Foundations of Data Science & Analytics: Overfitting

Ezgi Siir Kibris

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

Classification Errors

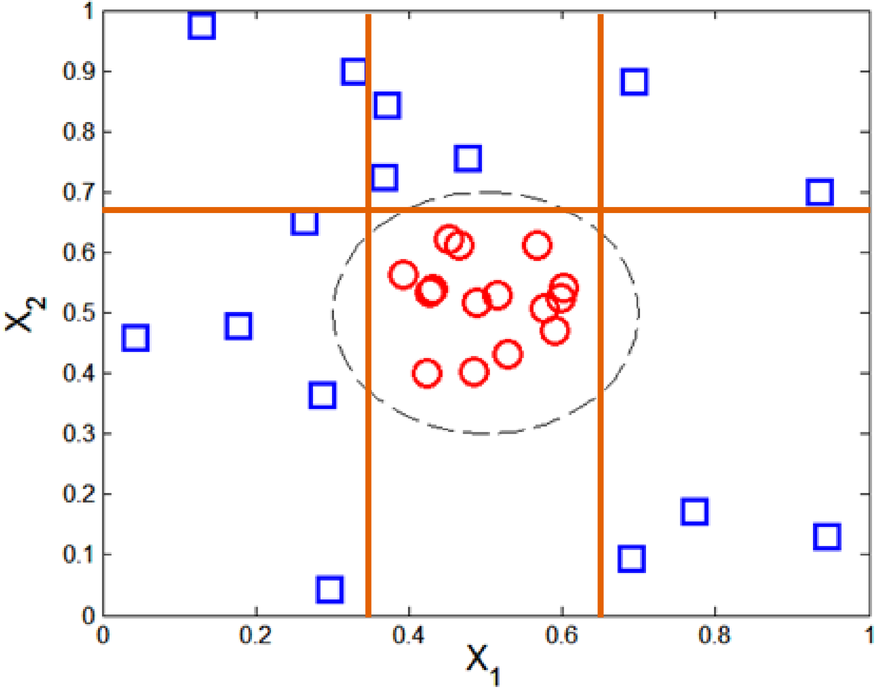
Training errors (apparent errors)

- Errors committed on the training set

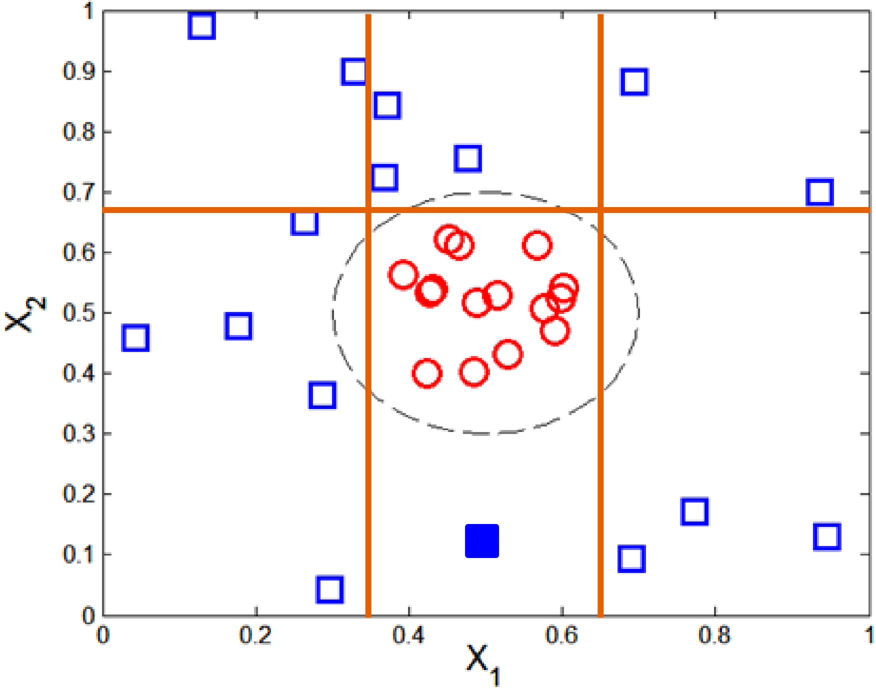
Test errors

- Errors committed on the test set

Training Errors



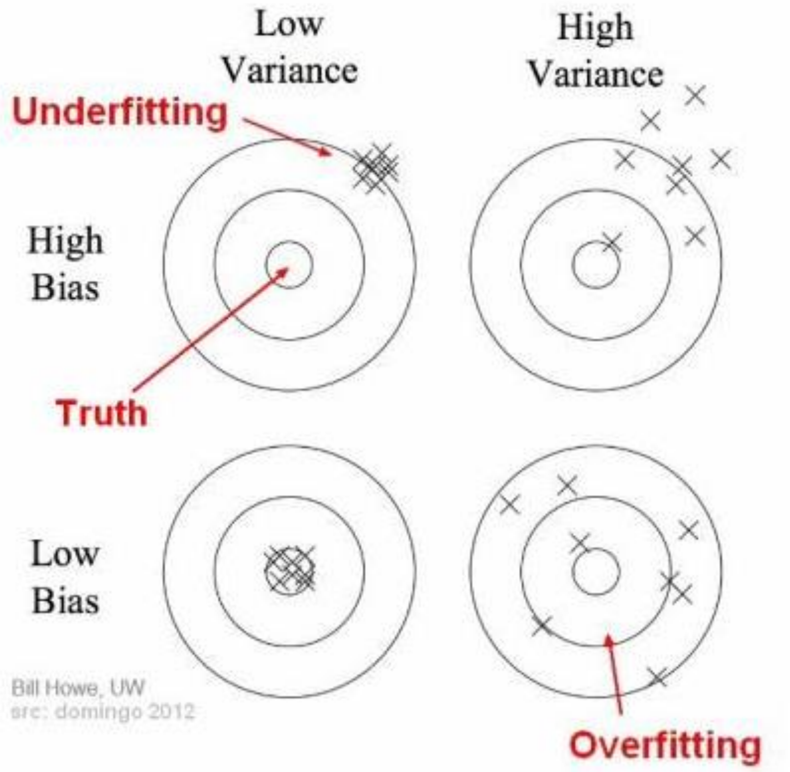
Test Errors



Bias and Variance

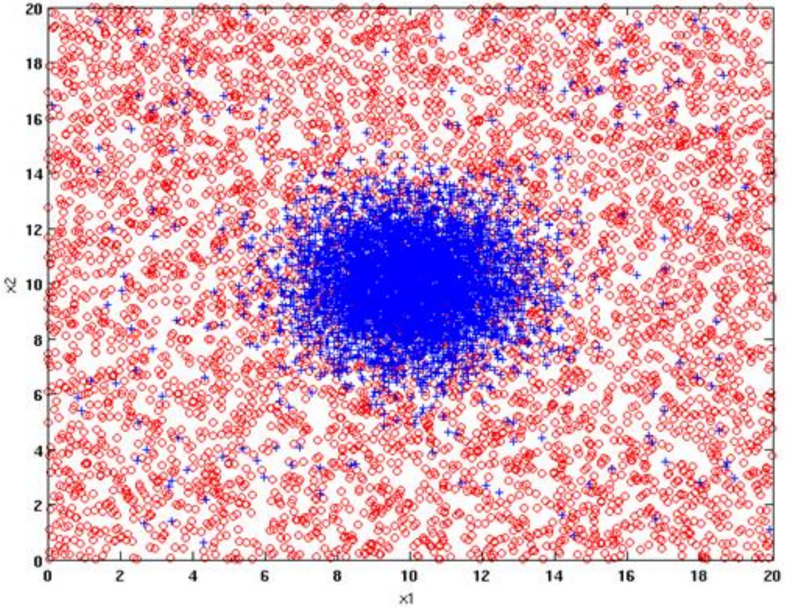
- **Underfitting:** Model with **high bias** pays very little attention to the training data and oversimplifies the model.

- **Overfitting:** Model with **high variance** pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.



Bill Howe, UW
src: domingo 2012

Example Data



Two class problem:

+ : 5400 instances

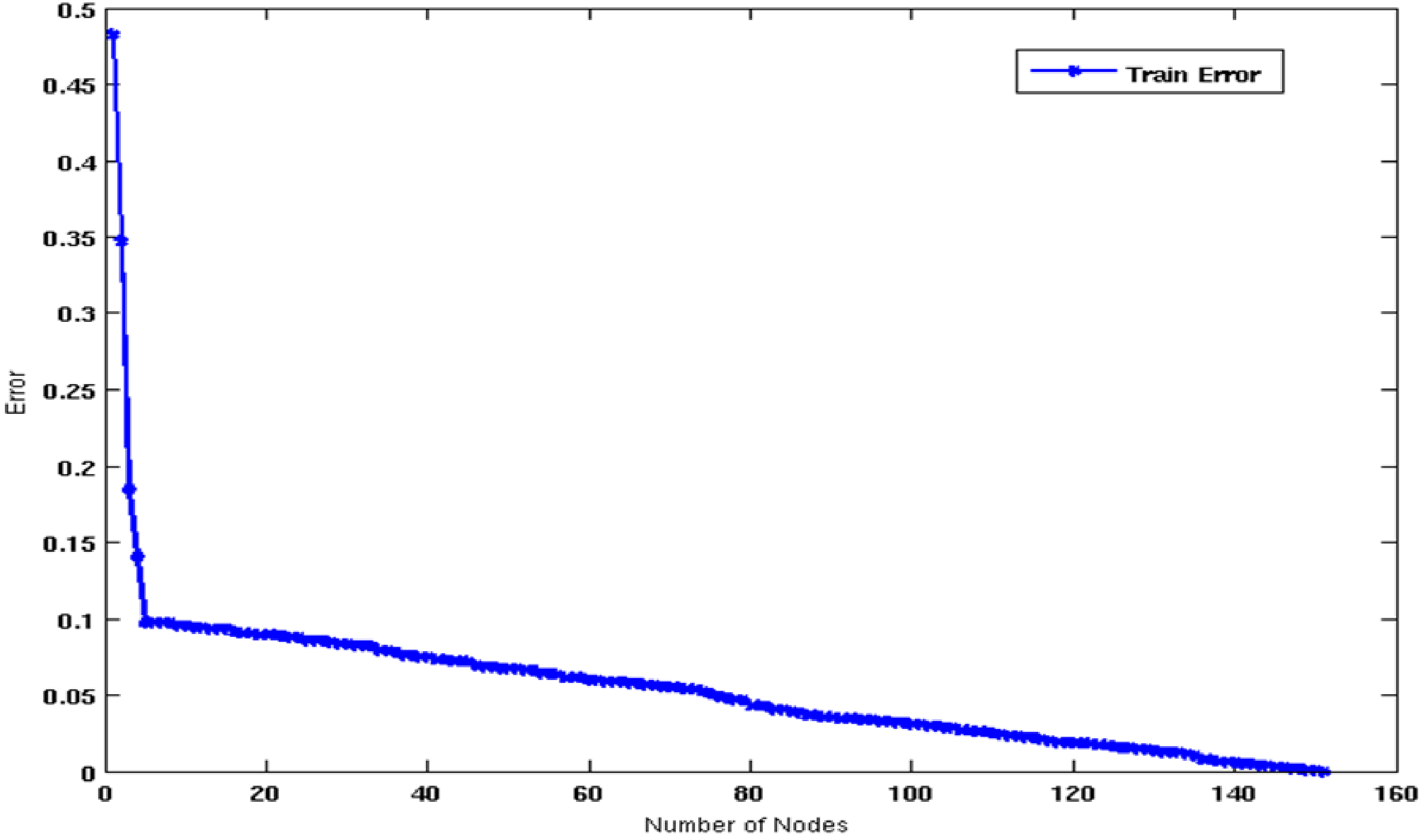
- 5000 instances generated from a Gaussian centered at (10,10)
- 400 noisy instances added

o : 5400 instances

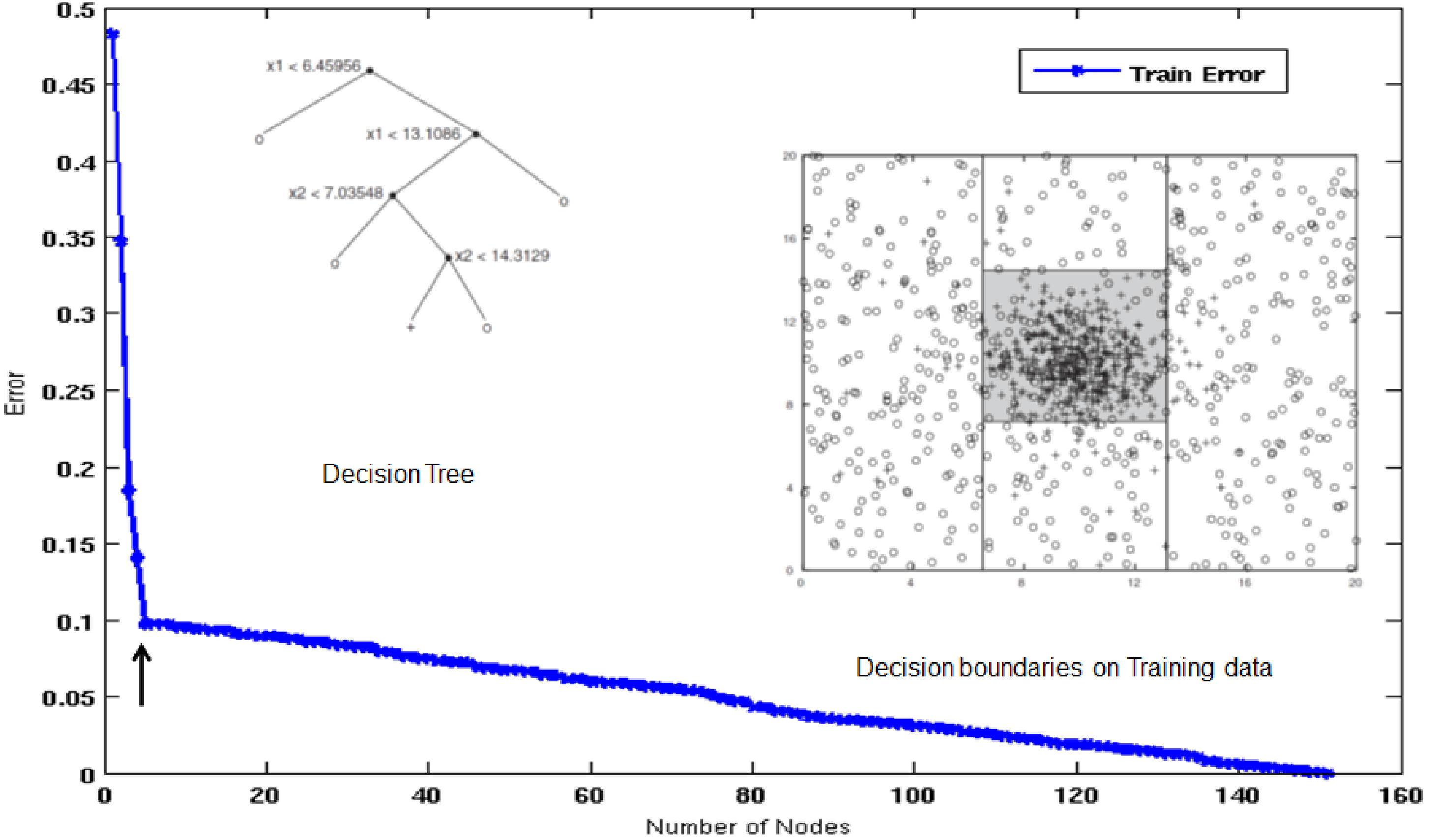
- Generated from a uniform distribution

10 % of the data used for training and 90% of the data used for testing

Decision Tree



Decision Tree

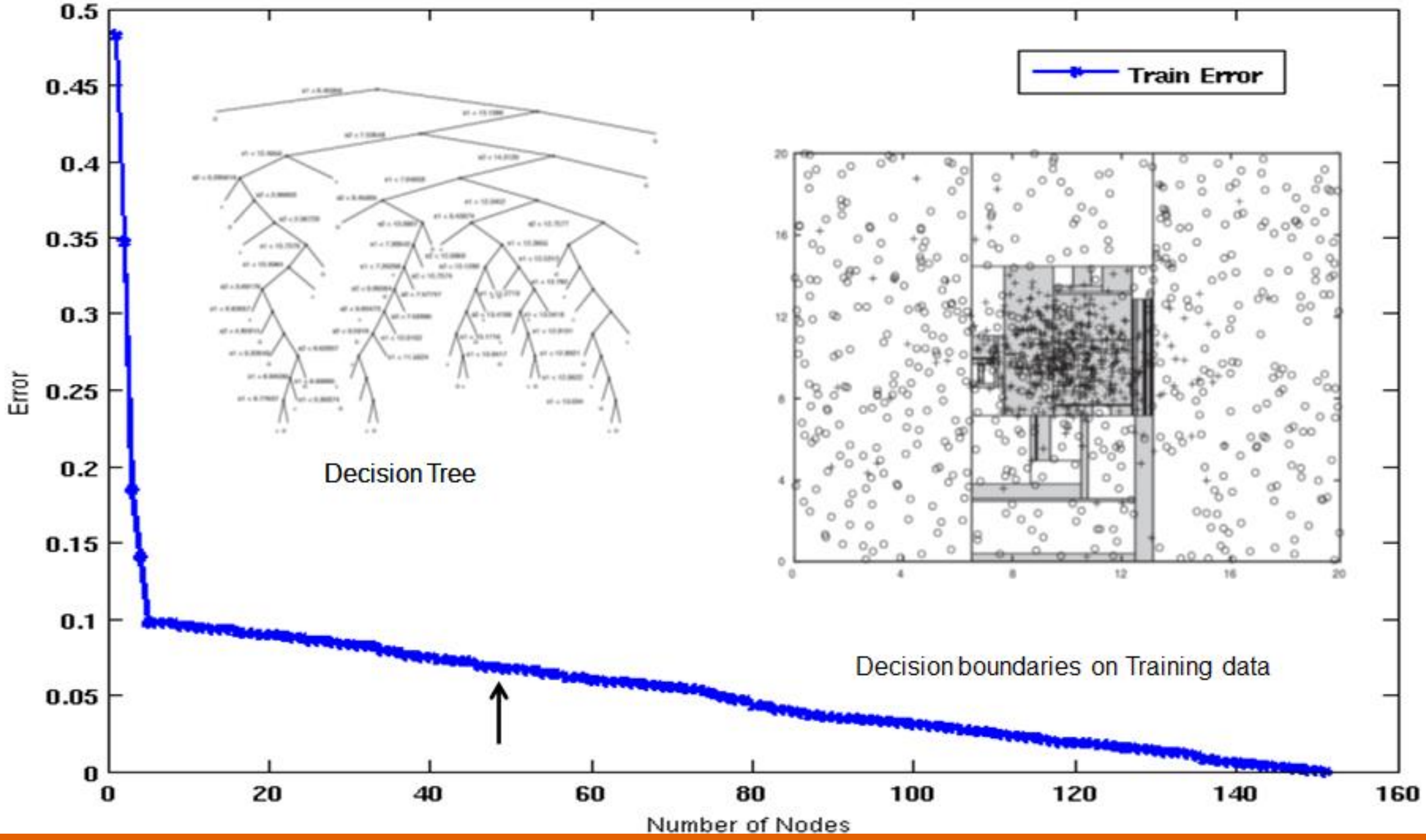


Decision Tree

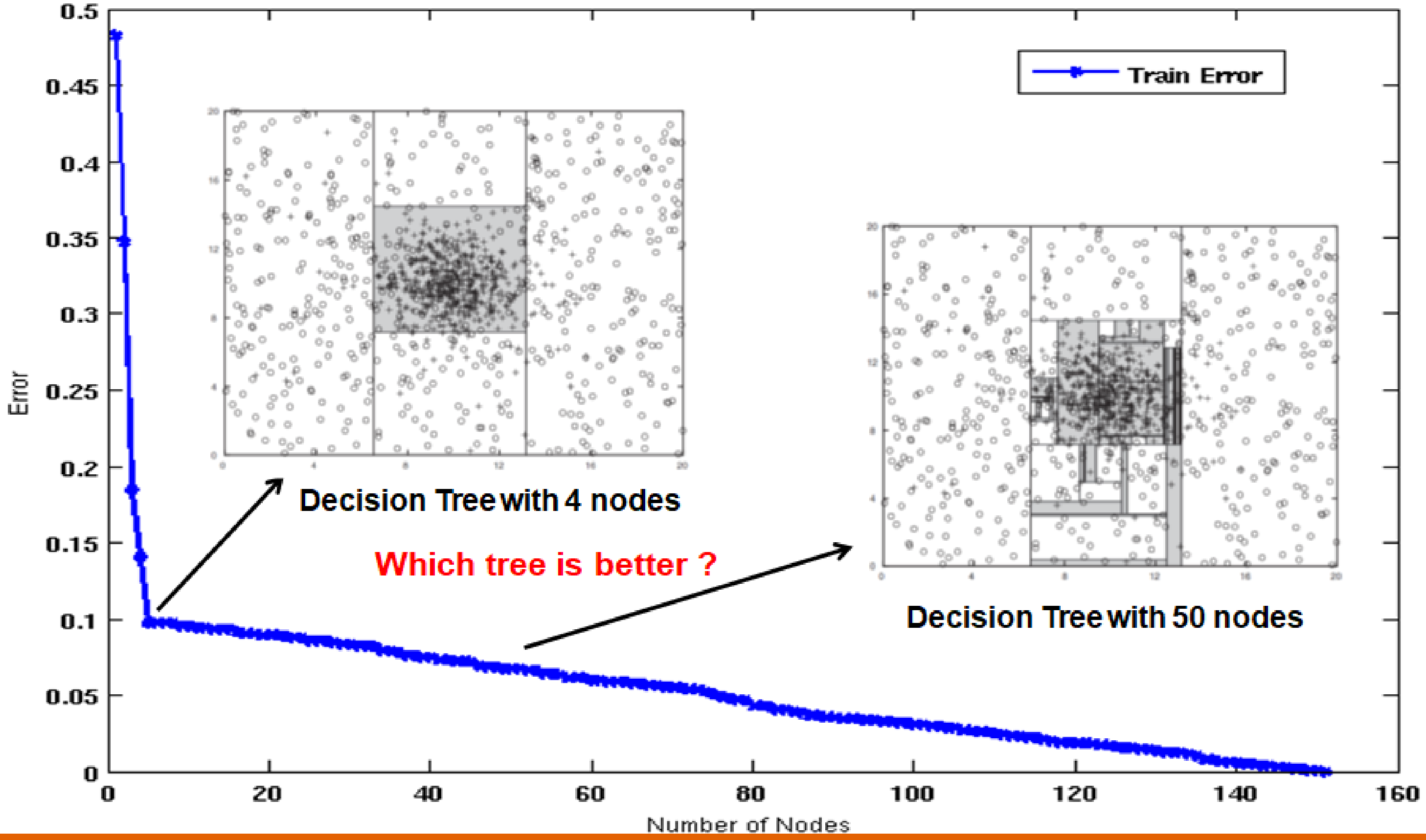
Decision boundaries on Training data

Overfitting

Decision Tree



Decision Tree

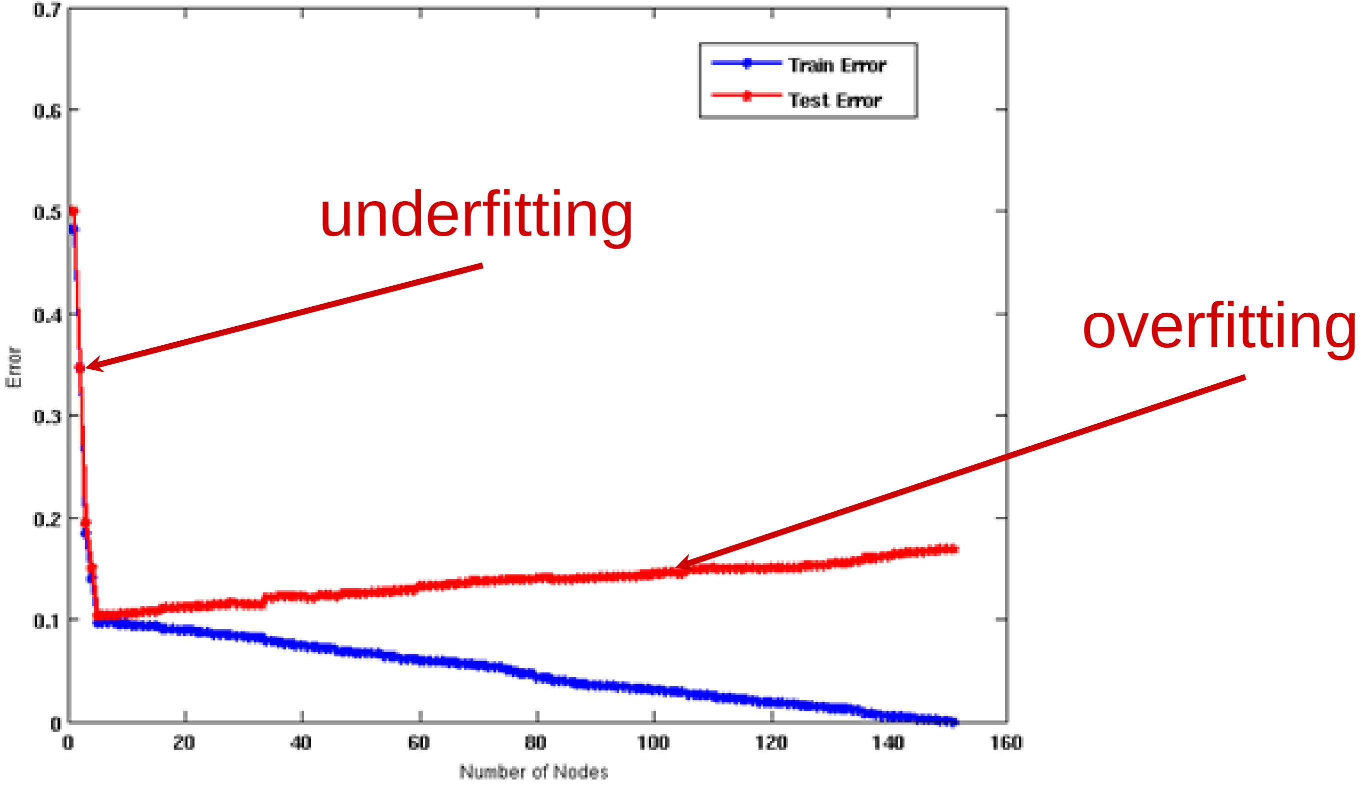


Decision Tree with 4 nodes

Which tree is better ?

Decision Tree with 50 nodes

Overfitting



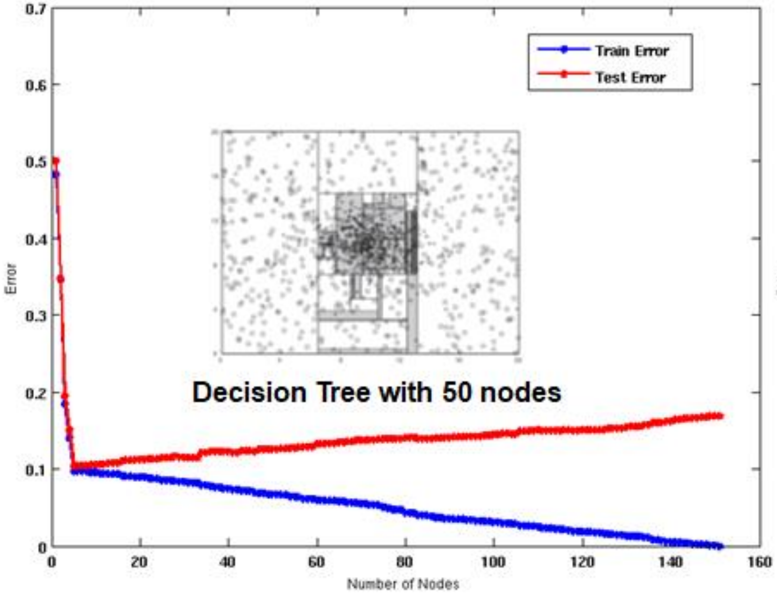
underfitting

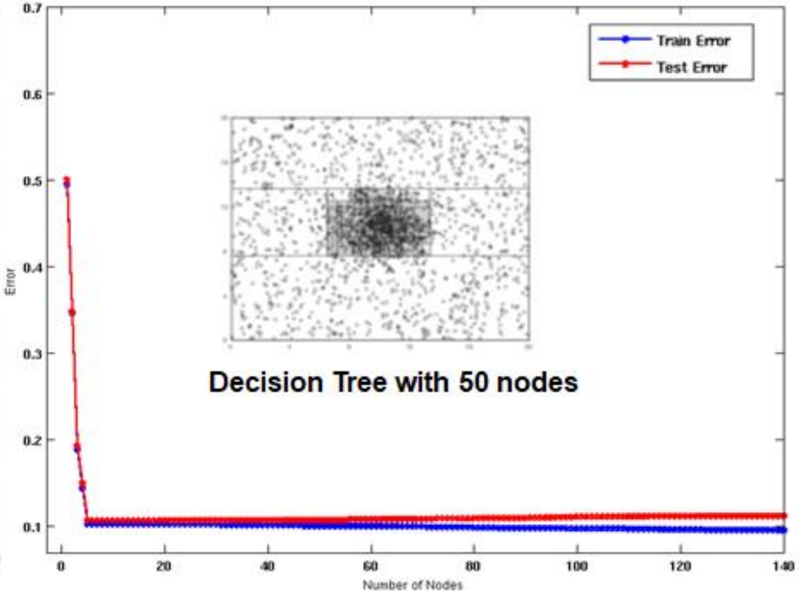
overfitting

Overfitting

Reasons for overfitting

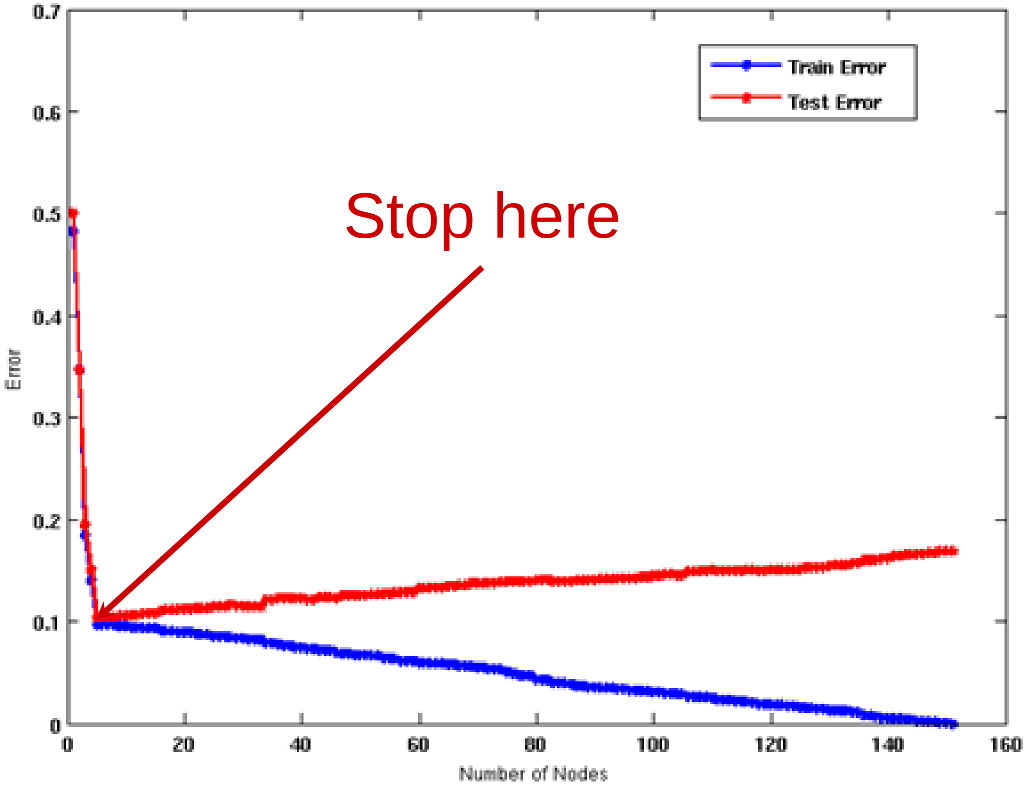
- **Limited training data size**
- **High model complexity**
  - E.g. too many nodes in a decision tree





Using twice the number of data instances

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model



Stop here

Overfitting

Avoiding overfitting

- **Avoid highly complex models**
    - Early stopping rules
    - Simplify model after training

Avoiding overfitting in decision trees

- **Pre-Pruning (Early Stopping Rule)**
  - Stop if number of instances is less than some user-specified threshold (**min_samples_split**).
  - Stop if the depth of the tree is reaches the user-specified maximum number (**max_depth**).
  - Stop if expanding the current node does not improve impurity measures (e.g., Gini or information gain) over a user-specified threshold (**min_impurity_decrease**).

Avoid overfitting in decision trees

- **Post-pruning (simplify the model)**
    - Grow decision tree to its entirety
    - Subtree replacement
        - Trim the nodes of the decision tree in a bottom-up fashion
        - If **generalization error** improves after trimming, replace sub-tree by a leaf node
        - Class label of leaf node is determined from majority class of instances in the sub-tree