Bagging – Random Forest

Overfitting

- You can perfectly fit to any training data
- Zero bias, high variance

Two approaches used to solve this for Decision Trees:

1. Stop growing the tree when further splitting the data does not yield an improvement
2. Grow a full tree, then prune the tree, by eliminating nodes.

Yet another approach to reduce variance

Use an ensemble of classifiers.

Two ensemble methods

Bagging: This is known to reduce variance.

Boosting: A weak method is progressively made in to a stronger one. This can reduce bias.

Bagging

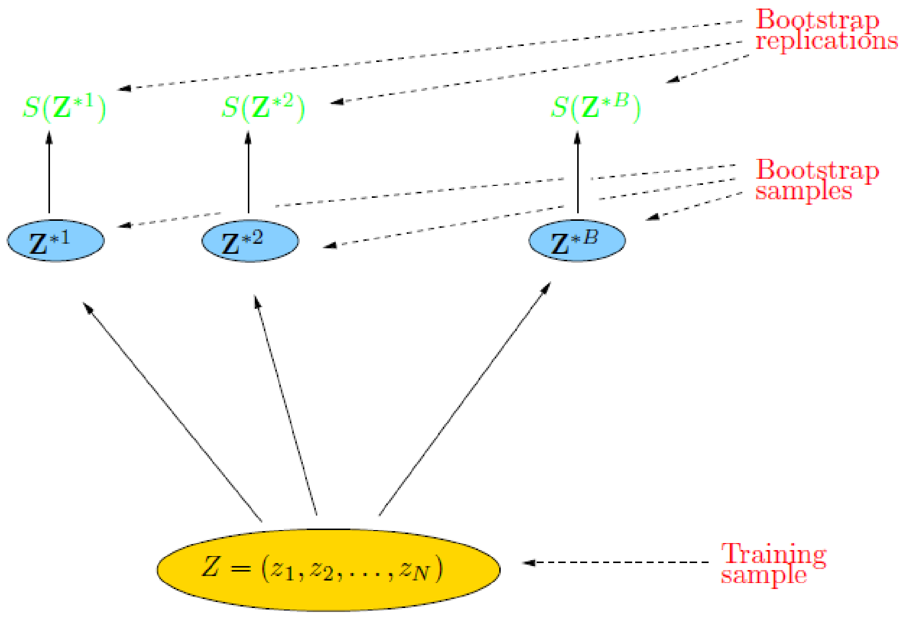
- Bagging or *bootstrap aggregation* a technique for reducing the variance of an estimated prediction function.

- For classification, a *committee* of trees each cast a vote for the predicted class.

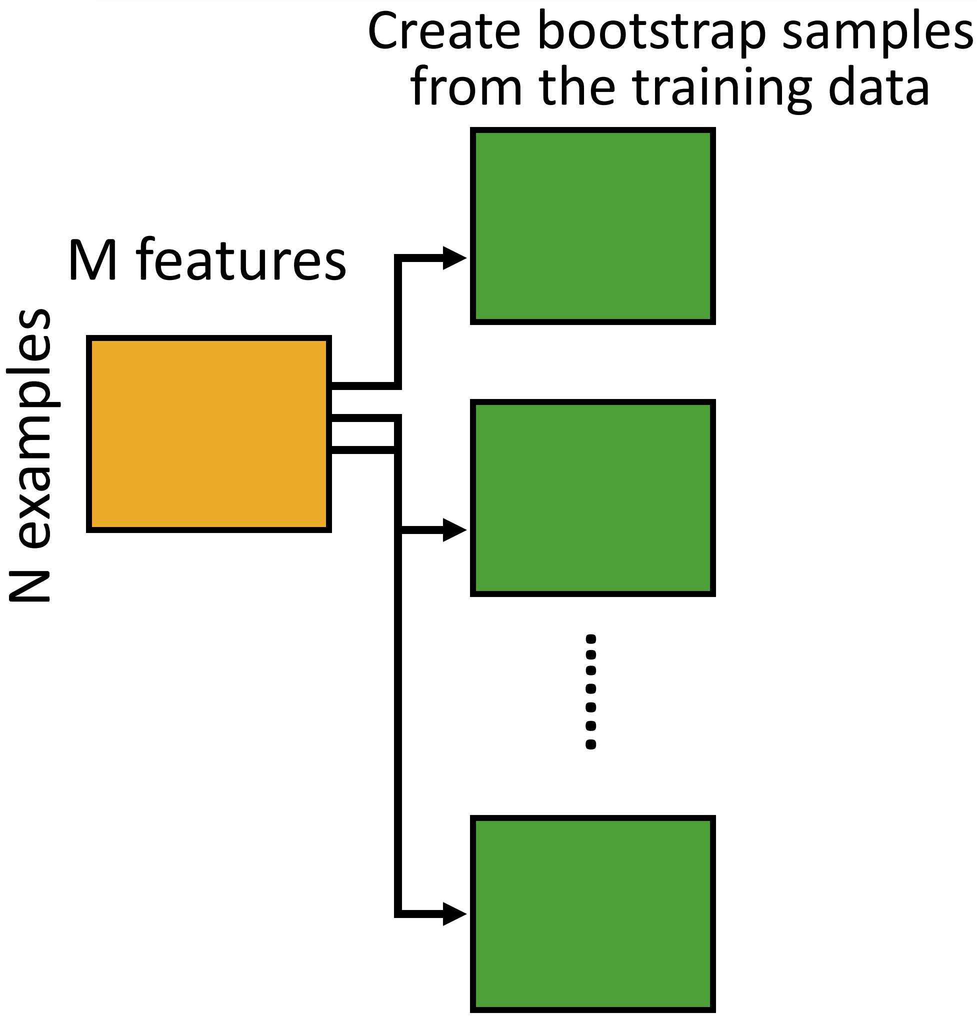
Bootstrap

The basic idea:

randomly draw datasets *with replacement* from the training data, each sample *the same size as the original training set*

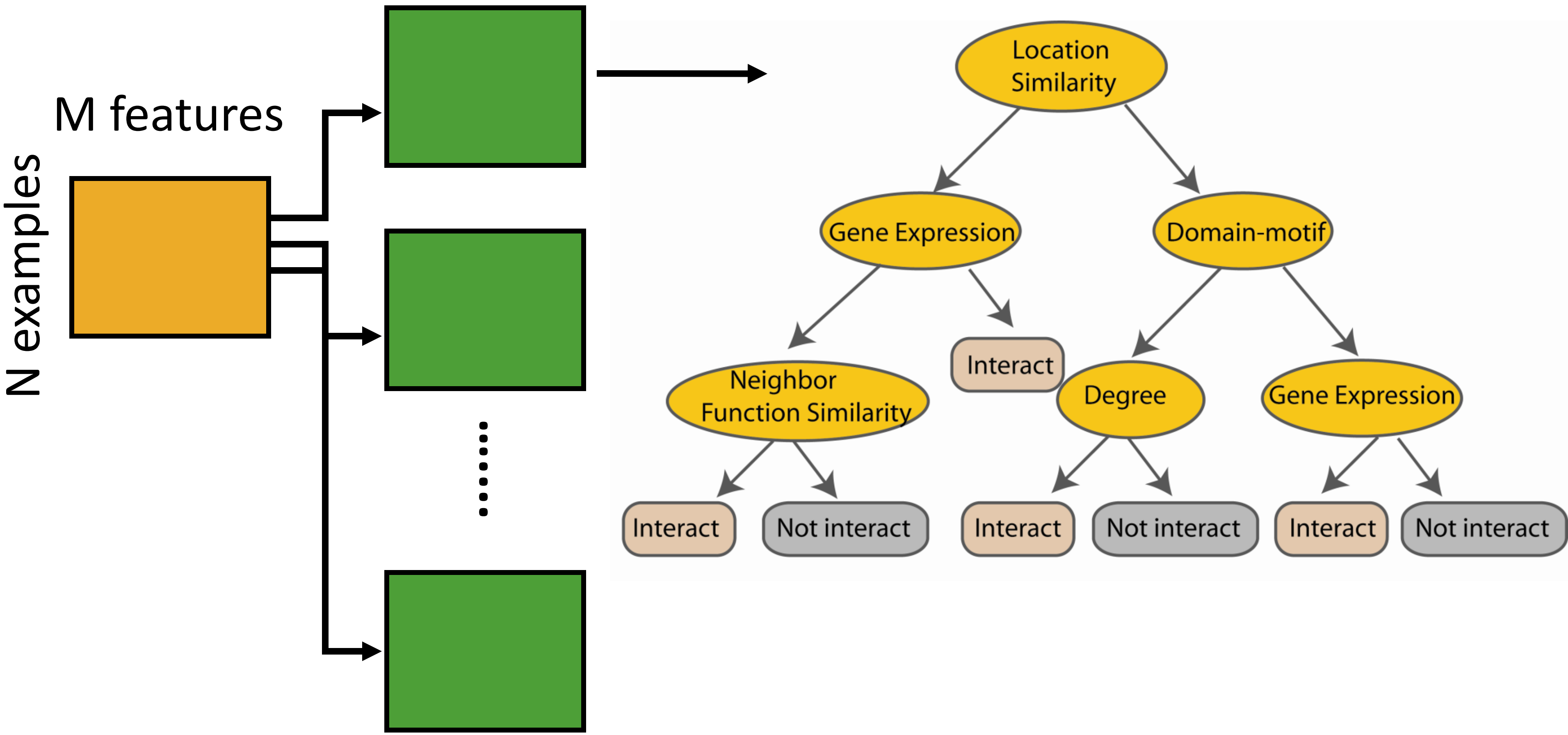


Bagging

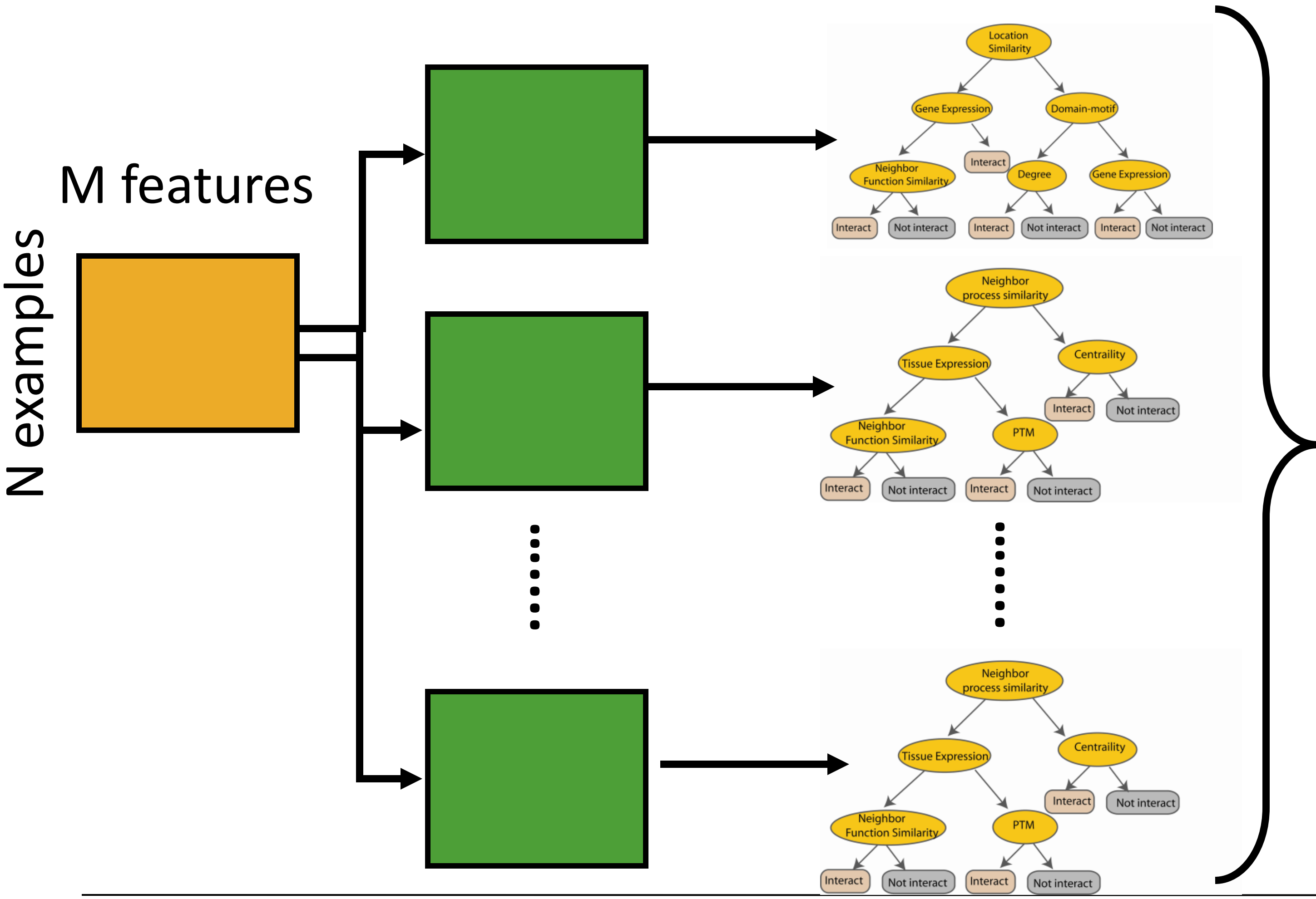

Create bootstrap samples
from the training data
M features
N examples

Random Forest Classifier

Construct a decision tree



Random Forest Classifier



Take the majority vote

Bagging

Z = {(x_1, y_1), (x_2, y_2), . . . , (x_N, y_N)}

Z^{*b} where = 1,.., B..

The prediction at input x when bootstrap sample *b* is used for training

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

http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf (Chapter 8.7)

Bagging : an simulated example

Generated a sample of size $N = 30$, with two classes and $p = 5$ features, each having a standard Gaussian distribution with pairwise Correlation 0.95.

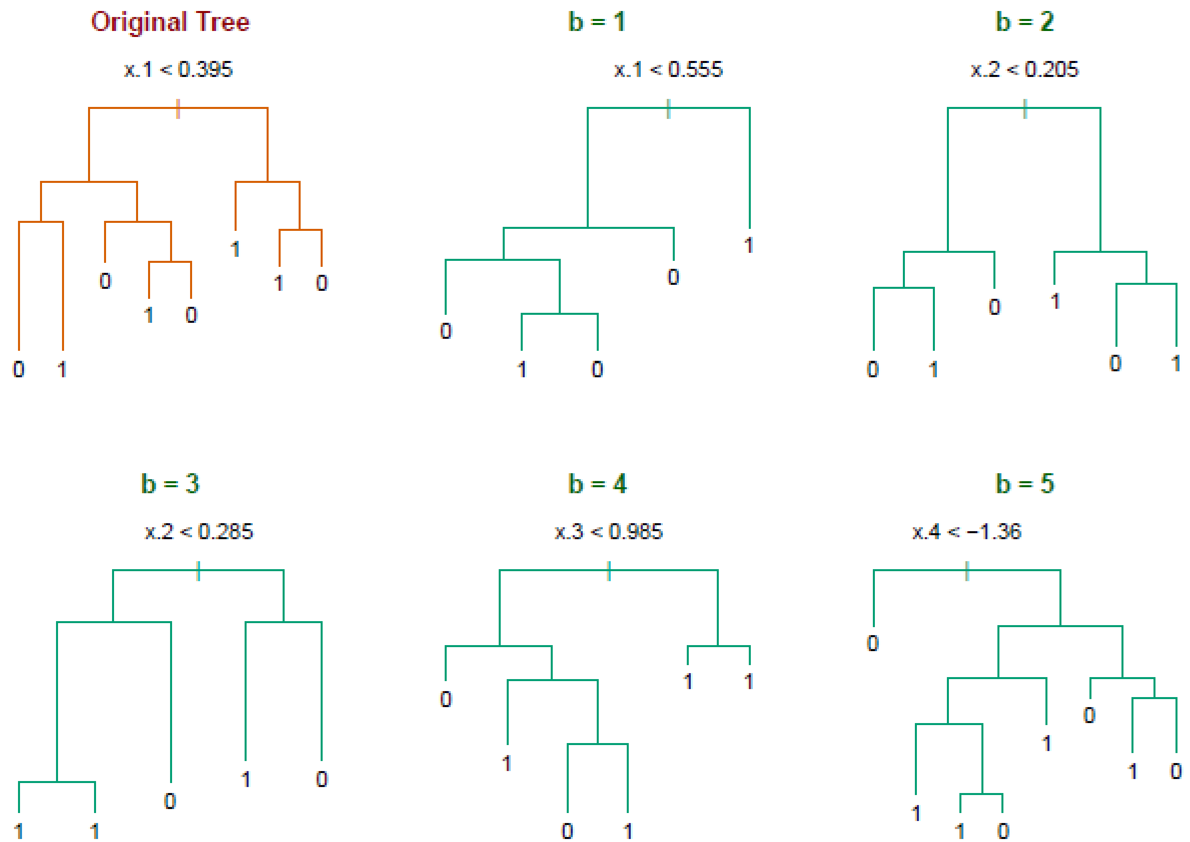
The response Y was generated according to

$\Pr(Y = 1/x1 \leq 0.5) = 0.2$,

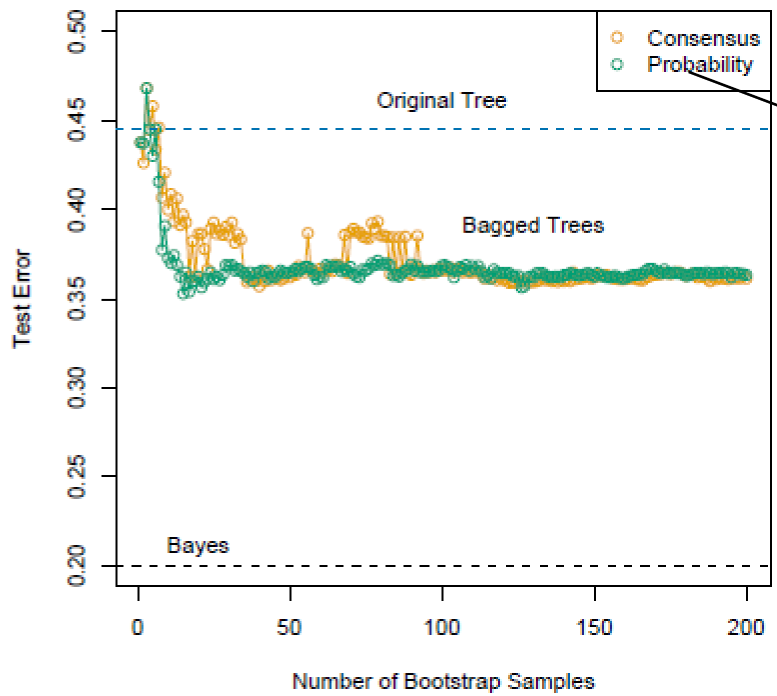
$\Pr(Y = 0/x1 > 0.5) = 0.8$.

Bagging

Notice the bootstrap trees are different than the original tree



Bagging



Treat the voting Proportions as probabilities

FIGURE 8.10. *Error curves for the bagging example of Figure 8.9. Shown is the test error of the original tree and bagged trees as a function of the number of bootstrap samples. The orange points correspond to the consensus vote, while the green points average the probabilities.*

bagging helps under squared-error loss, in short because averaging reduces

Hastie

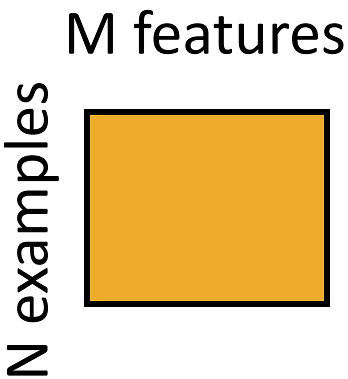
http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf Example 8.7.1

Random forest classifier

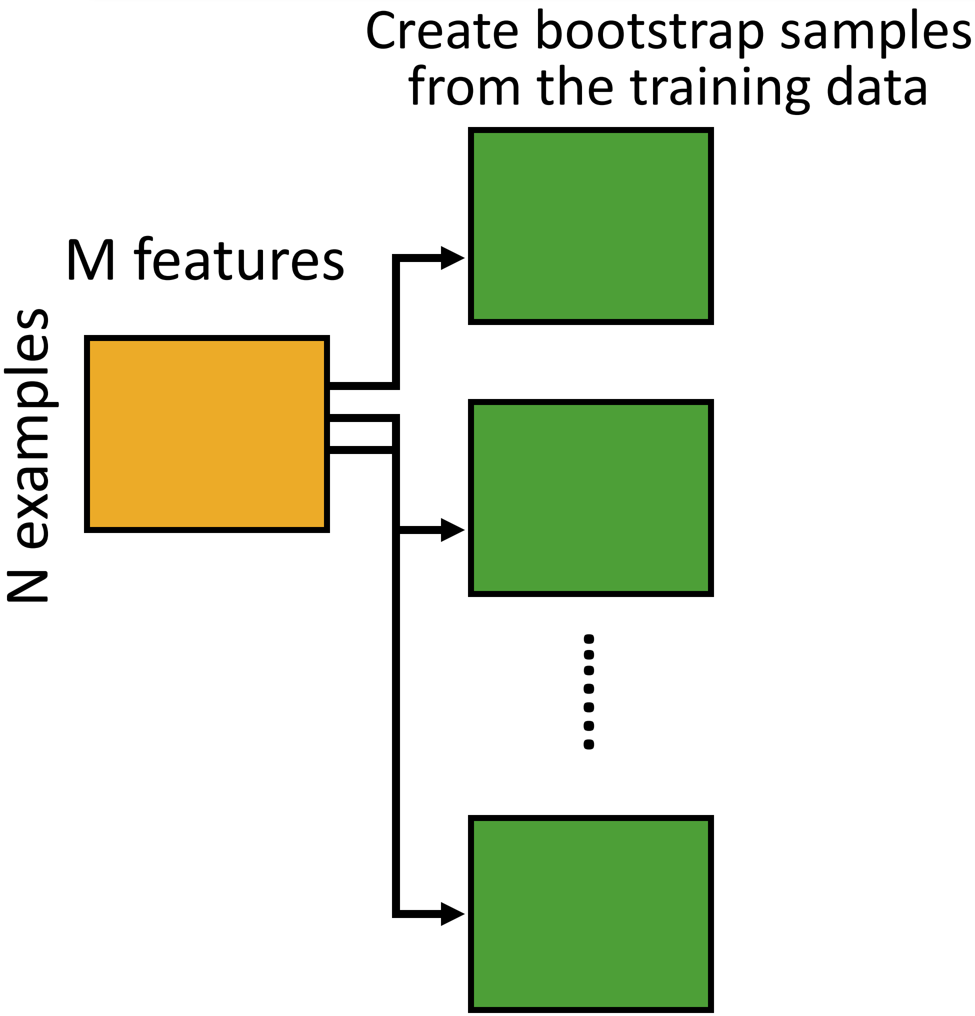
Random forest classifier, an extension to bagging which uses *de-correlated* trees.

Random Forest Classifier

Training Data

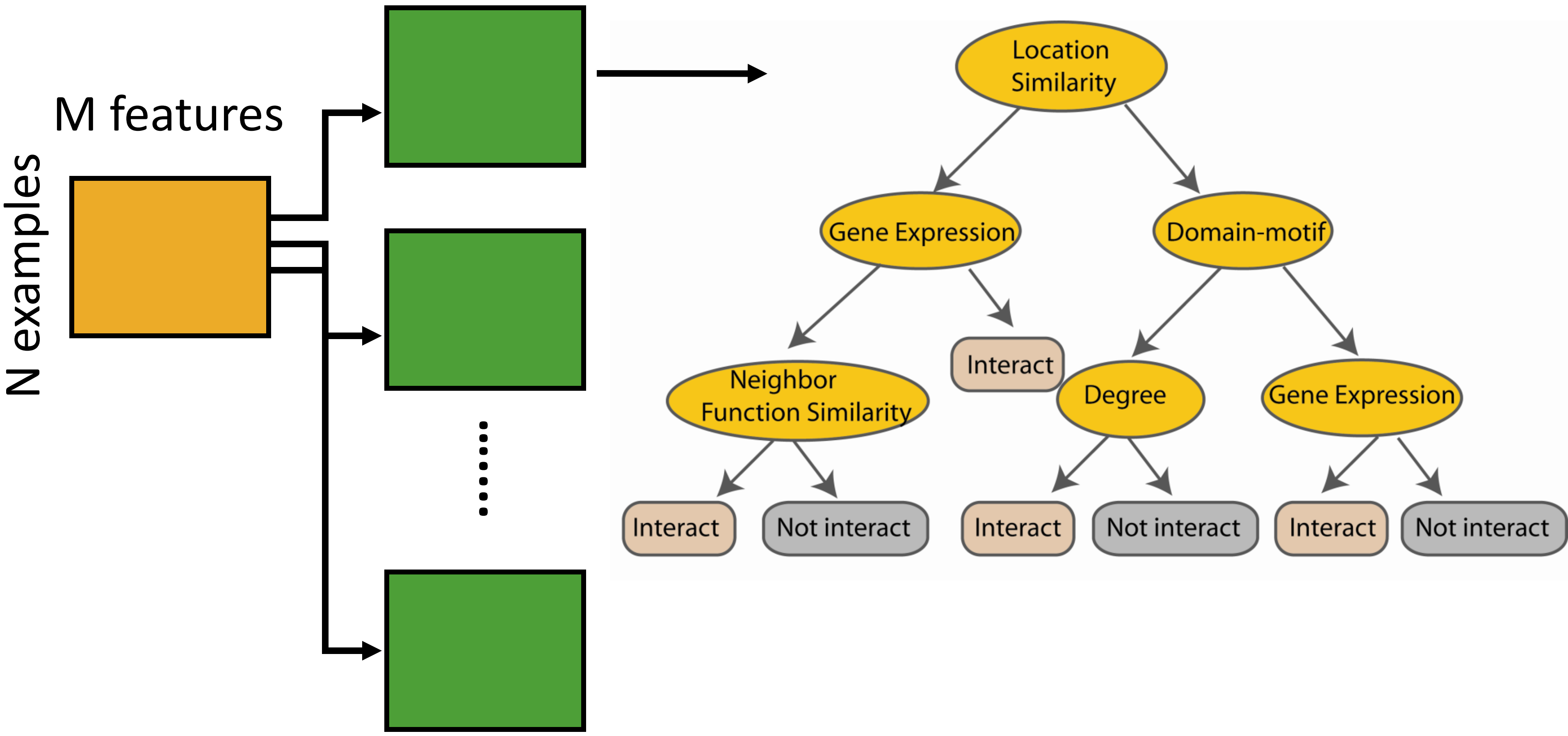


Random Forest Classifier



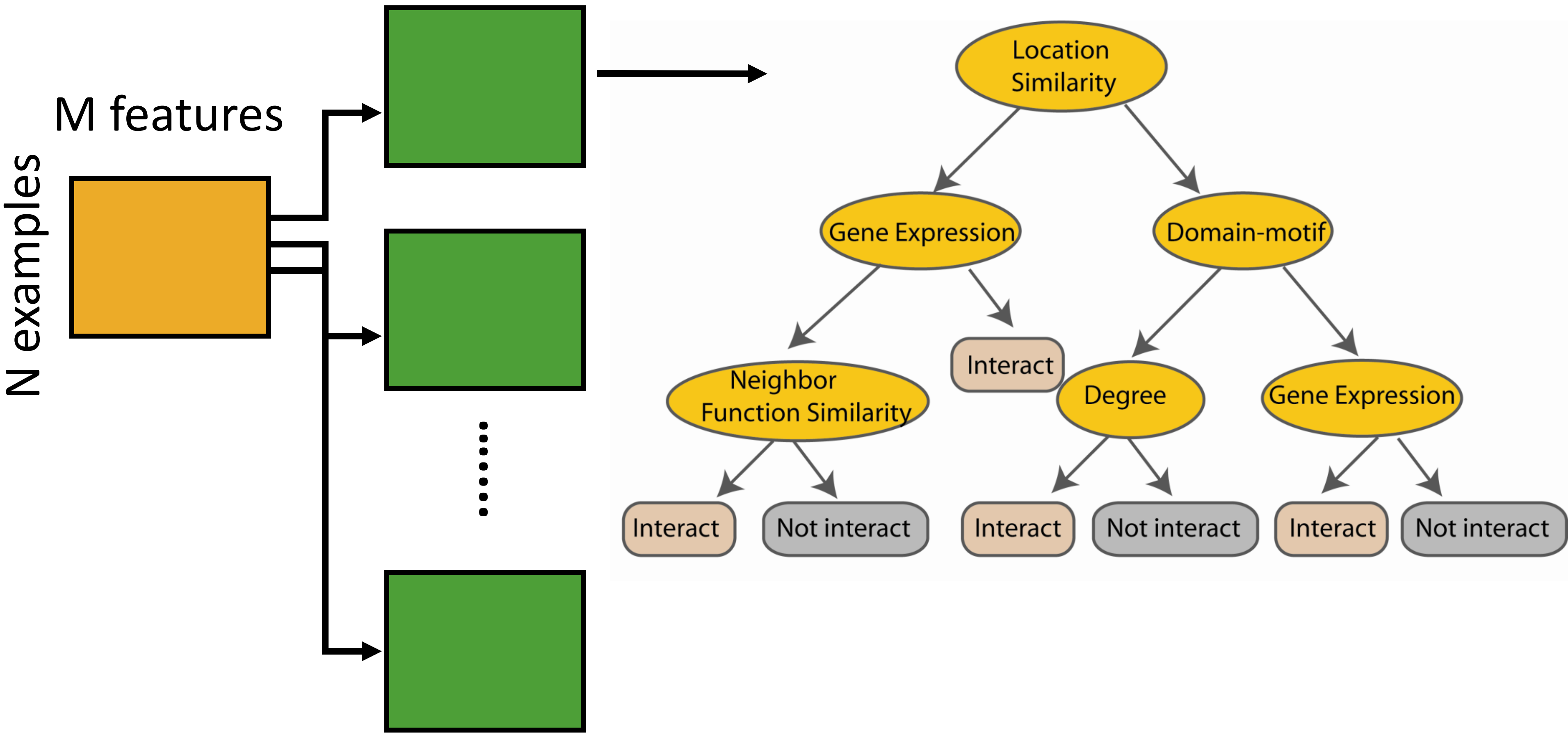
Random Forest Classifier

Construct a decision tree



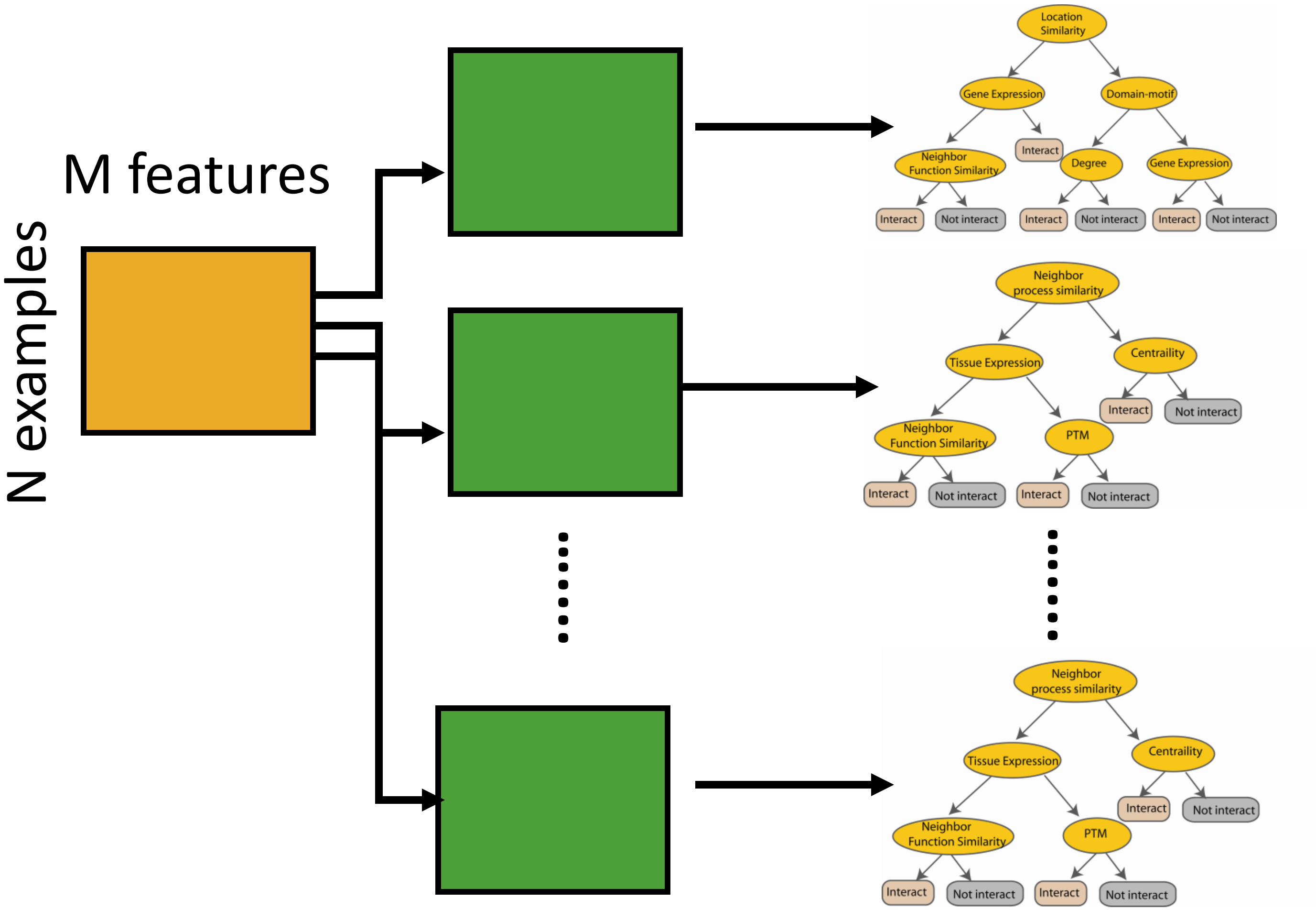
Random Forest Classifier

At each node in choosing the split feature choose only among *m*<*M* features

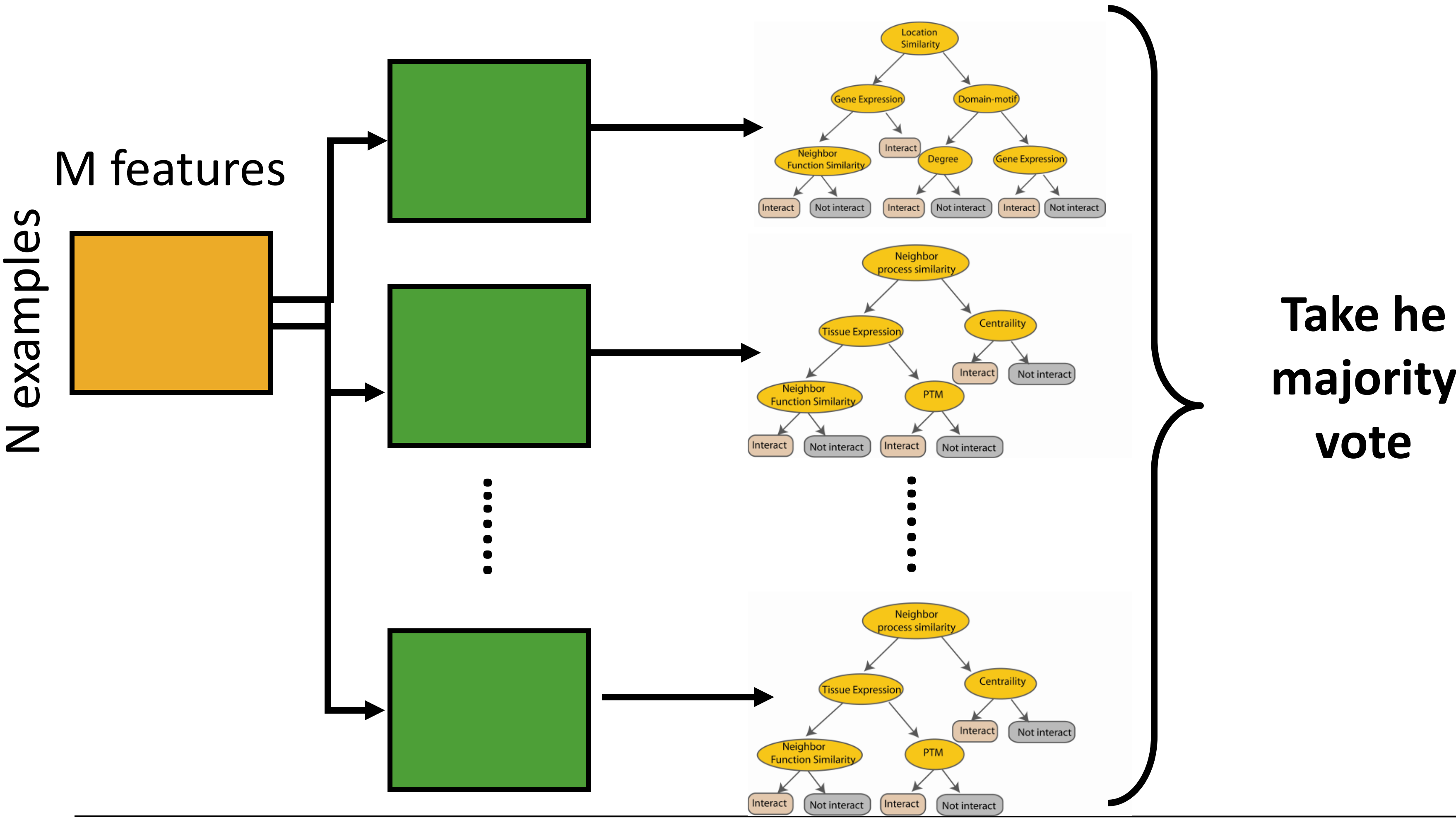


Random Forest Classifier

Create decision tree from each bootstrap sample



Random Forest Classifier



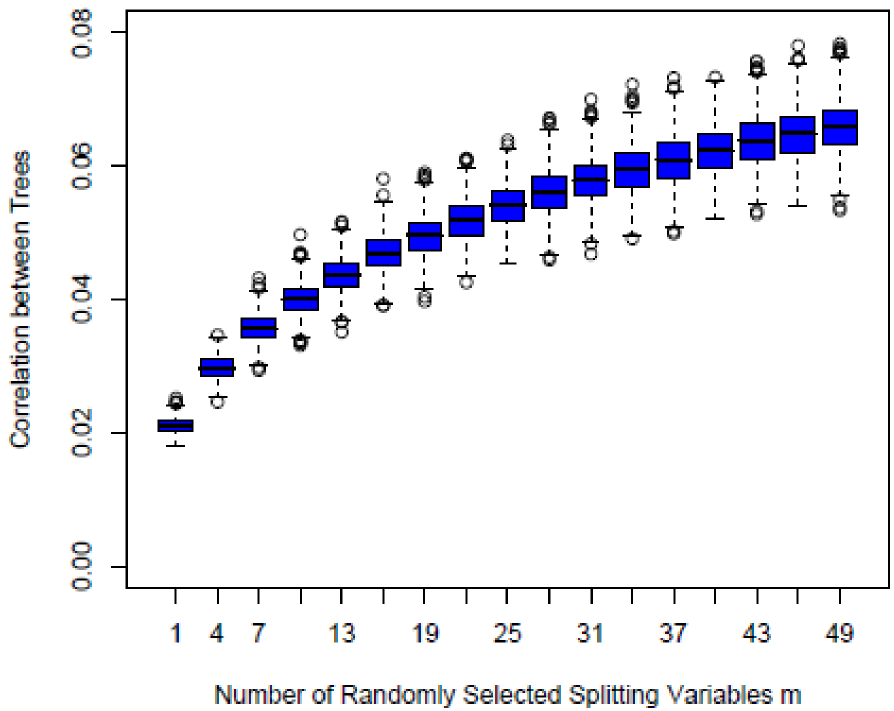


FIGURE 15.9. *Correlations between pairs of trees drawn by a random-forest regression algorithm, as a function of m. The boxplots represent the correlations at 600 randomly chosen prediction points x.*

Random forest

Available package:

http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

To read more:

http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf