# Ensemble Methods, including Random Forests

Alexander Ioannidis
ioannidis@stanford.edu

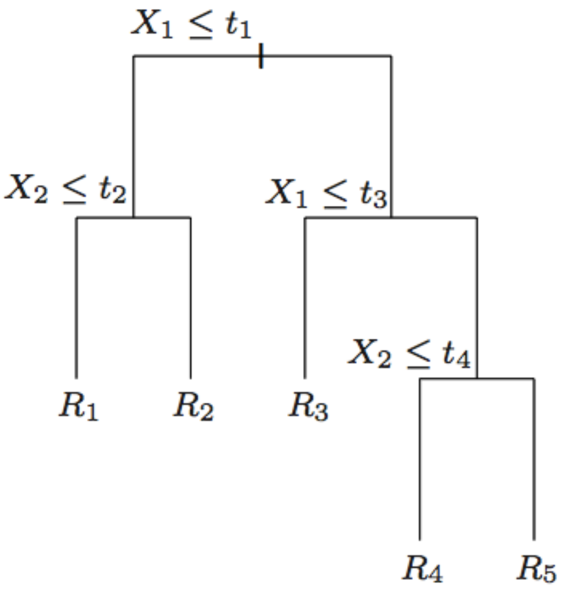
Institute for Computational and Mathematical Engineering,
Stanford University

# Trees

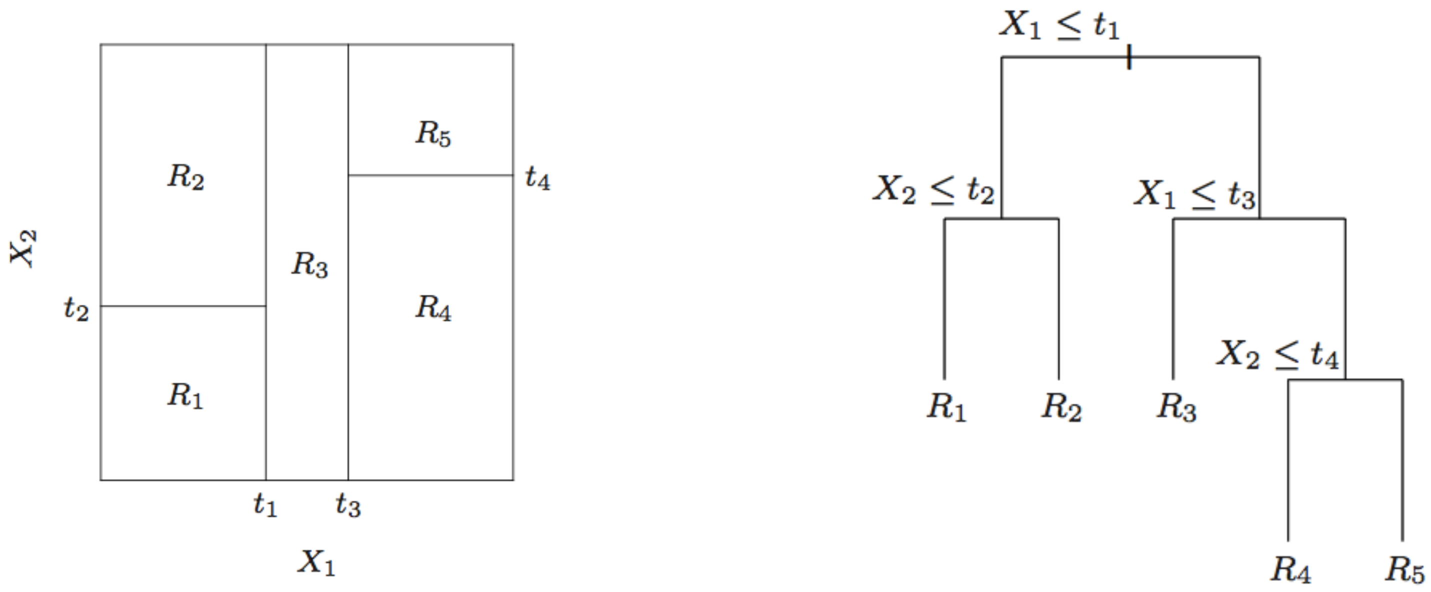
# Classification and Regression Trees

Recursive binary partitions of the feature space

# Classification and Regression Trees



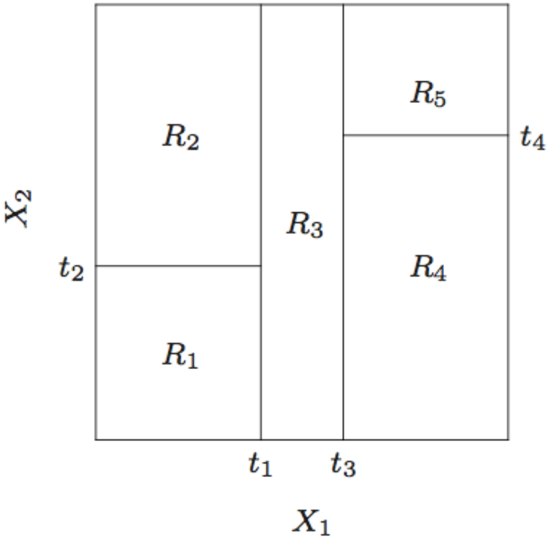
# Classification and Regression Trees



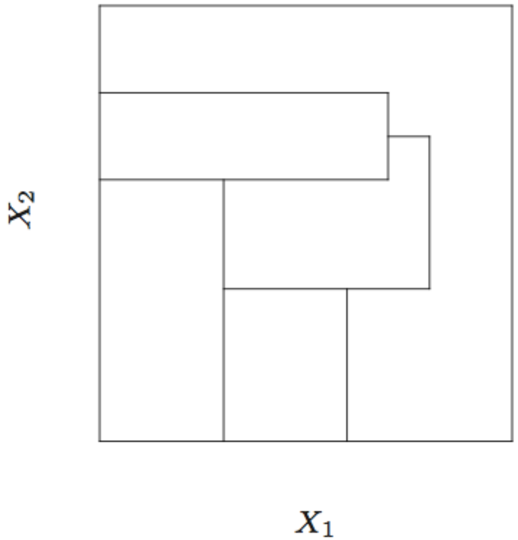

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

# Classification and Regression Trees

YES




NO




Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

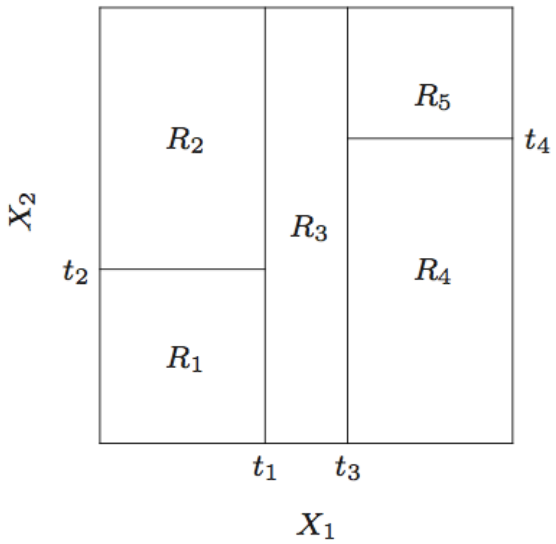
# Classification and Regression Trees

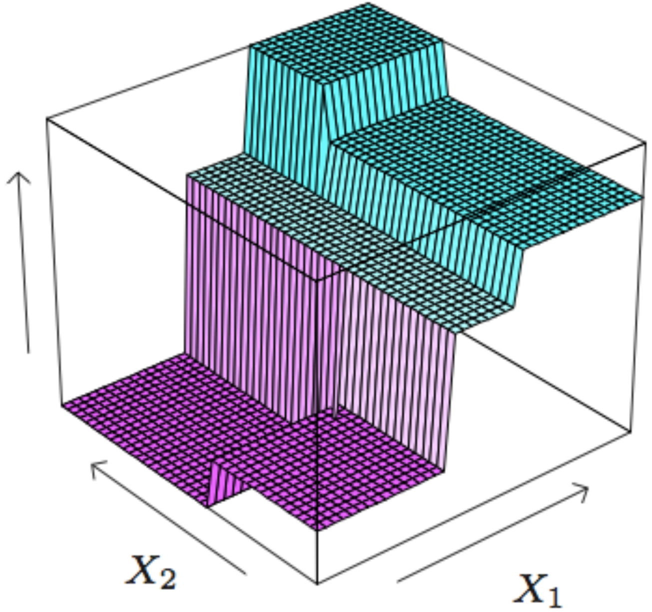
- Constant prediction within each region

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

# Classification and Regression Trees



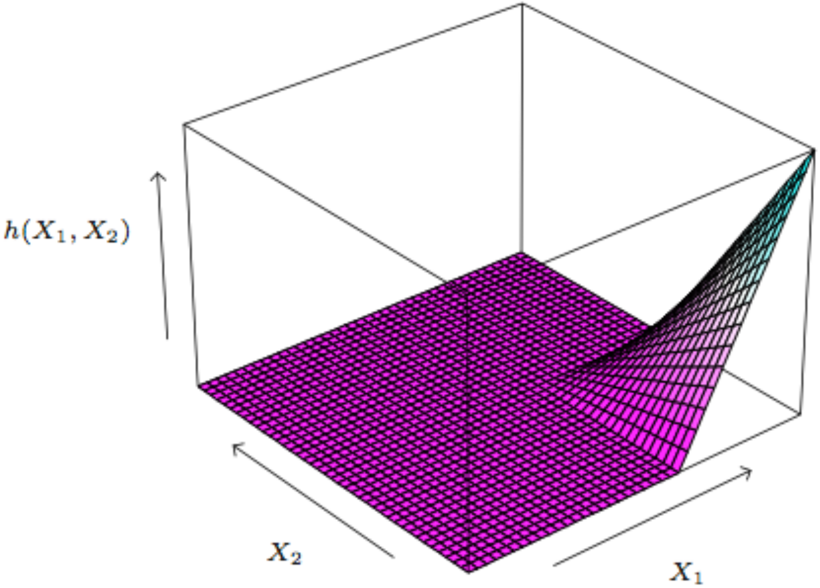

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

# Reminder: Machine Learning in one equation

$$\hat{f} = \underset{\tilde{f}}{\operatorname{argmin}} E[L(Y, \tilde{f}(X))]$$

# Aside: Piecewise linear model (MARS)

Multivariate Adaptive Regression Splines




Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

# Classification and Regression Trees

1 TYPE OF HOME
1. House
2. Condominium
3. Apartment
4. Mobile Home
5. Other

2 SEX
1. Male
2. Female

3 MARITAL STATUS
1. Married
2. Living together, not married
3. Divorced or separated
4. Widowed
5. Single, never married

4 AGE
1. 14 thru 17
2. 18 thru 24
3. 25 thru 34
4. 35 thru 44

5 EDUCATION
1. Grade 8 or less
2. Grades 9 to 11
3. Graduated high school
4. 1 to 3 years of college
5. College graduate
6. Grad Study

6 OCCUPATION
1. Professional/Managerial
2. Sales Worker
3. Factory Worker/Laborer/Driver
4. Clerical/Service Worker
5. Homemaker
6. Student, HS or College
7. Military
8. Retired
9. Unemployed

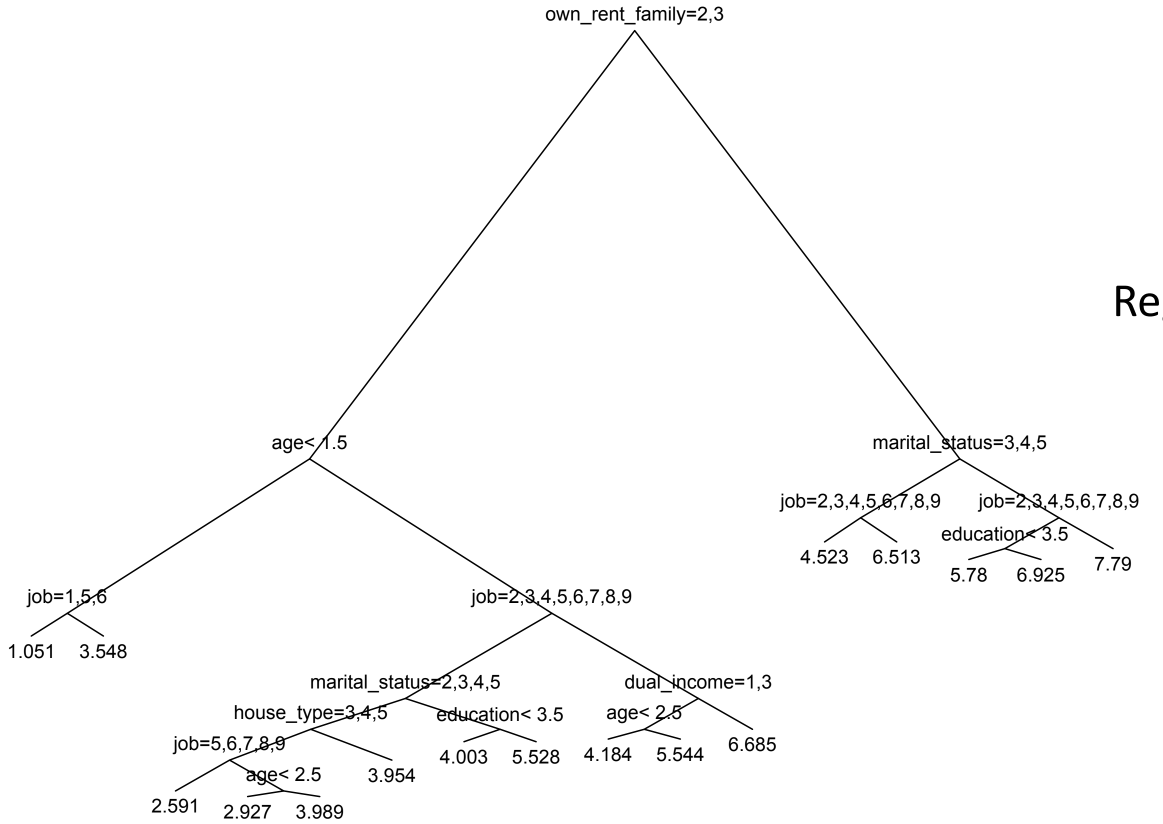
7 ANNUAL INCOME OF HOUSEHOLD (PERSONAL INCOME IF SINGLE
1. Less than $10,000
2. $10,000 to $14,999
3. $15,000 to $19,999
4. $20,000 to $24,999
5. $25,000 to $29,999
6. $30,000 to $39,999

8 HOW LONG HAVE YOU LIVED IN THE SAN FRAN./OAKLAND/SAN JOSE AREA?
1. Less than one year
2. One to three years
3. Four to six years
4. Seven to ten years
5. More than ten years

9 DUAL INCOMES (IF MARRIED)
1. Not Married
2. Yes
3. No

10 PERSONS IN YOUR HOUSEHOLD
1. One
2. Two
3. Three
4. Four
5. Five
6. Six
7. Seven
8. Eight
9. Nine or more

# Classification and Regression Trees




Regression

# Regression trees

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

$$\hat{c}_m = \mathrm{ave}(y_i | x_i \in R_m)$$

# Classification trees

$$\text{class } k(m) = \arg\max_k \hat{p}_{mk}$$

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

# Greedy algorithm

- Find splitting variable $j$ and split point $s$ that minimize prediction error

# Greedy algorithm: classification

- Multiple metrics for prediction error (node impurity)

Misclassification error: $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

# Categorical predictors, too many combinations

- $2^{q-1} - 1$ possible splits of $q$ unordered categories
- Improve computation time by ordering the categories based on their mean outcome values

# Missing predictor values

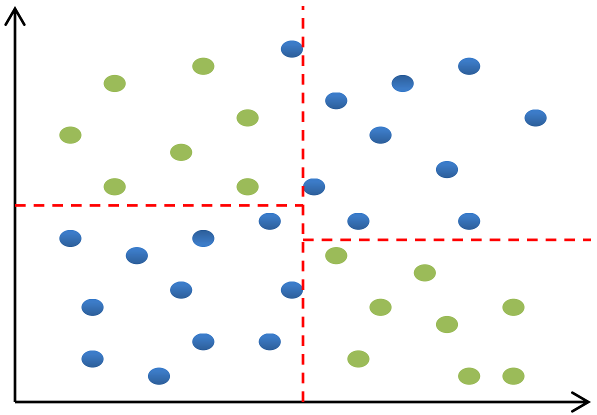
- Store "surrogate" predictors and split points, since predictors are often correlated

- Don't throw data away when building tree!

# Avoiding overfitting

- Stopping criterion? e.g. minimum decrease in prediction error

Problem:



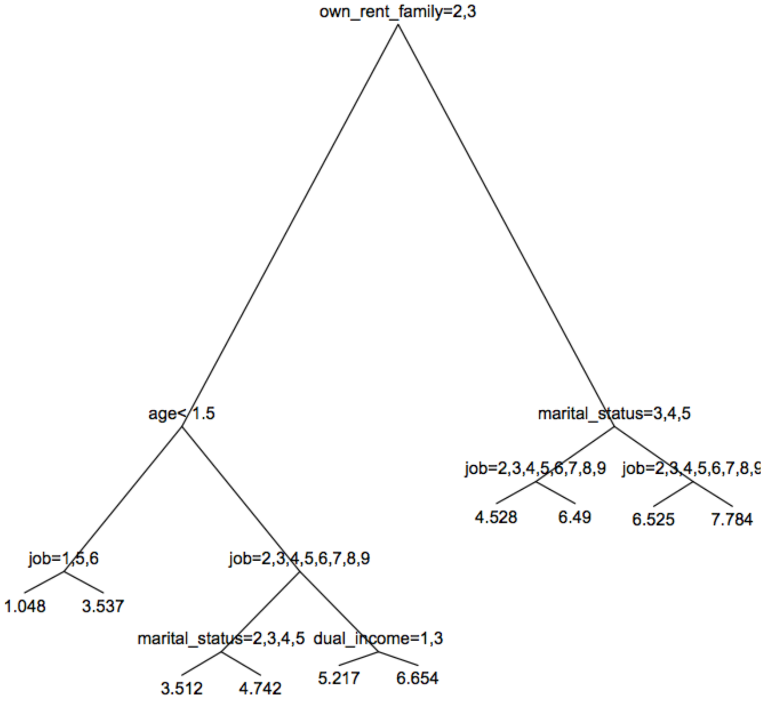
# Avoiding overfitting

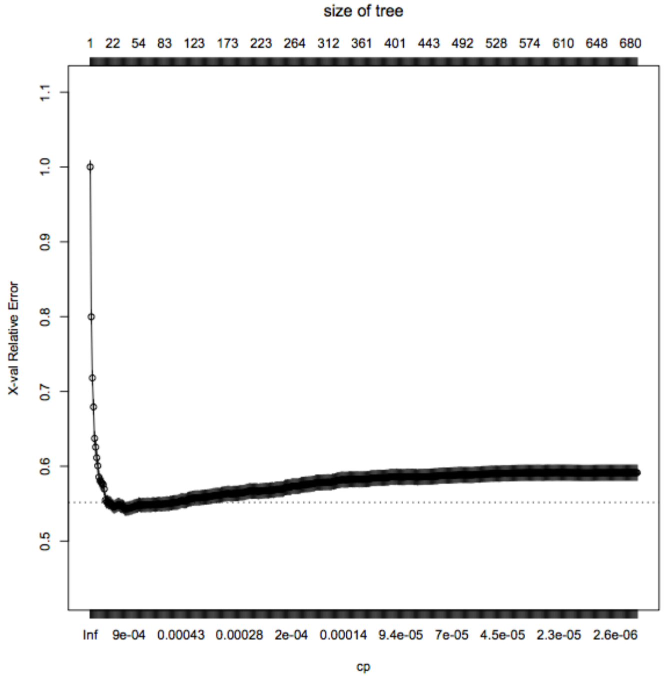
- Pruning:

  Grow large tree $T_0$

  Prune to some subtree $T \subset T_0$

# Cost-complexity pruning



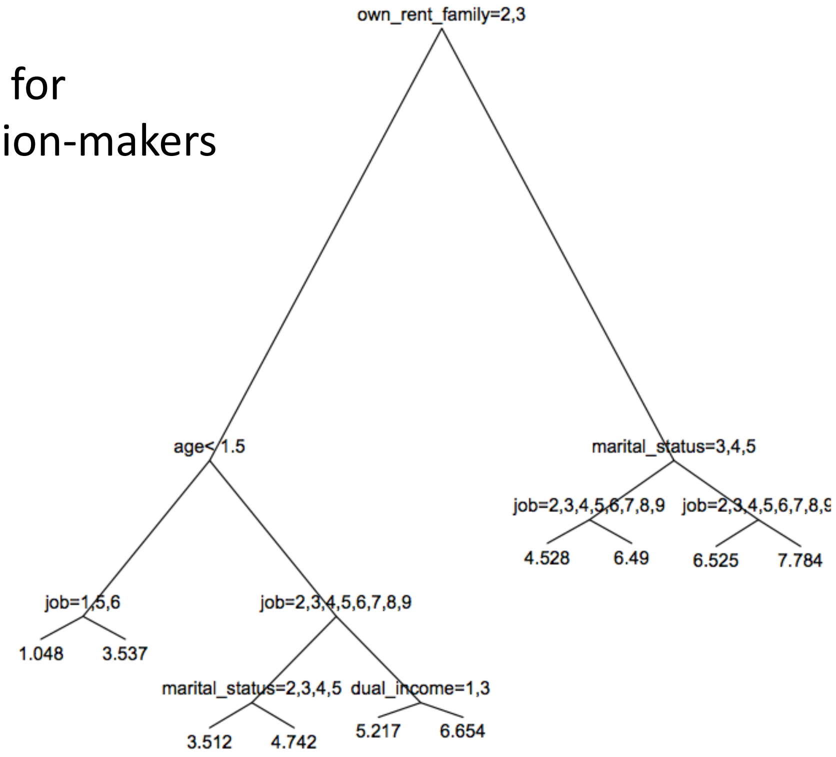



# Advantages of CART

1. Handles missing data easily (surrogate splits)
2. Robust to non-informative data
3. Automatic variable selection
4. Easily interpretable, ideal for explaining “why” to decision-makers
5. Captures high order interactions

# Advantages of CART

Easily interpretable, ideal for explaining "why" to decision-makers



# Advantages of CART

Captures high order interactions

$$Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2\ \ldots$$

$$Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \gamma_1\, x_1\, x_2 + \gamma_2\, x_1\, x_3 + \gamma_3\, x_2\, x_3 + \zeta_1\, x_1\, x_2\, x_3\ \ldots$$

Y = 3.5 if ((1<marital_status<6) AND (1<job<9)) AND (age<1.5) OR …

# Disadvantages of CART

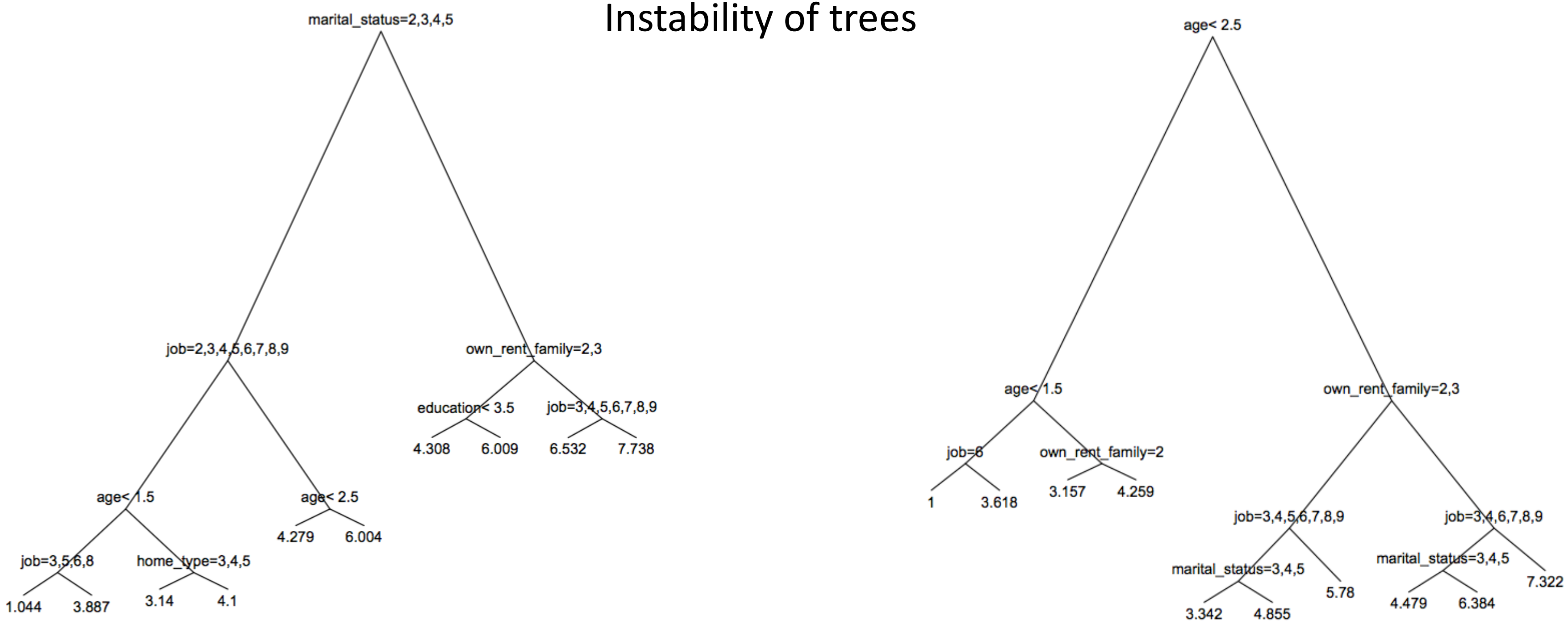
1. Instability of trees

2. Lack of smoothness

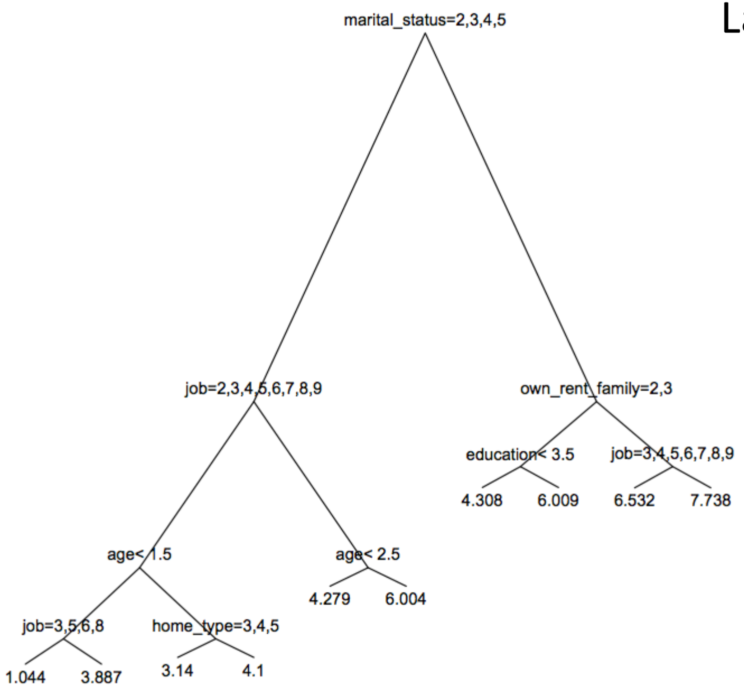
3. Hard to capture additivity

# Disadvantages of CART

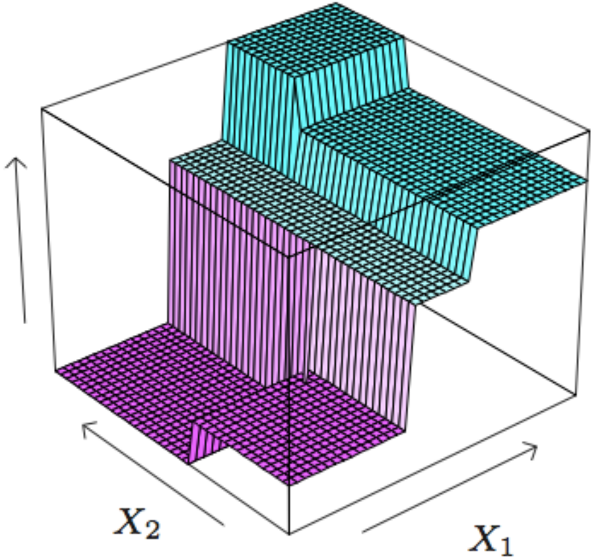
Instability of trees



# Disadvantages of CART



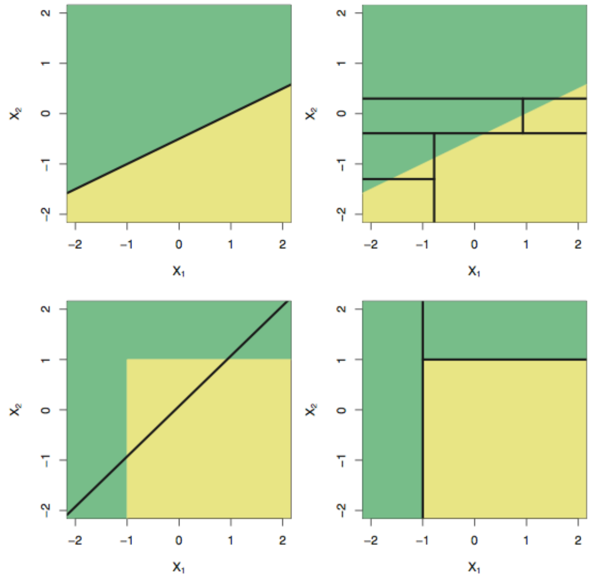

Lack of Smoothness



# Disadvantages of CART

Hard to capture additivity

$$Y = c_1 I(X_1 < t_1) + c_2 I(X_2 < t_2) + e$$




Hastie, Trevor, et al. Introduction to statistical learning.

# Disadvantages of CART

1. Instability of trees

   *Solution - Random Forests*

2. Lack of smoothness

   *Solution - MARS*

3. Hard to capture additivity

   *Solution – MART* or *MARS*

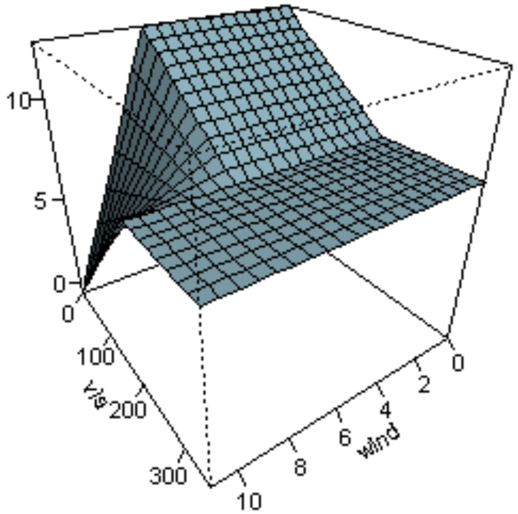
# Extensions

- MART – “Multiple Additive Regression Trees”

- MARS – “Multivariate Adaptive Regression Splines”

# MARS – “Multivariate Adaptive Regression Splines”

- Invented by Jerome Friedman in 1991



# Ensemble Methods: Bagging, and Random Forests

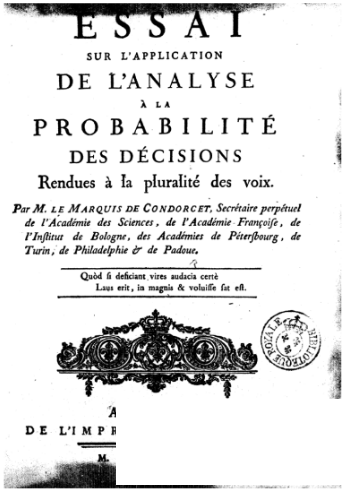
Alexander Ioannidis
ioannidis@stanford.edu

Institute for Computational and Mathematical Engineering,
Stanford University

# Ensemble Methods

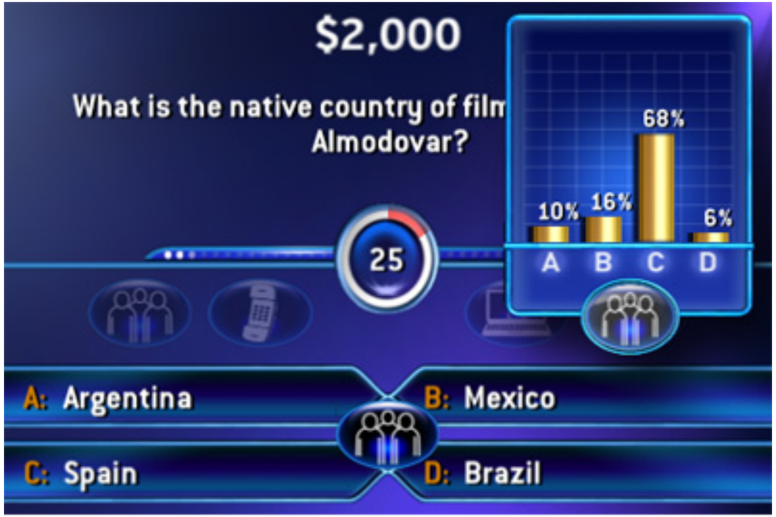
# The strength of weak classifiers

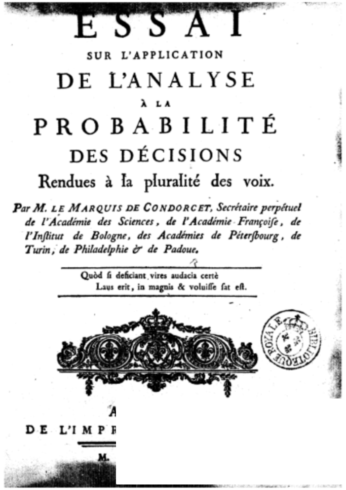
Condorcet's Jury Theorem - If p is greater than 1/2 (each voter is more likely to vote correctly), then adding more voters increases the probability that the majority decision is correct. In the limit, the probability that the majority votes correctly approaches 1 as the number of voters increases.



# The strength of weak classifiers

**Condorcet's Jury Theorem** - If p is greater than 1/2 (each voter is more likely to vote correctly), then adding more voters increases the probability that the majority decision is correct. In the limit, the probability that the majority votes correctly approaches 1 as the number of voters increases.



# The strength of weak classifiers

- Averaging reduces variance without raising bias (bias remains unchanged)

$$\mathrm{Var}[\bar{Y}] = \sigma^2/n$$

# The strength of weak classifiers

- Averaging reduces variance without raising bias (bias remains unchanged)

$$\mathrm{Var}[\bar{Y}] = \sigma^2/n$$

- The votes of correlated classifiers don't help as much

**THE CHOICE OF A CANDIDATE**

THE NEW YORK TIMES supported Franklin D. Roosevelt for the Presidency in 1932 and again in 1936. In 1940 it will support Wendell Willkie.

# The strength of weak classifiers

- Averaging reduces variance without raising bias (bias remains unchanged)
  $\mathrm{Var}[\bar{Y}] = \sigma^2/n$

    - The votes of correlated classifiers don't help as much
      $\mathrm{Var}[\bar{Y}] = \sigma^2/n + (\rho\sigma^2)(n-1)/n$

# The strength of weak classifiers

- Averaging reduces variance without raising bias (bias remains unchanged)
  $\mathrm{Var}[\bar{Y}] = \sigma^2/n$

- The votes of correlated classifiers don’t help as much → Random Forest

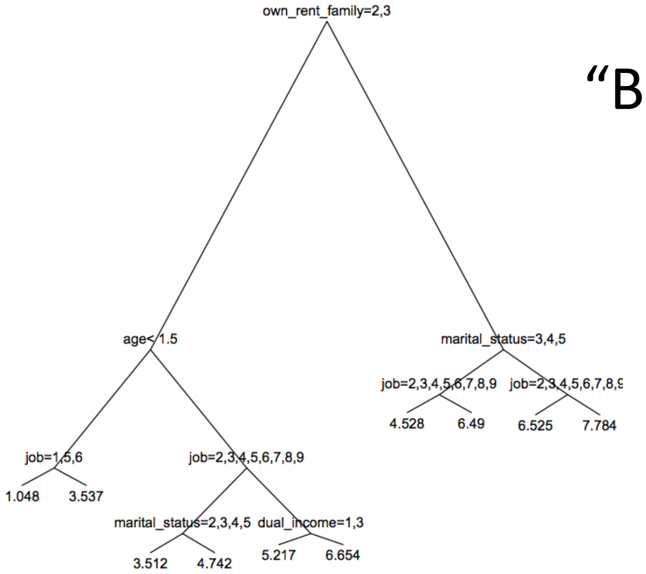
# Be Creative

$$\alpha \cdot \{CART\} + (1-\alpha) \cdot \{LinearModel\}$$

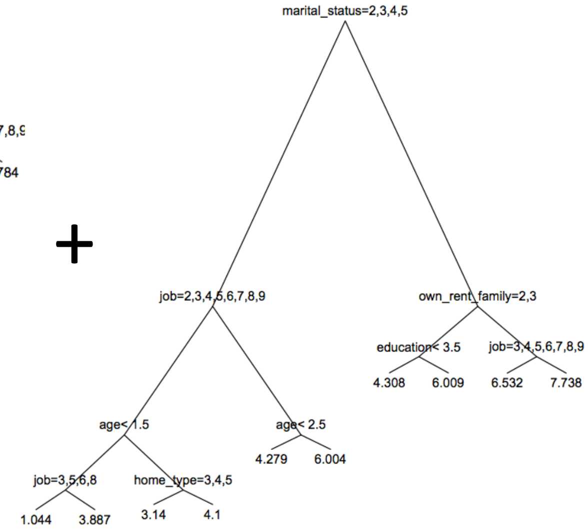
# Ensemble Methods: Bagging

# What is bagging?

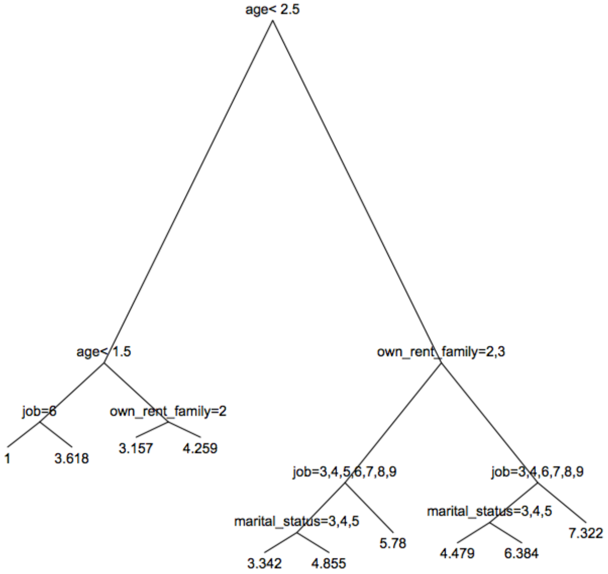
## “Bootstrap Aggregation”




+




+



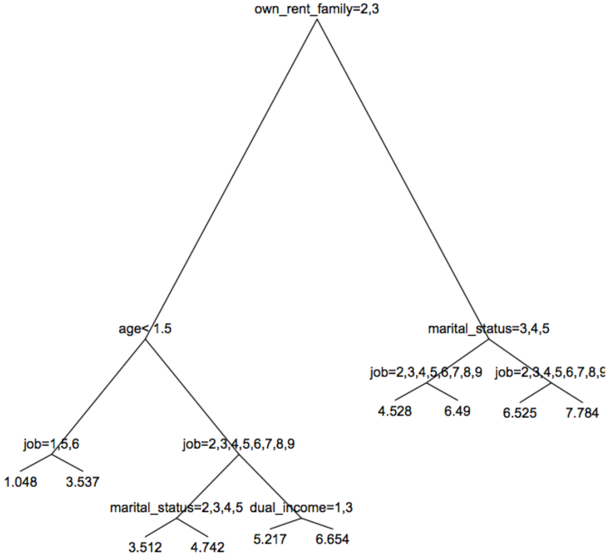
# What is bagging?

"Bootstrap Aggregation"

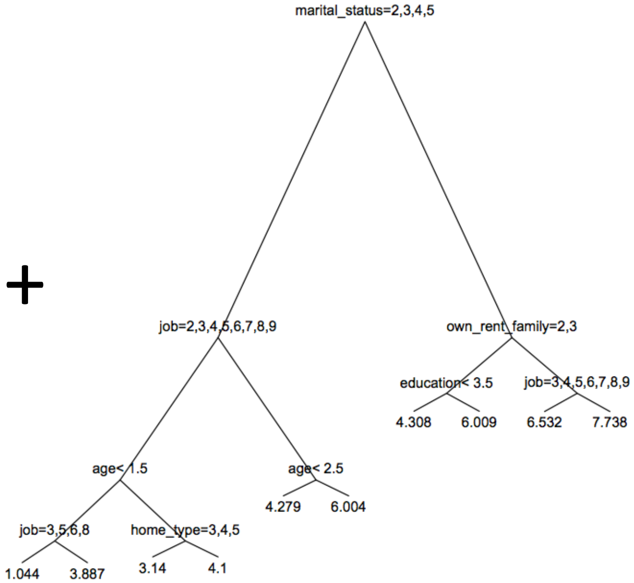
$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

# Bagging

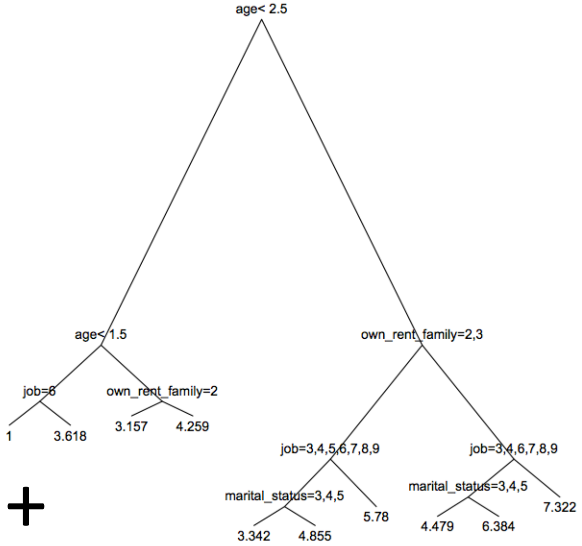
Address the instability of CART




+

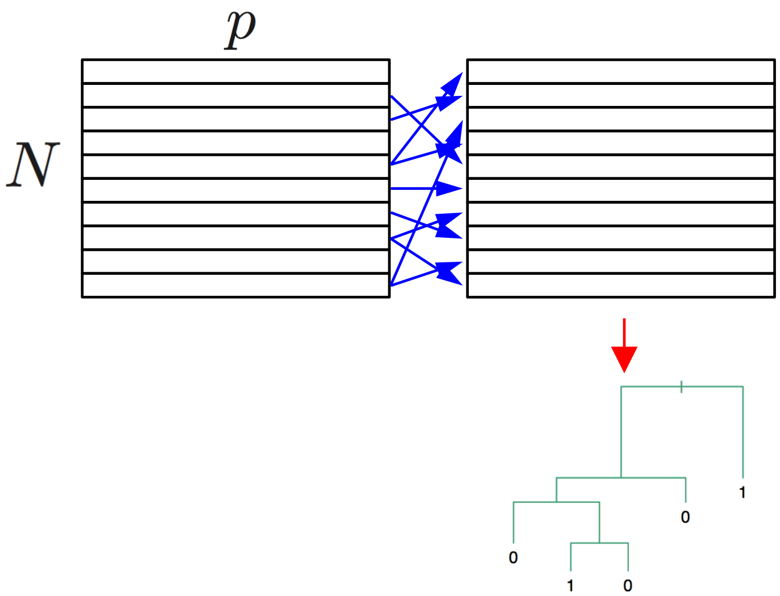



+



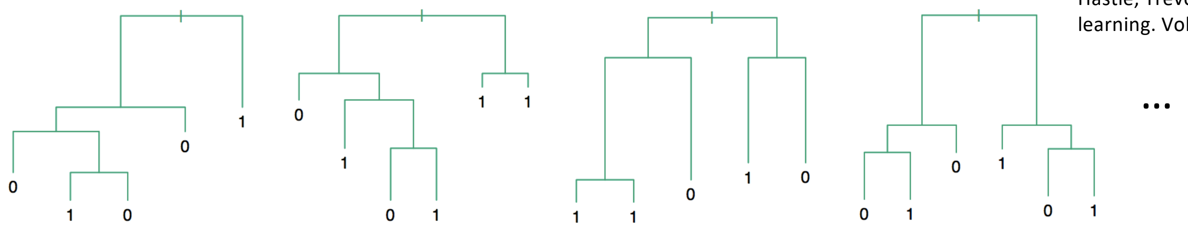
# Bagging

- Bootstrap the training data samples to build an ensemble of predictors.



# Bagging

- Bootstrap the training data samples to build an ensemble of predictors.




Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

- Average (or majority vote) the individual predictions.
- Bagging reduces variance and maintains bias.

# Bagging

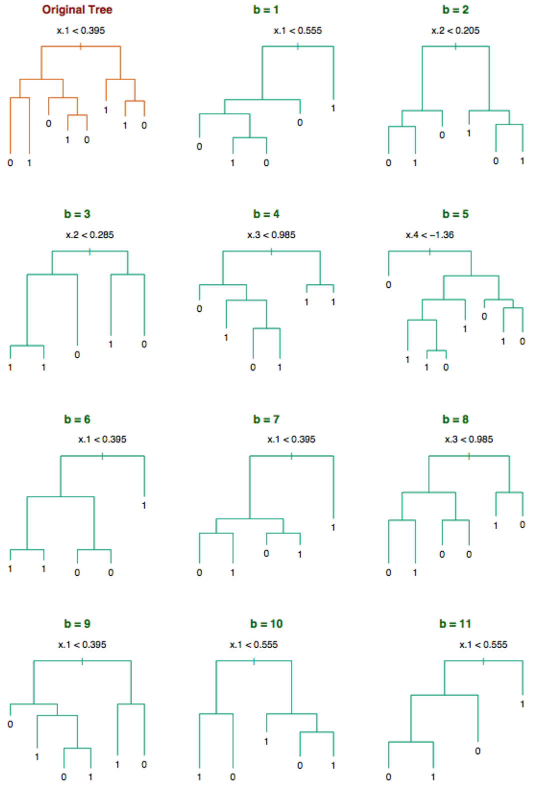
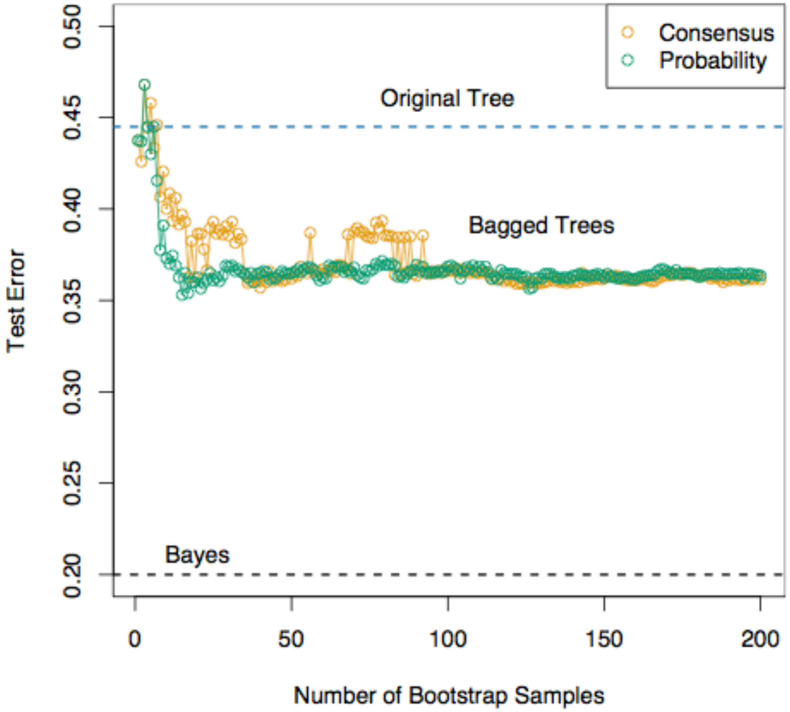



**FIGURE 8.9.** *Bagging trees on simulated dataset. The top left panel shows the original tree. Eleven trees grown on bootstrap samples are shown. For each tree, the top split is annotated.*

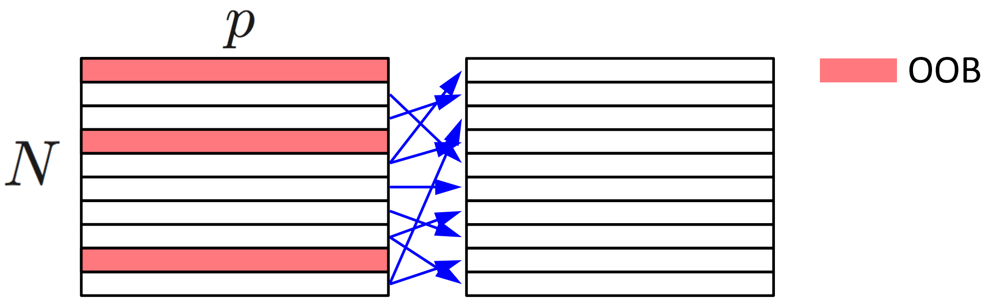
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.



# Bonus! Out-of-bag cross-validation

# Out-of-bag (OOB) samples

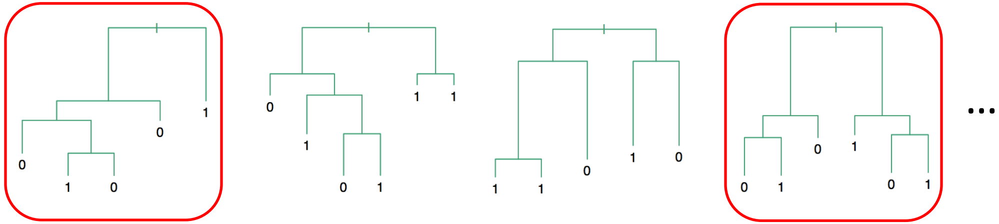
- Bootstrapping process:




- Each tree uses only a subset of the training samples (~2/3 of samples on average).
- Each sample is OOB for ~1/3 of trees.

# Predictions for OOB samples

- For each sample, find the trees for which it is OOB.




- Predict its value from each of those trees.
- Estimate prediction error of the bagged trees using all of the OOB predictions.
- Similar to cross-validation.

# Random Forests

# Bagged trees vs. random forests

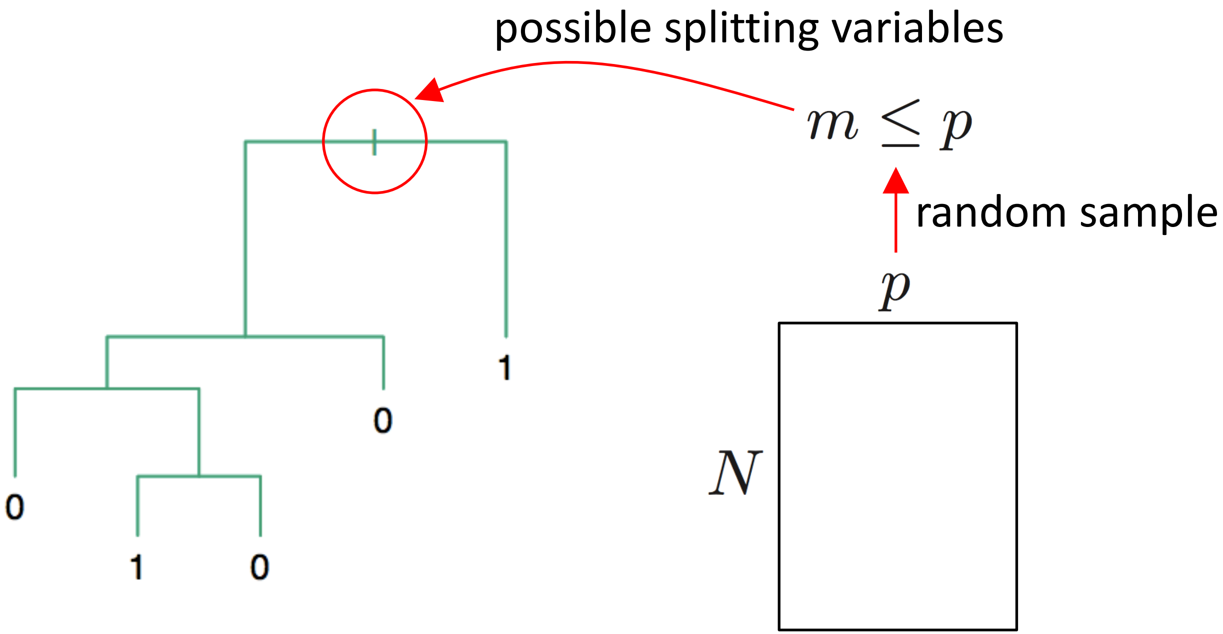
- Bagging introduces variability between trees by random selection of training data.

- Bagged trees can still be correlated, limiting the reduction in variance.

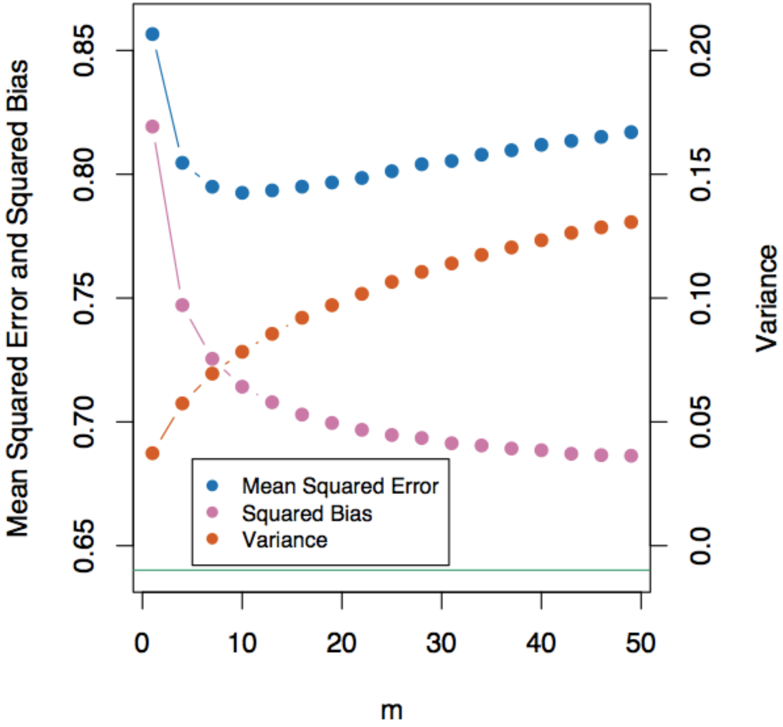
**Random forests introduce additional randomness:**

- Reduce correlation between trees by randomizing the variables considered for splitting at each node.

# Candidate splitting variables



# Candidate splitting variables




Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

# Important parameters

**Parameters of random forests:**

- # of candidate splitting variables at each node ($m$)
- Depth of each tree (minimum node size)
- # of trees

# Number of splitting variables

| | Default values |
|---|---|
| Classification | $m = \lfloor \sqrt{p} \rfloor$ |
| Regression | $m = \lfloor p/3 \rfloor$ |

# Important parameters

**Parameters of random forests:**

- # of candidate splitting variables at each node ($m$)
- Depth of each tree (minimum node size)
- # of trees

# Tree depth (minimum node size)

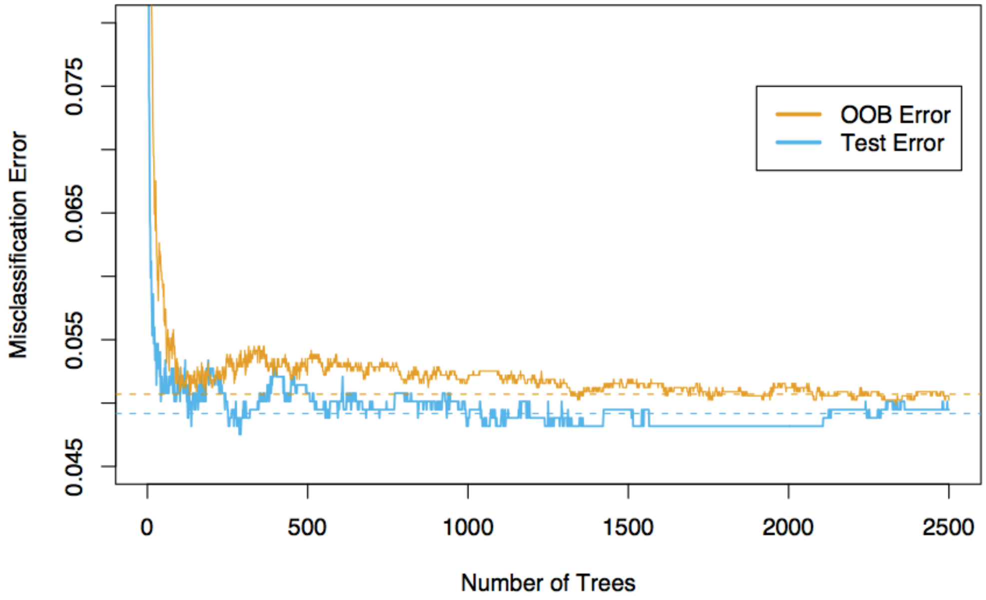
| | Default values |
|---|---|
| Classification | 1 |
| Regression | 5 |

# Important parameters

**Parameters of random forests:**

- # of candidate splitting variables at each node ($m$)
- Depth of each tree (minimum node size)
- # of trees

# Number of trees




Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

- Adding more trees does not cause overfitting.

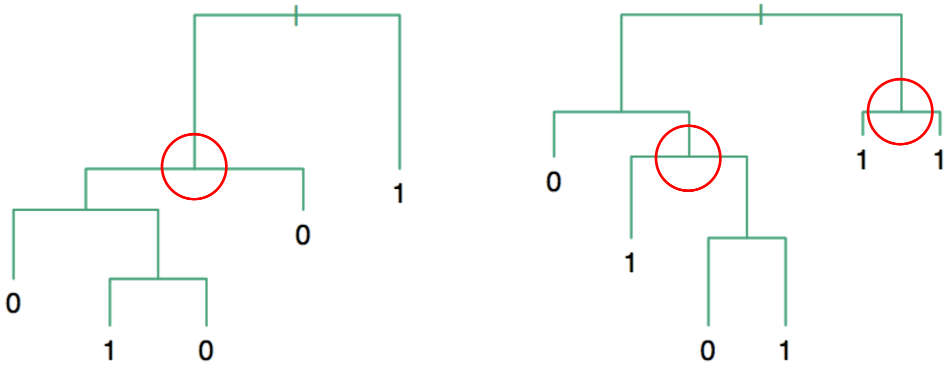
# Other features of random forests

- Out-of-bag (OOB) samples
- Variable importance measurements

# Variable importance

**Metric 1:**

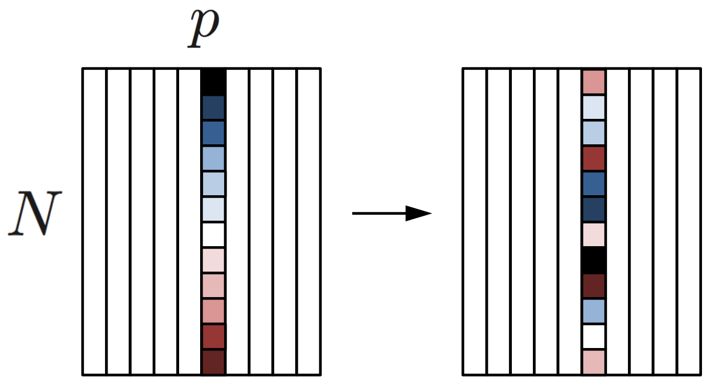
Decrease in prediction error or impurity from all splits involving that variable, averaged over trees.



# Variable importance

**Metric 2:**

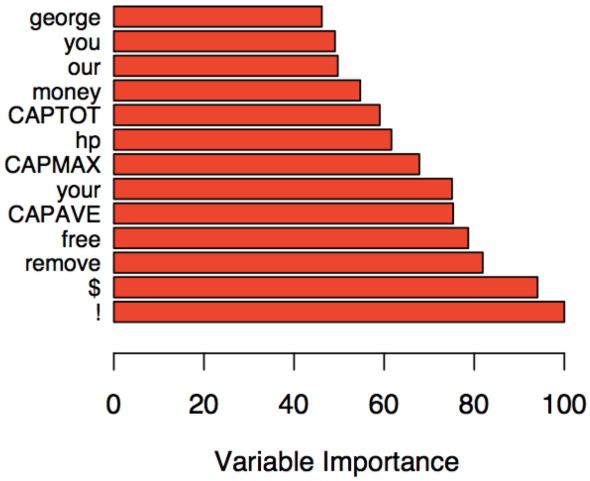
Increase in overall prediction error when the values of that variable are randomly permuted between samples.



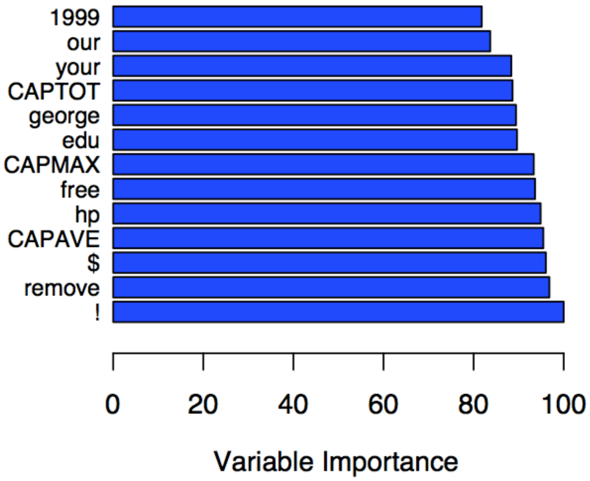
# Variable importance example

- The metrics give similar but not identical rankings:

Metric 1




Metric 2




Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

# Advantages of random forests

**Similar to CART:**

- Relatively robust to non-informative variables (built-in variable selection)
- Capture high-order interactions between variables
- Low bias
- Naturally handle mixed predictors (quantitative and categorical)

# Advantages of random forests

**Advantages over CART:**

- Lower variance (more robust to choice of training data due to bootstrapping)
- Less prone to overfitting
- No need for pruning
- Built-in cross-validation (using OOB samples)

# Disadvantages of random forests

**Similar to CART:**

- Hard to capture additive effects

**Disadvantages relative to CART:**

- Hard to interpret/explain the model predictions

# Questions?