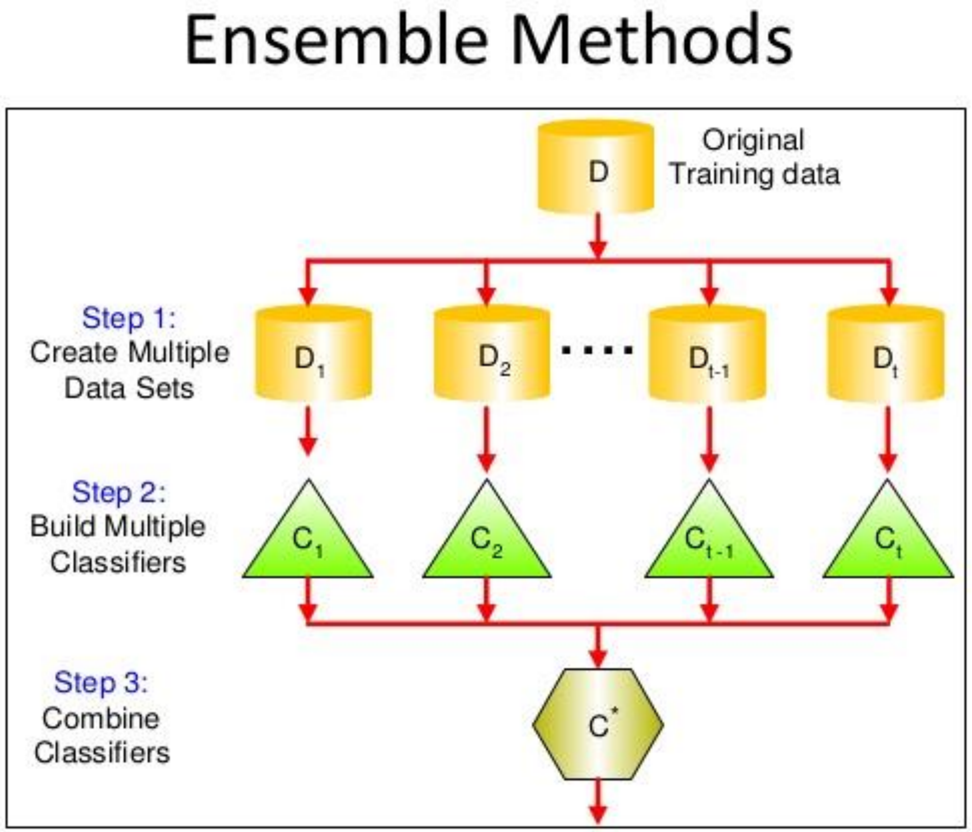
Ensemble
Dominik Lewy

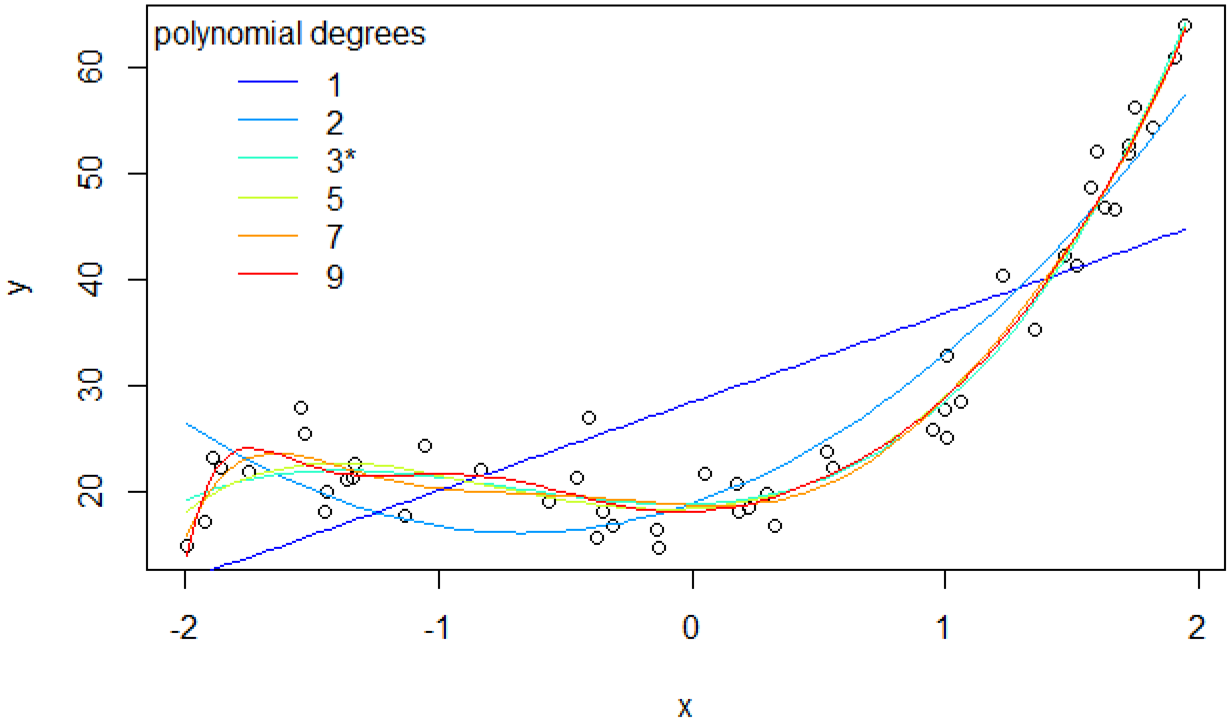
Ensemble - definition

An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.



Introduction –
Bias & Variance

Bias vs. Variance



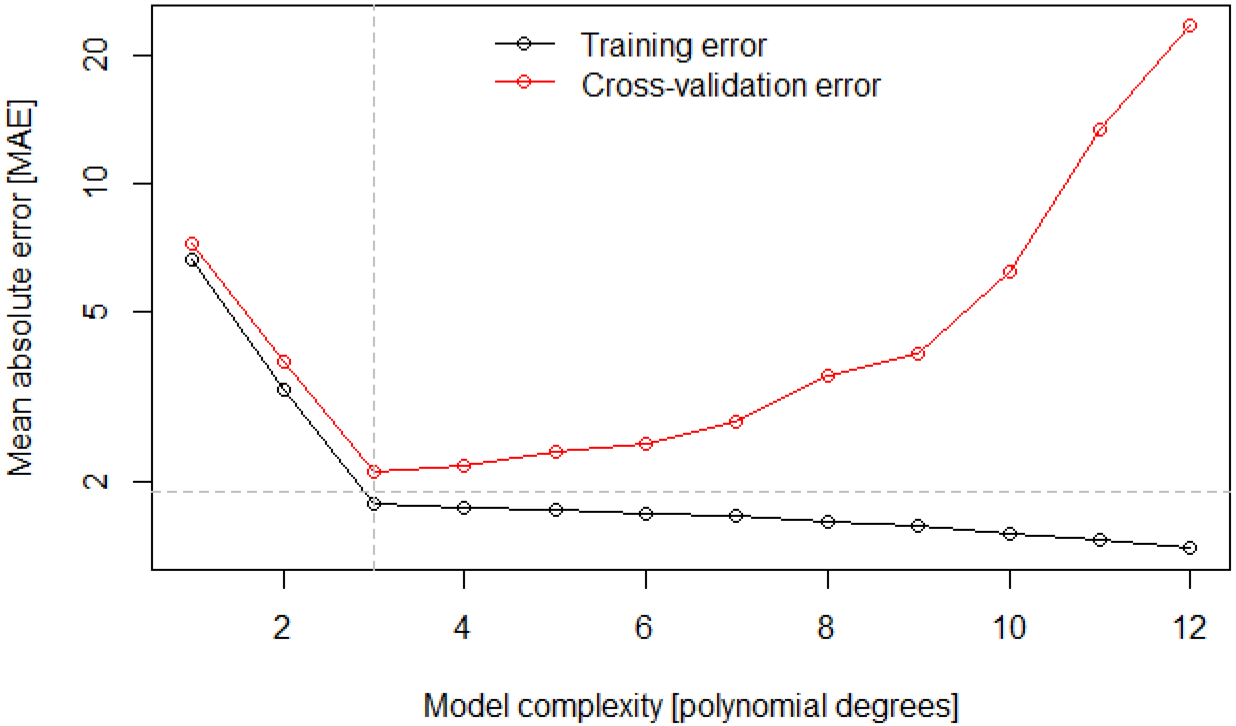
http://menugget.blogspot.com/2014/05/evaluating-model-performance-practical.html

Bias vs. Variance

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y-\hat{f}(x_0))^2|X=x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance.}
\end{aligned}$$

http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

Bias vs. Variance



http://menugget.blogspot.com/2014/05/evaluating-model-performance-practical.html

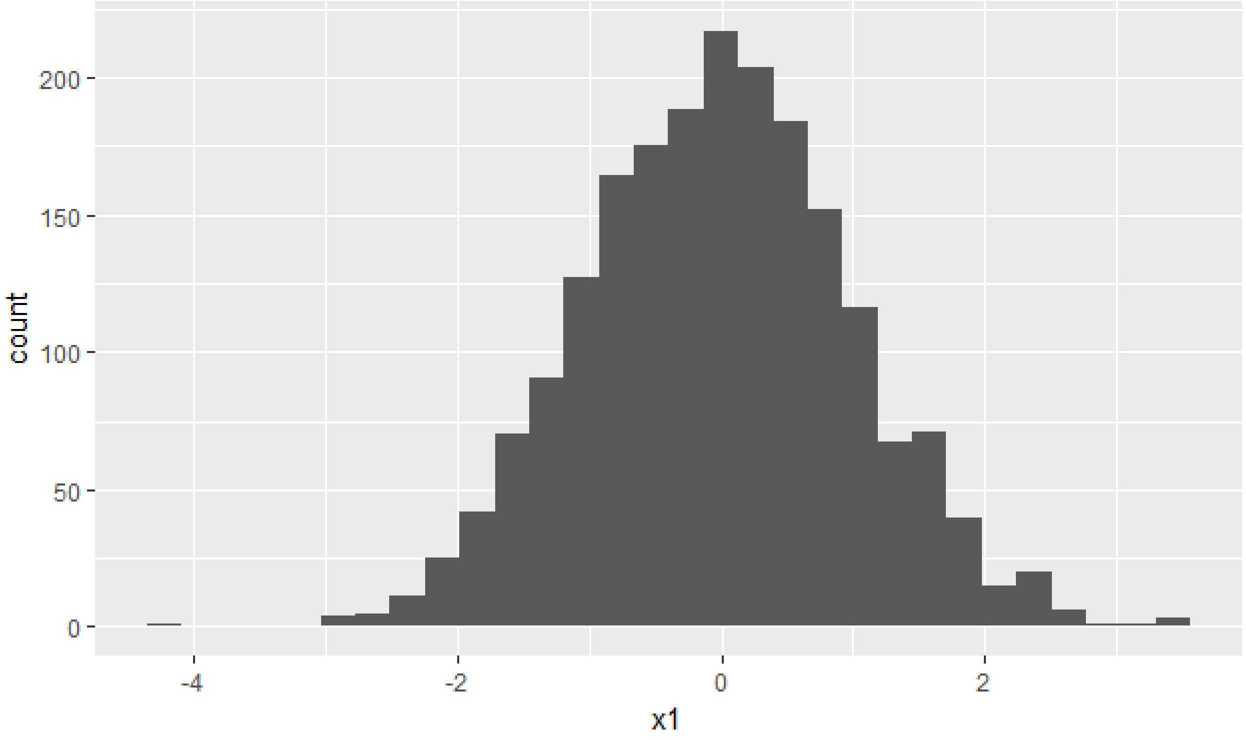
Ensemble –
Simple Voting, Bagging, Boosting

Data set used in this presentation

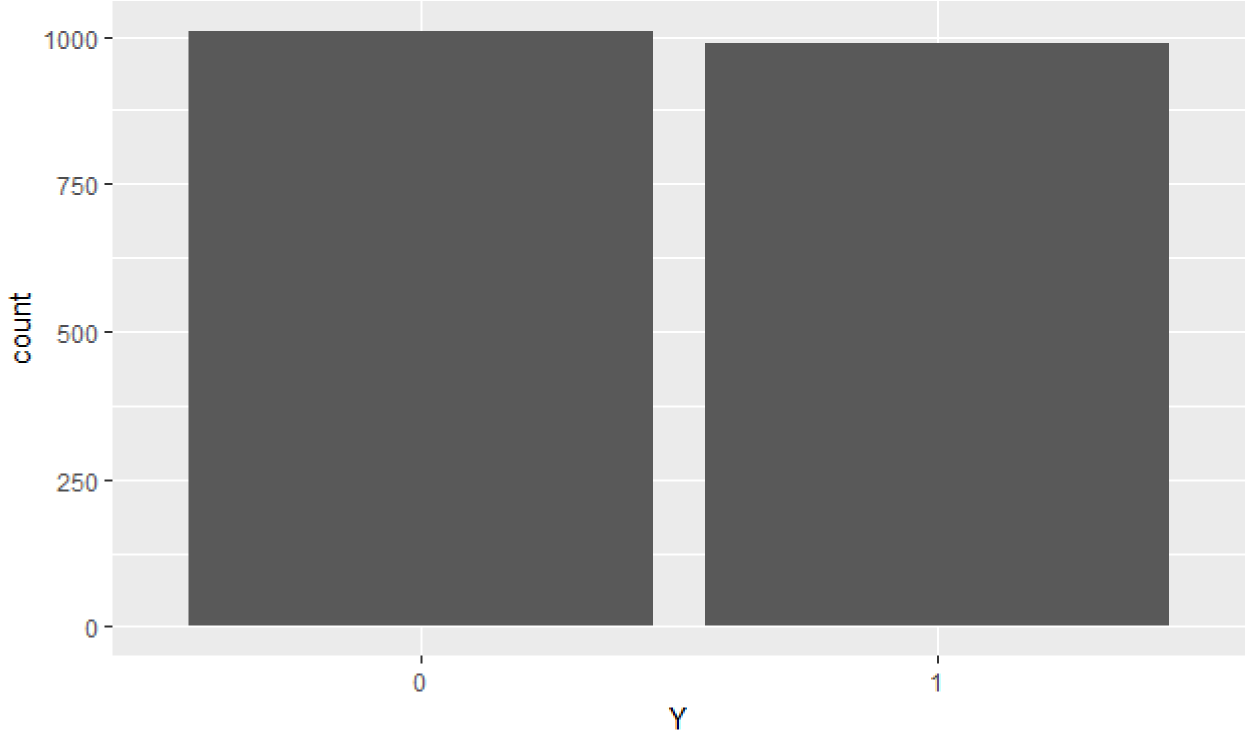
- X1 … X10 – standard independent Gaussian

$$Y = \begin{cases} 1 & \text{if } \sum_{j=1}^{10} X_j^2 > \chi_{10}^2(0.5), \\ -1 & \text{otherwise.} \end{cases}$$

X1 … X10

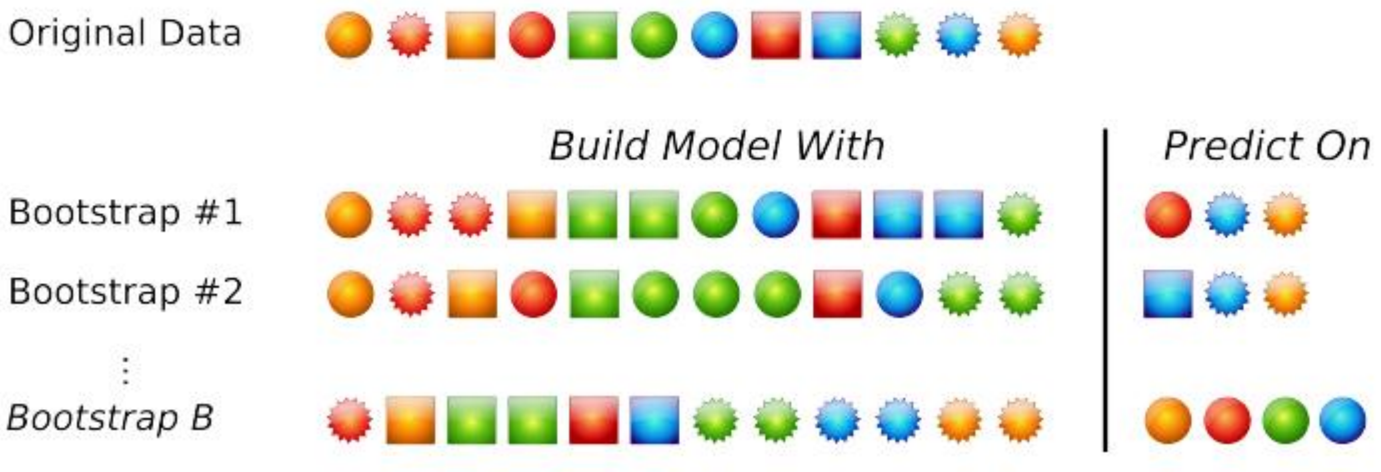


Y



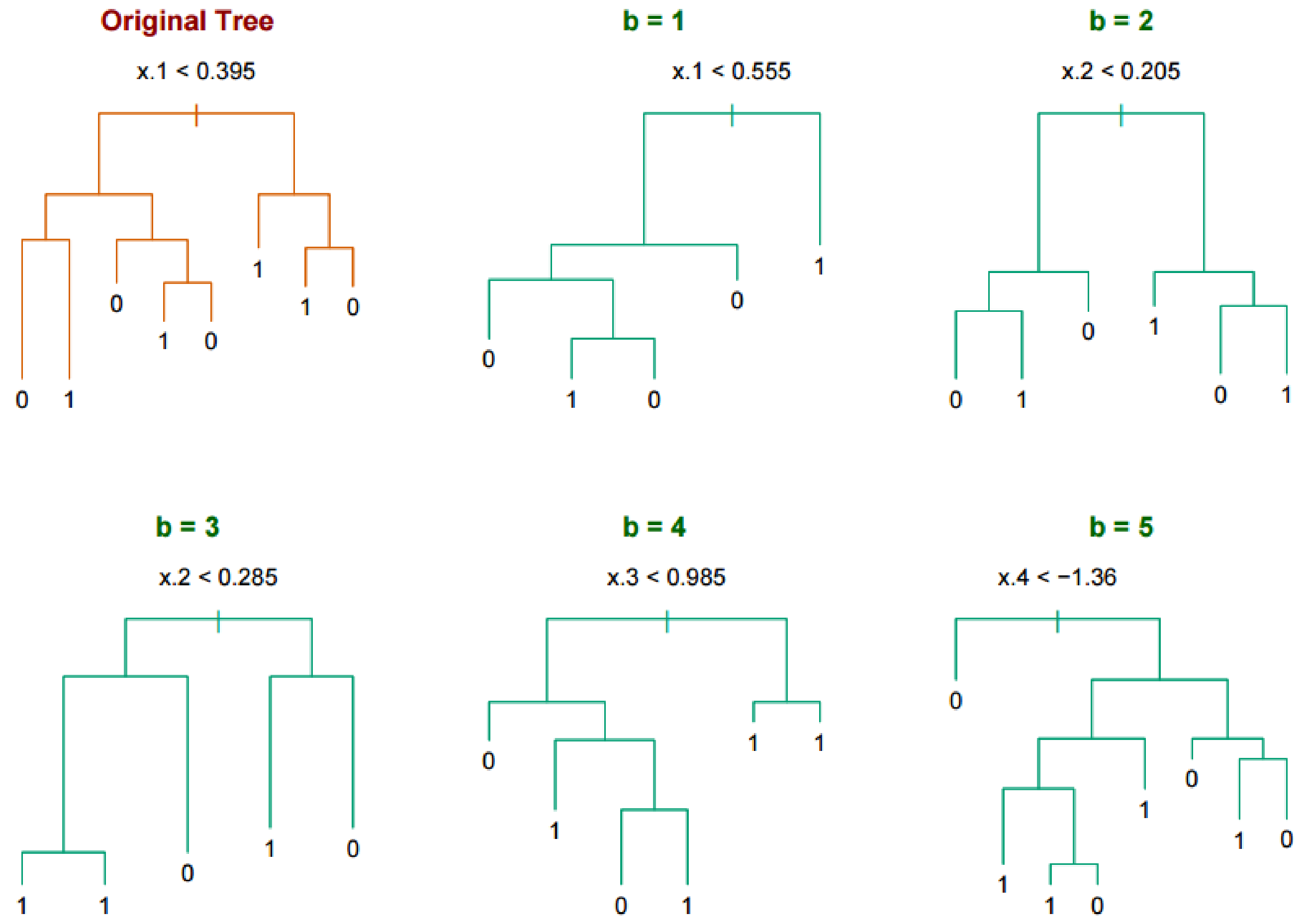
Bootstrap - theory

- Bootstrapping takes a random sample with replacement.
- The random sample is the same size as the original data set.
- Samples may be selected more than once and each sample has a 63.2% chance of showing up at least once.
- Some samples won’t be selected and these samples will be used to predict performance.

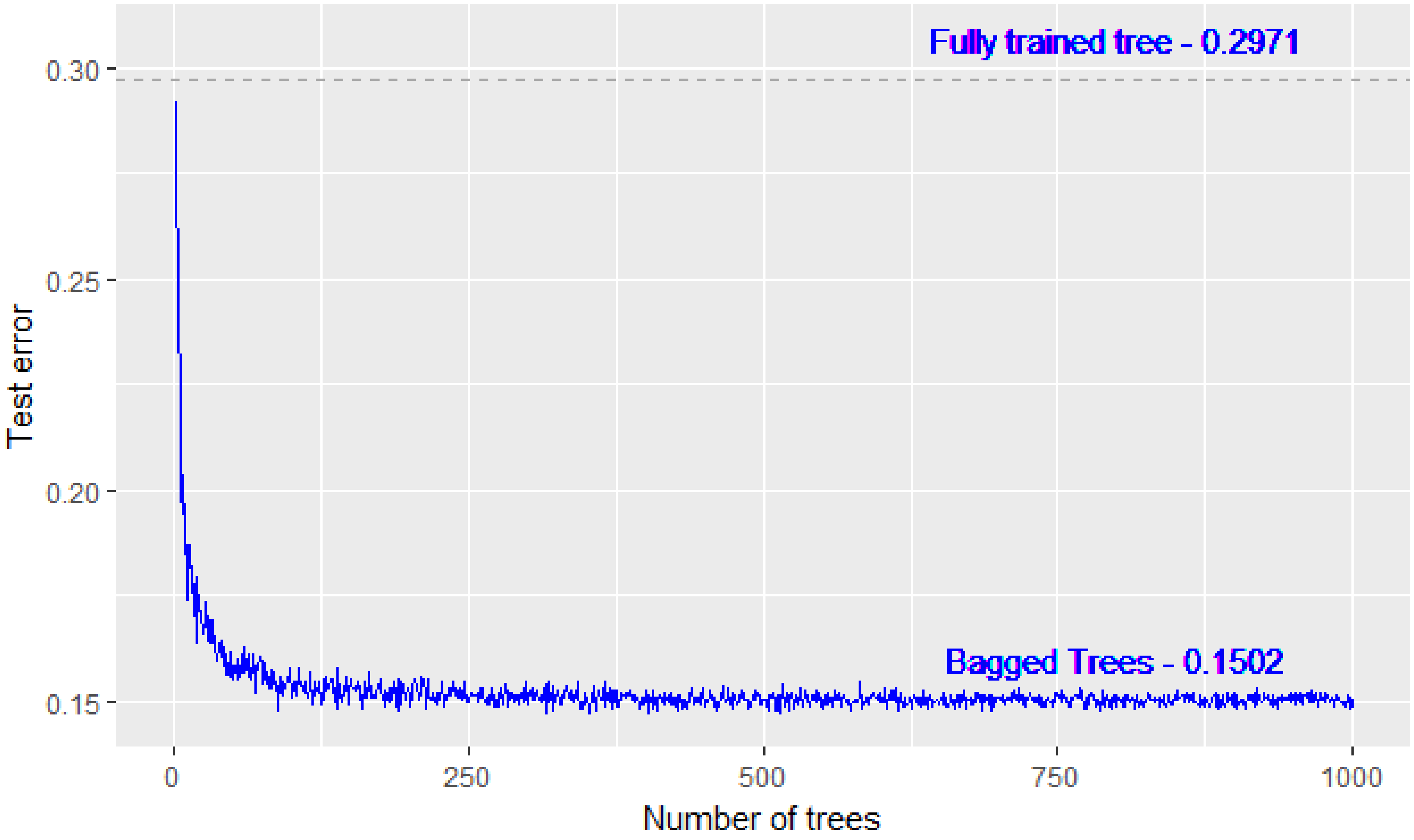


Bagging - theory

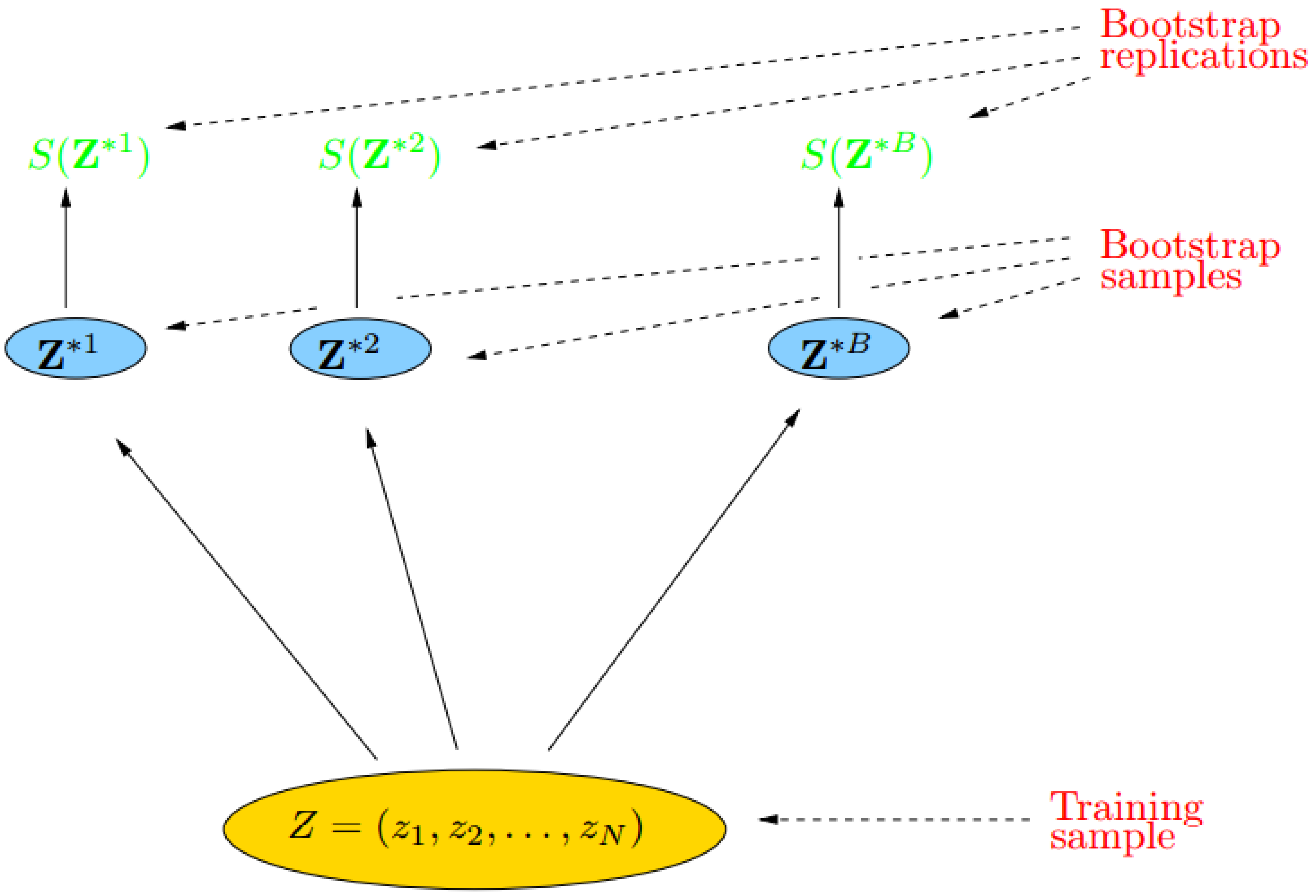
$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x).$$



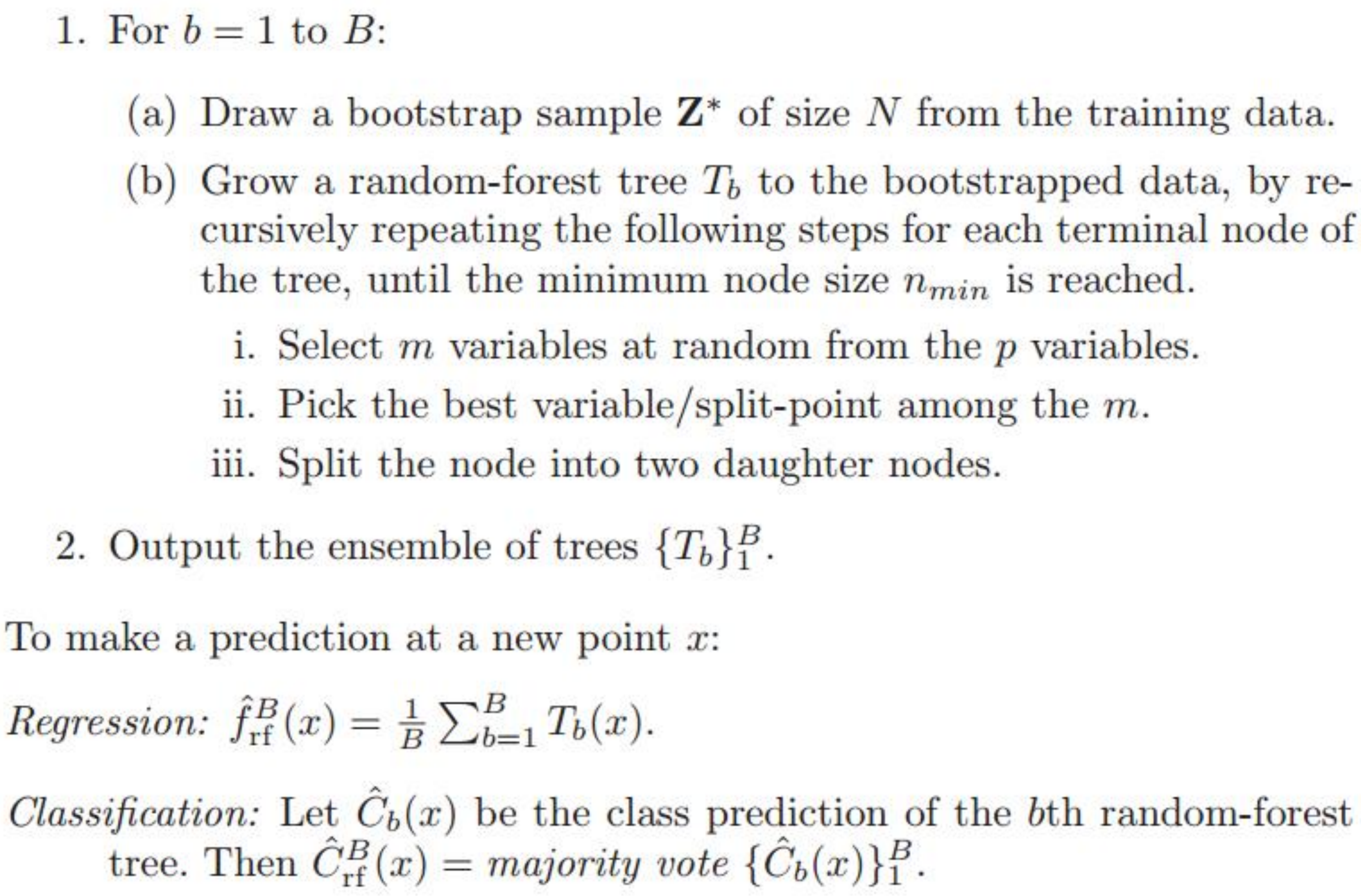
Bagging - example



Random Forest - theory



Random Forest - theory



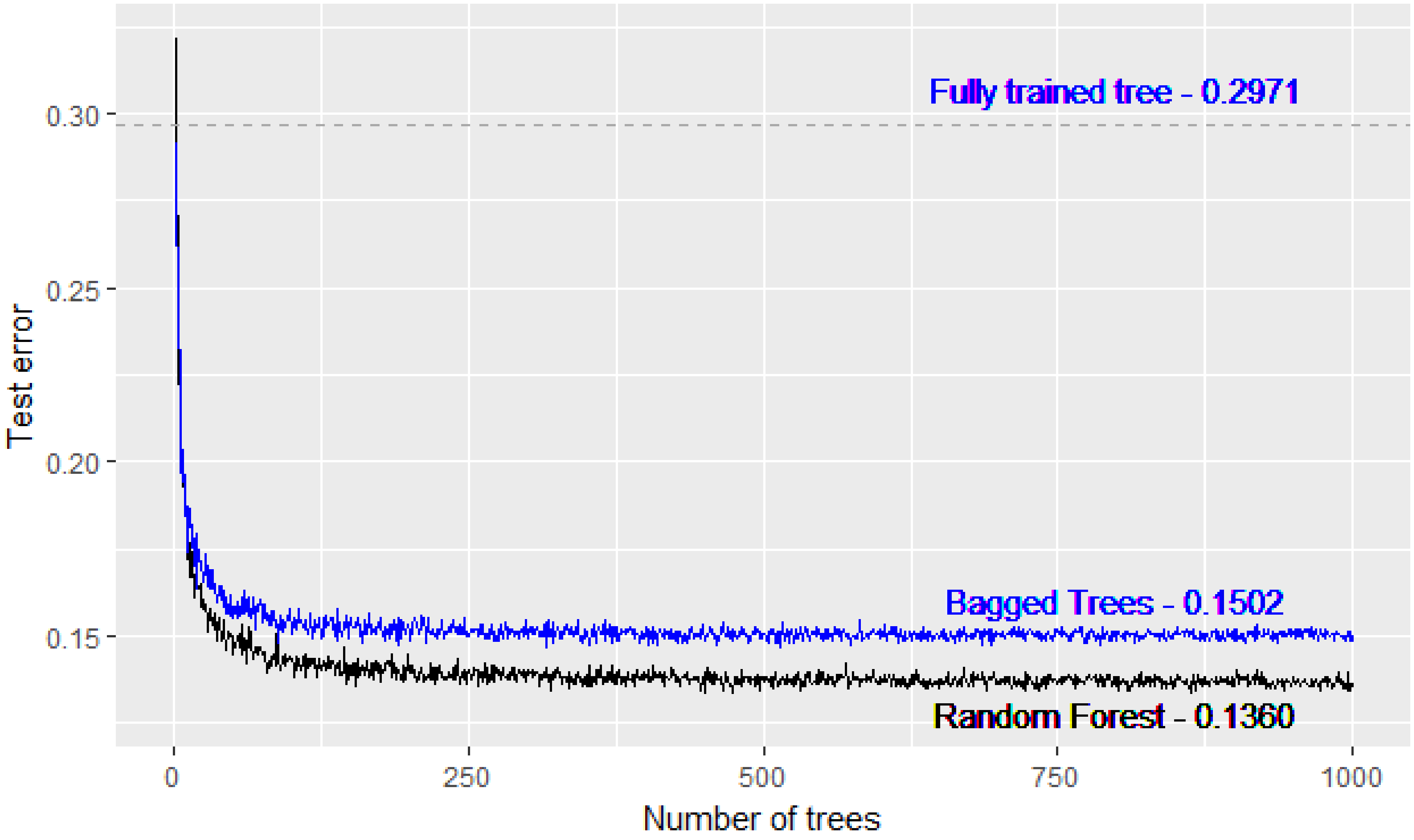
```
1. For $b = 1$ to $B$:
    (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size $N$ from the training data.
    (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached.
        i. Select $m$ variables at random from the $p$ variables.
        ii. Pick the best variable/split-point among the $m$.
        iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
```

To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

Random Forest - example



Boosting - theory

1. Initialize the observation weights $w_i = 1/N,\ i = 1, 2, \ldots, N$.

2. For $m = 1$ to M:

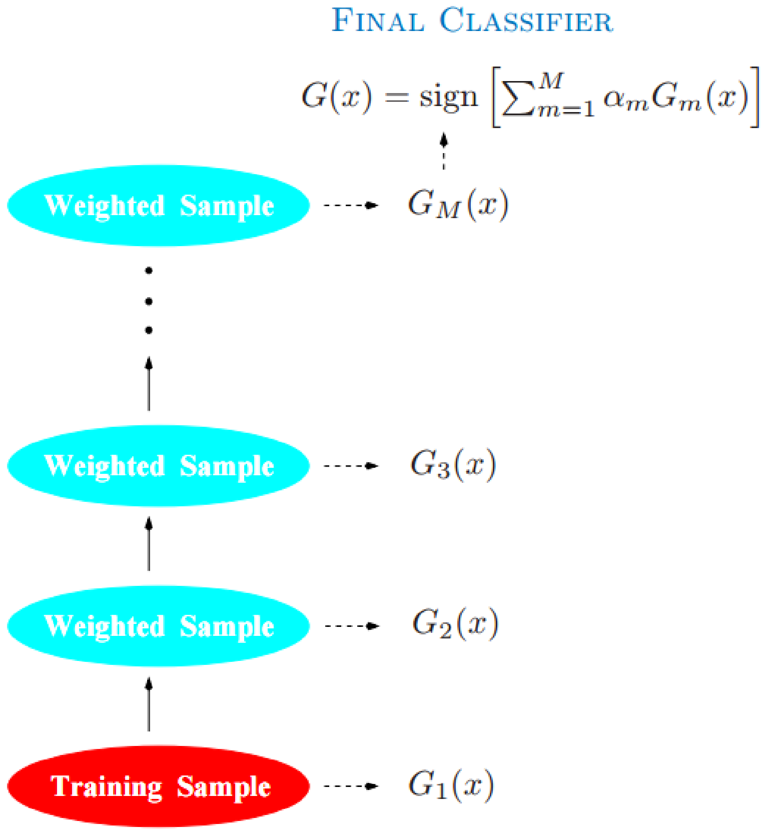
   (a) Fit a classifier $G_m(x)$ to the training data using weights w_i.

   (b) Compute
   $$\text{err}_m = \frac{\sum_{i=1}^N w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i}.$$

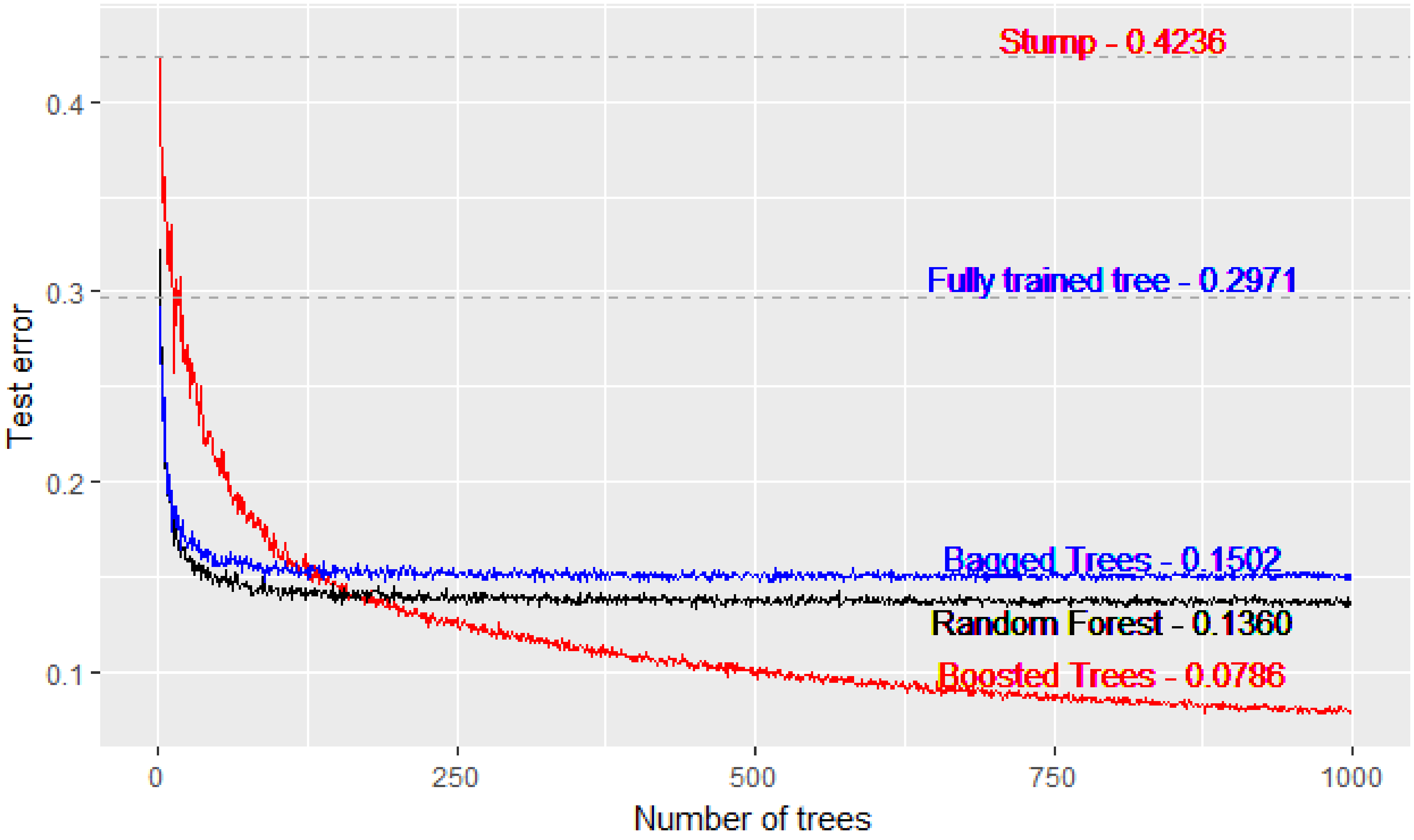
   (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.

   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))],\ i = 1, 2, \ldots, N$.

3. Output $G(x) = \text{sign}\left[\sum_{m=1}^M \alpha_m G_m(x)\right]$.



Boosting - example



Random Forest vs Boosting - comparison

- The Bias of bagged trees is the same as that of an individual tree, and the only hope of improvement is through variance reduction. This is in contrast to boosting, where the trees are grown in an adaptive way to remove bias, and hence are not i.d.

Random Forest vs Bagging - comparison

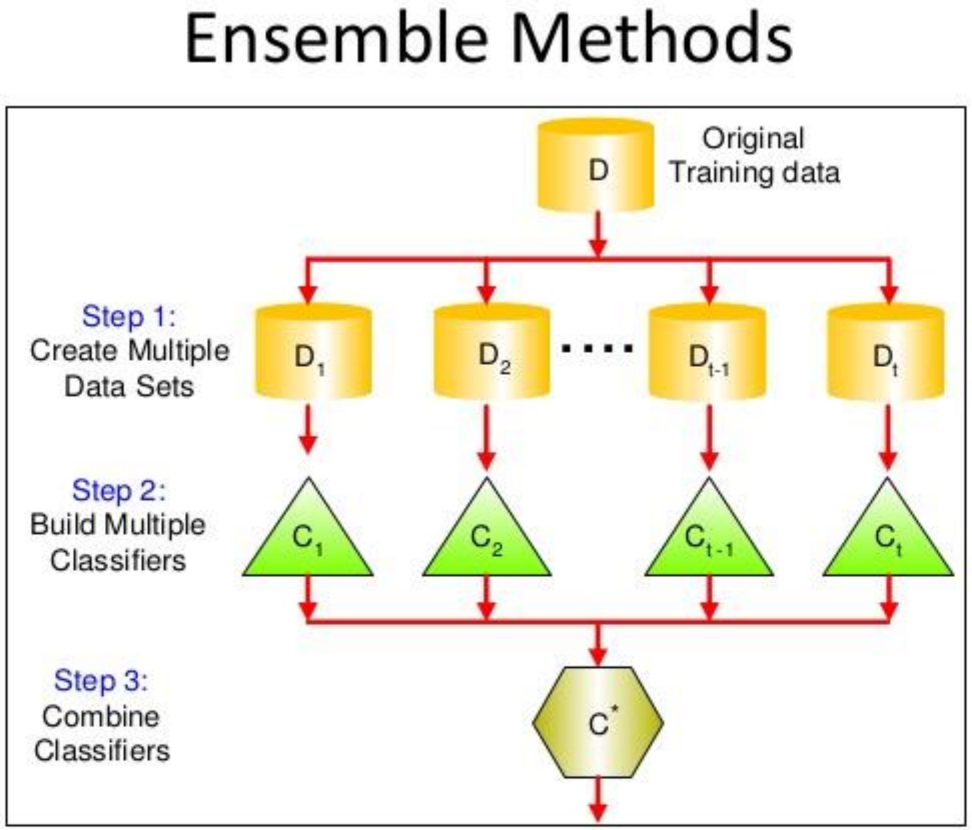
- Bagging uses only one parameter, which is the number of trees. Trees are fully grown using all the variables to perform splits.
- Random Forest uses two parameters:
  - First, same as in bagging the number of trees
  - Second, is the number of features to search over to find the best feature to perform split
- Trees in general choose which variable to split on using a greedy algorithm that minimizes error. As such, even with Bagging, the decision trees can have a lot of structural similarities and in turn have high correlation in their predictions. Random Forest produces estimates with lower variance.

Random Forest, Bagging, Boosting – common features

- This algorithms will not overfit to training data.

Ensemble – other methods

- Simple average/majority voting
- Weighted average
- Stacking – applying meta-level model to individual models



The end.