Data Mining with Weka

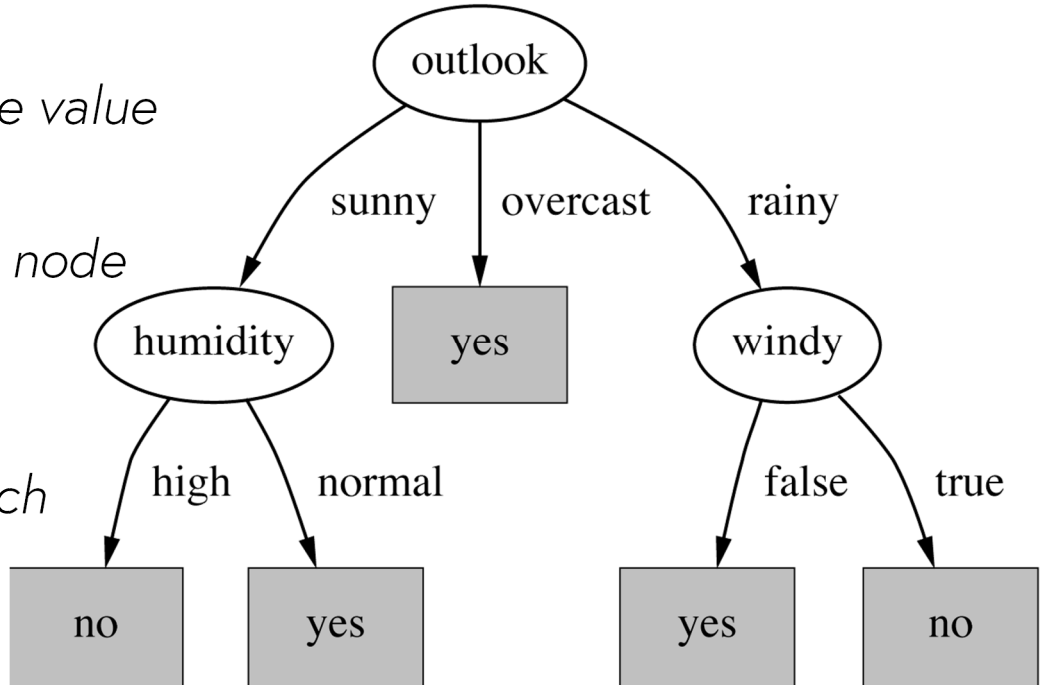
Decision trees

Ian H. Witten

Decision trees

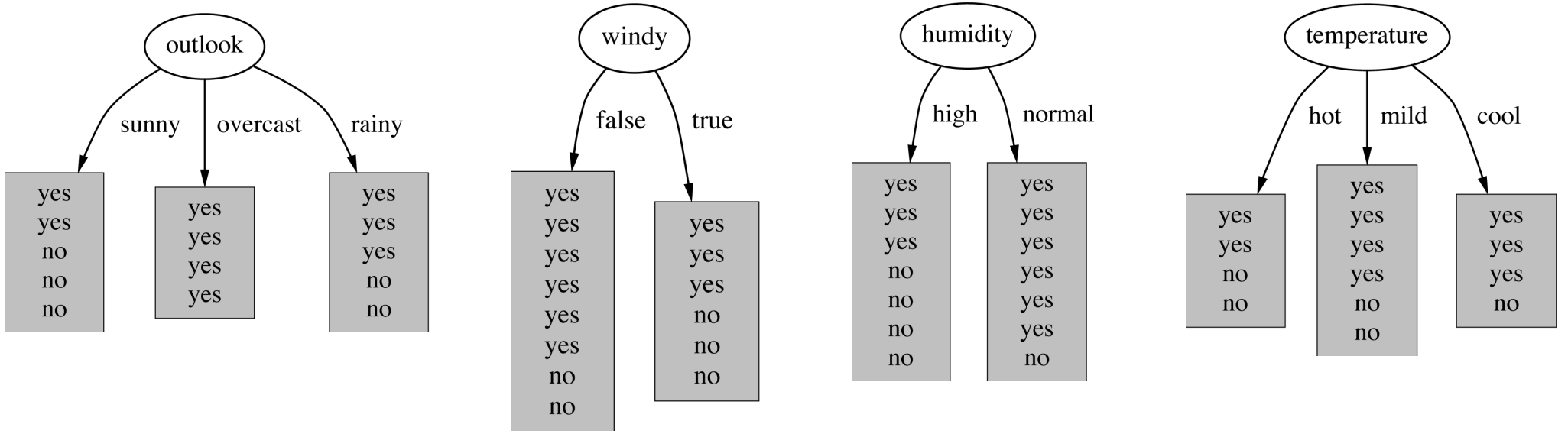
Top-down: recursive *divide-and-conquer*

- **Select** attribute for root node
  - *Create branch for each possible attribute value*
- **Split** instances into subsets
  - *One for each branch extending from the node*
- **Repeat** recursively for each branch
  - *using only instances that reach the branch*
- **Stop**
  - *if all instances have the same class*



Decision trees

Which attribute to select?

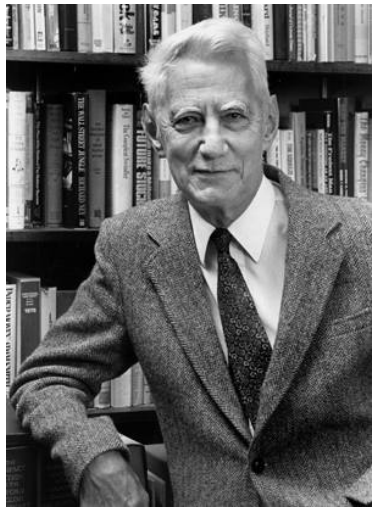


Decision trees

Which is the best attribute?

- Aim: to get the smallest tree
- Heuristic
  - *choose the attribute that produces the "purest" nodes*
  - *I.e. the greatest information gain*
- Information theory: measure information in bits

$$\text{entropy}(p_1, p_2, \ldots, p_n) = -p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n$$



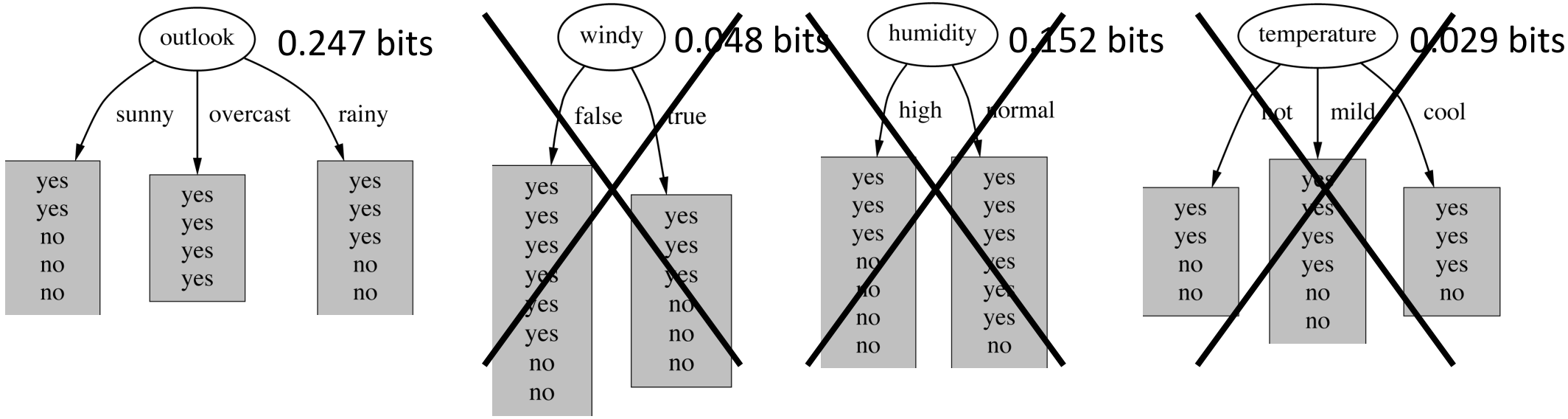
Information gain

- Amount of information gained by knowing the value of the attribute
- (Entropy of distribution before the split) – (entropy of distribution after it)

Claude Shannon, American mathematician and scientist 1916–2001

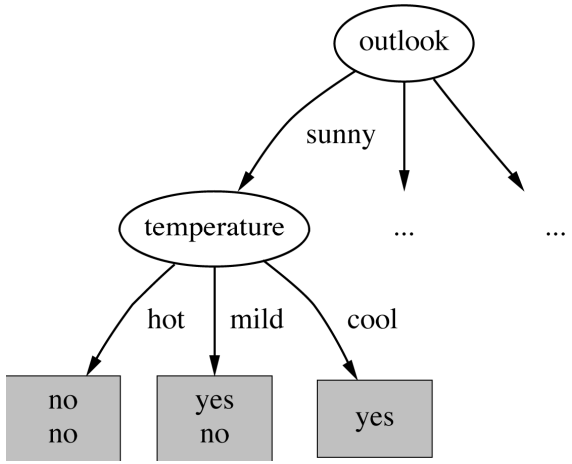
Decision trees

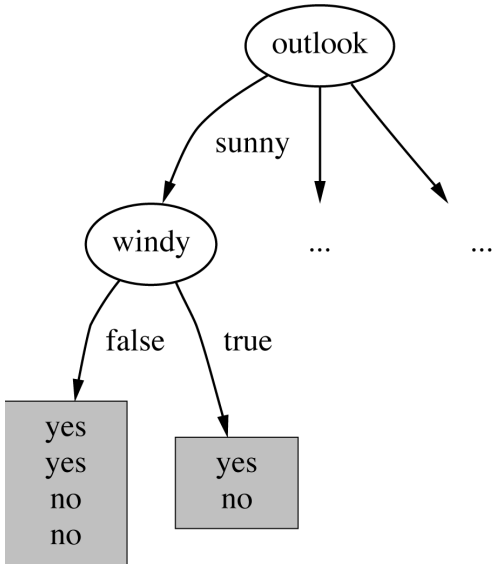
Which attribute to select?

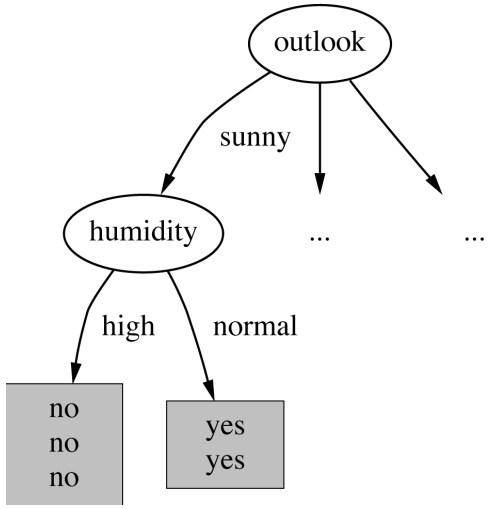


Decision trees

Continue to split ...







gain(*temperature*) = 0.571 bits

gain(*windy*) = 0.020 bits

gain(*humidity*) = 0.971 bits

Decision trees

Use J48 on the weather data

- Open file weather.nominal.arff
- Choose J48 decision tree learner (trees>J48)
- Look at the tree
- Use right-click menu to visualize the tree

Decision trees

- J48: "top-down induction of decision trees"
- Soundly based in information theory
- Produces a tree that people can understand
- Many different criteria for attribute selection
  - rarely make a large difference
- Needs further modification to be useful in practice
  (next lesson)