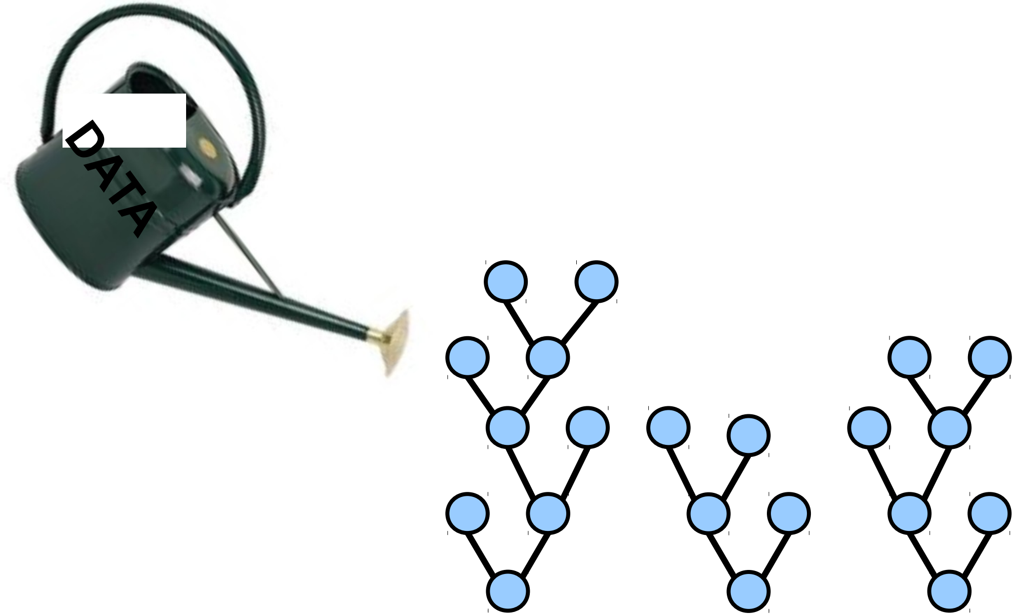
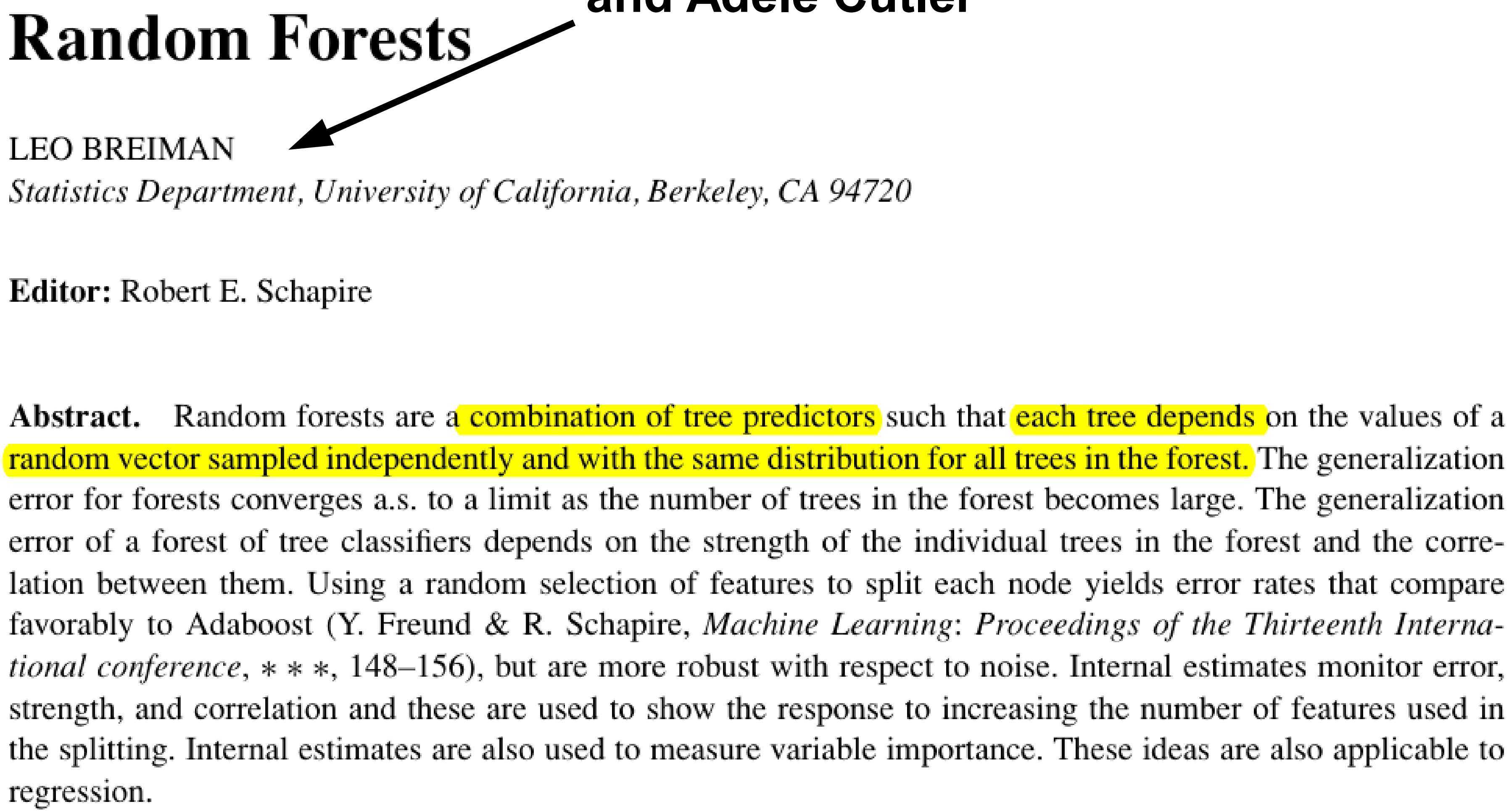
Learning by growing trees:
An introduction to Random Forests



and Adele Cutler



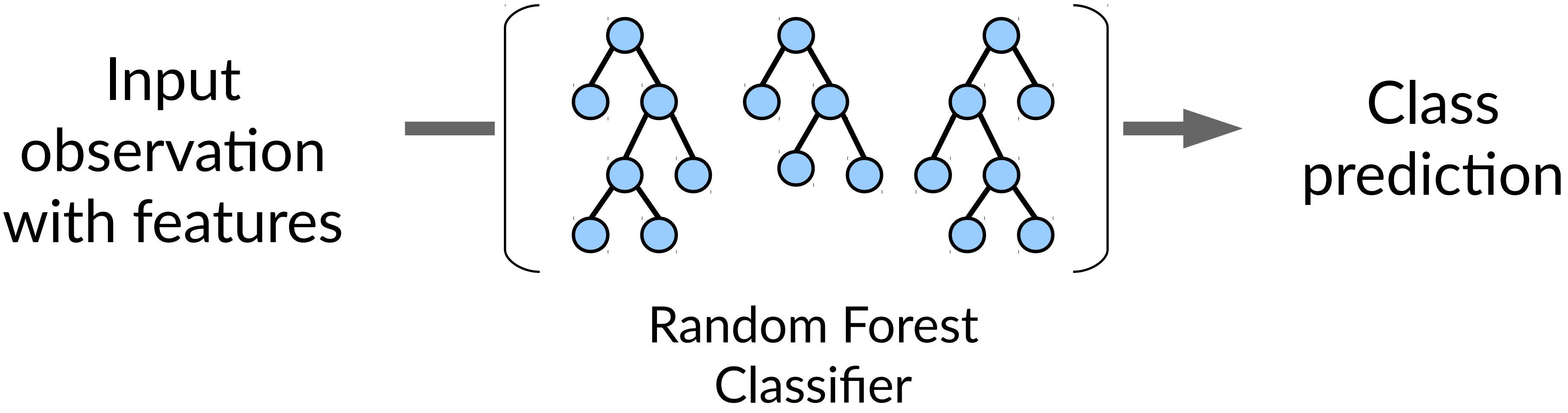
Random Forests

LEO BREIMAN
Statistics Department, University of California, Berkeley, CA 94720

Editor: Robert E. Schapire

Abstract. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Y. Freund & R. Schapire, *Machine Learning*: *Proceedings of the Thirteenth International conference*, ∗ ∗ ∗, 148–156), but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

Random forests for classification

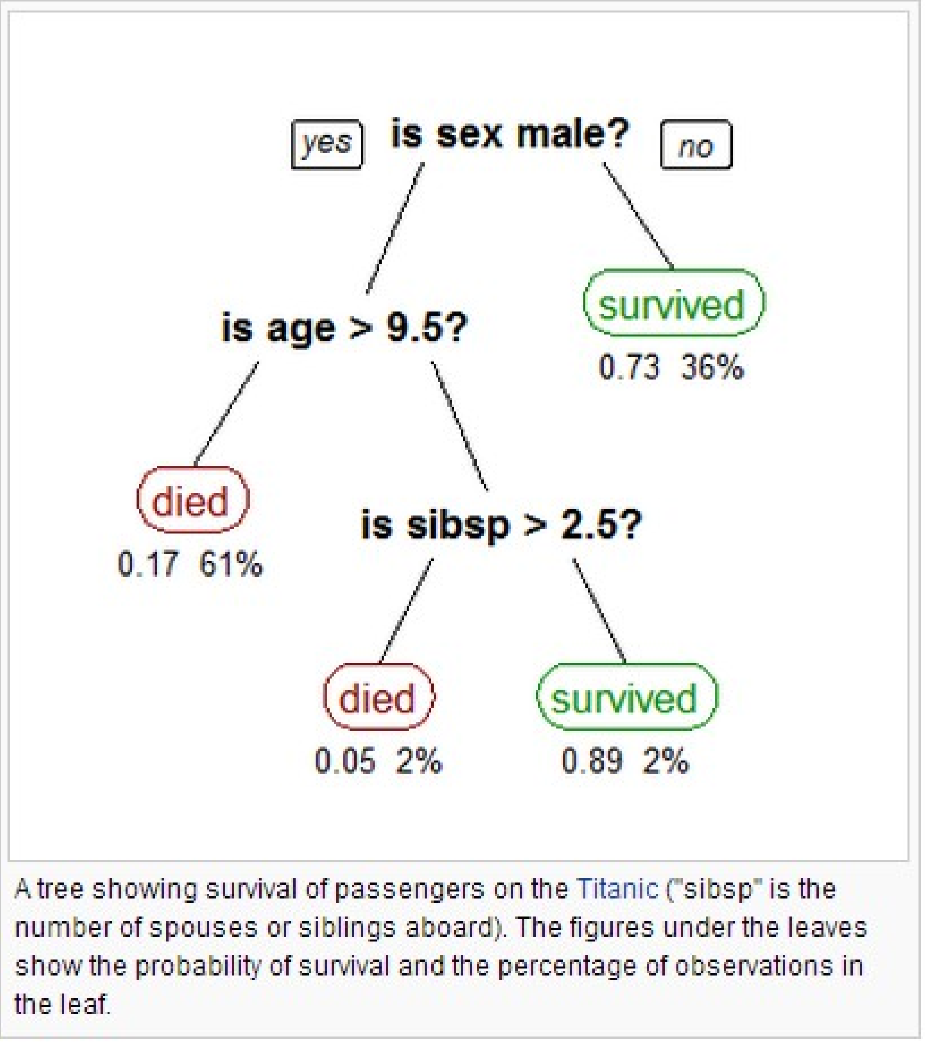


Input observation with features

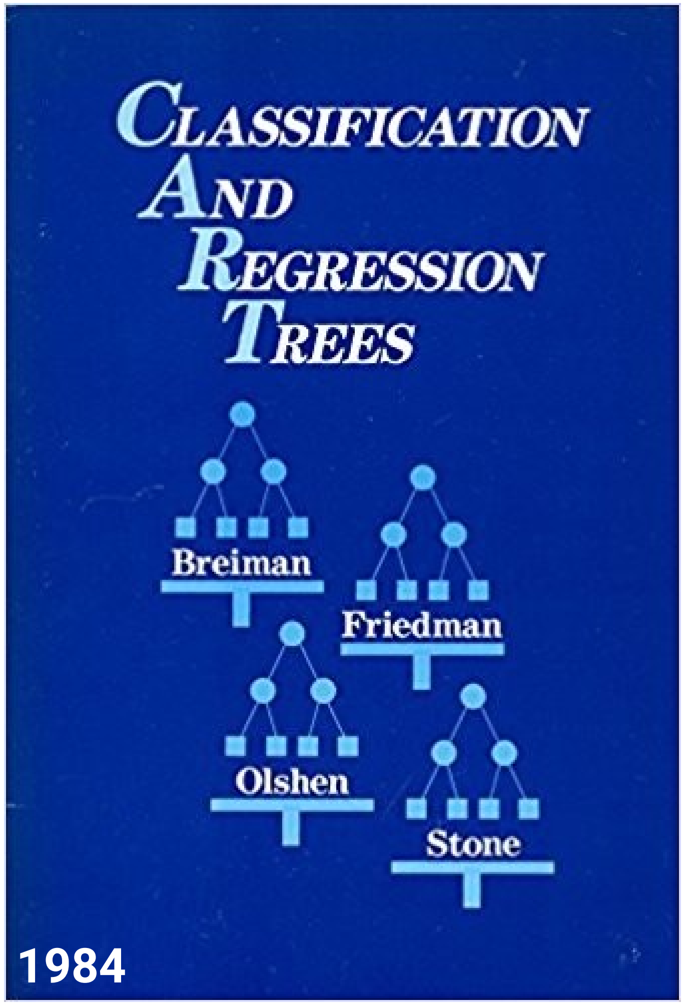
Class prediction

Random Forest Classifier

What's a decision tree?

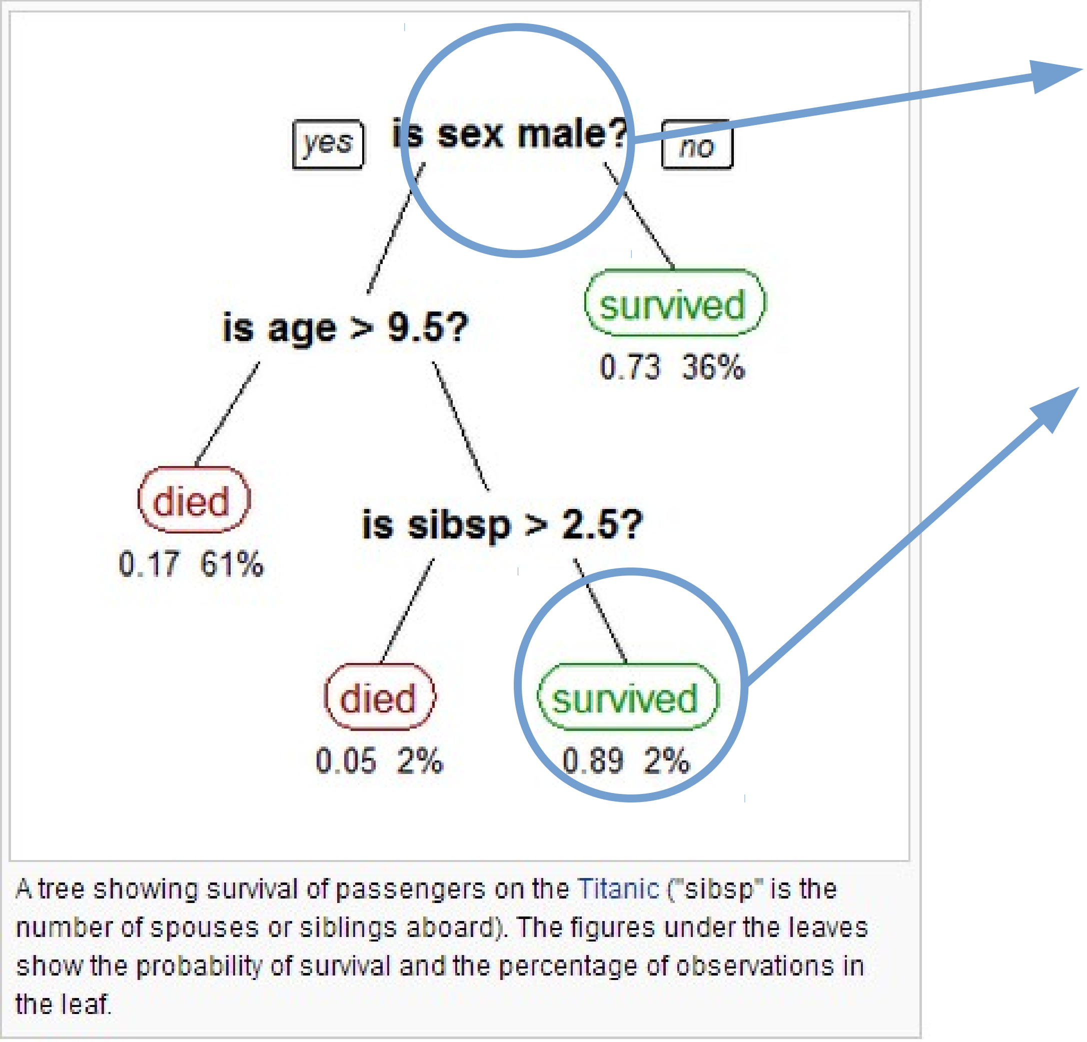


A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.



1984

What's a decision tree?



A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

- How to choose a feature at a node?
- What makes a node a leaf?
- How to define the maximum depth of the tree?

Gini impurity criterion

Gini impurity is a measure of how often a **randomly chosen element** from the set would be **incorrectly labeled** if it was **randomly labeled** according to the distribution of labels in the subset.

$$I_G(f) = \sum_{i=1}^{J} f_i(1 - f_i)$$

$$I_G(f) = \sum_{i \neq k} f_i f_k$$

Decision trees characteristics

- Easy to understand / interpret
- Require little data preparation
- Performs well with large amount of data

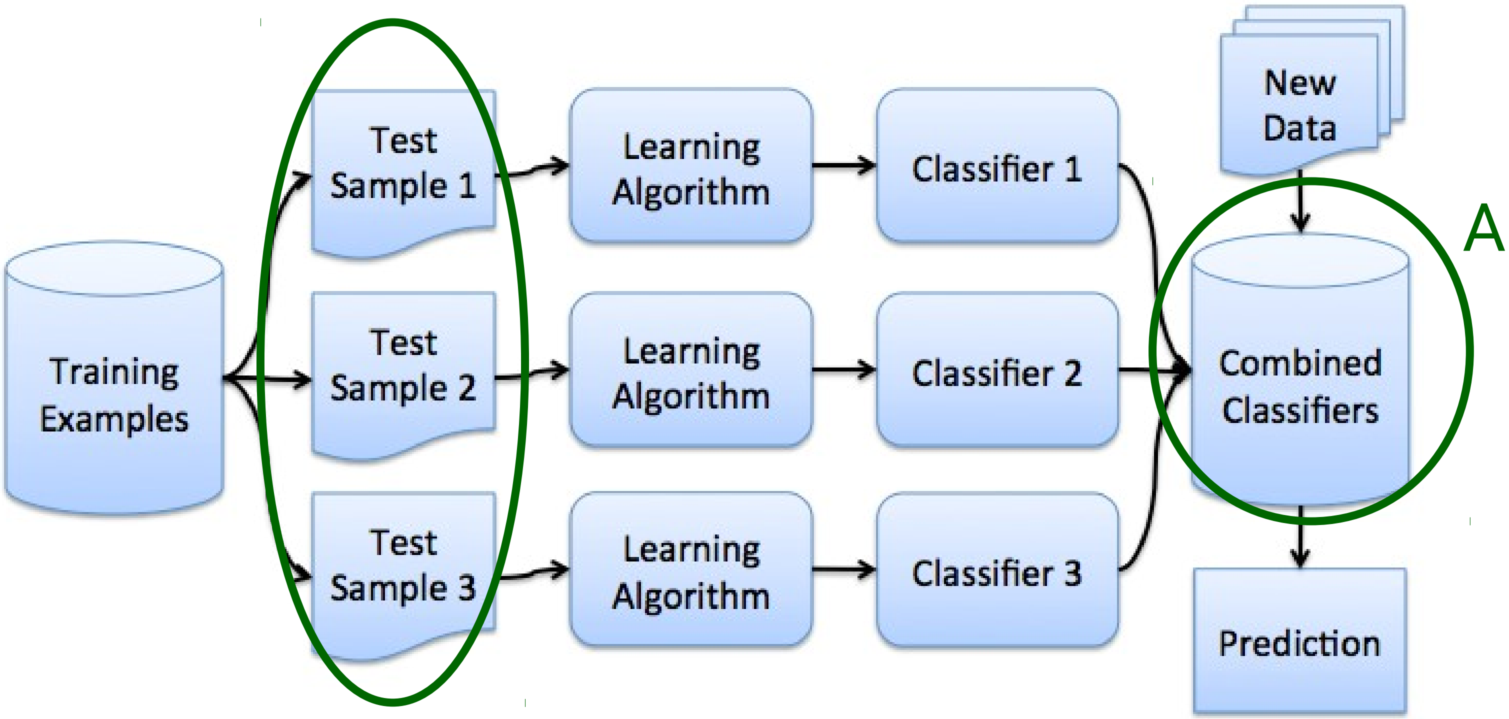
Decision trees characteristics

- Easy to understand / interpret
- Require little data preparation
- Performs well with large amount of data

- Prone to overfitting
- Not robust to changes in the training dataset

Growing a random forest

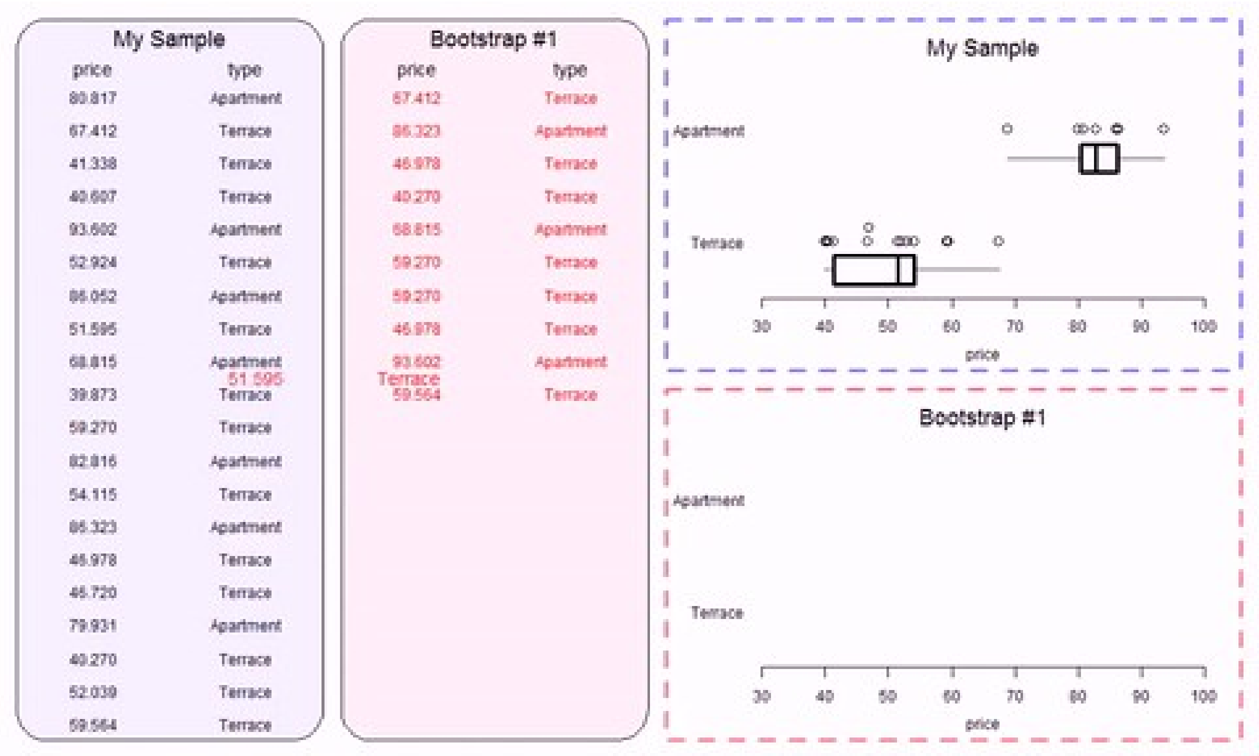
- Idea: combine multiple decision trees
- Randomness is introduced by bagging and feature selection.



Aggregating

Boostrapping

Reminder on Boosting



My Sample

price	type
80.817	Apartment
67.412	Terrace
41.338	Terrace
40.607	Terrace
93.602	Apartment
52.924	Terrace
86.052	Apartment
51.595	Terrace
68.815	Apartment 51.595
39.873	Terrace
59.270	Terrace
82.816	Apartment
54.115	Terrace
86.323	Apartment
46.978	Terrace
46.720	Terrace
79.931	Apartment
40.270	Terrace
52.039	Terrace
59.564	Terrace

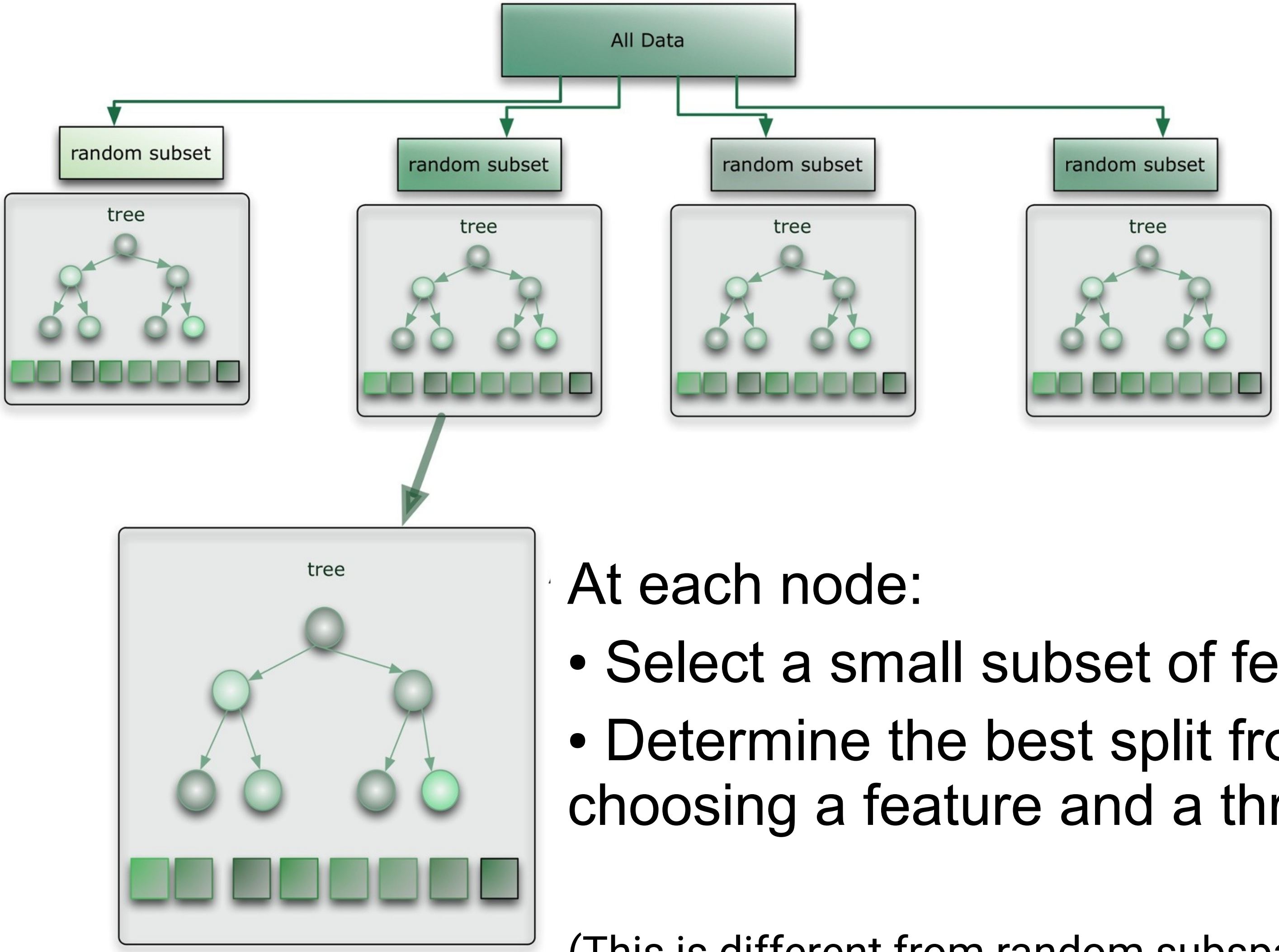
Bootstrap #1

price	type
67.412	Terrace
86.323	Apartment
46.978	Terrace
40.270	Terrace
68.815	Apartment
59.270	Terrace
59.270	Terrace
46.978	Terrace
93.602 Terrace	Apartment
59.564	Terrace

Boosting allows each tree to consider a subset of the initial observations.

Sampling with replacement is equivalent to assigning weights.

Random forests detailed growth



At each node:

- Select a small subset of features at random.
- Determine the best split from this subset, by choosing a feature and a threshold.

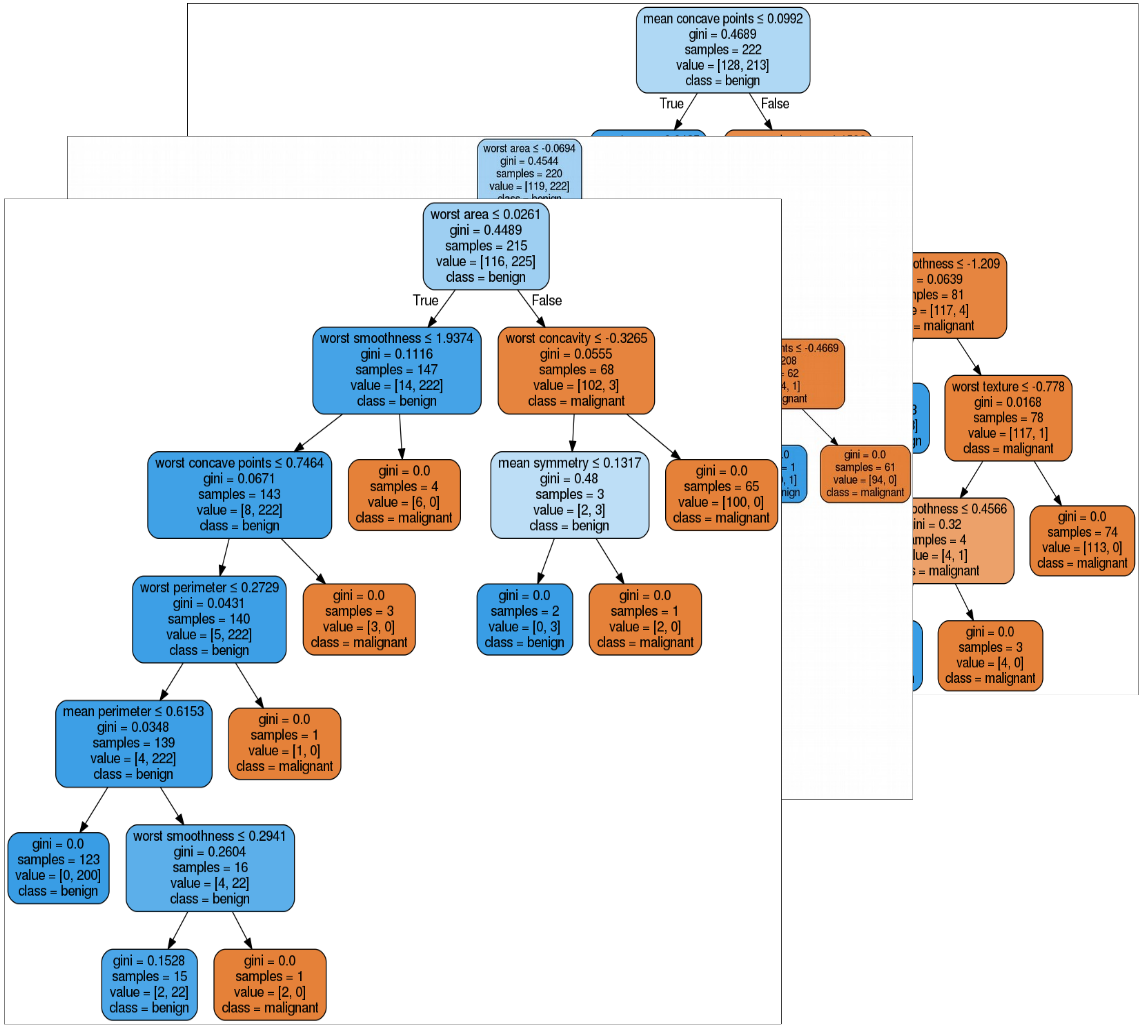
(This is different from random subspacing)

Random forest parameters

- Number of trees
- Number of features selected per node
- Depth of the tree
- Out-of-bag error computation:

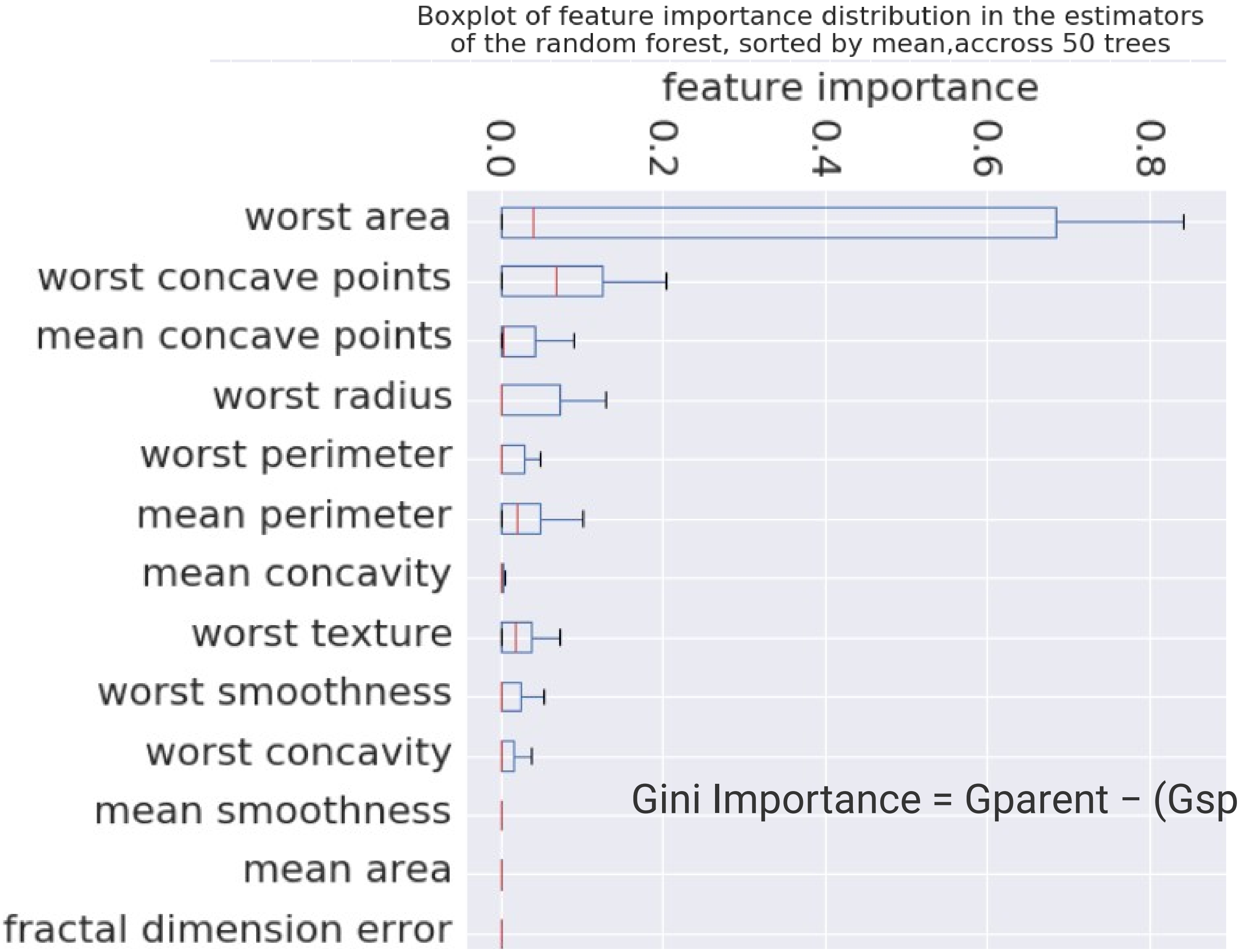
  Allows to control for generalization error while building the forest.

Advantages of the forest over the tree



- Multiple trees = multiple combination of samples and features explored
- No over-fitting anymore …
- … as long as the correlation between trees is not too high.
- Lowering the number of features taken at random lowers this correlation, but also the strength of each tree.

Random forests and feature selection



Gini Importance = Gparent − (Gsplit1+Gsplit2)

Summary on random forests

- Quite easy to understand
- Solves the overfitting problem of single decision trees
- Allows for feature selection
- Quick to use and train for acceptable results
- Exploring high dimension data requires more trees to explore the subspaces, thus quickly increasing the computational time...

Resources

- https://en.wikipedia.org/wiki/Random_forest
- https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
- Tin Kam Ho (1998). The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 832–844.
- http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html