POLITECNICO DI MILANO

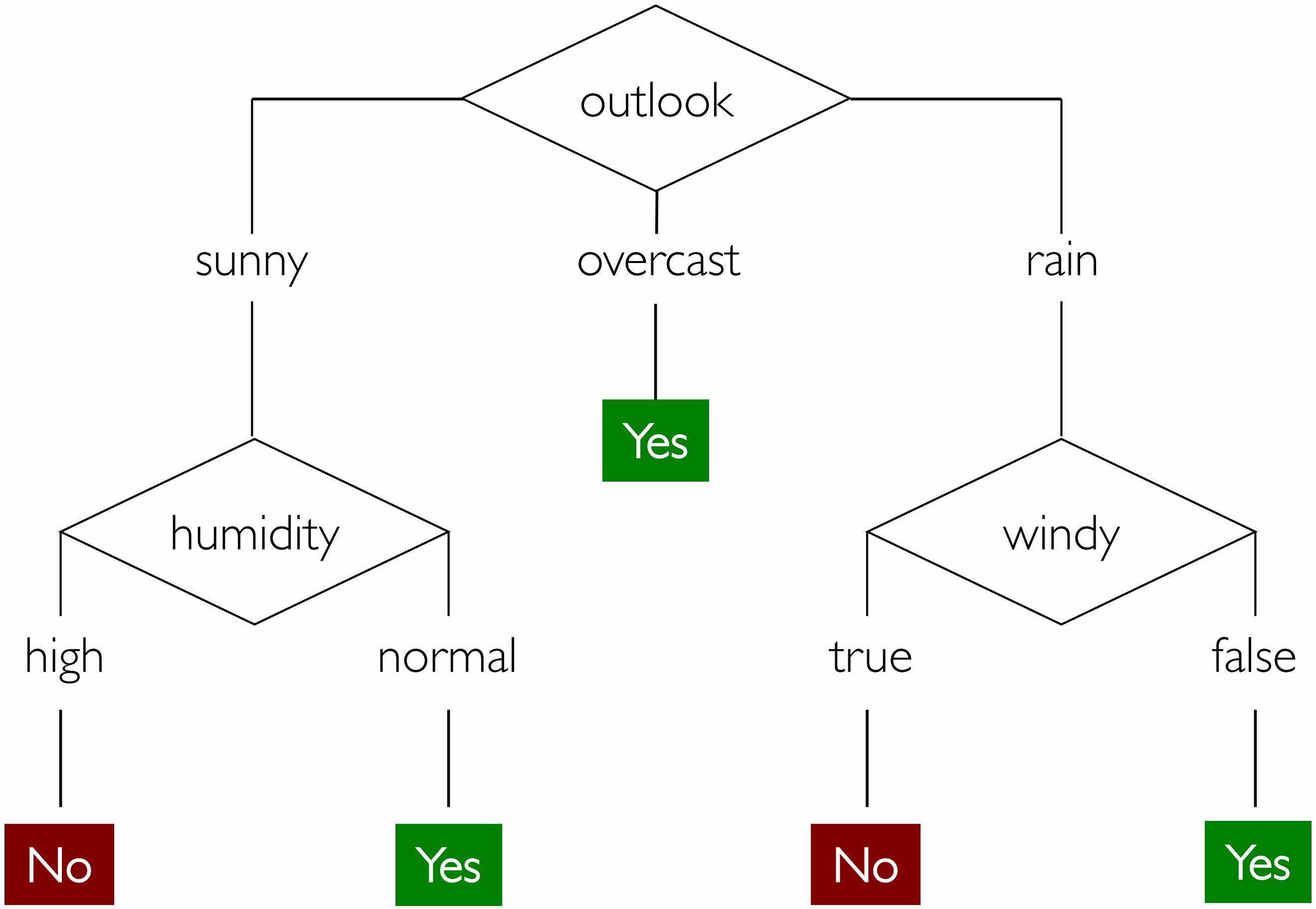
Decision Trees

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

The Weather Dataset

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

The Decision Tree for the Weather Dataset



What is a Decision Tree?

- An internal node is a test on an attribute
- A branch represents an outcome of the test, e.g., outlook=windy
- A leaf node represents a class label or class label distribution
- At each node, one attribute is chosen to split training examples into distinct classes as much as possible
- A new case is classified by following a matching path to a leaf node

Computing Decision Trees

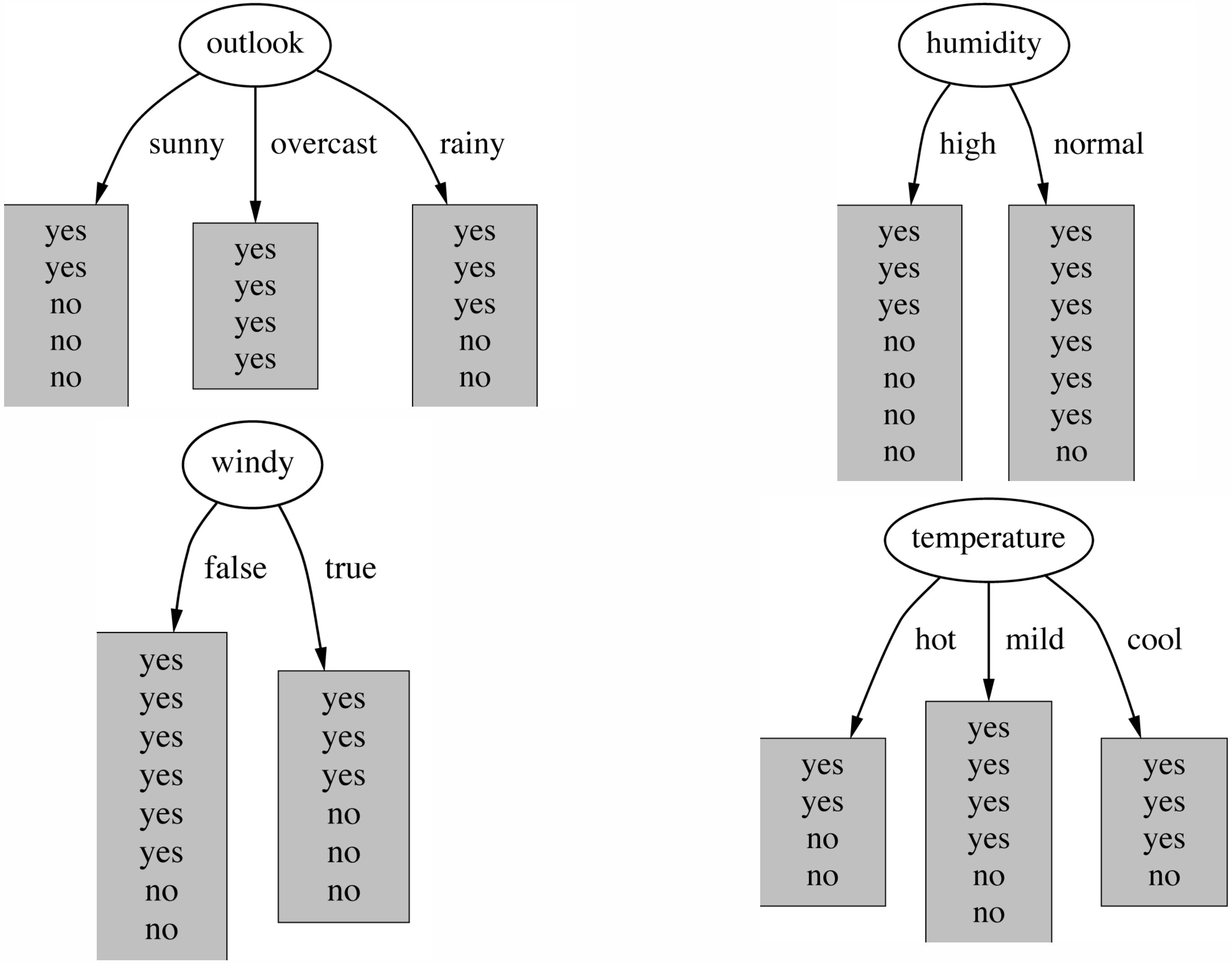
- Top-down Tree Construction
  - Initially, all the training examples are at the root
  - Then, the examples are recursively partitioned, by choosing one attribute at a time

- Bottom-up Tree Pruning
  - Remove subtrees or branches, in a bottom-up manner, to improve the estimated accuracy on new cases.

Which Attribute for Splitting?

- At each node, available attributes are evaluated on the basis of separating the classes of the training examples
- A purity or impurity measure is used for this purpose
- Splitting Strategy: choose the attribute that results in greatest purity gain
- Typical goodness functions: information gain (ID3), information gain ratio (C4.5), gini index (CART)

Which Attribute Should We Select?



Gini Index

Another Splitting Criteria: The Gini Index

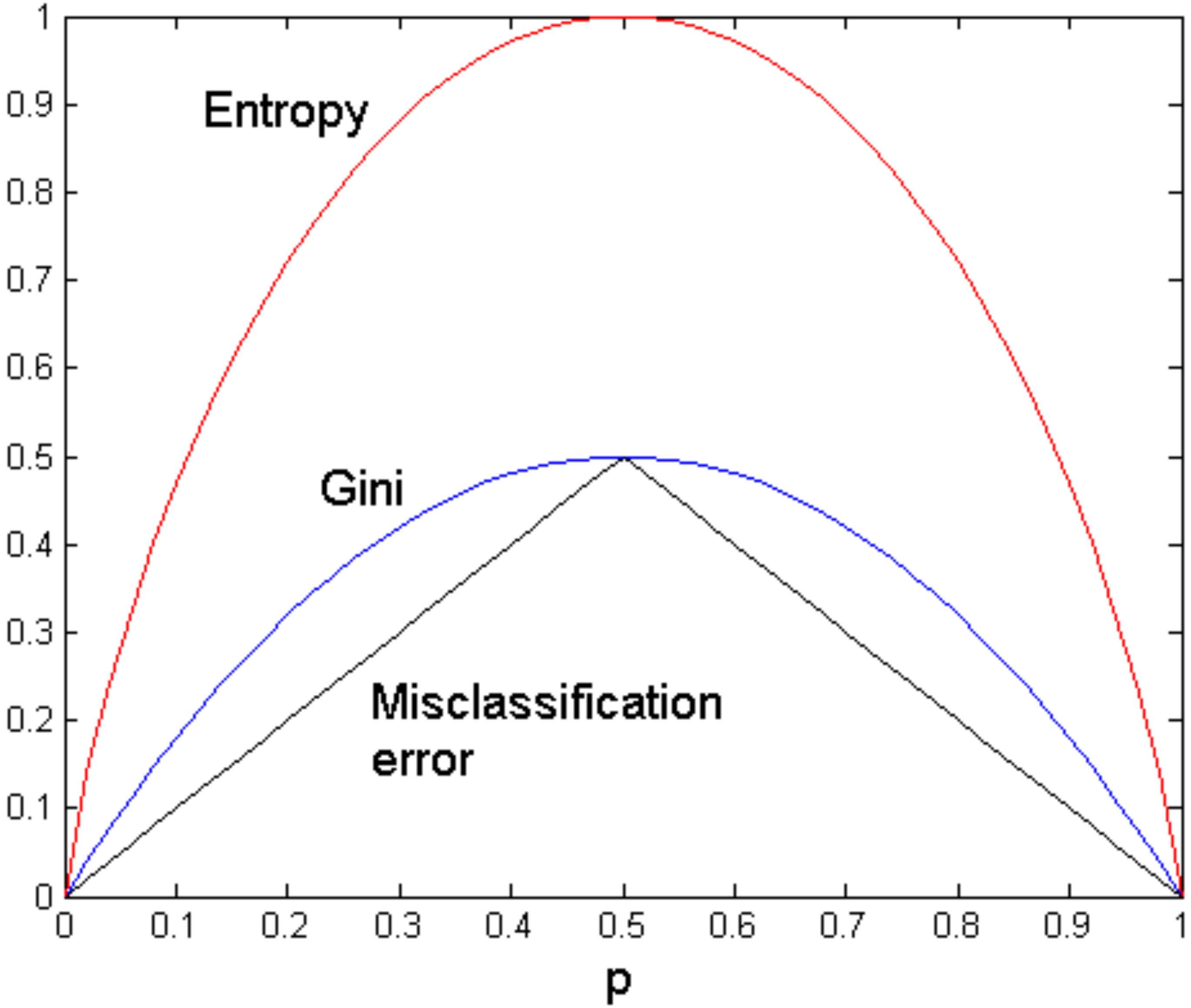
- The Gini index, for a data set T contains examples from n classes, is defined as

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

  where p_j is the relative frequency of class j in T

- gini(T) is minimized if the classes in T are skewed

The Gini Index vs Entropy



Gini index is applied
to produce binary splits

The Gini Index for the Outlook Attribute

- The dataset has 9 tuples labeled "yes" and 5 labeled "no"

$$gini(D) = 1 - (\frac{9}{14})^2 - (\frac{5}{14})^2 = 0.459$$

- The outlook attribute has three values (overcast, rainy, sunny), thus we have to evaluate three possible partitions
  - {overcast, rainy} and {sunny}
  - {sunny, rainy} and {overcast}
  - {sunny, overcast} and {rainy}

When Should Building Stop?

- There are several possible stopping criteria
- All samples for a given node belong to the same class
- If there are no remaining attributes for further partitioning, majority voting is employed
- There are no samples left
- Or there is nothing to gain in splitting

We Can Always Build a 100% Accurate Tree

But do we want it?

Generalization and overfitting in trees

Avoiding Overfitting in Decision Trees

- The generated tree may overfit the training data
- Too many branches, some may reflect anomalies due to noise or outliers
- Result is in poor accuracy for unseen samples
- Two approaches to avoid overfitting
  - Prepruning
  - Postpruning

Pre-pruning vs Post-pruning

- Prepruning
  - Halt tree construction early
  - Do not split a node if this would result in the goodness measure falling below a threshold
  - Difficult to choose an appropriate threshold

- Postpruning
  - Remove branches from a "fully grown" tree
  - Get a sequence of progressively pruned trees
  - Use a set of data different from the training data to decide which is the "best pruned tree"

Decision trees can also be used to predict the value of a numerical target variable

Regression and model trees work similarly to decision trees

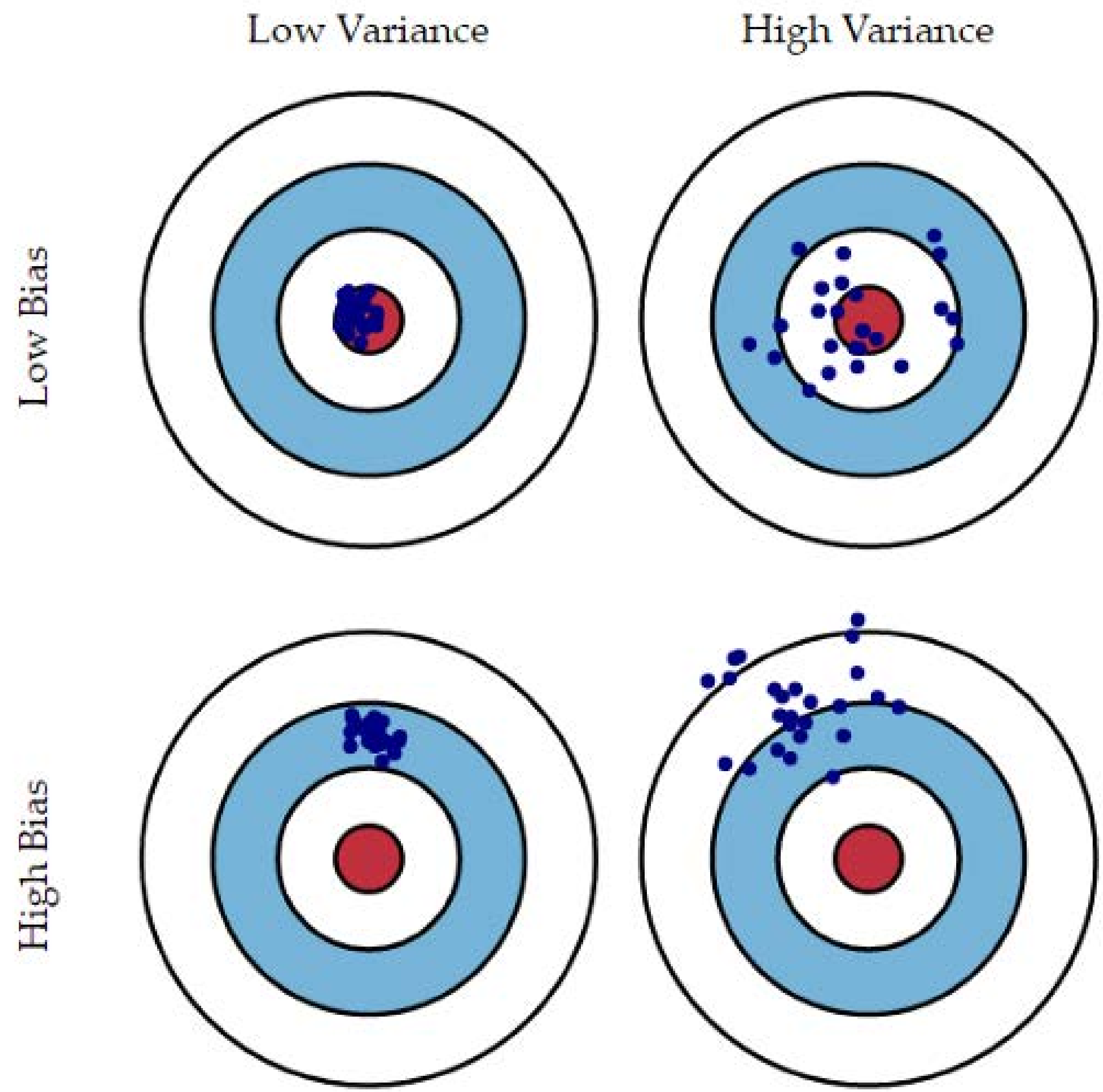
They search for the best split that minimizes an impurity measure

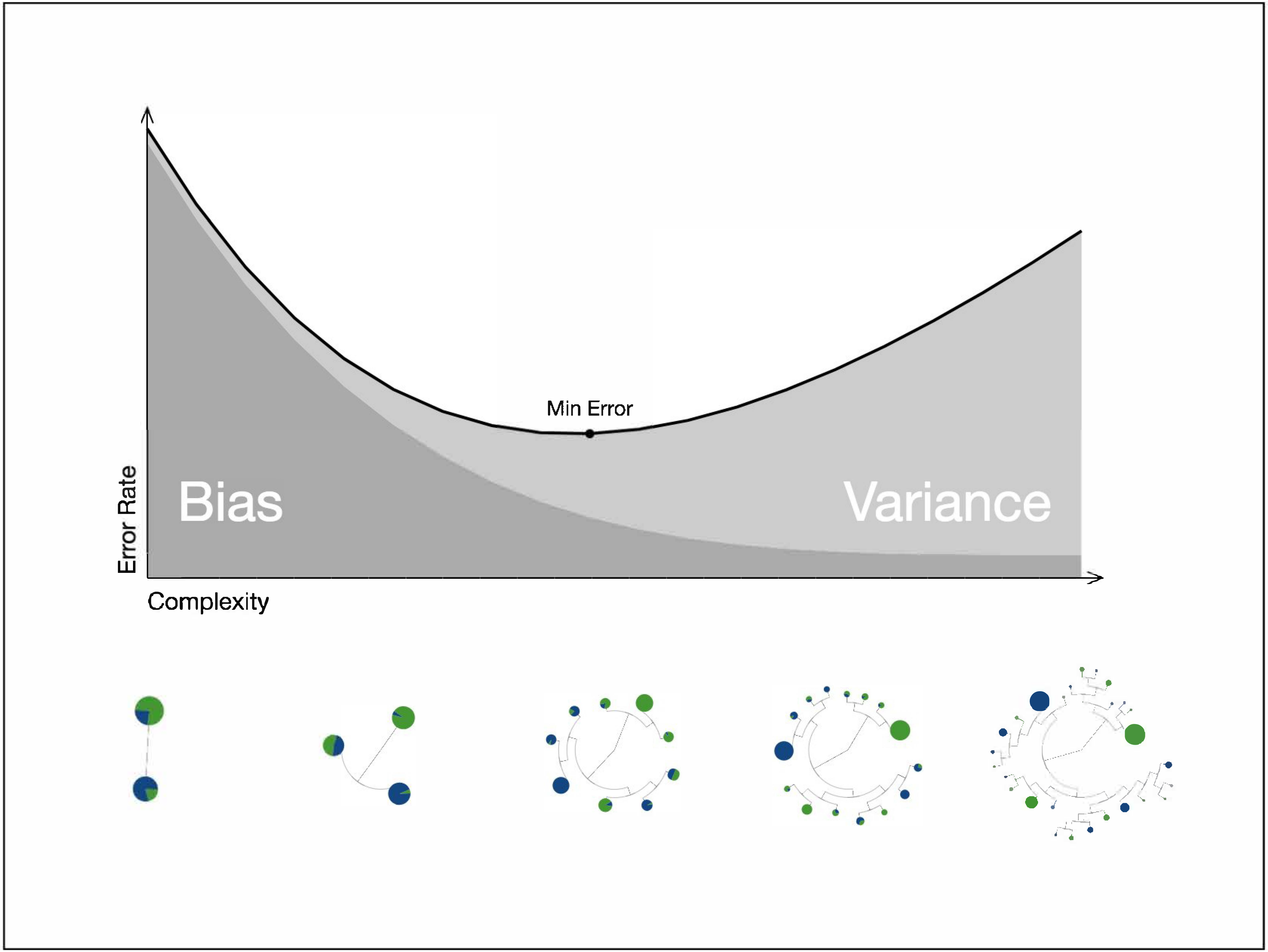
Summary

Summary

- Decision tree construction is a recursive procedure involving
  - The selection of the best splitting attribute via a purity measure
  - Early stopping or pruning to avoid overfitting
- Trees are very easy to interpret and explain
- Their predictive power is limited by their tendency to overfit the data

Bias-Variance Decomposition


Low Variance
High Variance
Low Bias
High Bias


Min Error
Bias
Variance
Error Rate
Complexity