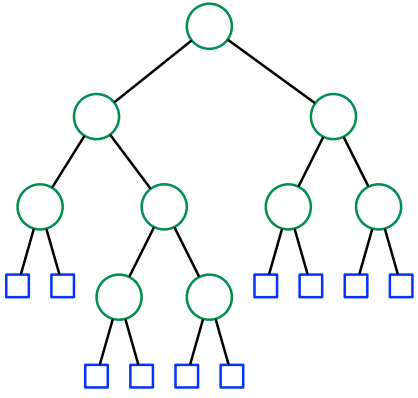
Random forests

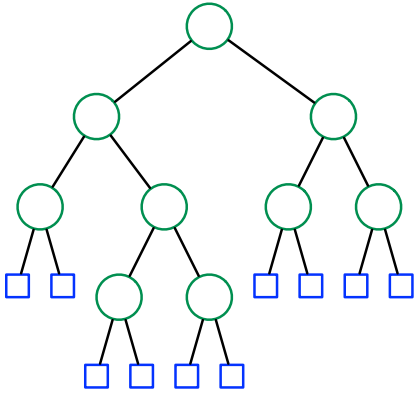
DSE 220

From tree to forest



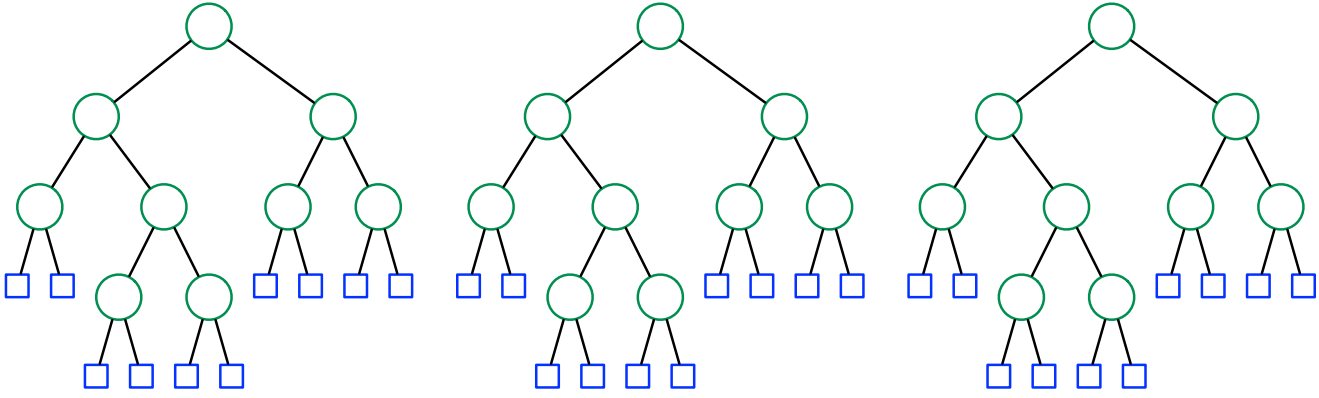
- Decision tree.

From tree to forest



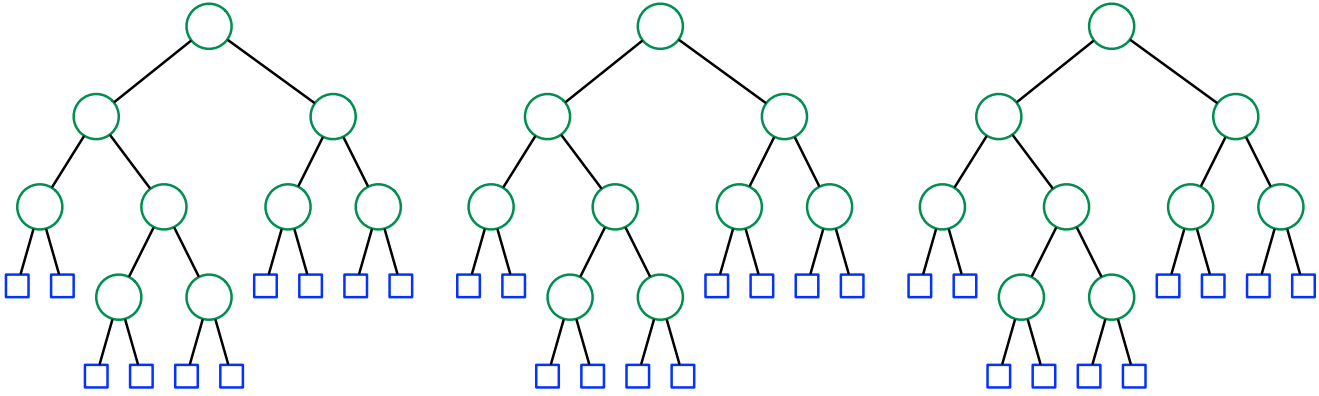
- Decision tree. Starts overfitting beyond a point.

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.

Can we build multiple trees in parallel? Need to make sure they're DIFFERENT.
Idea: Inject randomness into the tree-building process.

Random forests

E.g. 4 points: 1 2 3 4

Pick 4 points with replacement: 3 1 3 2

Two types of randomization:

Given a data set S of n labeled points:

Build T trees

- For $t = 1$ to T:
  - Choose n' points randomly, with replacement, from S. (1) Typically $n' = n$.
  - Fit a decision tree h_t to these points.
    - At each node restrict to one of k features chosen at random. (2)

Forces trees to be diverse, to make predictions based on different reasoning... robustness.

Example settings:

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

Final predictor: majority vote of $h_1, \ldots, h_T$. (equal weight)

Ecological prediction problem: "covertype" data

Predict forest type:

- Spruce-fir
- Lodgepole pine
- 5 other classes

overwhelming majority of points are these

← ignore

54 cartographic/geological features:

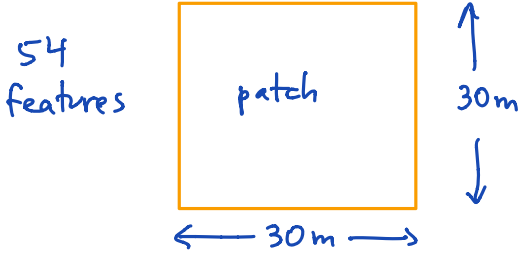
- Elevation, slope, amount of shade, ...
- Distance to water, road, ...
- Soil type

Data set details:

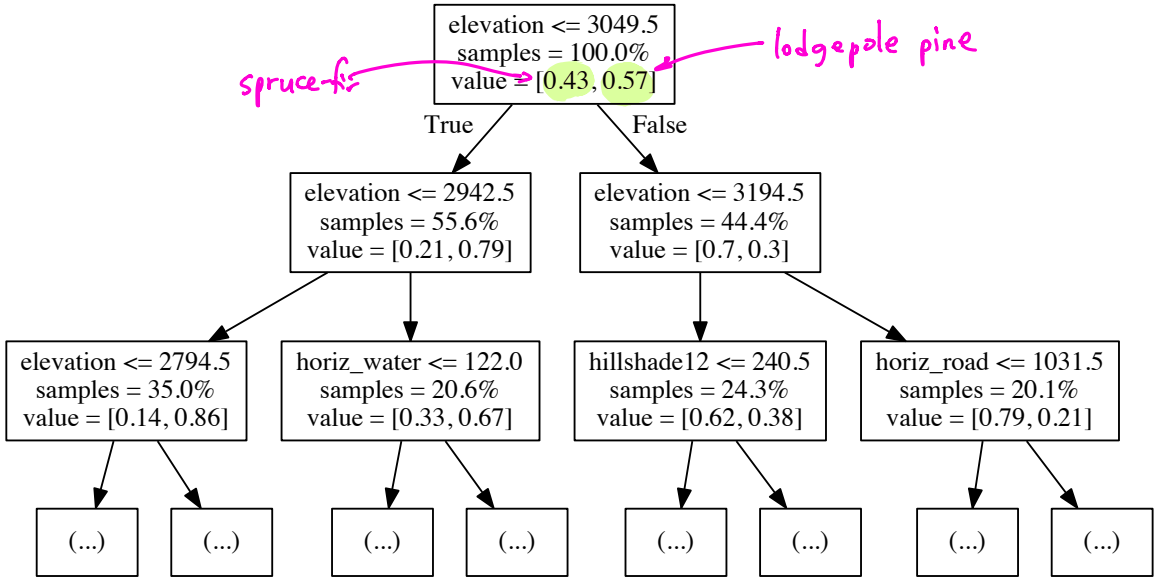
- 49,514 training points
- 445,627 test points ←

Data from different types of forest in US national parks

Each data pt:



Decision tree

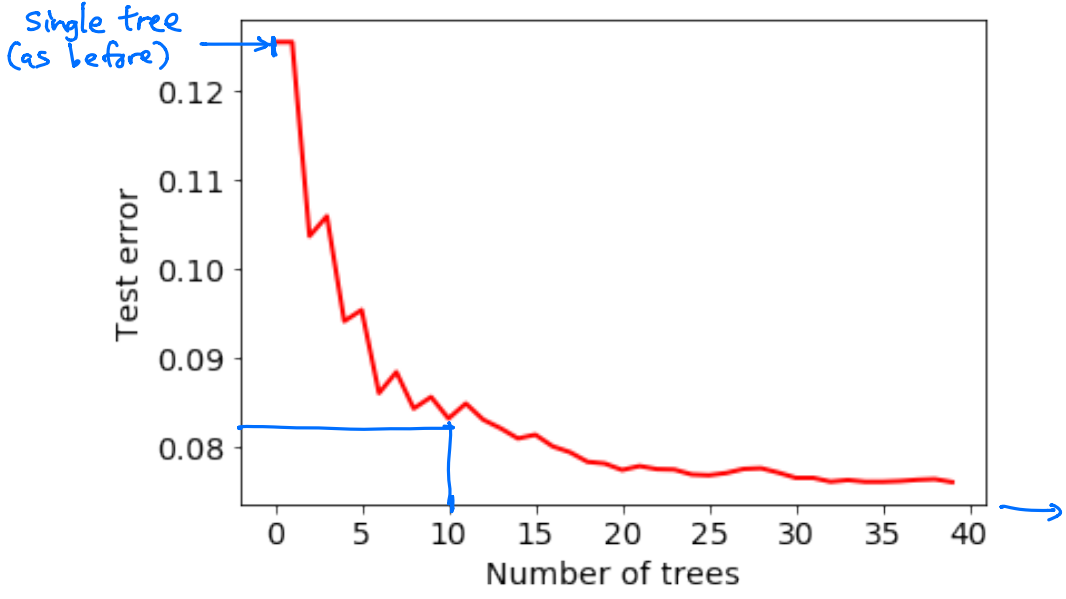


Depth 20: training error 1%, test error 12.6%

Growing tree further doesn't help

Boosted decision trees

Trees of depth 20.



Random forest

Lab 3 # 1, 2

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%