

ONLINE MASTERS IN **DATA SCIENCE**

DSC 255 - MACHINE LEARNING FUNDAMENTALS

RANDOM FORESTS

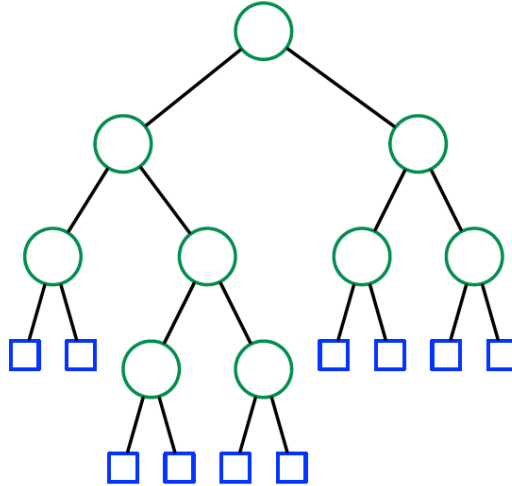
SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

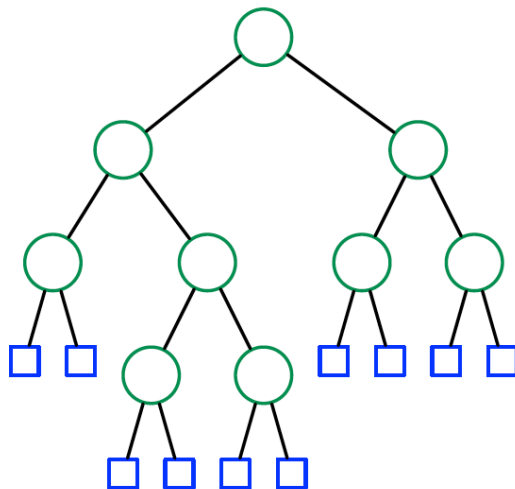


From Tree to Forest



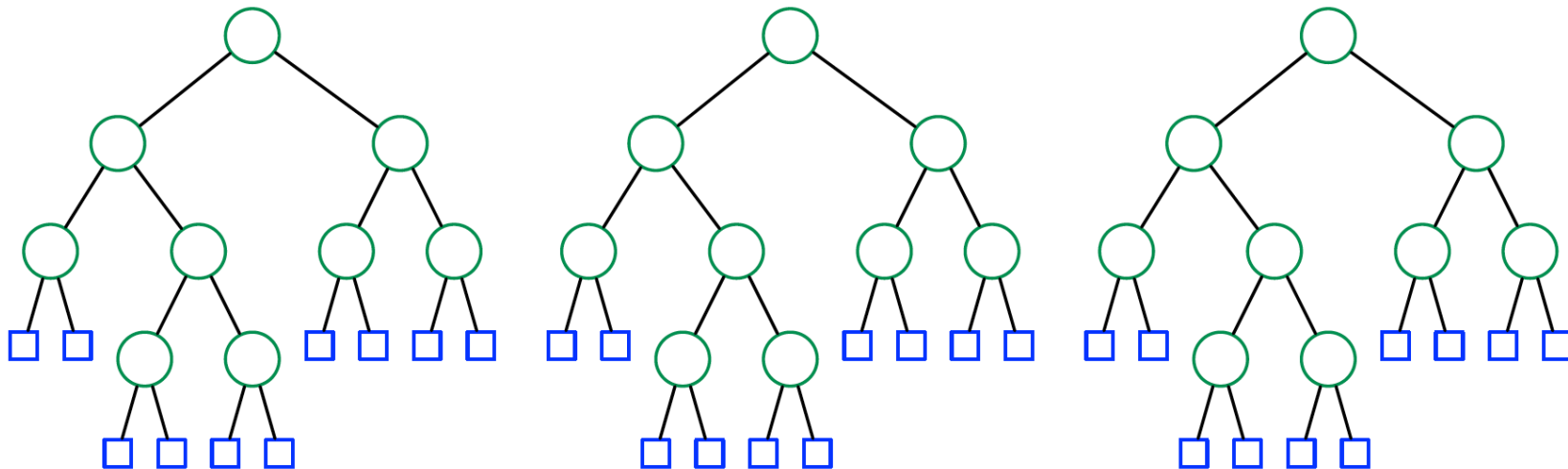
- Decision tree.

From Tree to Forest



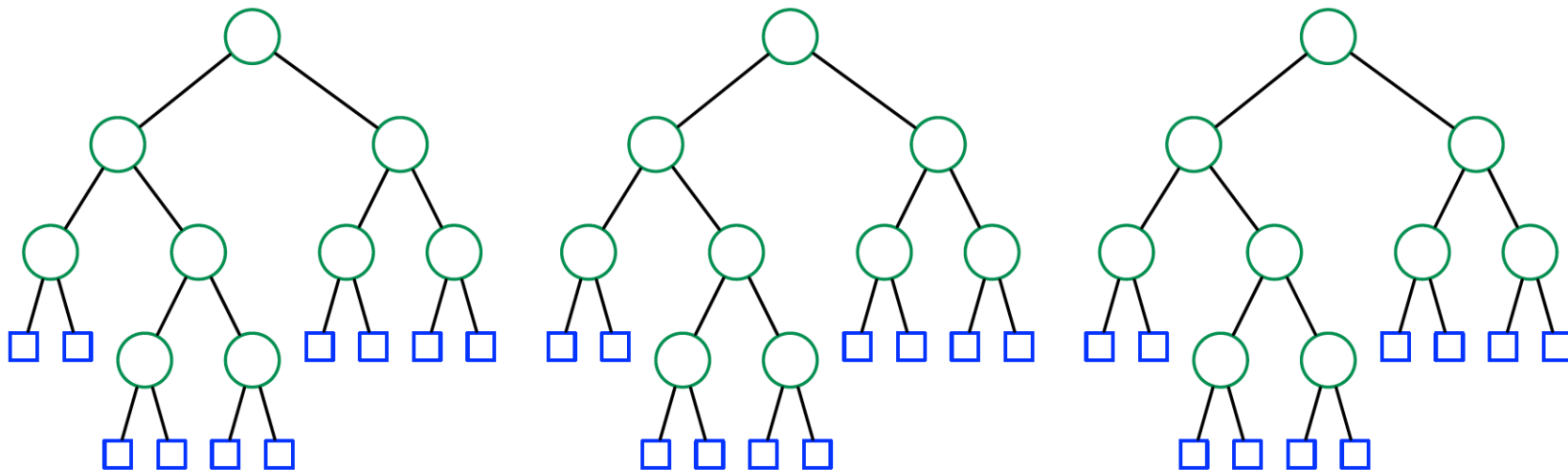
- **Decision tree.** Starts overfitting beyond a point.

From Tree to Forest



- **Decision tree.** Starts overfitting beyond a point.
- **Boosted decision trees.**

From Tree to Forest



- **Decision tree.** Starts overfitting beyond a point.
- **Boosted decision trees.** Learning is sequential, slow.

Random Forests

Given a data set S of n labeled points:

- For $t = 1$ to T :
 - Choose n' points randomly, with replacement, from S .
 - Fit a decision tree h_t to these points.
 - At each node restrict to one of k features chosen at random.

Example settings:

- $n' = n$
- $k = \sqrt{d}$ for d -dimensional data

Final predictor: majority vote of h_1, \dots, h_T .

Ecological Prediction Problem: "Covertypes" Data

Predict forest type:

- Spruce-fir
- Lodgepole pine
- 5 other classes

54 cartographic/geological features:

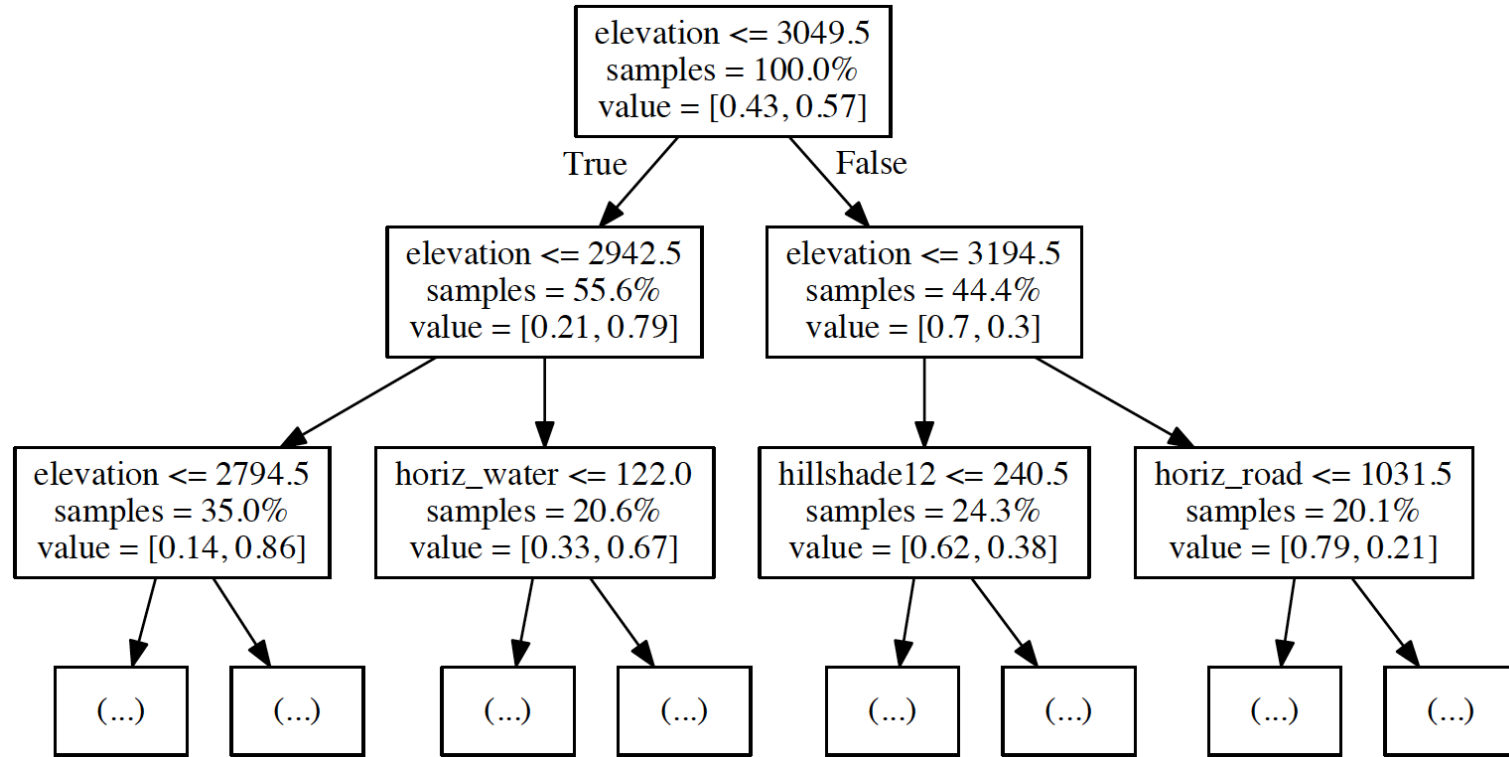
- Elevation, slope, amount of shade, ...
- Distance to water, road, ...
- Soil type

Data set details:

- 49,514 training points
- 445,627 test points



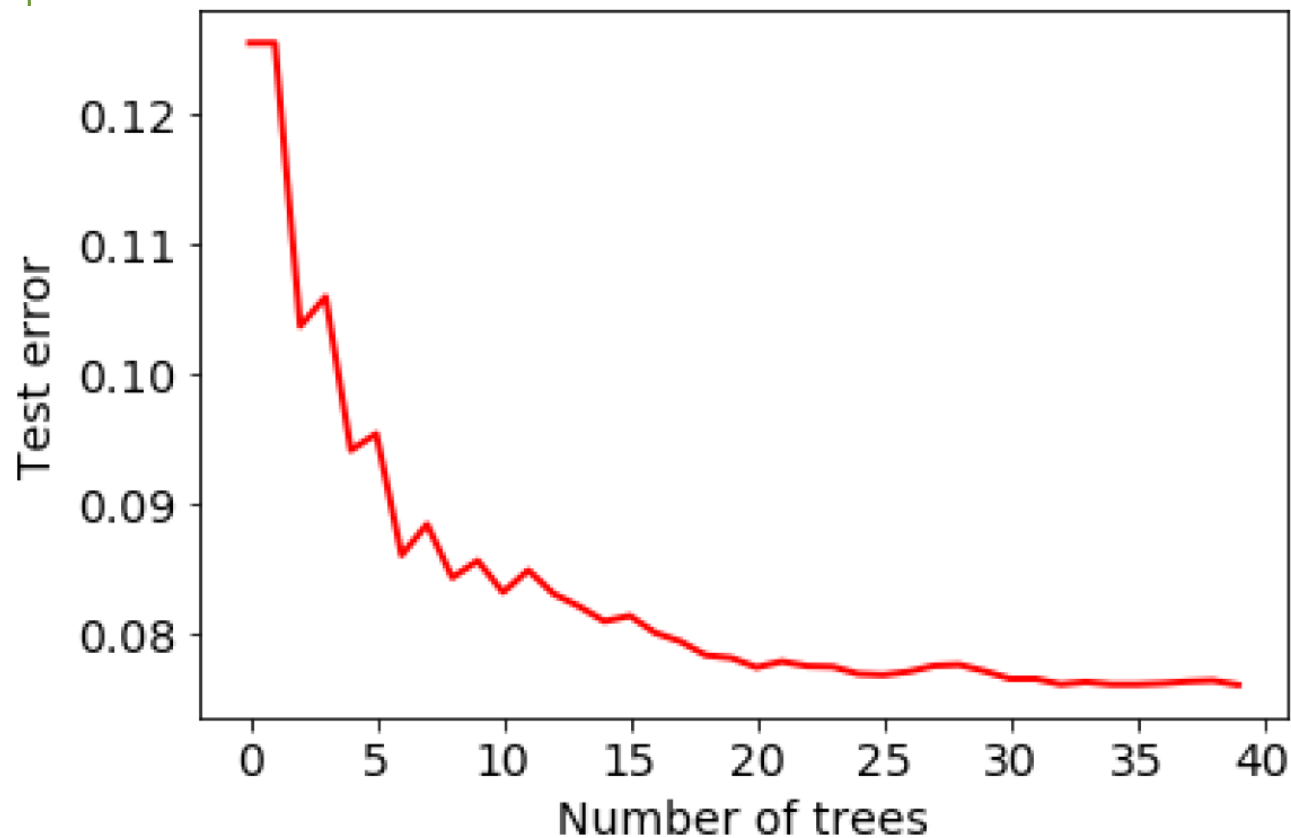
Decision Tree



Depth 20: training error 1%, test error 12.6%

Boosted Decision Trees

Trees of depth 20.



Random Forest

Recall:

- Decision tree: depth 20, test error 12.6%
- Boosted decision trees, 10 trees, depth 20: test error 8.7%

Random forest setting: 10 trees, 50% features dropped, depth 40.

- Each individual tree has test error 15% to 17%
- Forest test error: 8.8%