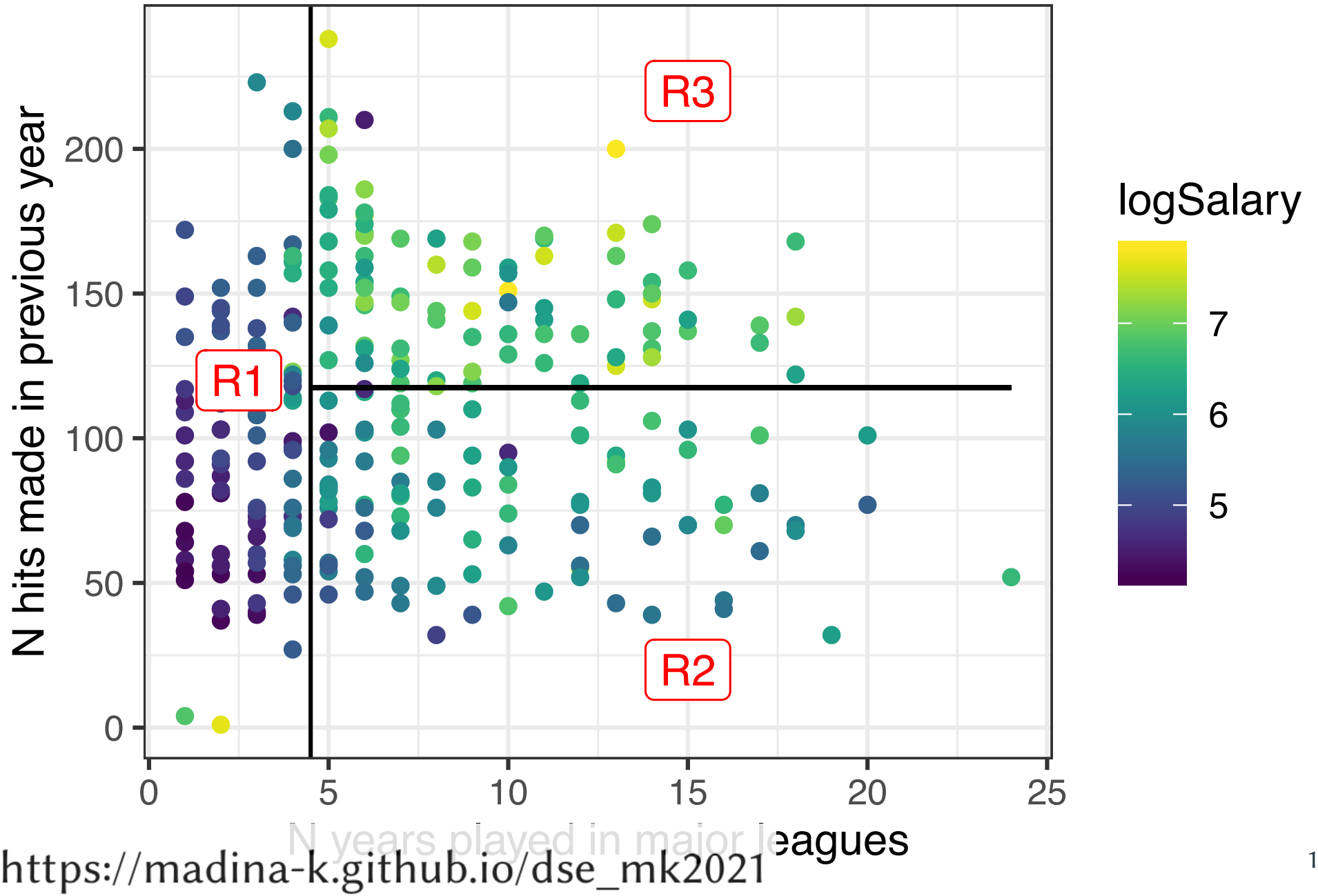


https://madina-k.github.io/dse_mk2021

Baseball players' salary: Decision tree

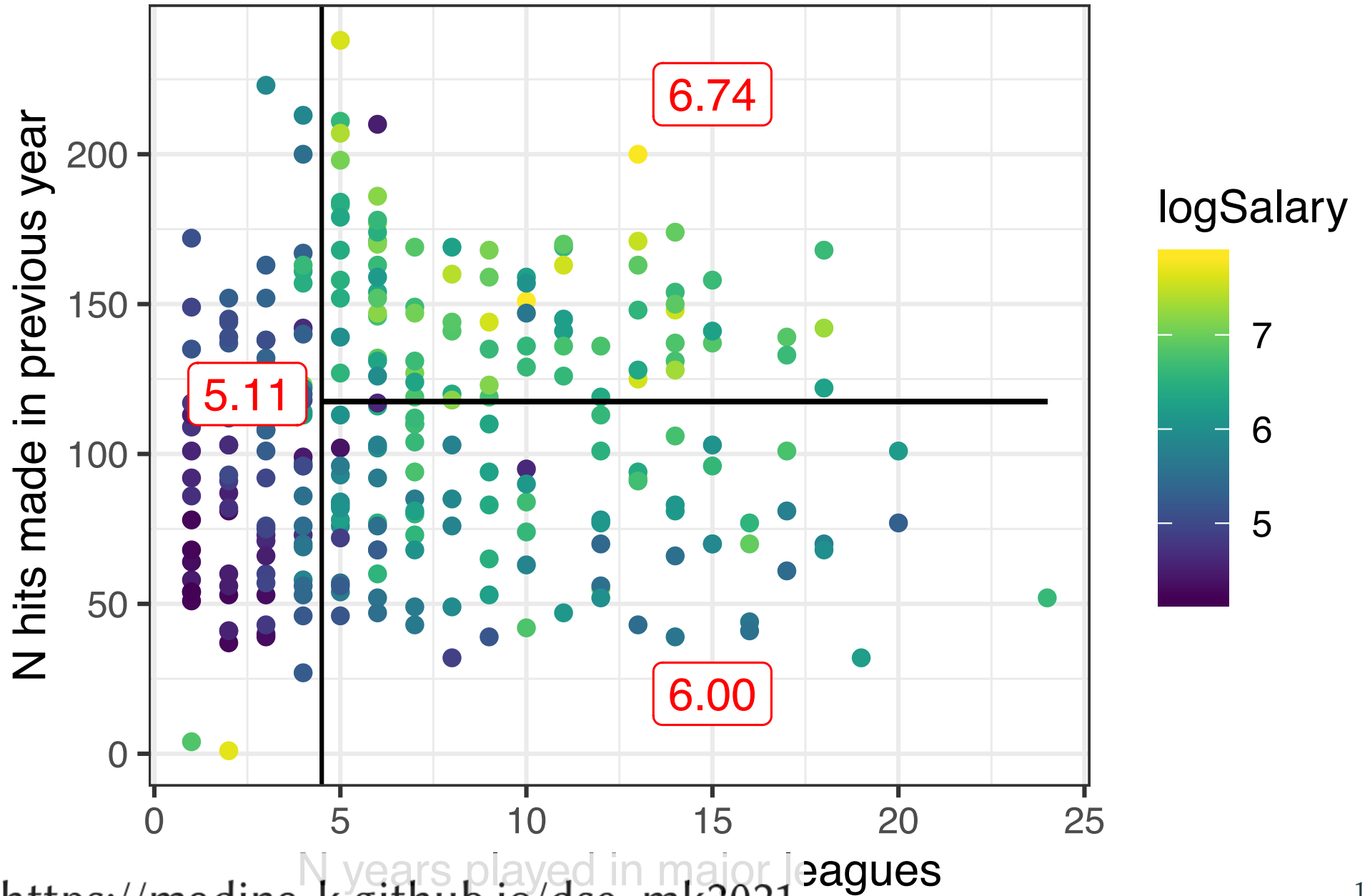
If we train a decision tree to predict the log(Salary) based on the player's characteristics, this is an example of tree that we may get:

Baseball players' salary: regions R_j, $j \in \{1, 2, 3\}$



https://madina-k.github.io/dse_mk2021

Baseball players' salary: predictions $\hat{y}_{R_j}, j \in \{1, 2, 3\}$



https://madina-k.github.io/dse_mk2021

Prediction

Hence, for any new player we predict his log-salary depending on which region (R_1, R_2, or R_3) he belongs to, i.e.,:

$$\hat{y}_i(X_i) = \hat{y}_{R_j} \text{ such that } X_i \in R_j, j \in \{1, 2, 3\} \tag{1}$$

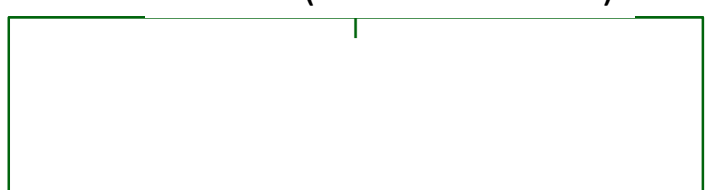
How does the tree decide to split data into regions?

The goal is to find the best possible partition of X-space into $R_1, \ldots, R_J$ that minimizes the RSS:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \underbrace{\hat{y}_{R_j}}_{\text{prediction for } R_j})^2 \tag{2}$$

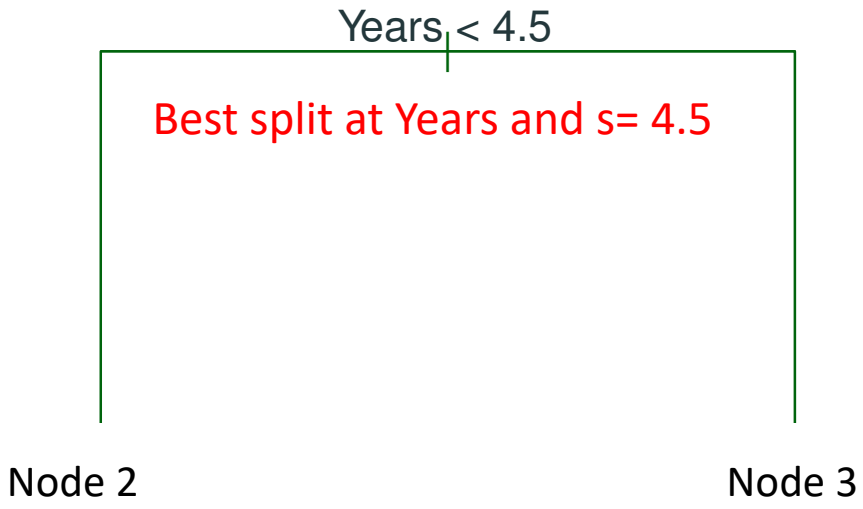
We search for the best partition using **recursive binary splitting.**

STEP 1

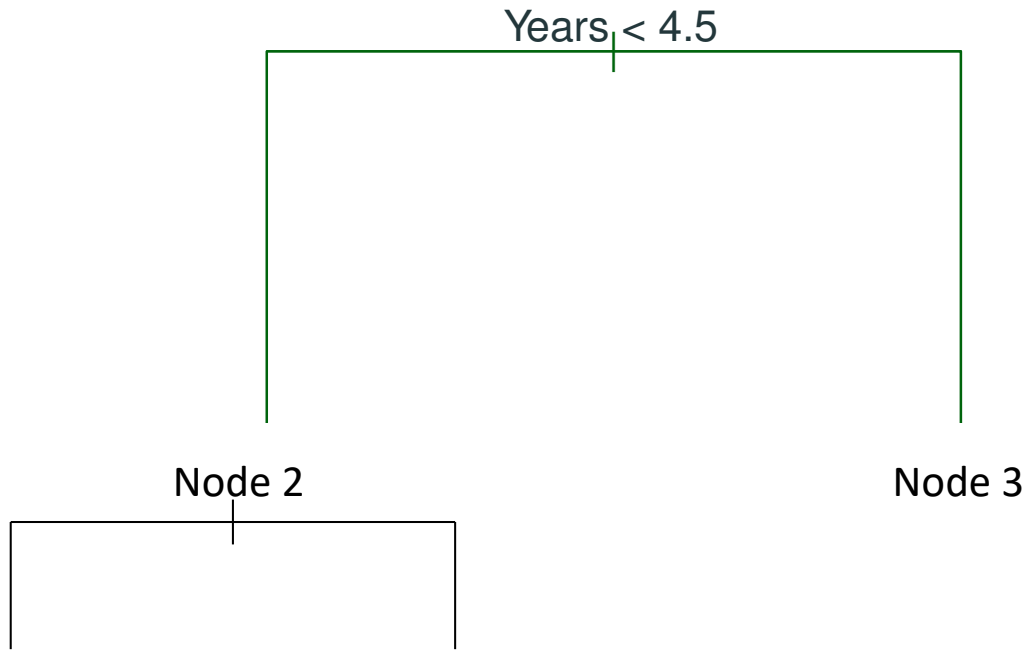


Best split at Years or at Hits? At which cutpoint?

STEP 1

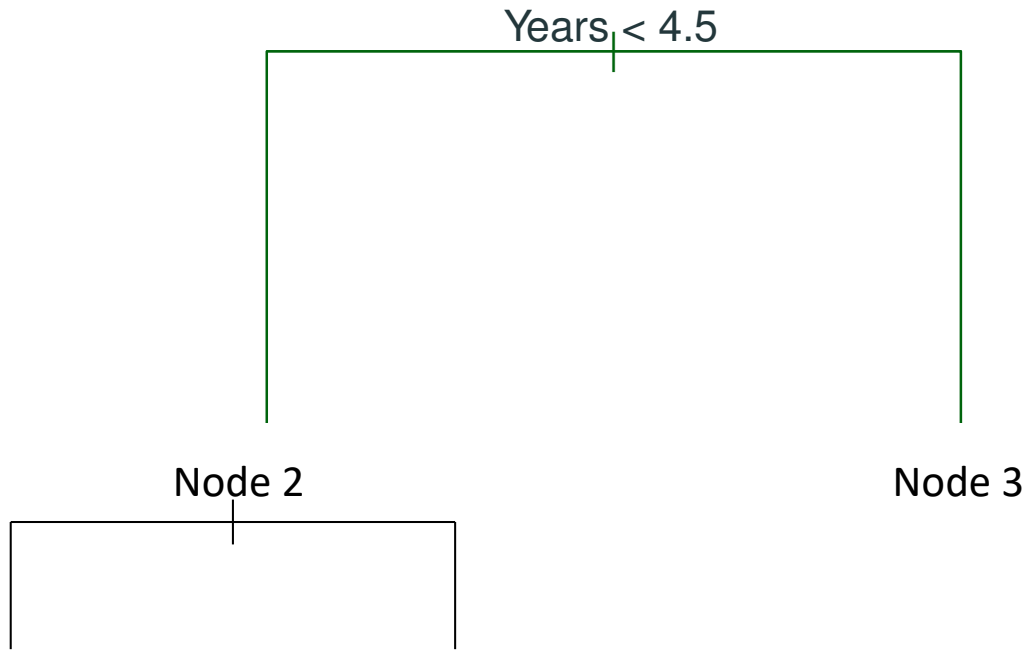


STEP 2



Best split at Years or at Hits? At which cutpoint?

STEP 2

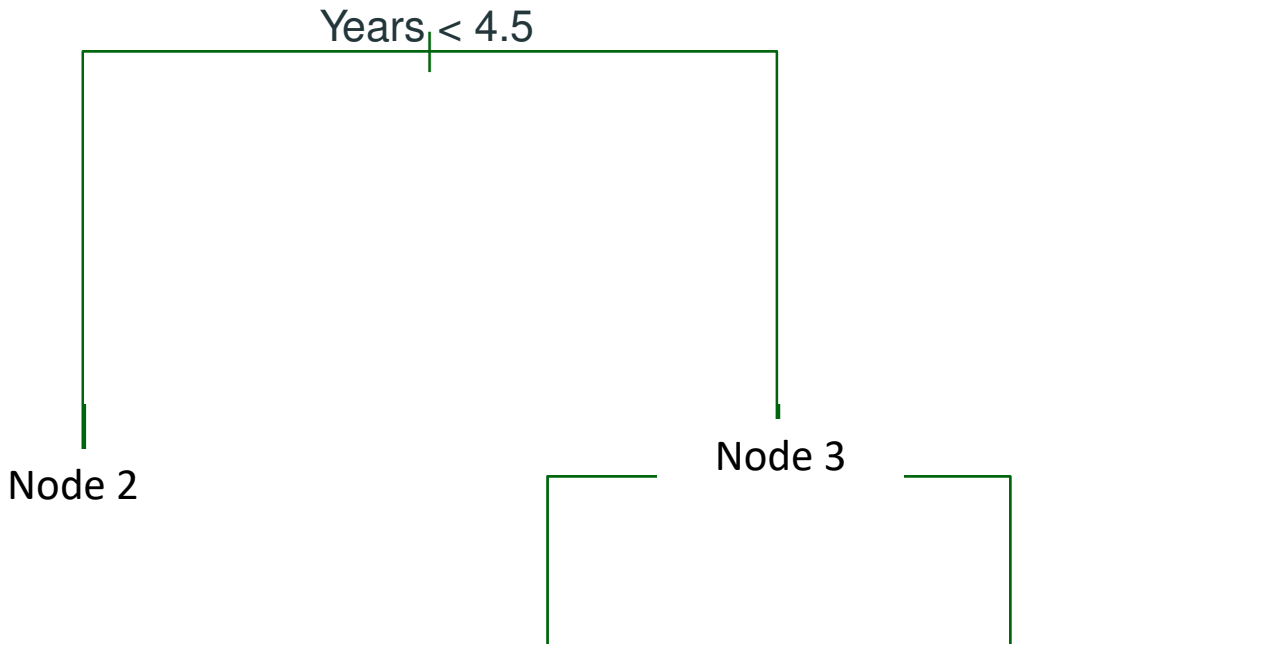


The best split does not pass through the stopping criteria, Node 2 becomes a leaf

STEP 2

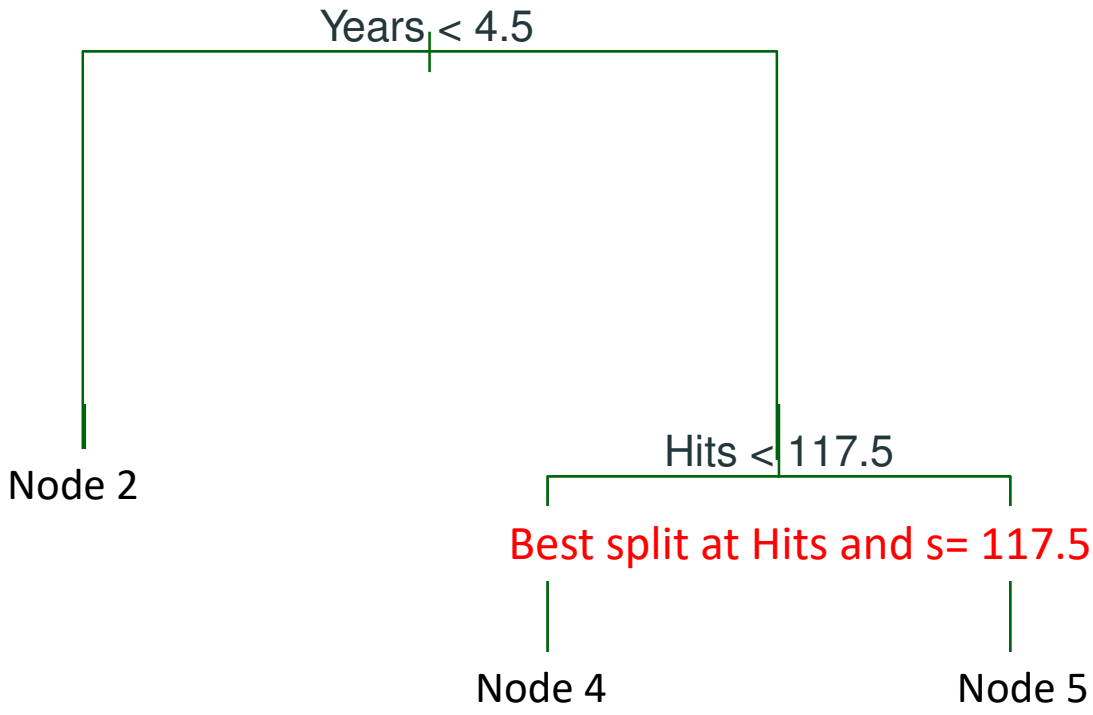
Terminal node

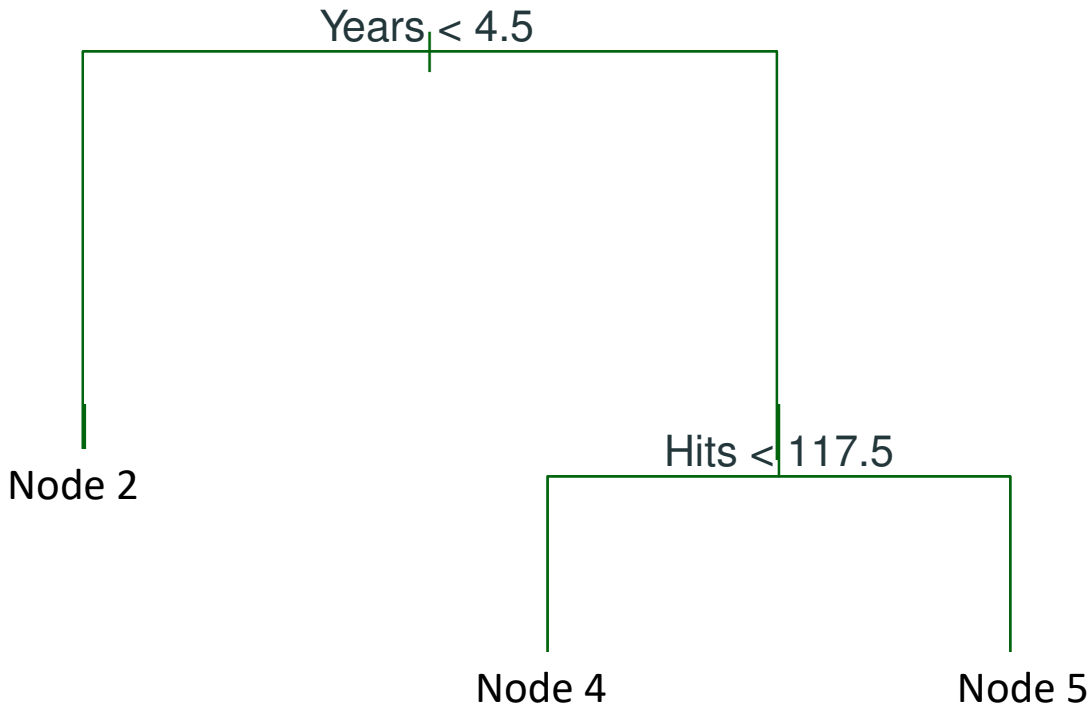
STEP 3



Best split at Years or at Hits? At which cutpoint?

STEP 3


Years < 4.5
Node 2
Hits < 117.5
Best split at Hits and s= 117.5
Node 4
Node 5

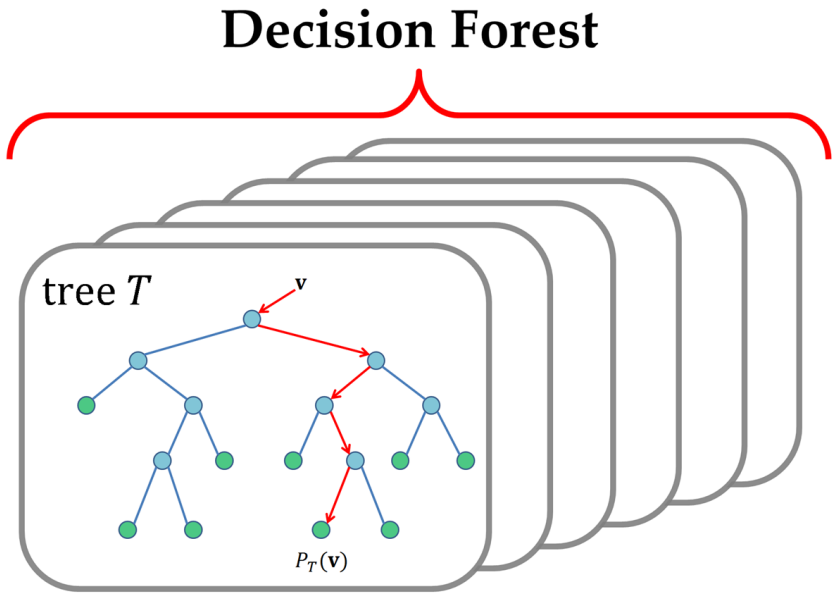


All end nodes are now terminal, algorithm stops

Calculate predictions

Finally, for each terminal node calculate the mean response value (target variable) for the observations in the training sample which fall into that terminal node.

Random Forest



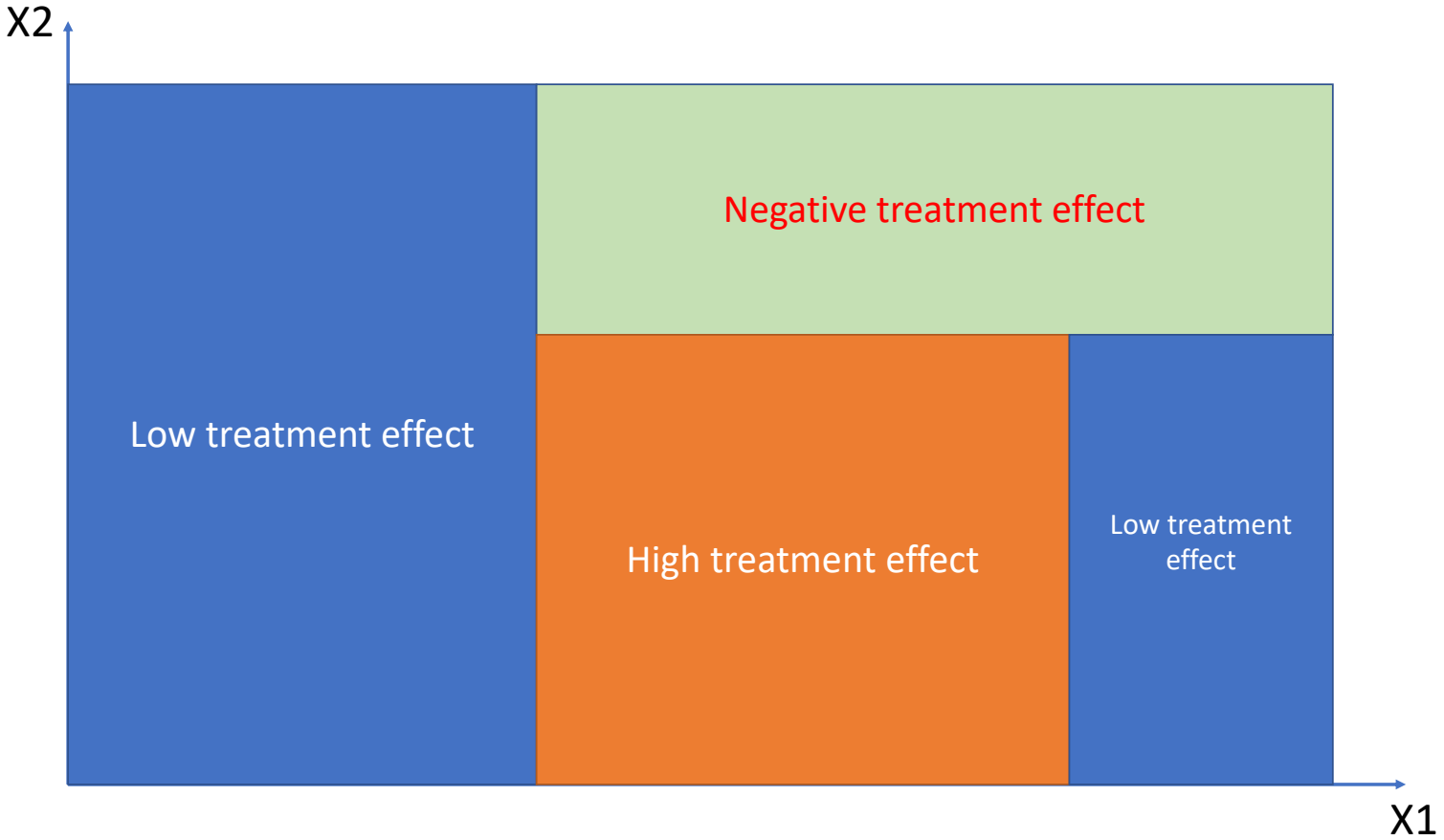
$$\hat{y}^{(RF)} = \underbrace{\sum_{t=1}^{T} \hat{y}_t}_{\text{average of decision trees}}$$

- Bootstrap: Draw T **random subsamples** of the training data $\Longrightarrow$ T trees
- At each split, consider only **some randomly-chosen predictors** as splitting candidates

A random forest for causal inference?

https://madina-k.github.io/dse_mk2021

Random forest and Causal inference



Suppose we want to partition covariate space based on the **heterogeneity in treatment effects**/CATE given X. Can we use a random forest?

https://madina-k.github.io/dse_mk2021

Cannot use Random Forest for treatment effects

If we had been able to observe the **counterfactuals** directly, ...

	$Y^{W=1}$	$Y^{W=0}$	τ_i	W_i	X_i
1	$y_1^{W=1}$	$y_1^{W=0}$	τ_1	0	x_1
2	$y_2^{W=1}$	$y_2^{W=0}$	τ_2	1	x_2
3	$y_3^{W=1}$	$y_3^{W=0}$	τ_3	0	x_3
...	...	...	...	...	...
n	$y_n^{W=1}$	$y_n^{W=0}$	τ_n	1	x_n

...then we would have been able to simply plug τ_i as the target variable in a Random forest.

Reality

In reality, the **counterfactuals are missing**

	$Y^{W=1}$	$Y^{W=0}$	τ_i	W_i	X_i
1	?	$y_1^{W=0}$	?	0	x_1
2	$y_2^{W=1}$	?	?	1	x_2
3	?	$y_3^{W=0}$	?	0	x_3
...	...	...	...	...	...
n	$y_n^{W=1}$	?	?	1	x_n

So we cannot use RF, we do not observe τ_i!

Naive and wrong approach

Some people wrongly use a RF to simply learn two predictions and take the difference

$$\hat{\tau}^{(RF)}(x) \equiv \underbrace{\hat{y}^{(RF)}(W = 1, X = x) - \hat{y}^{(RF)}(W = 0, X = x)}_{\textbf{WRONG}}$$

Even under an RCT, $\hat{\tau}^{(RF)}(x)$ estimator will be **biased**

Why? Because **we used the same data** to decide how **to partition** into leaves and **to estimate** mean values within each leaf.

In general

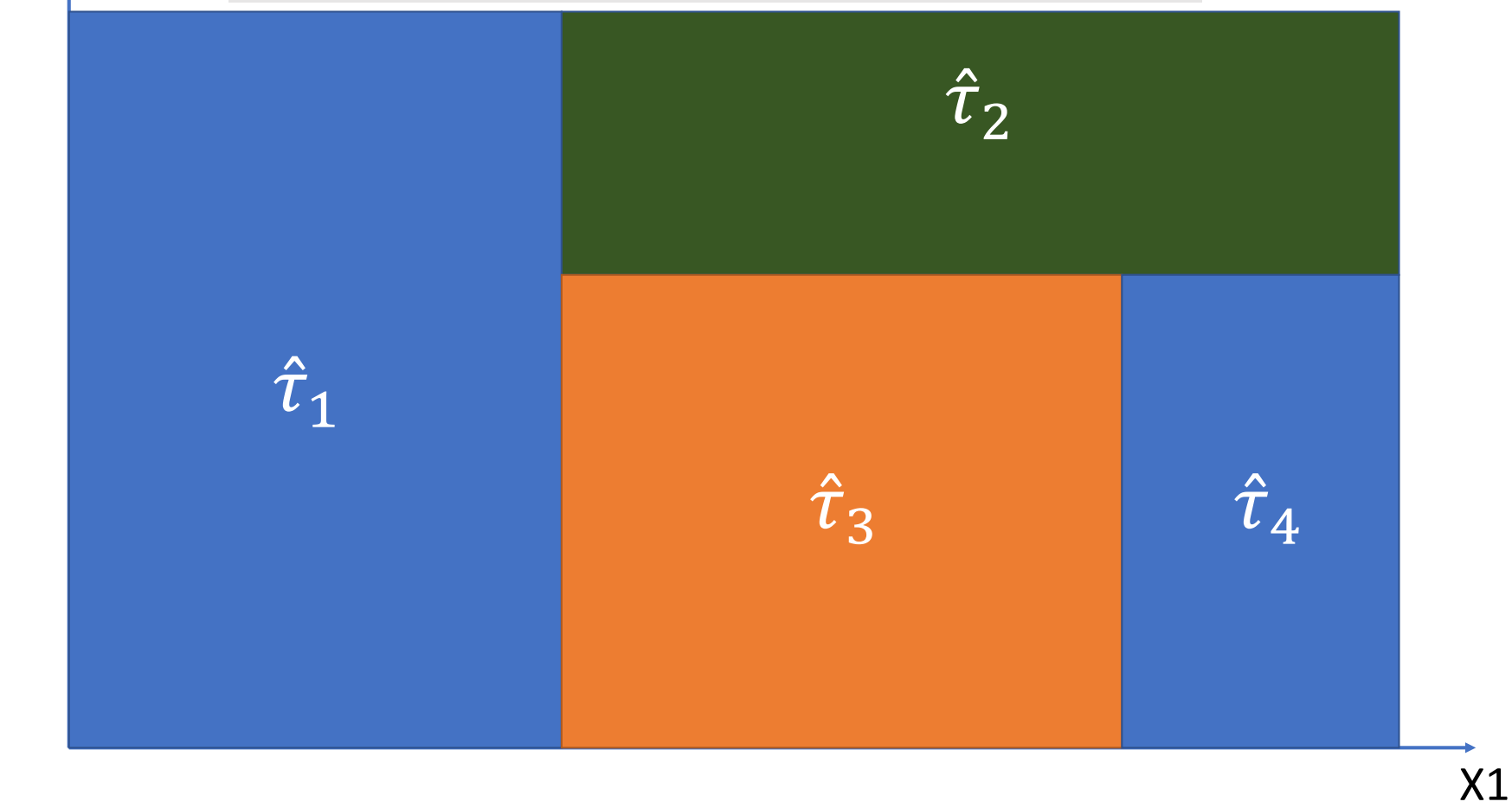
Prediction ≠ Causal inference

What do we want

We want to have a partition of X-space such that:

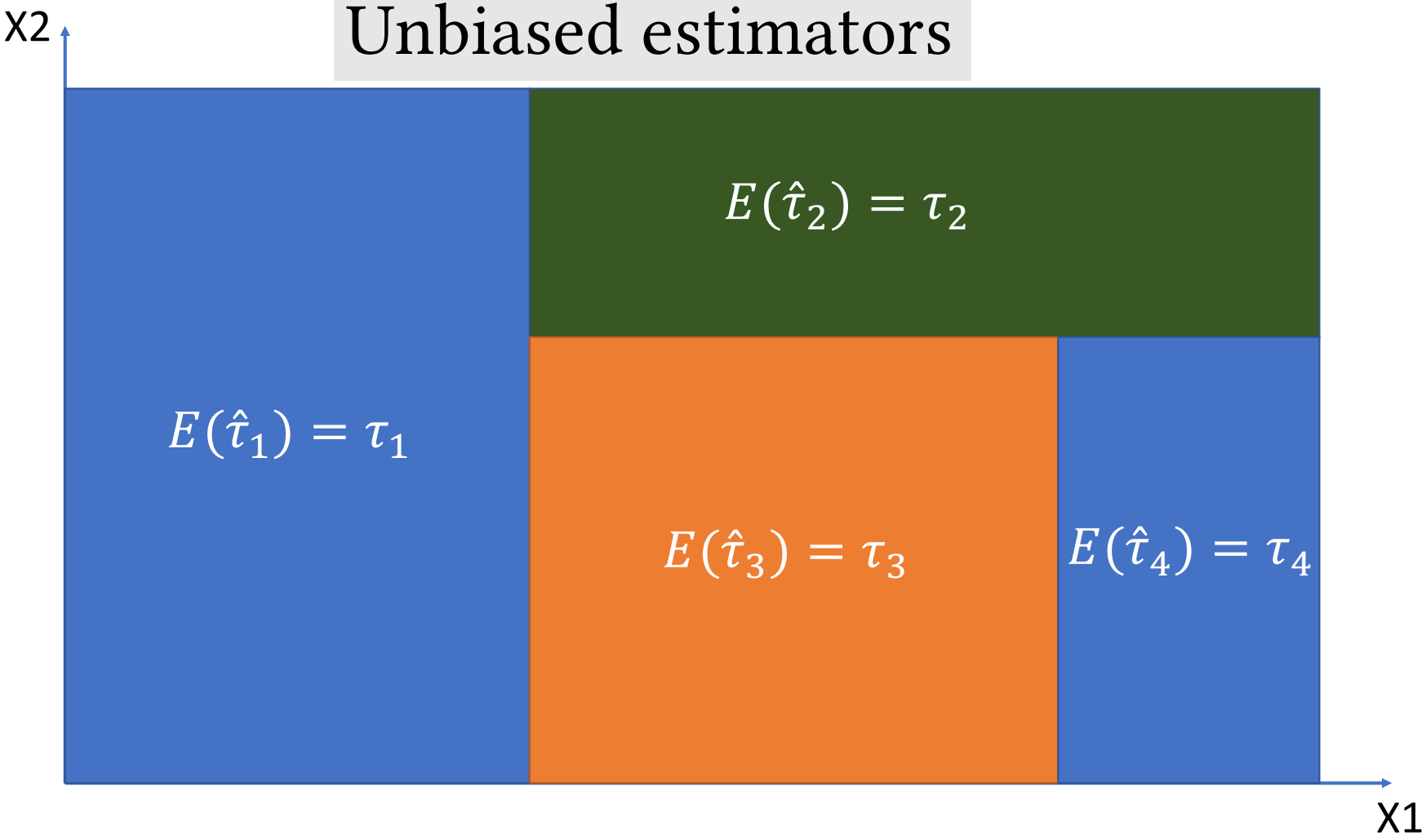
- captures **heterogeneity** in treatment effects
- **unbiased** estimate
- accounts for the need to estimate **standard errors** (not too narrow partitions)

Captures heterogeneity in TE

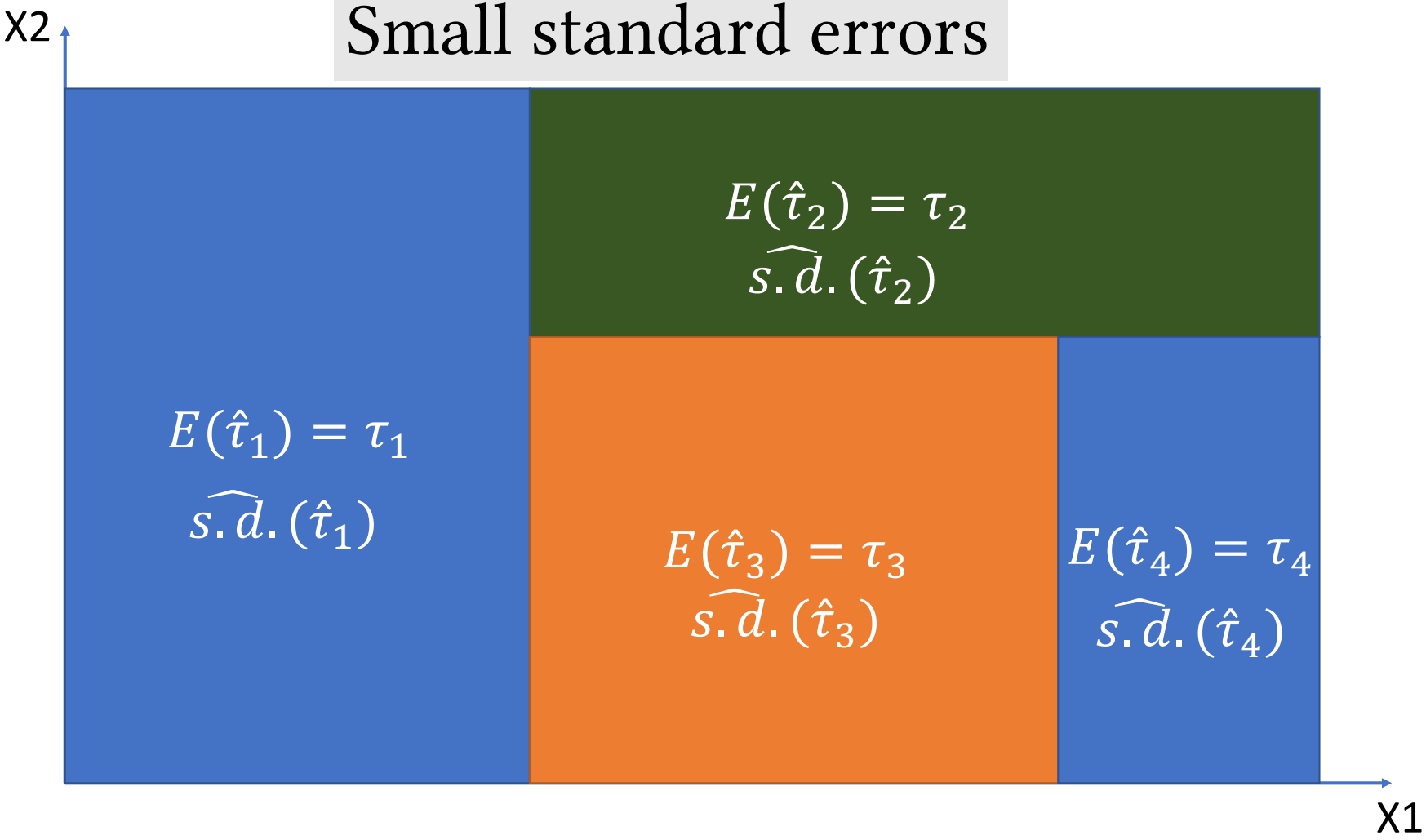


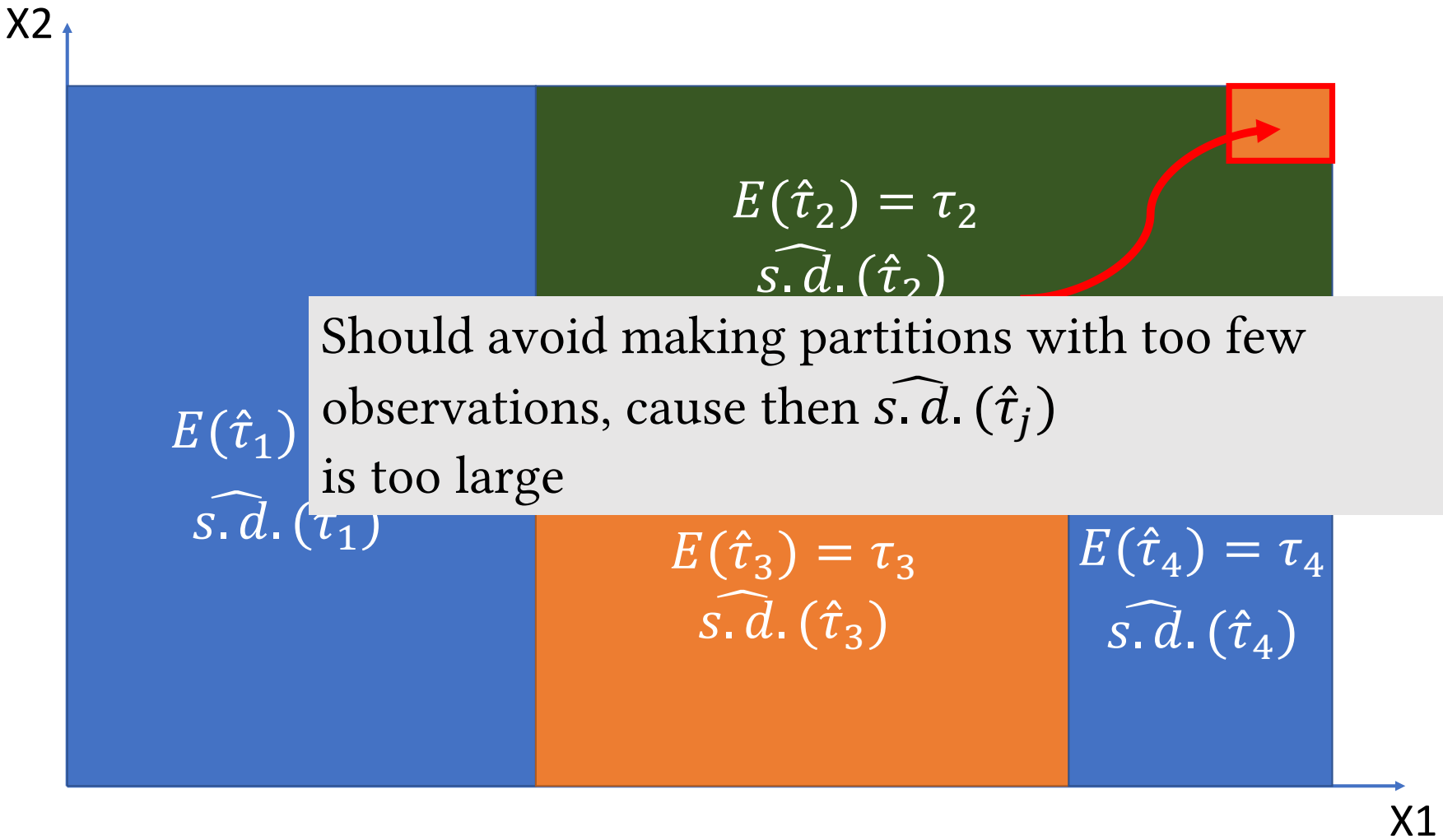
https://madina-k.github.io/dse_mk2021

Unbiased estimators



Small standard errors




X2
$E(\hat{\tau}_2) = \tau_2$
$\widehat{s.d.}(\hat{\tau}_2)$
Should avoid making partitions with too few observations, cause then $\widehat{s.d.}(\hat{\tau}_j)$ is too large
$E(\hat{\tau}_1)$
$\widehat{s.d.}(\hat{\tau}_1)$
$E(\hat{\tau}_3) = \tau_3$
$\widehat{s.d.}(\hat{\tau}_3)$
$E(\hat{\tau}_4) = \tau_4$
$\widehat{s.d.}(\hat{\tau}_4)$
X1

Causal Trees: RCT, binary treatment (Athey and Imbens, PNAS 2016)

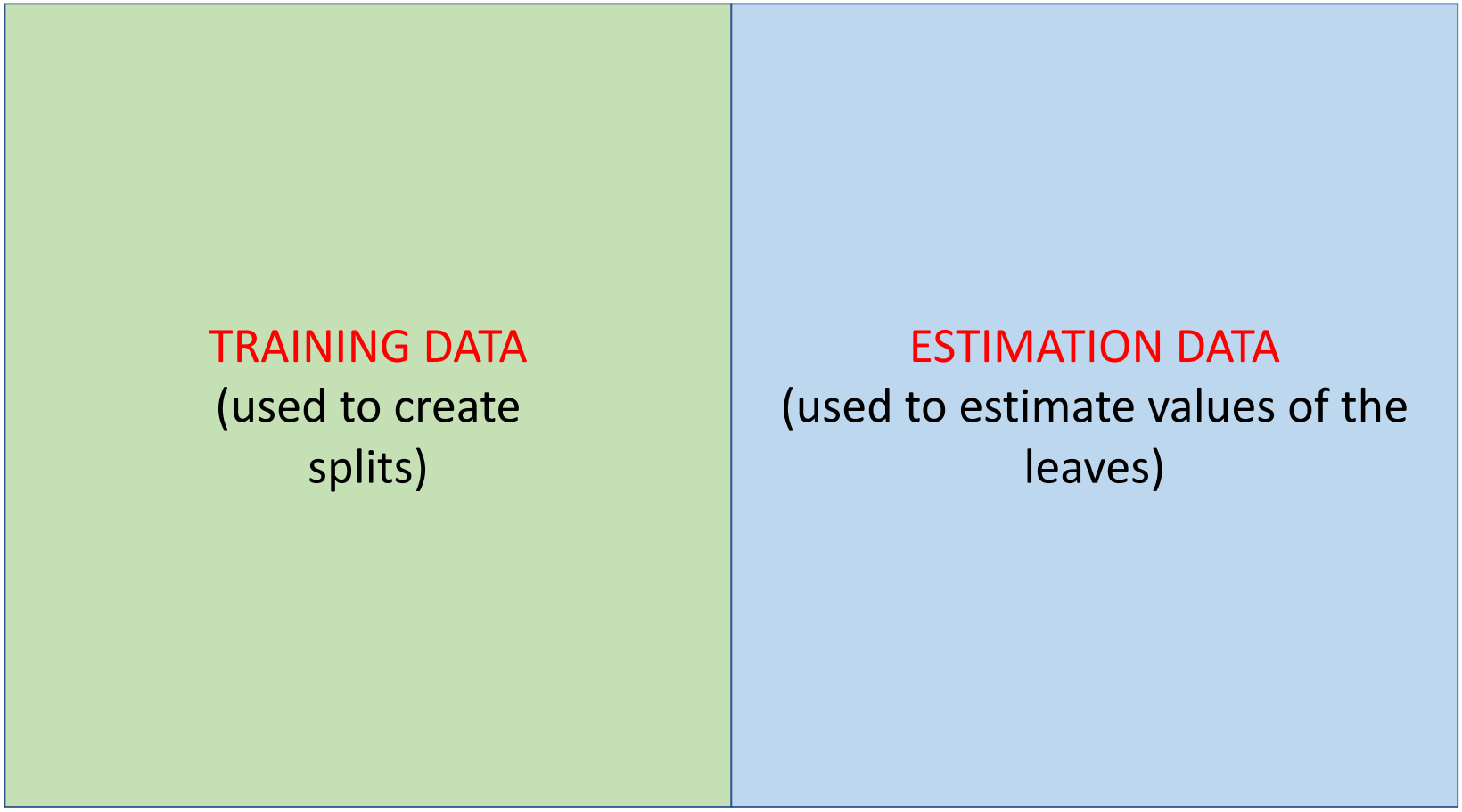
https://madina-k.github.io/dse_mk2021

Preview

Causal trees solve the bias problem by adopting Honest Splitting:

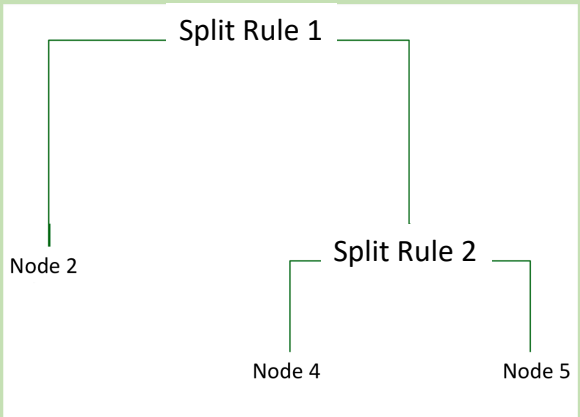
- uses an **independent sample to estimate** leaf means
- modifies the splitting functions to generate **unbiased estimates** of treatment effects, but
- accounting for the fact that finer leaves:
  - more unbiased treatment effects
  - higher standatd errors

Building block 1. Honest splitting



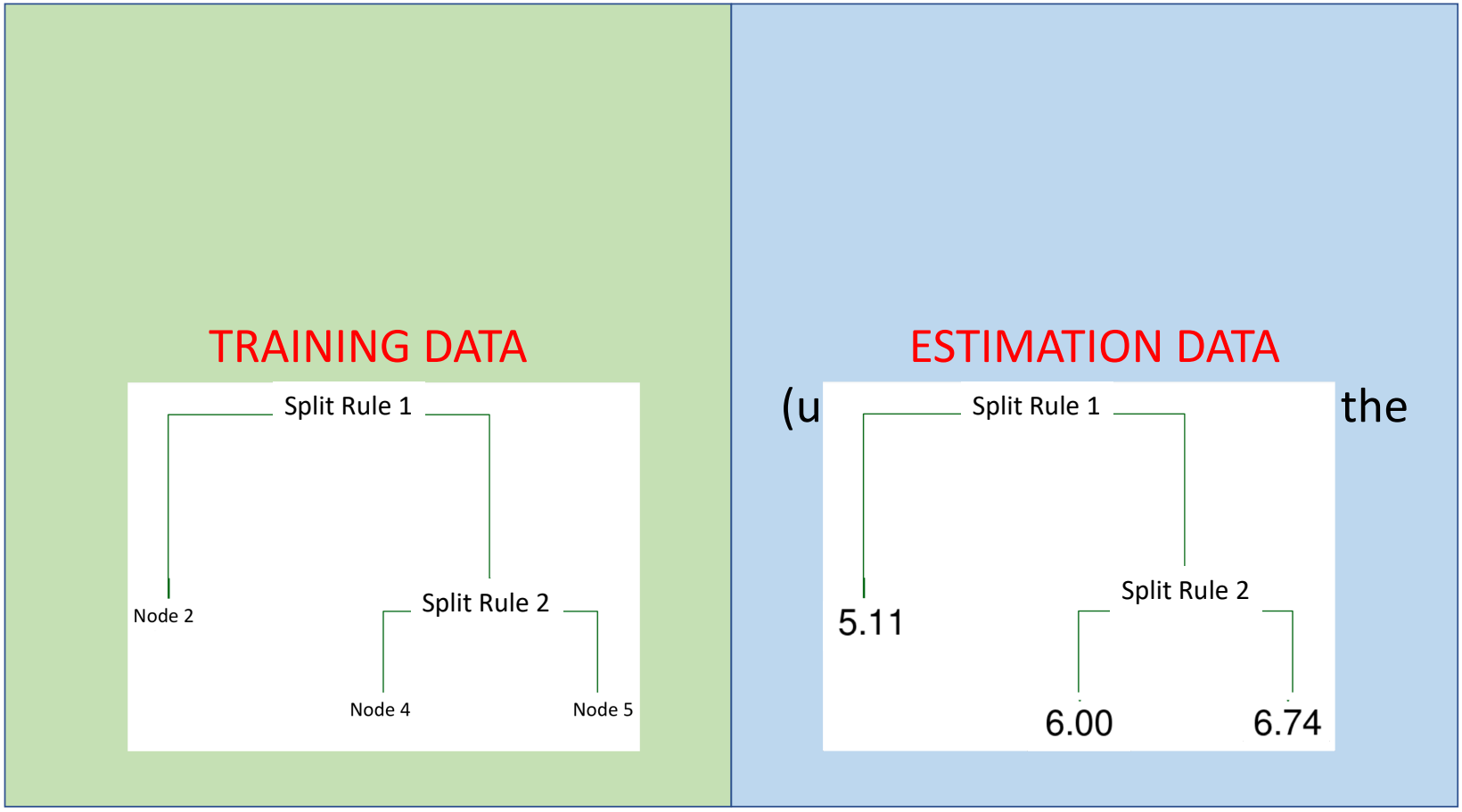
Building block 1. Honest splitting

TRAINING DATA



ESTIMATION DATA
(used to estimate values of the leaves)

Building block 1. Honest splitting



Normal decision trees' splitting criteria

$$\min \widehat{MSE}_{\mu}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi^{tr}) \equiv - \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})$$

Normal decision trees' splitting criteria

$$\min \widehat{MSE}_{\mu}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi^{tr}) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})$$

test set

Normal decision trees' splitting criteria

$$\min \widehat{MSE}_{\mu}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi^{tr}) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})$$

test set

training set

Normal decision trees' splitting criteria

$$\min \widehat{MSE}_{\mu}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi^{tr}) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})$$

test set

training set

tree partitions based on training set

Normal decision trees' splitting criteria

equivalent to maximizing the sum of the final nodes' squared predictions

$$\min \widehat{MSE}_{\mu}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi^{tr}) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})$$

test set

training set

tree partitions based on training set

Normal decision trees' splitting criteria

equivalent to maximizing the sum of the final nodes' squared predictions

$$\min \widehat{MSE}_{\mu}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi^{tr}) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})$$

test set

training set

tree partitions based on training set

It needs only the training sample and the number of observations in the training sample.

Building block 2. New objective function

Causal Trees:

$$\widehat{EMSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi^{tr}) \equiv \underbrace{-\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})}_{\text{rewards high heterogeneity}}$$

$$\underbrace{+ \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi^{tr}} \left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p}\right)}_{\text{Penalizes splits leading to small leafs}}$$

Building block 2. New objective function

Causal Trees:

$$\widehat{EMSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi^{tr}) \equiv \underbrace{-\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi^{tr})}_{\text{rewards high heterogeneity}}$$

$$+ \underbrace{\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi^{tr}} \left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p}\right)}_{\text{Penalizes splits leading to small leafs}}$$

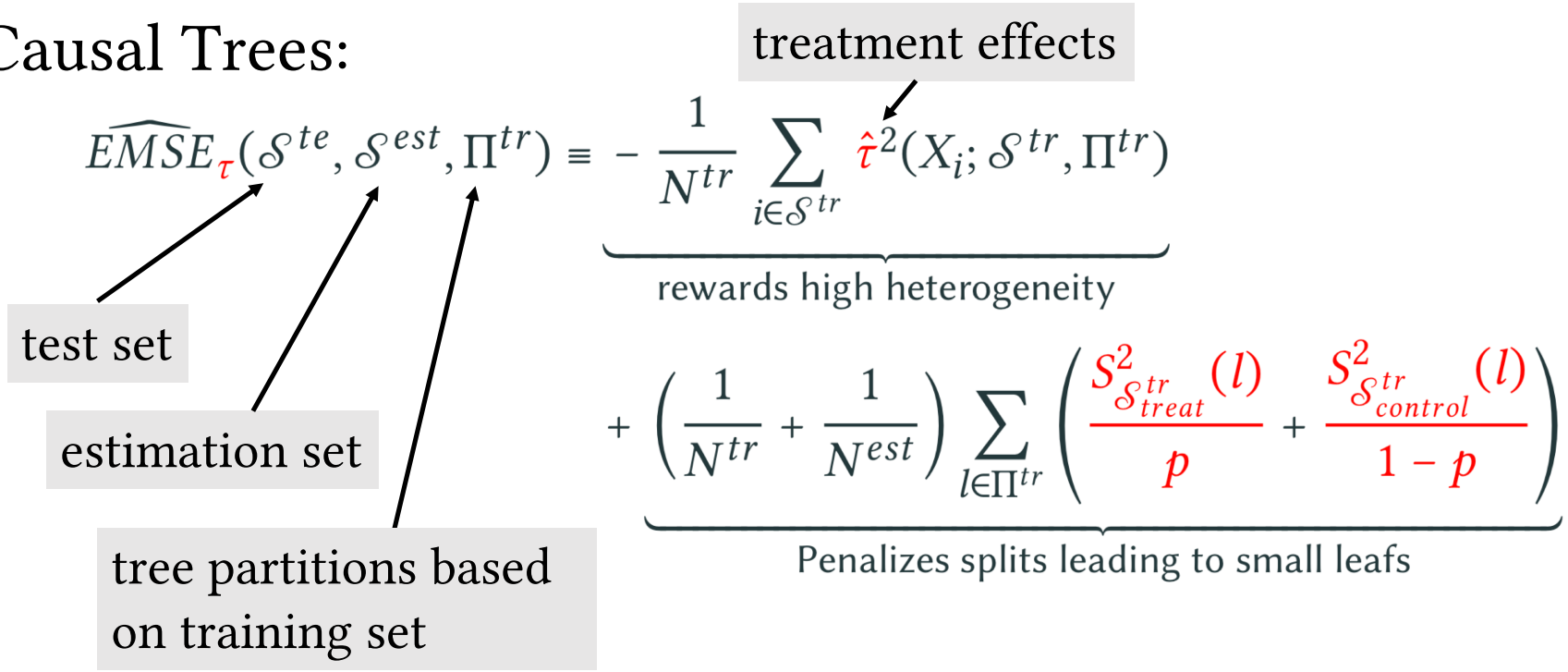
test set

estimation set

tree partitions based on training set

Building block 2. New objective function

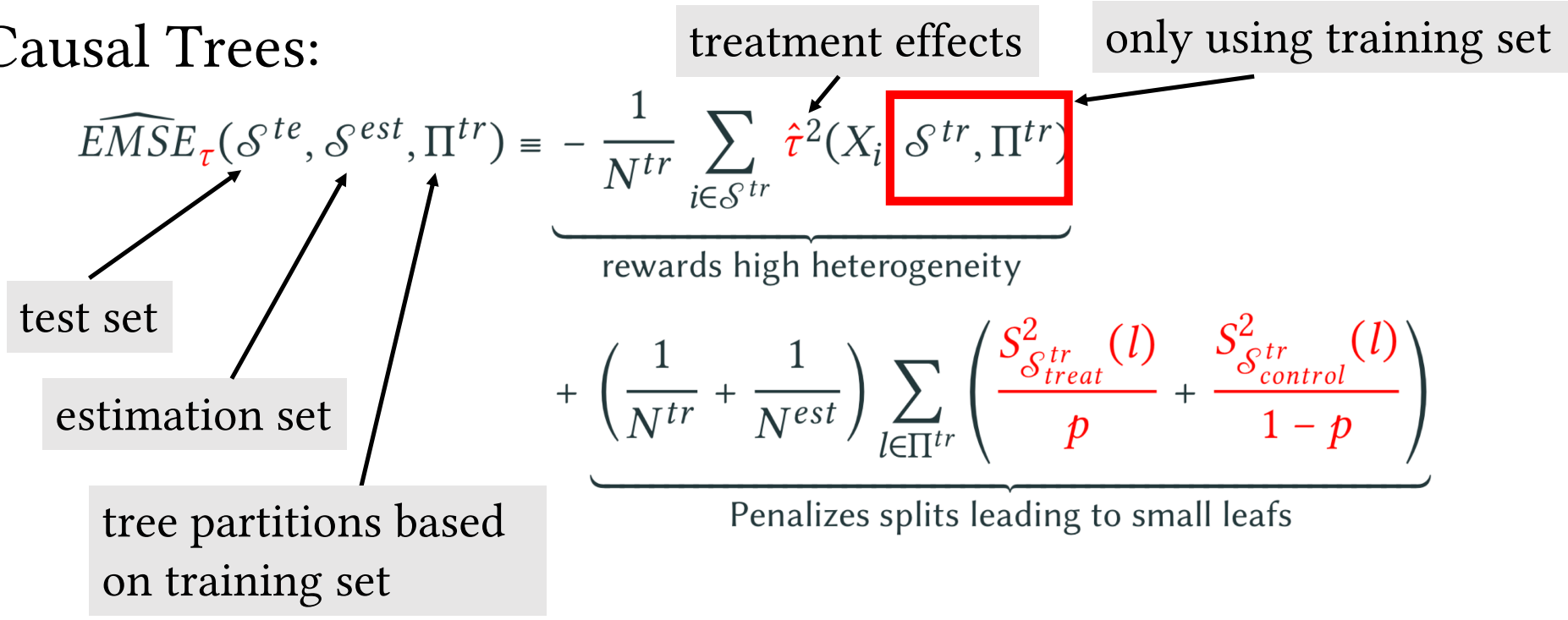
Causal Trees:



https://madina-k.github.io/dse_mk2021

Building block 2. New objective function

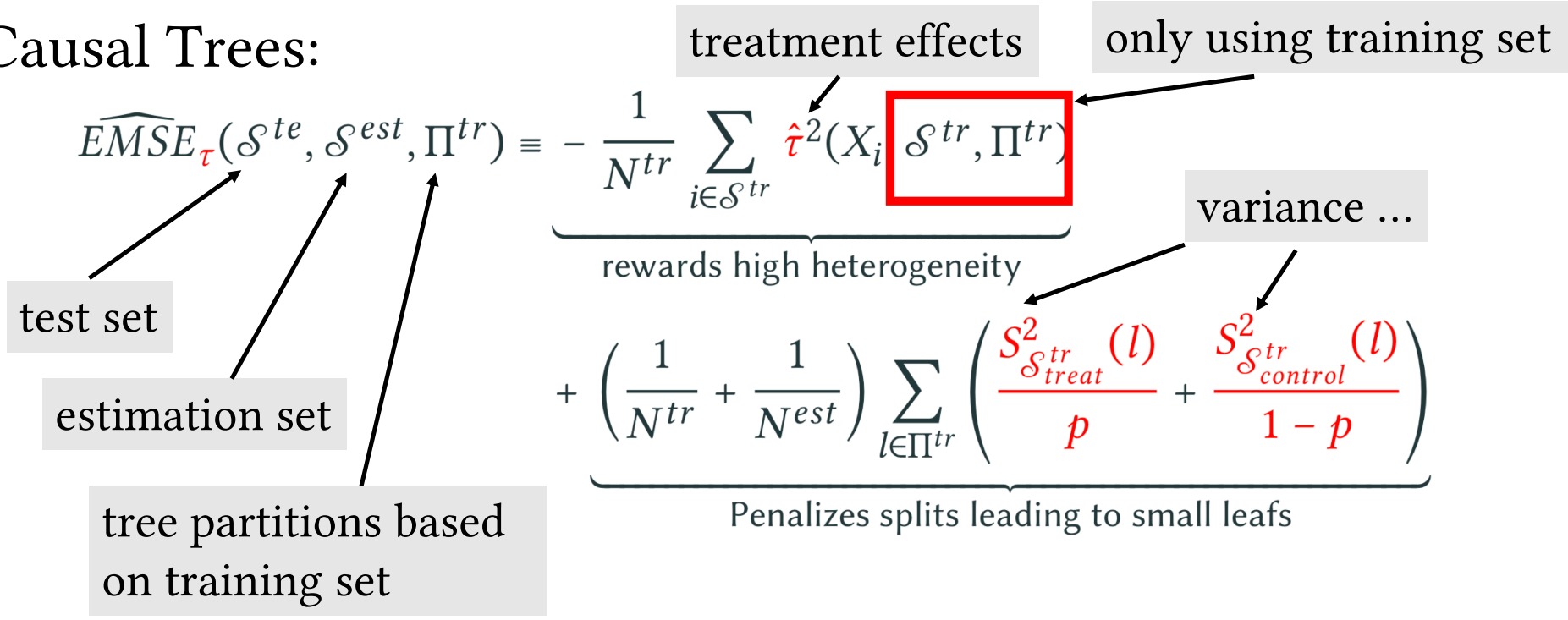
Causal Trees:



$$\widehat{EMSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi^{tr}) \equiv \underbrace{-\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i | \mathcal{S}^{tr}, \Pi^{tr})}_{\text{rewards high heterogeneity}} + \underbrace{\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi^{tr}} \left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p}\right)}_{\text{Penalizes splits leading to small leafs}}$$

Building block 2. New objective function

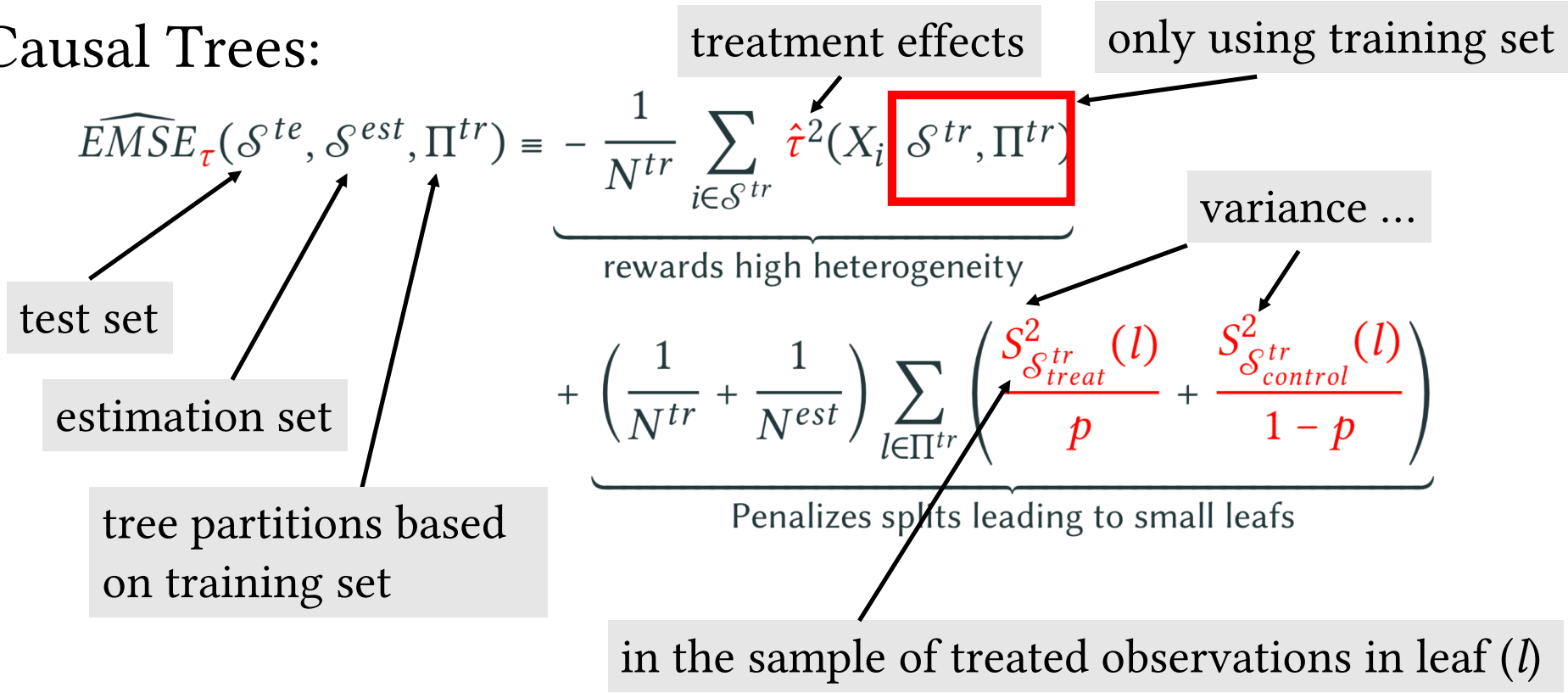
Causal Trees:



$$\widehat{EMSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi^{tr}) \equiv \underbrace{-\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i | \mathcal{S}^{tr}, \Pi^{tr})}_{\text{rewards high heterogeneity}} + \underbrace{\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi^{tr}} \left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p}\right)}_{\text{Penalizes splits leading to small leafs}}$$

Building block 2. New objective function

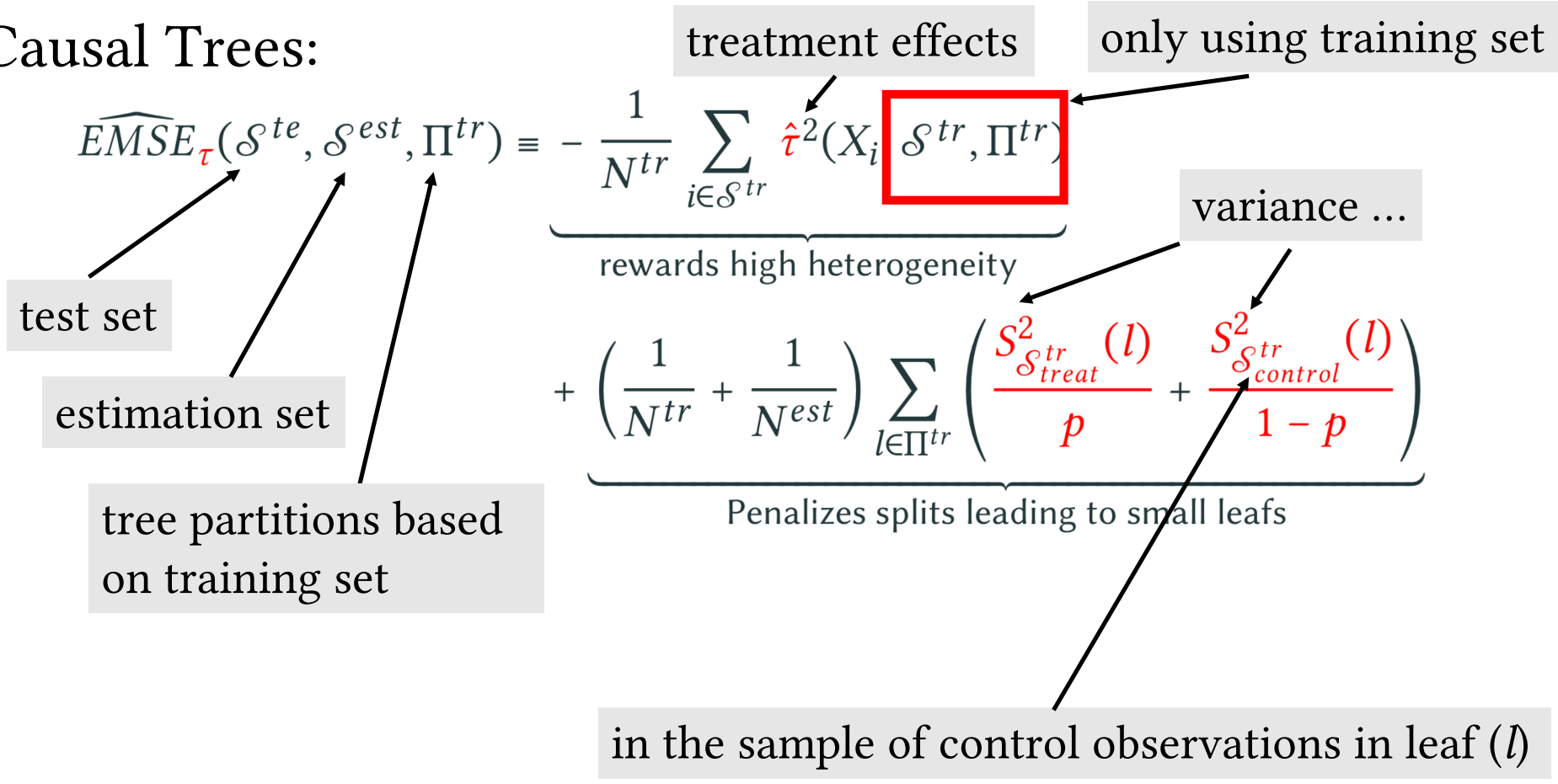
Causal Trees:



$$\widehat{EMSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi^{tr}) \equiv \underbrace{-\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i | \mathcal{S}^{tr}, \Pi^{tr})}_{\text{rewards high heterogeneity}} + \underbrace{\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi^{tr}} \left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p}\right)}_{\text{Penalizes splits leading to small leafs}}$$

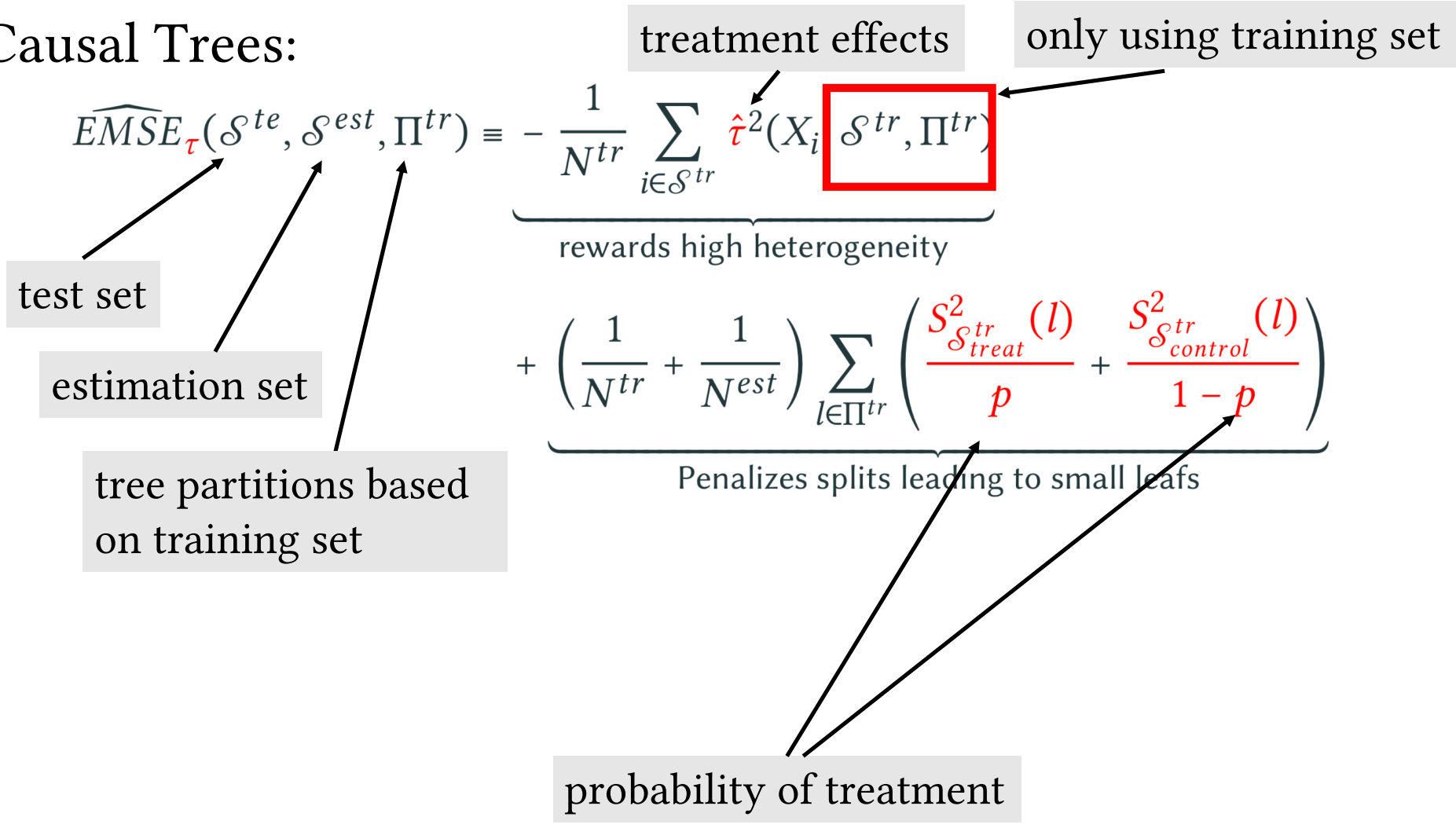
Building block 2. New objective function

Causal Trees:



Building block 2. New objective function

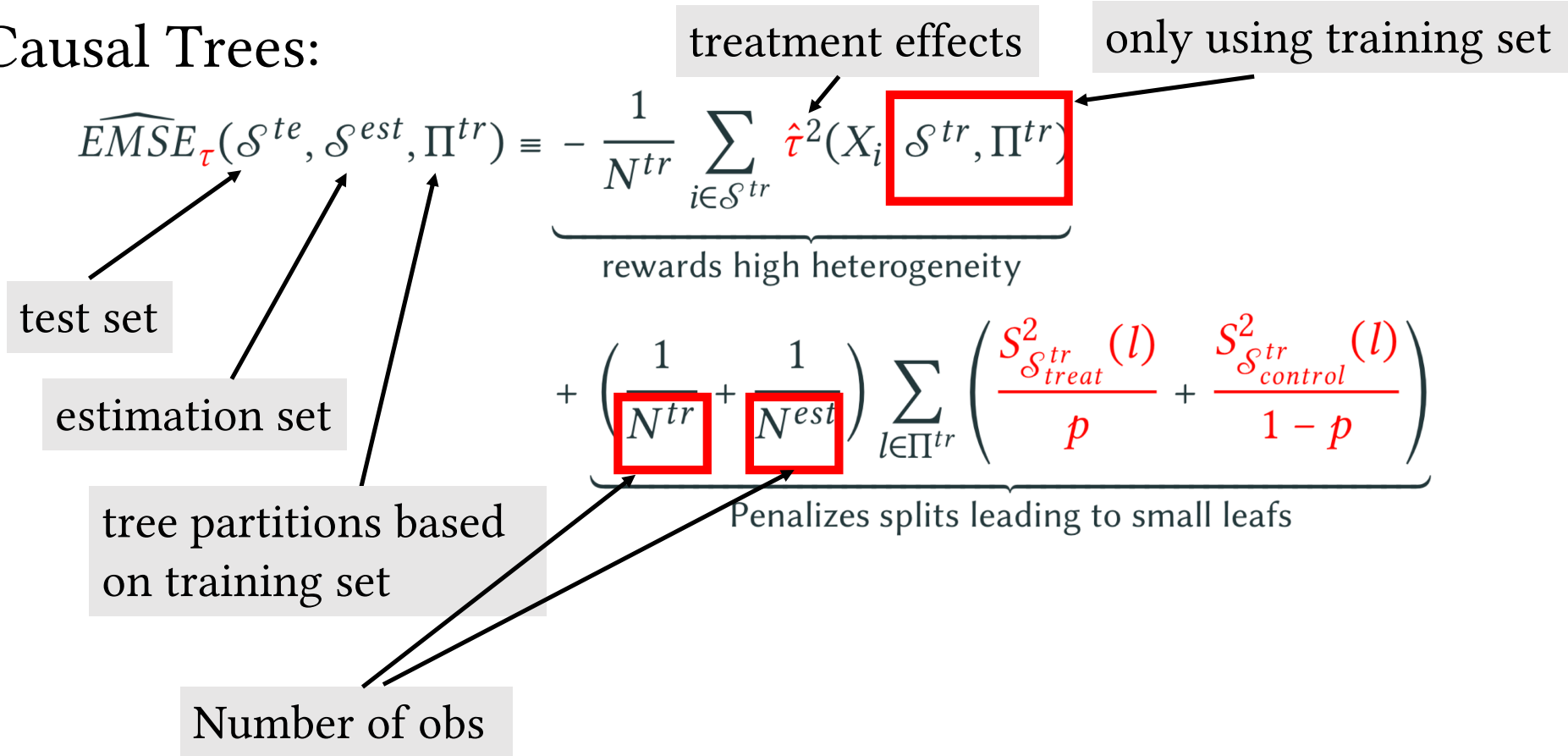
Causal Trees:



https://madina-k.github.io/dse_mk2021

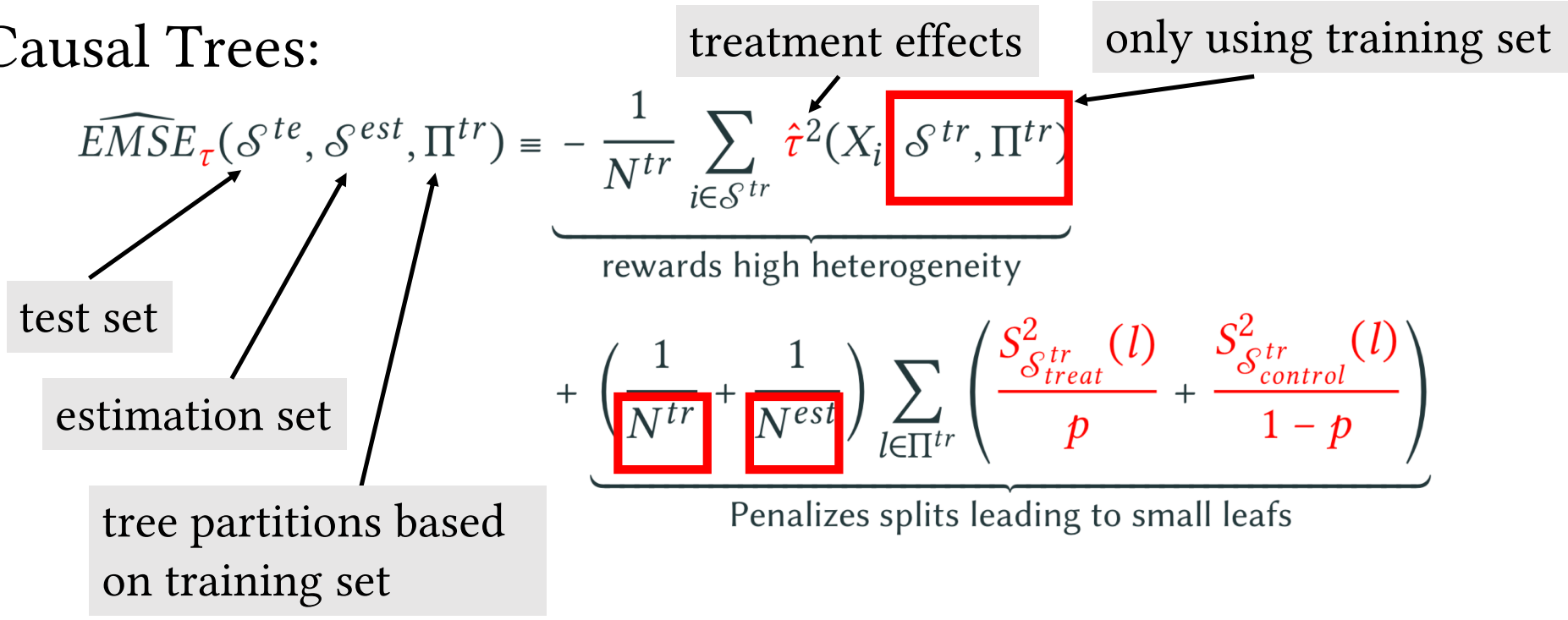
Building block 2. New objective function

Causal Trees:



Building block 2. New objective function

Causal Trees:



It needs only the training sample and the number of observations in the training and the estimation sample.

Building block 2. New objective function

Causal Trees:

$$\widehat{EMSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi^{tr}) \equiv \underbrace{-\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i | \mathcal{S}^{tr}, \Pi^{tr})}_{\text{rewards high heterogeneity}}$$

$$+ \underbrace{\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi^{tr}} \left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p}\right)}_{\text{Penalizes splits leading to small leafs}}$$

Estimated treatment effect is the difference in the average y between treated and control observations:

$$\hat{\tau}(x; \Pi^{tr}) \equiv \hat{\mu}(W = 1, x; \mathcal{S}^{tr}, \Pi^{tr}) - \hat{\mu}(W = 0, x; \mathcal{S}^{tr}, \Pi^{tr})$$

Summary: Decision trees vs Causal Trees

	Regression Tree	Causal Tree
Predictions based on	training sample	estimation sample
Splitting rule minimizes in-sample	RSS	Honest Target
Segments X for heterogeneity	in outcomes	in treatment effects

In DML

The last step:

$$Y_i = \beta W_i + \epsilon_i \tag{3}$$

Generalized Random Forest

Random Effects Model:

$$Y_i = b_i W_i + \epsilon_i; \quad \beta(x) = E(b_i | X_i = x); \tag{4}$$

(Note: under non-RCT, we first partial out X_i from Y_i and W_i: 3-steps procedure)

Generalized Random Forest

Random Effects Model:

$$Y_i = b_i W_i + \epsilon_i; \quad \beta(x) = E(b_i | X_i = x); \tag{4}$$

(Note: under non-RCT, we first partial out X_i from Y_i and W_i: 3-steps procedure)

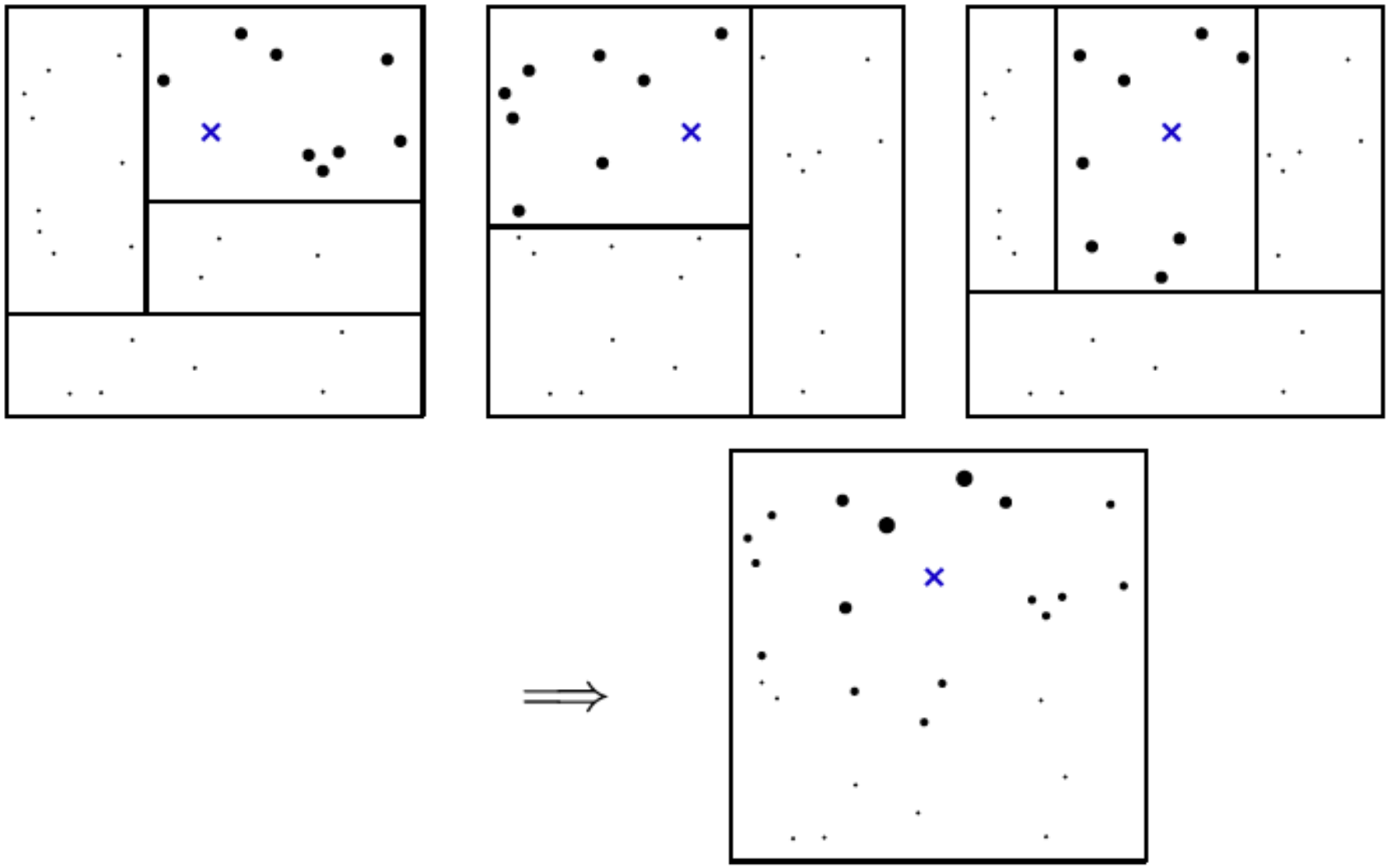
In particular, the generalized random forest estimates $\beta(x)$ non-parametrically:

$$\hat{\beta}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)(W_i - \bar{W}_\alpha)(Y_i - \bar{Y}_\alpha)}{\sum_{i=1}^{n} \alpha_i(x)(W_i - \bar{W}_\alpha)^2} \tag{5}$$

where

- $\alpha_i(x)$ is a weight determined by Causal Forest
- $\bar{W}_\alpha = \sum \alpha_i(x) W_i$ is a weighted average treatment
- $\bar{Y}_\alpha = \sum \alpha_i(x) Y_i$ is a weighted average outcome

Weights from Causal Forest



https://madina-k.github.io/dse_mk2021

Weights from Causal Forest

$$\alpha_{bi}(x) = \frac{\mathbf{1}(\{X_i \in L_b(x)\})}{|L_b(x)|}, \qquad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x).$$

Back to the National Study of Learning Mindsets (Athey and Wager, 2019)

https://madina-k.github.io/dse_mk2021

grf algorithm

```
cf <- causal_forest(X, Y, W, clusters = school.id,
                    equalize.cluster.weights = TRUE)
```

This code performs the following steps:

1. Predict Y from X using Random Forest, collect out-of-sample $\hat{Y}$
2. Predict W from X using Random Forest, collect out-of-sample $\hat{W}$
3. Run a causal forest of:
    - $Y - \hat{Y}$ on $W - \hat{W}$
    - using X as potential mediators
    - and school ids as clusters
    - drawing the same # of obs per cluster

Cluster-robust random/causal forests

To account for clusters, random/causal forests:

- At bootstrapping stage:
  1. First draw a subsample of clusters
  2. Then random samples of data within the clusters
- For out-of-bag predictions:
  - consider an observation i to be out-of-bag if its cluster was not drawn at the boostrapping stage

(Conditional) average treatment effect

```
average_treatment_effect(cf, subset = X1>0)
```

calculates augmented inverse-propensity weighting CATE (Robins, Rotnitzky, and Zhao, 1994) for the subset of observations where X1>0

$$\hat{\tau}_j = \frac{1}{n_j} \sum_{\{i:A_i=j\}} \hat{\Gamma}_i, \quad \hat{\tau} = \frac{1}{J}\sum_{j=1}^{J} \hat{\tau}_j, \quad \hat{\sigma}^2 = \frac{1}{J(J-1)}\sum_{j=1}^{J} (\hat{\tau}_j - \hat{\tau})^2,$$

$$\hat{\Gamma}_i = \hat{\tau}^{(-i)}(X_i) + \frac{W_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left(1 - \hat{e}^{(-i)}(X_i)\right)} \left(Y_i - \hat{m}^{(-i)}(X_i) - \left(W_i - \hat{e}^{(-i)}(X_i)\right)\hat{\tau}^{(-i)}(X_i)\right).$$

In the study:

95% CI for the ATE : 0.247 +/- 0.04

Test for the presence of heterogeneity

```
test_calibration(cf)
```

tests for the presence of heterogeneity in treatment effects (Chernozhukov et al, 2018):

- Creates two synthetic variables:

$$C_i = \bar{\tau}(W_i - \hat{W}_i) \quad \text{prediction by ATE}$$

$$D_i = (\hat{\tau}^{cf} - \bar{\tau})(W_i - \hat{W}_i) \quad \text{additional pred. by CF}$$

- Runs regression:

$$Y_i - \hat{Y}_i = \gamma C_i + \delta D_i$$

- if $\delta = 1 \Longrightarrow$ treatment heterogeneity is well calibrated
- if $\delta = 0 \Longrightarrow$ cf captures no heterogeneity

In the study, p-value of $\delta = 0.294 \Longrightarrow$ No evidence of heterogeneity in TE

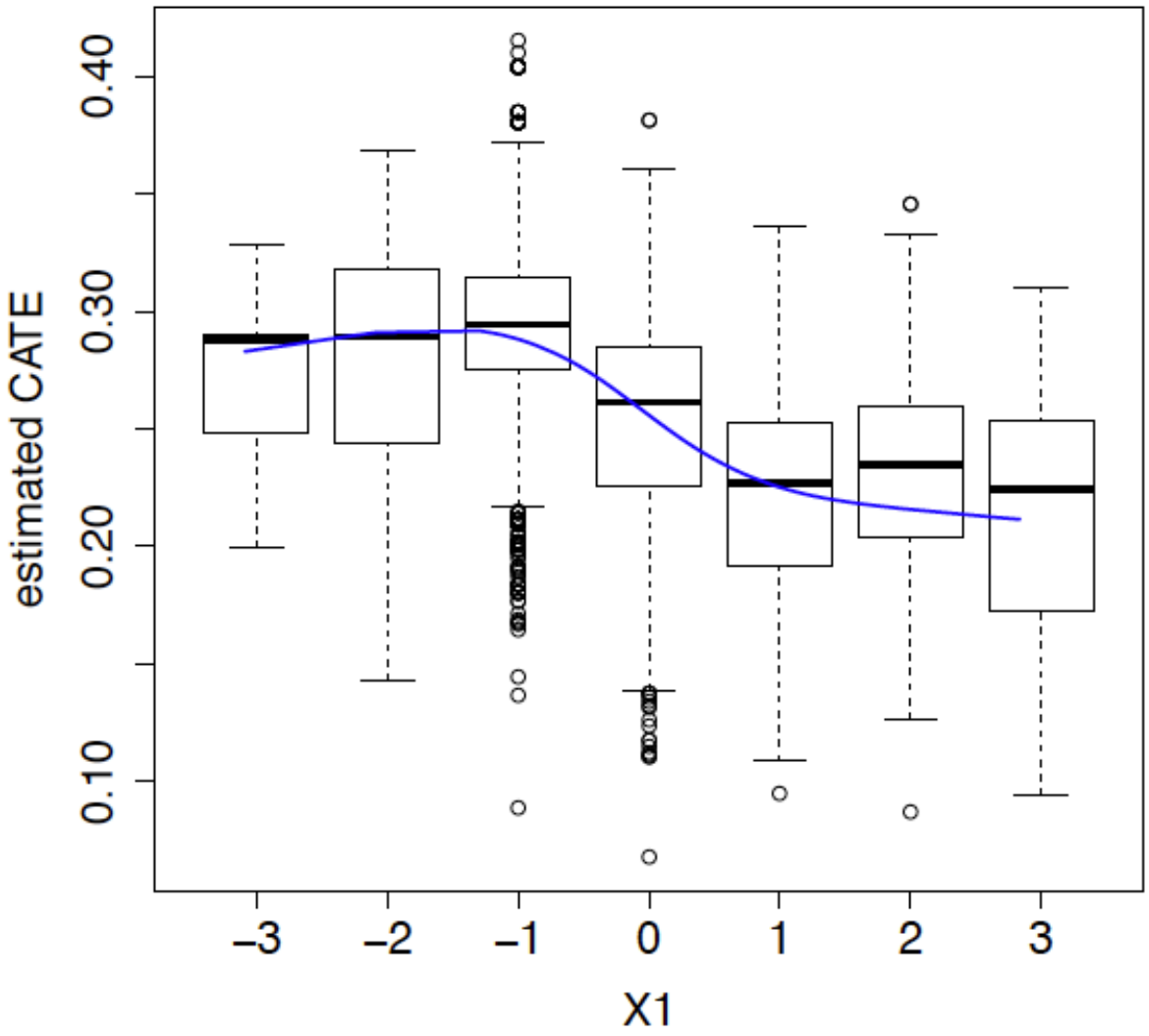
Variable importance

```
variable_importance(cf)
```

calculates a simple weighted sum of how many times feature *i* was split on at each depth in the forest.

In the study, 24% of splits were on X1

TE predictions by school-level mindset levels



There may be some heterogeneity within X1 levels, after performing formal tests (See page 8)

Summary

The GRF provides an elegant solution to:

- estimate ATE non-parametrically
- capture heterogeneity in TE and formally test it
- find which variables mediates the heterogeneous response the most
- all accounting for clustering and potential confounding

Moreover, the GRF can be used for any local moment equations estimation:

- quantile regressions
- IV estimation
- etc.