ML@LSE - Bootcamp 2: Tree-based methods

Ref. for today: Chapter 8, Hastie et al.

Warming up

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

Reminder: Machine learning can be subdivided into two categories

- Supervised learning: a set a features predict an outcome variable
- Unsupervised learning: no outcome variable, finding general patterns in data

Supervised learning can itself be subdivided into two categories

- Regression: the outcome variable is continuous
- Classification: the outcome variable is categorical

Today we're going to learn about tree based methods, which belongs to the supervised learning category.

What is a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

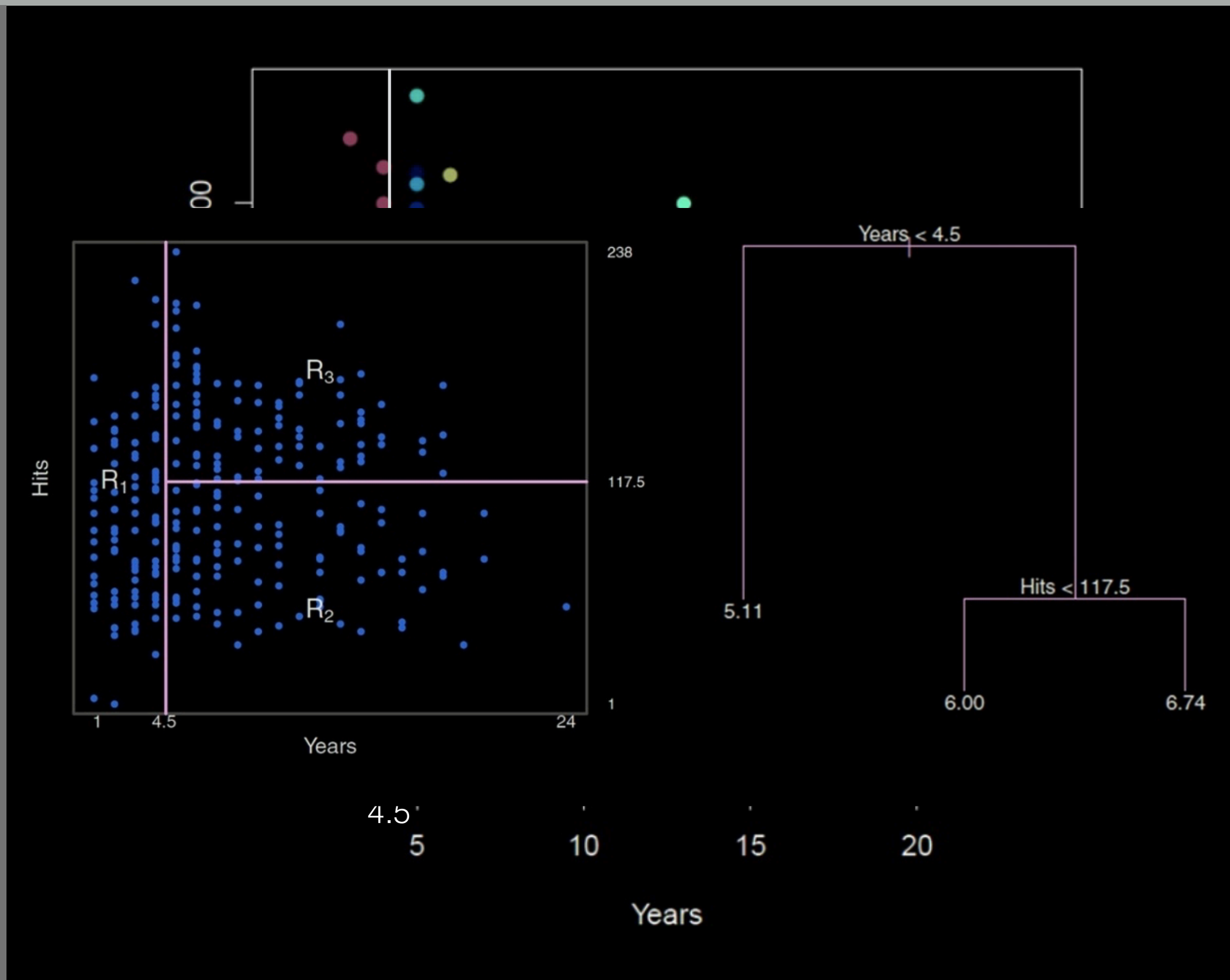
a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?



How to build a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

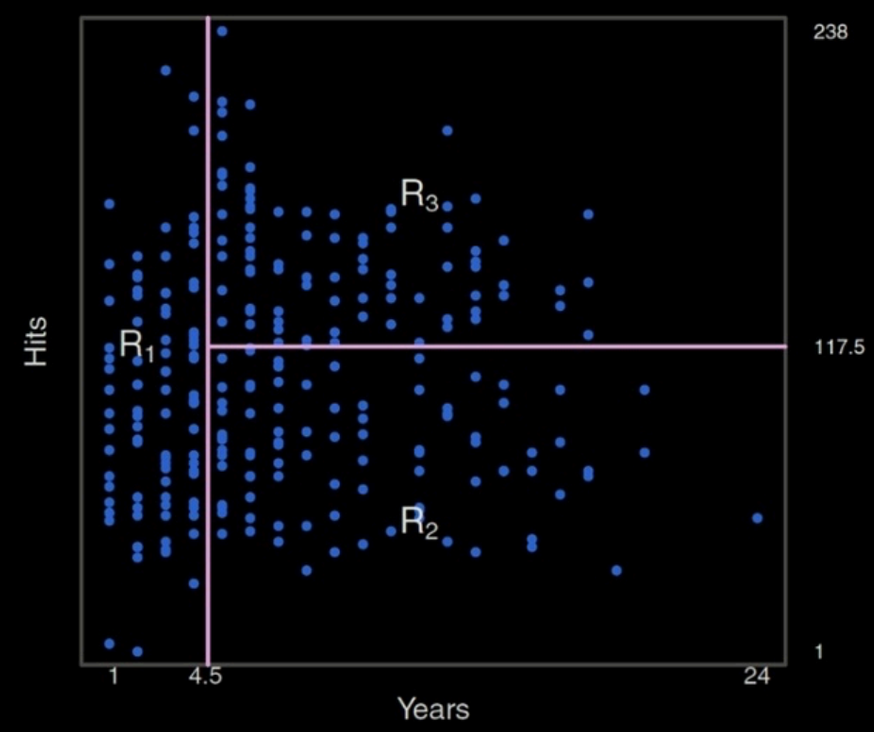
What is the exact procedure for building these trees?

➢ Recursive binary splitting

- We want to split our features space into several regions R_1, …, R_J
- The prediction will be the same within a given region, namely, we take the average of the training data $\hat{y}_{R_j}$.
- When choosing the delimitation of the regions R_1, …, R_J, we seek to minimize the sum of squares (RSS):

$$\frac{\sum_{j=1}^{J}\sum_{i\in R_j}(y_i-\hat{y}_{R_j})^2}{\text{Total sum of intra-region sum of squares = RSS}}$$

Intra-region sum of squares



How to build a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

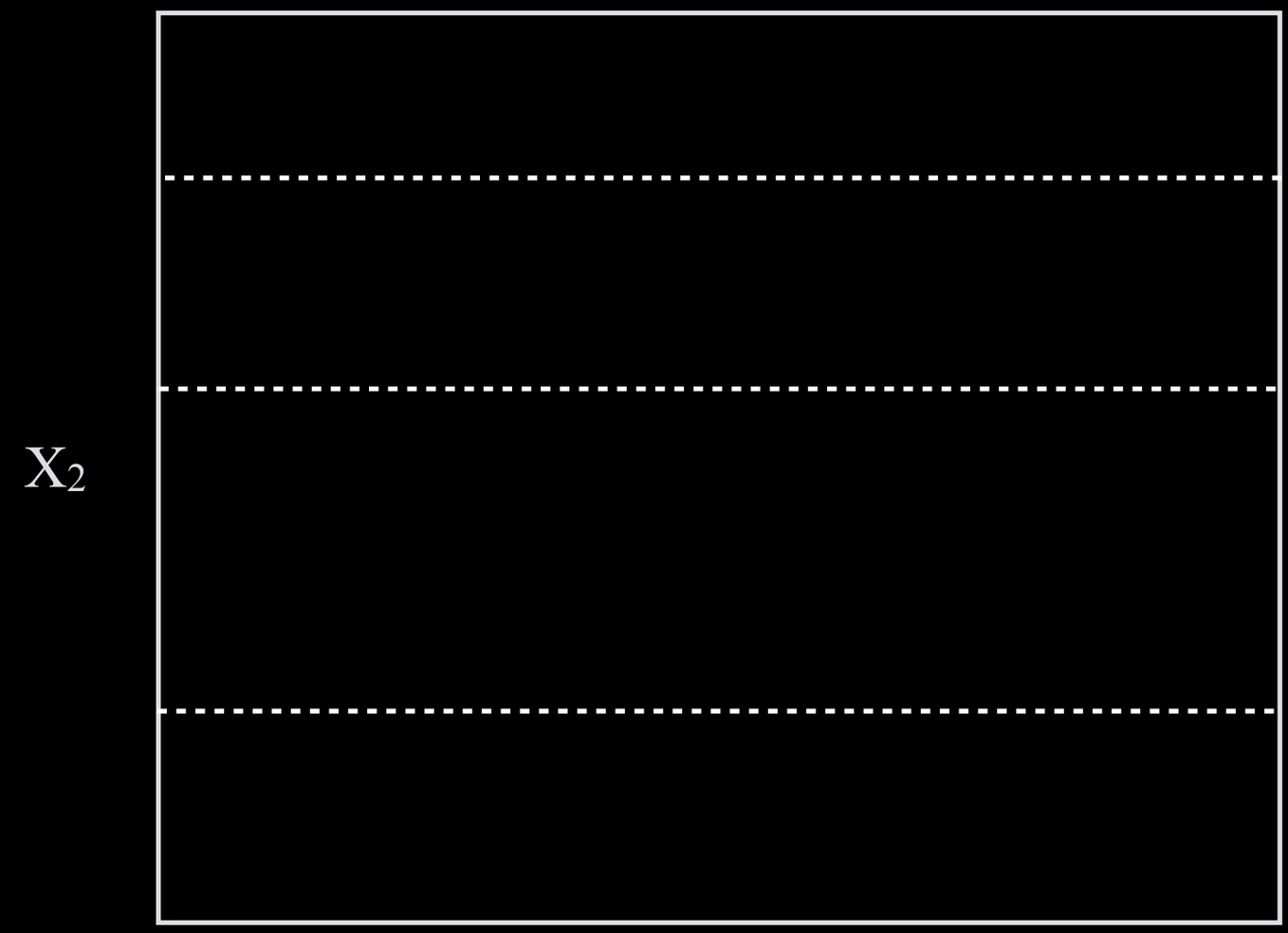
2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

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



How to build a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

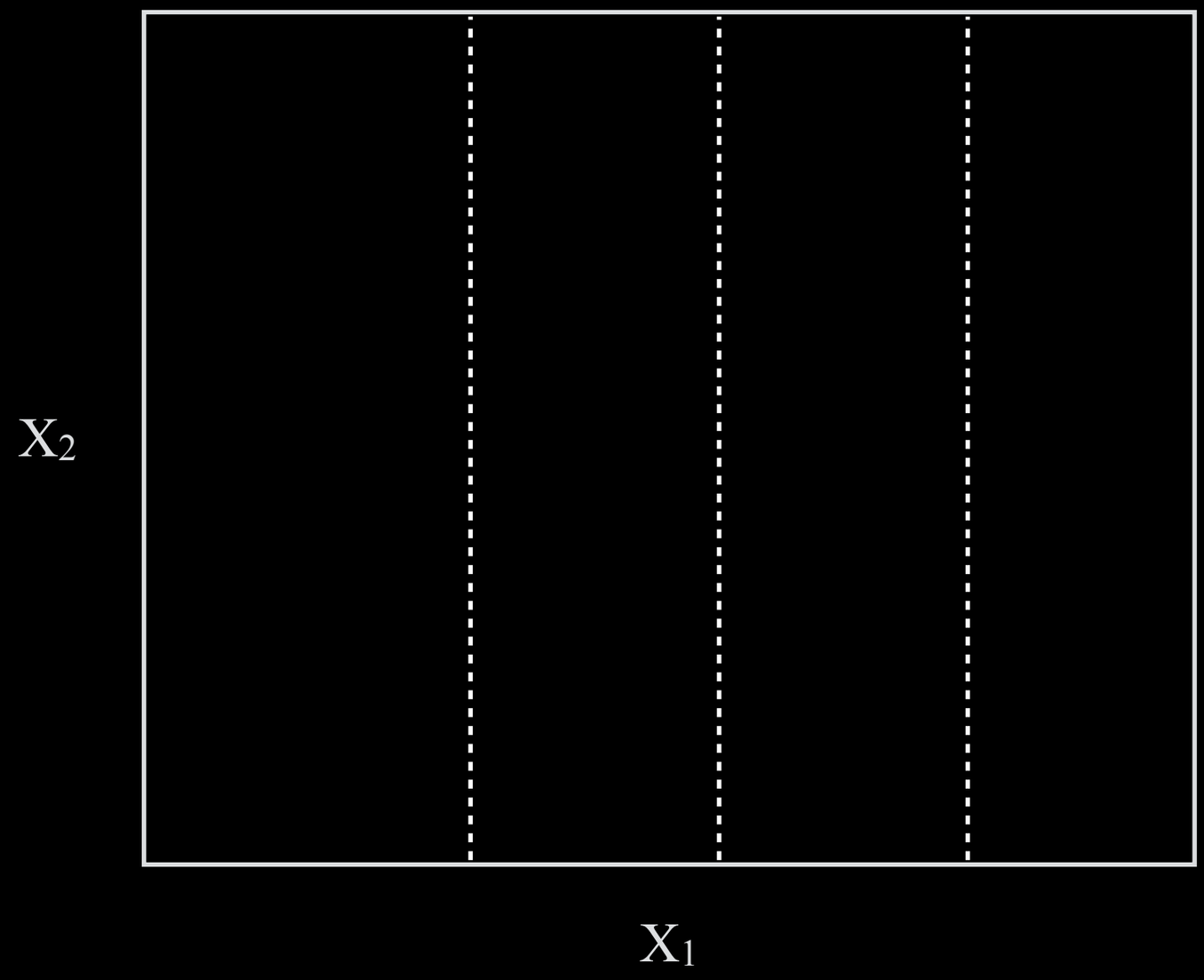
2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

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



How to build a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

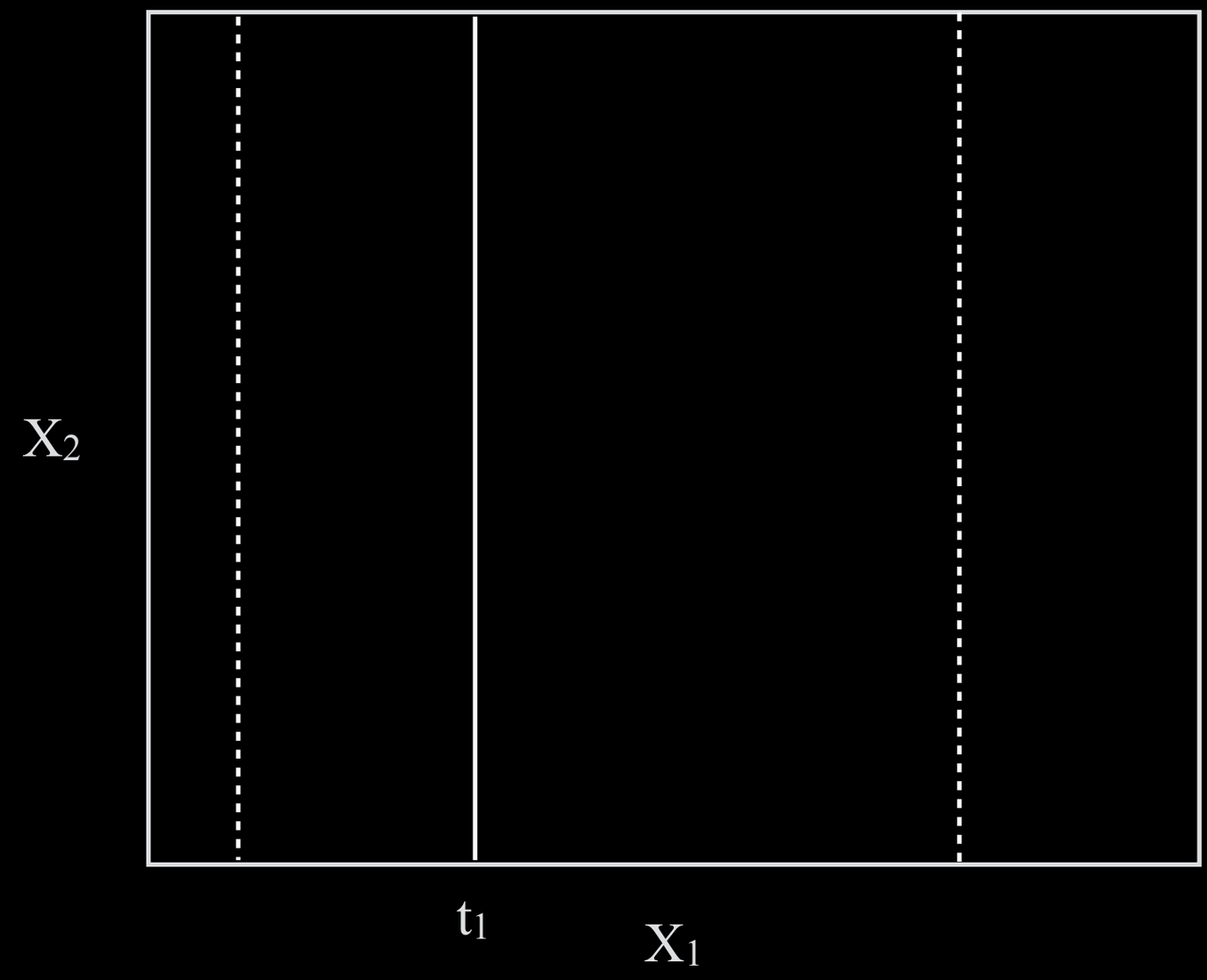
2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

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



How to build a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

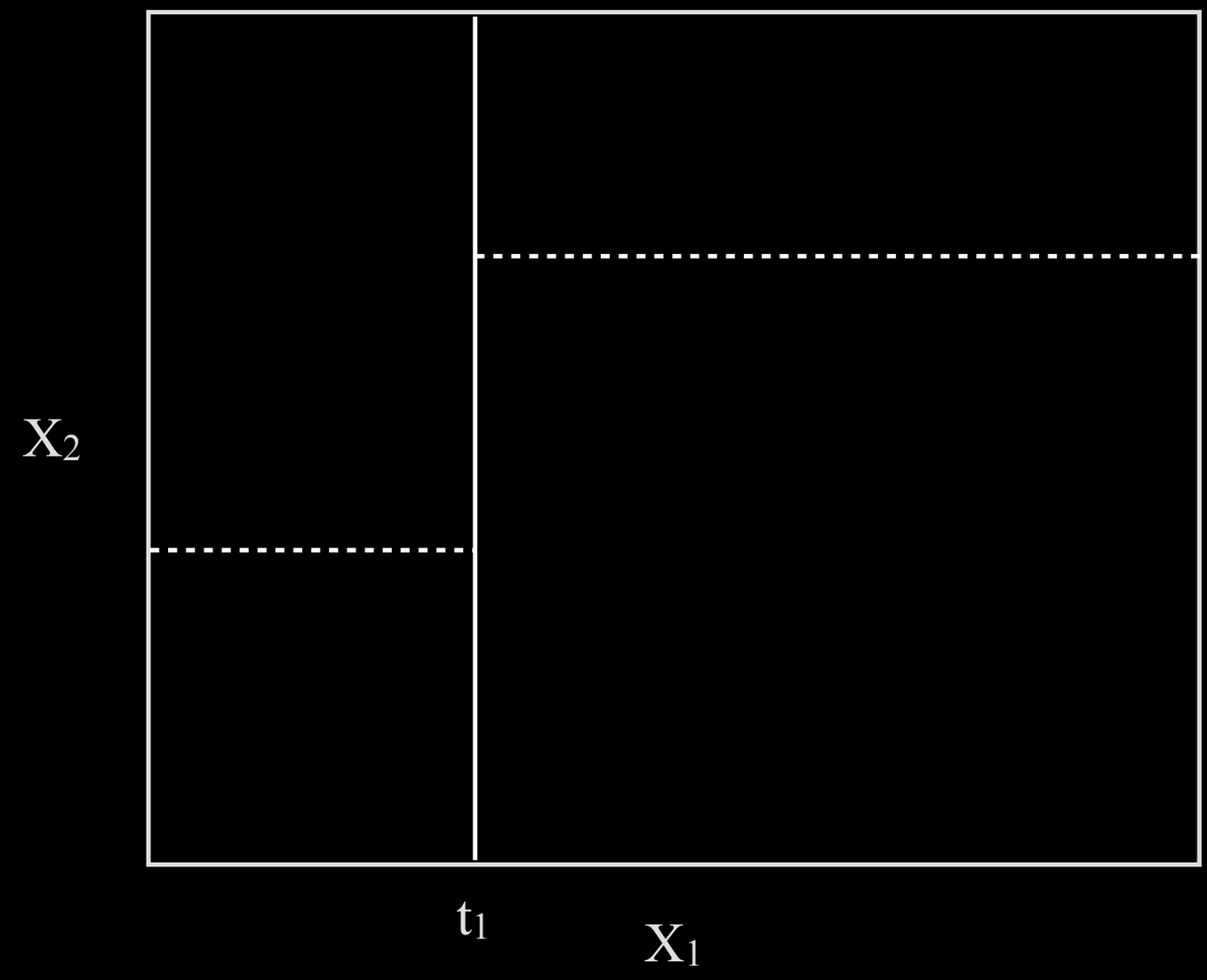
2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

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



How to build a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

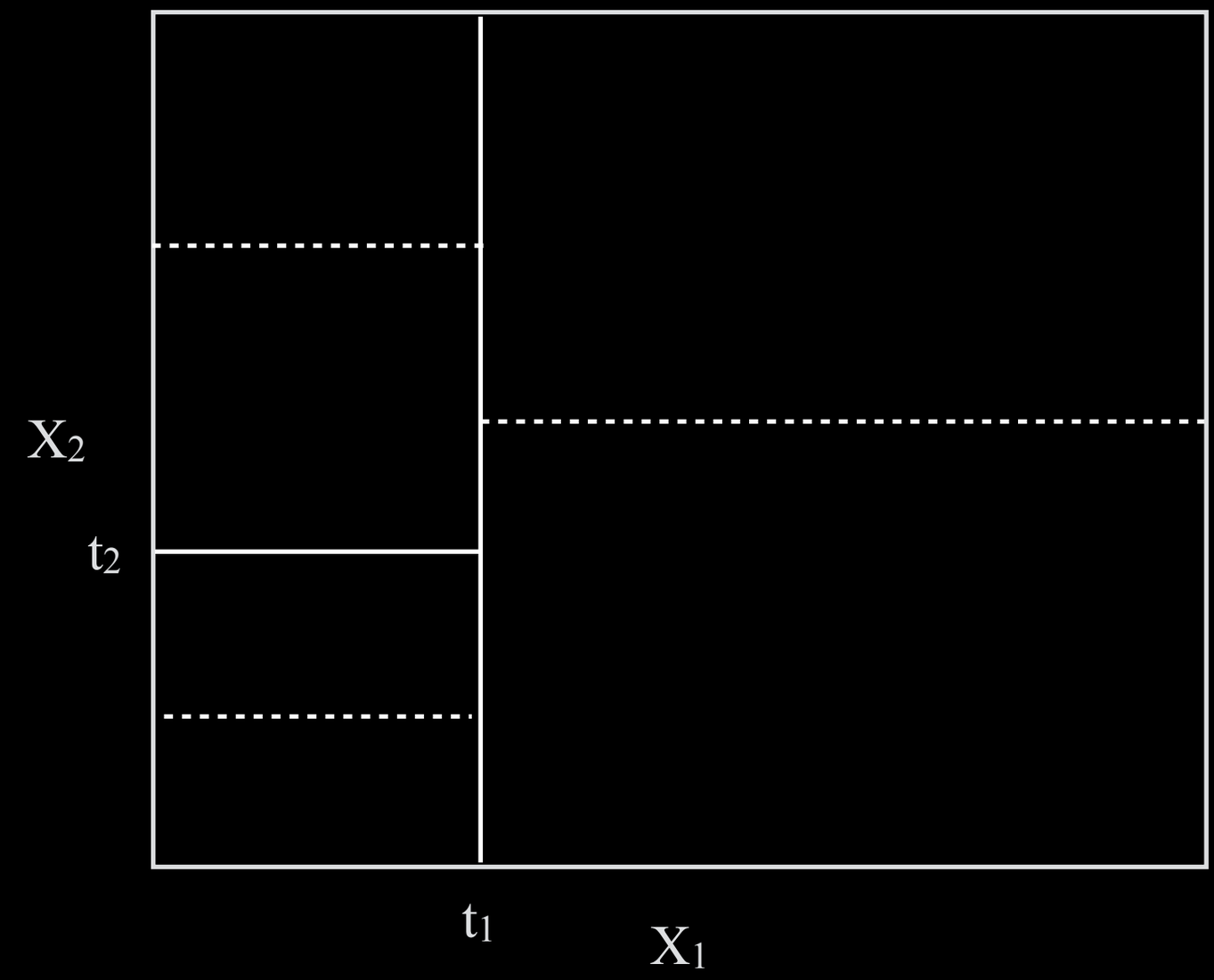
2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

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



How to build a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

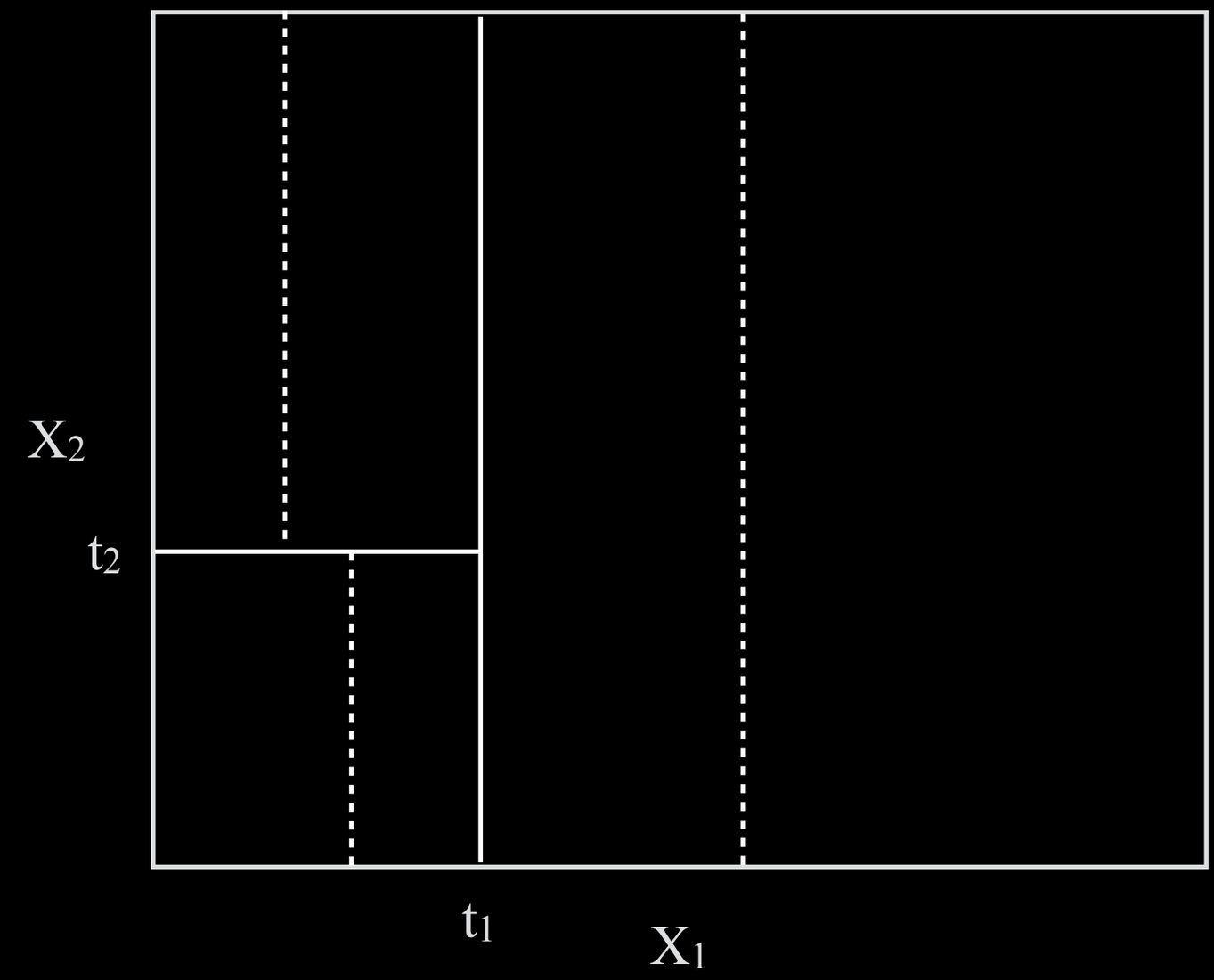
2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

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



How to build a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

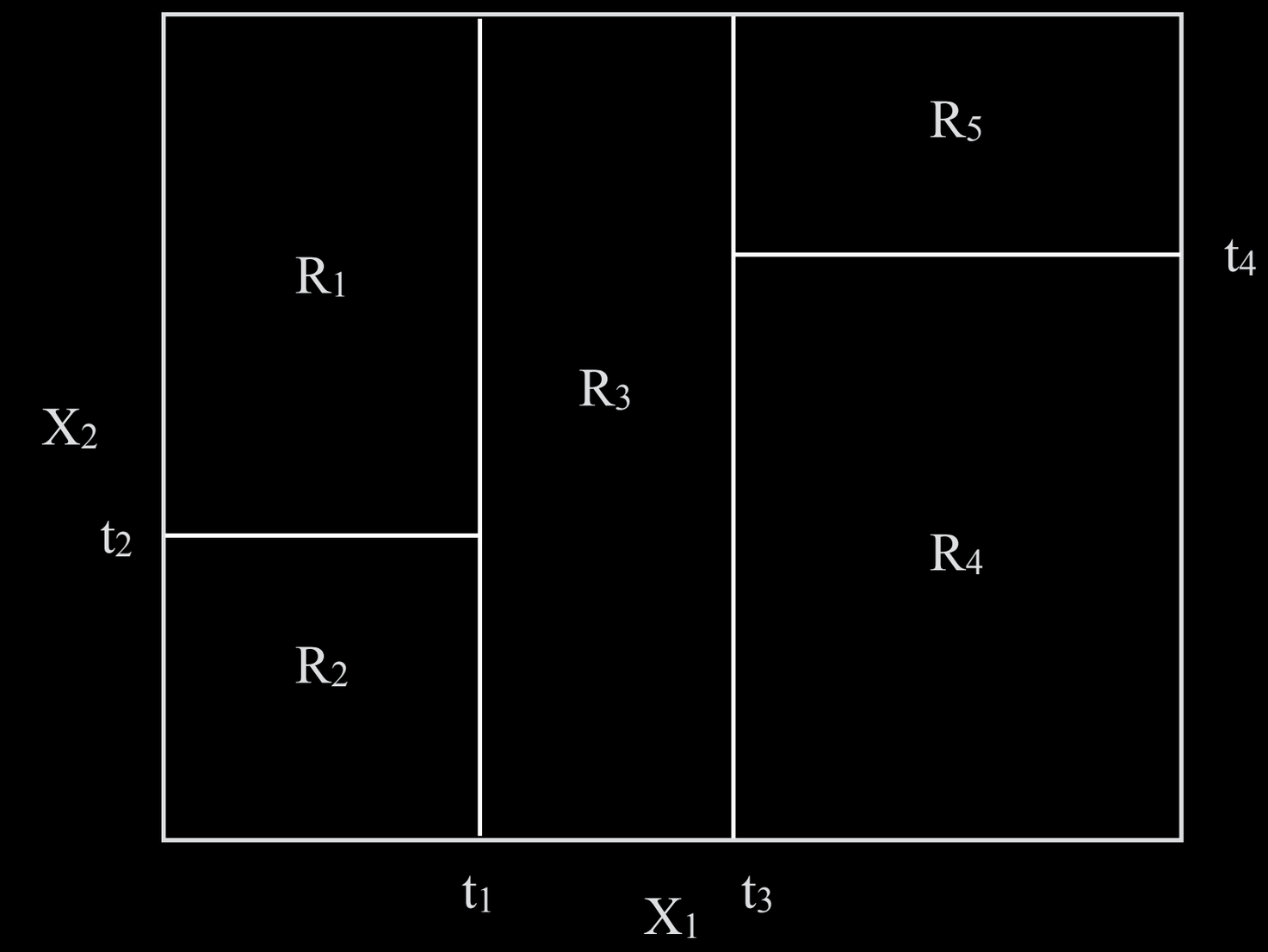
2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

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



What is a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

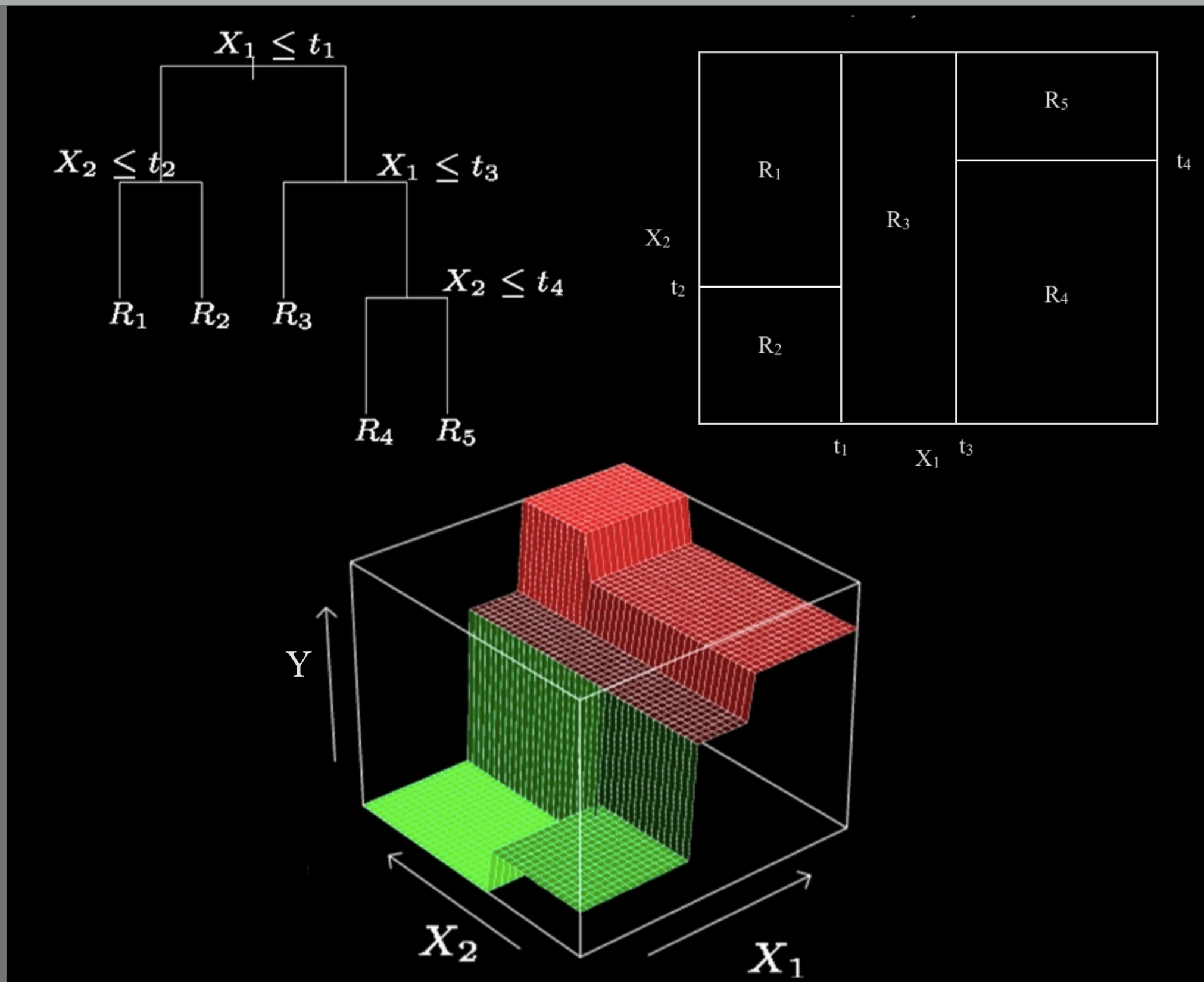
a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?



What is a decision tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

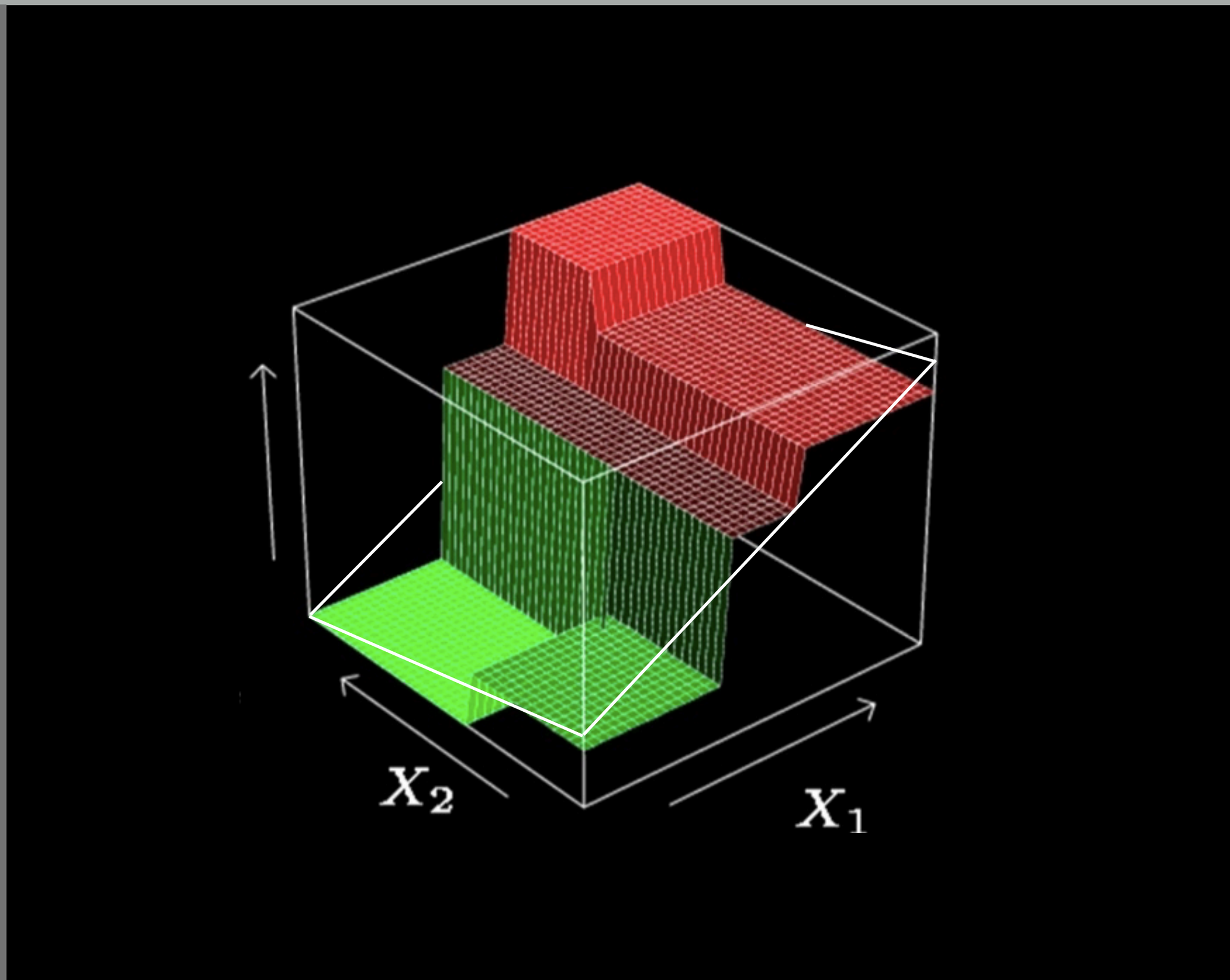
a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?



Advantages and disadvantages of trees

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

Advantages and disadvantages?

- High interpretability: trees are intuitive and easily visualizable.
- Non-parametric: we do not need to impose a structure f on the data.
- Trees can accommodate qualitative predictors without requiring dummy variables.

BUT

- Trees prediction accuracy is not as good as that of other supervised learning techniques.

Overfitting in decision trees

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

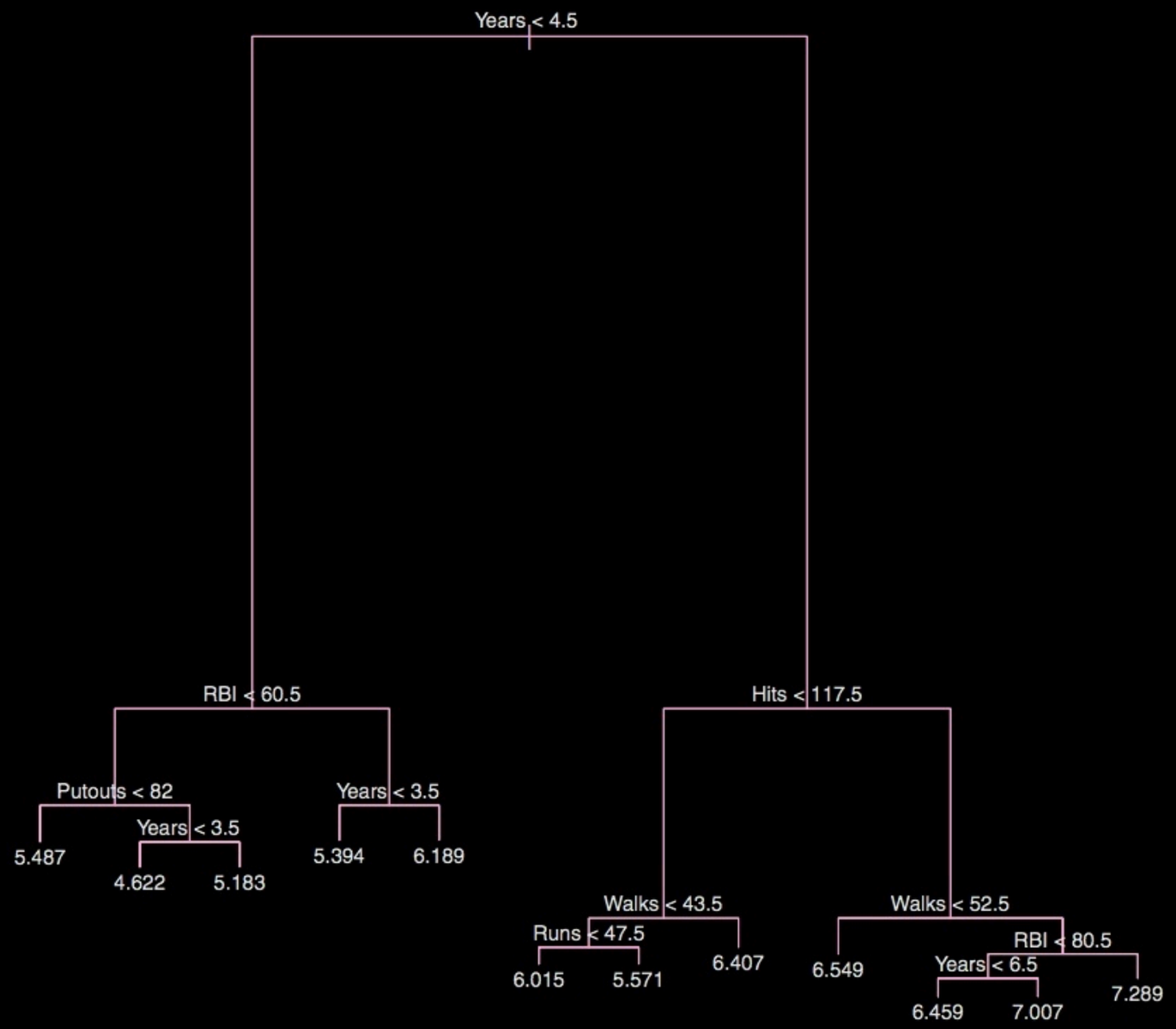
a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

Warning: Overfitting the tree to the training data!

➢ How to choose the optimal number of branches?



How to prune a tree?

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

Technique: Let the tree grow and then prune it optimally = pruning.

➢ How to prune a tree efficiently? (i.e. how to account for the bias-variance tradeoff).

1. Let the tree grow until each region contains a maximal number of observations.

2. Impose a cost to the number of terminal nodes in the tree, i.e. for a given α, pick up the tree that minimizes

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

3. Choose alpha using **Cross-Validation**

K-fold cross validation:

Bagging

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

How to reduce the variance of our method (i.e. make our prediction more accurate)?

Theorem: *If $X_1, \ldots, X_n$ are independent and identically distributed with variance σ^2, then their average $\bar{X}$ has variance σ^2/n. Hence, averaging a set of observations reduces variance, and increases prediction.*

Ex: If I want to know the average height of LSE students, which method yields a more accurate result:

- Picking one student randomly and measuring his/her height.
- Picking (randomly) 300 students and averaging their height.

Can you see how we can use this to reduce the variance of our predictions?

Bagging

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

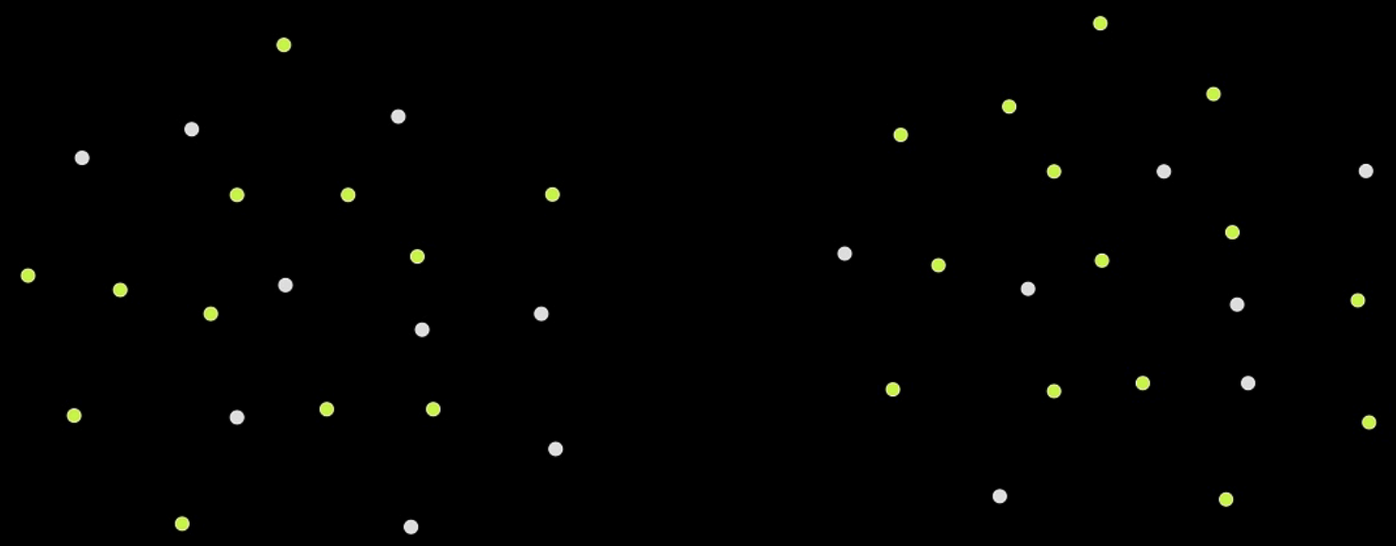
a. Boosting a regression tree
b. Why does this work well?

Answer: Bagging, or Bootstrap Aggregation!

Take several random subsamples of your training data, apply the model to them, and then take the average:

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x)$$

Illustration of bootstrapped data:



Bagging

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

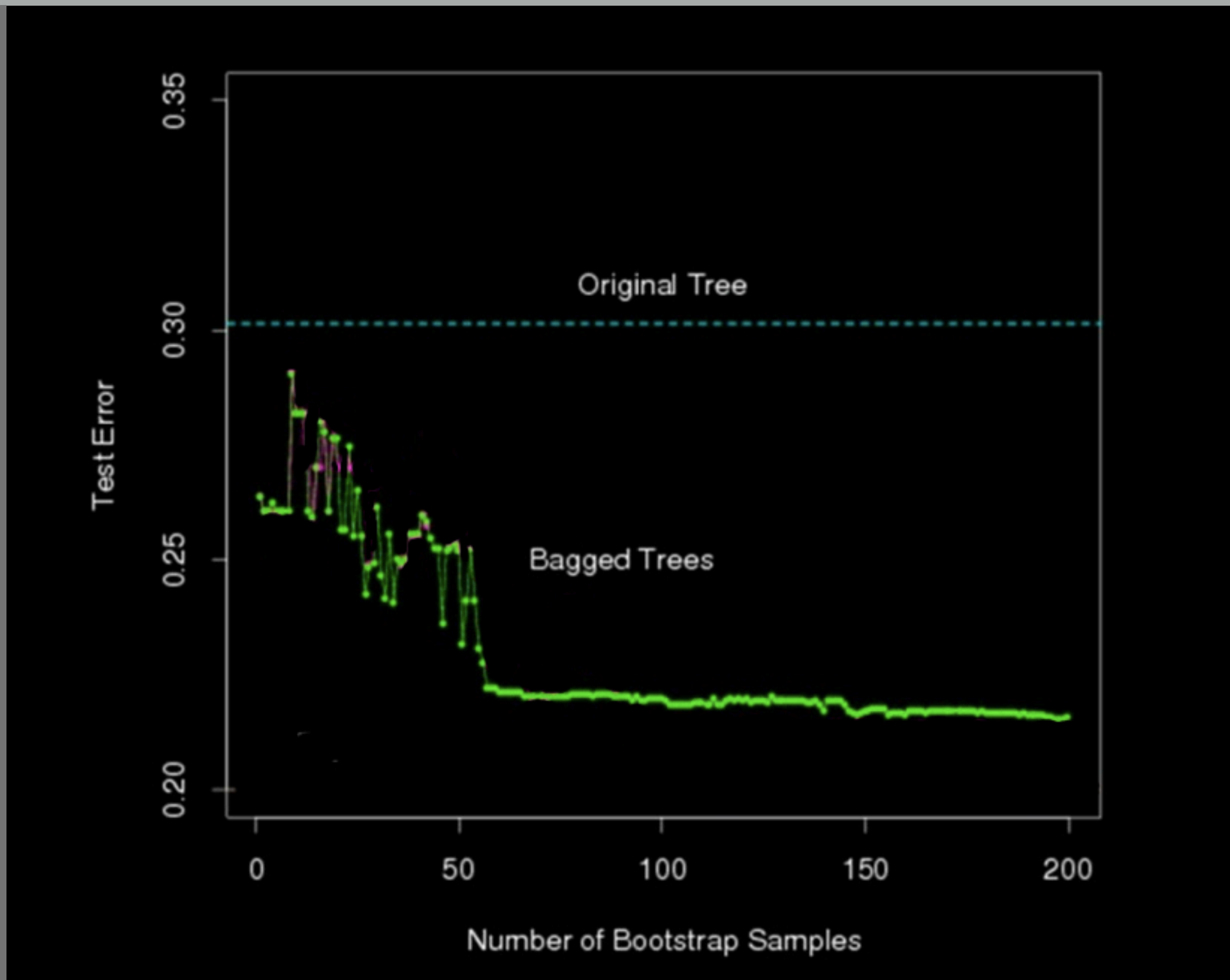
a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?



Random forest

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

Here is an better yet method: Random Forests.

Random forest algorithm:

1. Like previously, we construct a set of decision trees with bootstrapped data
2. But when constructing the trees we add a trick: each time we split the features space, only a set of m randomly selected predictors are considered to make the split, instead of the p predictors $X_1, \ldots, X_n$ and with $m < p$.
3. We usually choose $m \approx \sqrt{p}$.

Random forest

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

Why does this work better?

Theorem:

- *If $X_1, \ldots, X_n$ are **independent** and identically distributed with variance σ^2, then their average $\bar{X}$ has variance σ^2/n. Hence, averaging a set of observations reduces variance, and increases prediction.*
- *If $X_1, \ldots, X_n$ are **correlated** and identically distributed with variance σ^2, then σ^2/n is a lower bound on the variance of their average $\bar{X}$.*

Suppose there is a very strong predictor X_s, and that other predictors are not as strong. Then most of our bootstrapped trees will use the strong predictor in their first split, they will look similar and their predictions will be **correlated**.

Forcing each split to consider randomly selected features mitigates this problem, this **decorrelates** the trees, and our theorem can be fully applied.

Boosting

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

Last algorithm for today: Boosting a regression tree

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.

2. For $b = 1, 2, \ldots, B$, repeat:

   (a) Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r).

   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$

   (c) Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$

3. Output the boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

Boosting

A- Decision and classification trees

1. **Decision trees**

a. Intuition: what is a decision tree?
b. How to build a decision tree?

2. **Why using decision trees?**

a. Decision trees vs Linear models
b. Advantages and disadvantages

3. **An example of Cross Validation: Pruning**

a. Pruning
b. **How to prune a tree using Cost complexity?**

B- Ensemble Methods

1. **Bagging**

a. How to reduce the variance of our method?
b. Bagging in practice

2. **Random forest**

a. Can we do even better?
b. Implementing random forests

3. **Boosting**

a. Boosting a regression tree
b. Why does this work well?

Why does this work better?

Learning slowly: avoids overfitting to the training data.