

Decision trees

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 31, Spring 2019

Wednesday April 24

Plan

- Decision trees
- Engineering features
- Information entropy and feature selection
- ID3 algorithm
- Pruning

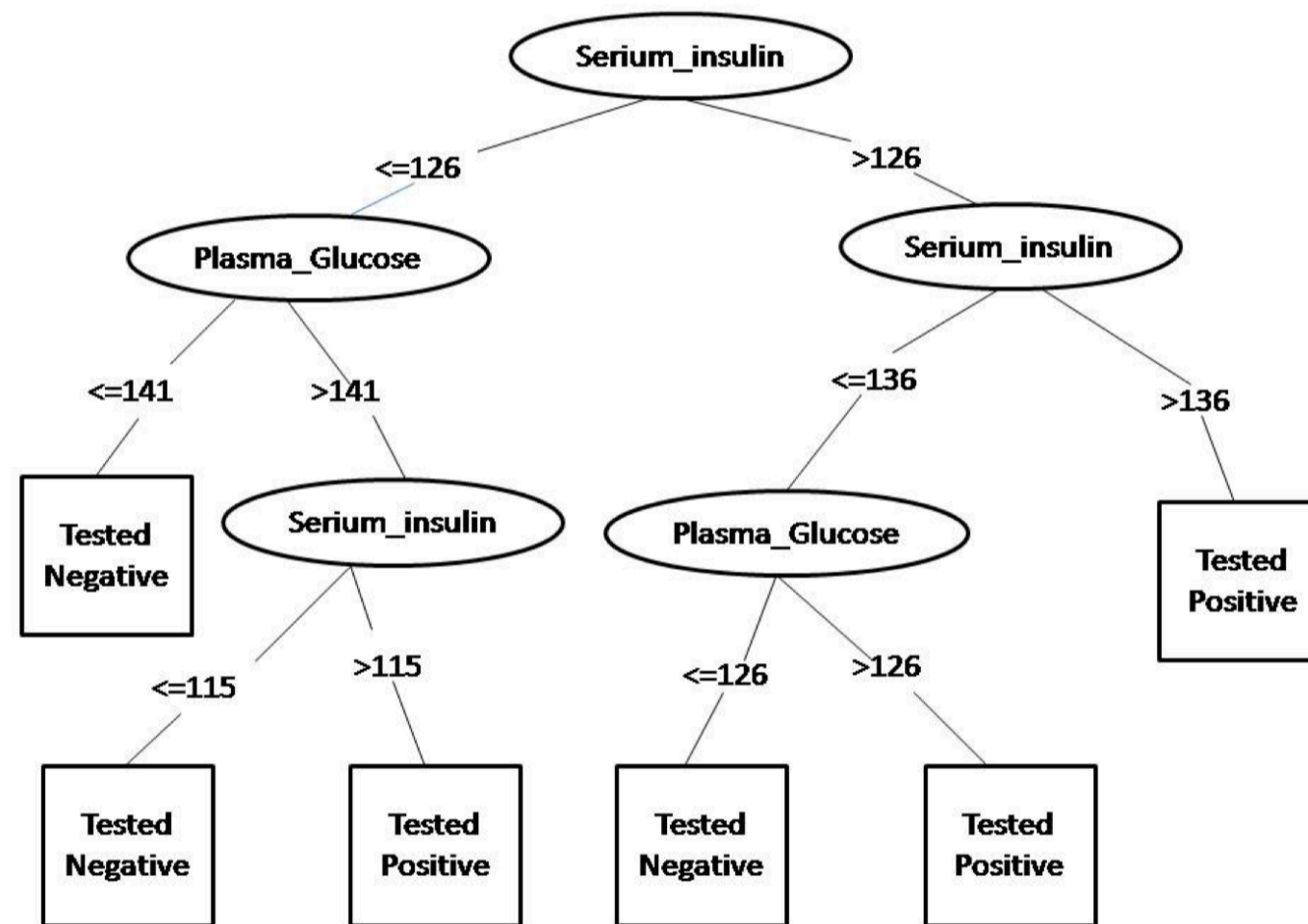
Diagnose diabetes (quasi-fictitious data)

Blood pressure	2hr serum insulin	2hr plasma glucose	Tested with diabetes?
Low	Low	Low	FALSE
Low	Low	High	FALSE
Low	High	Low	FALSE
Low	High	High	TRUE
High	Low	Low	FALSE
High	Low	High	TRUE
High	High	Low	TRUE
High	High	High	TRUE

From factors to quantitative data

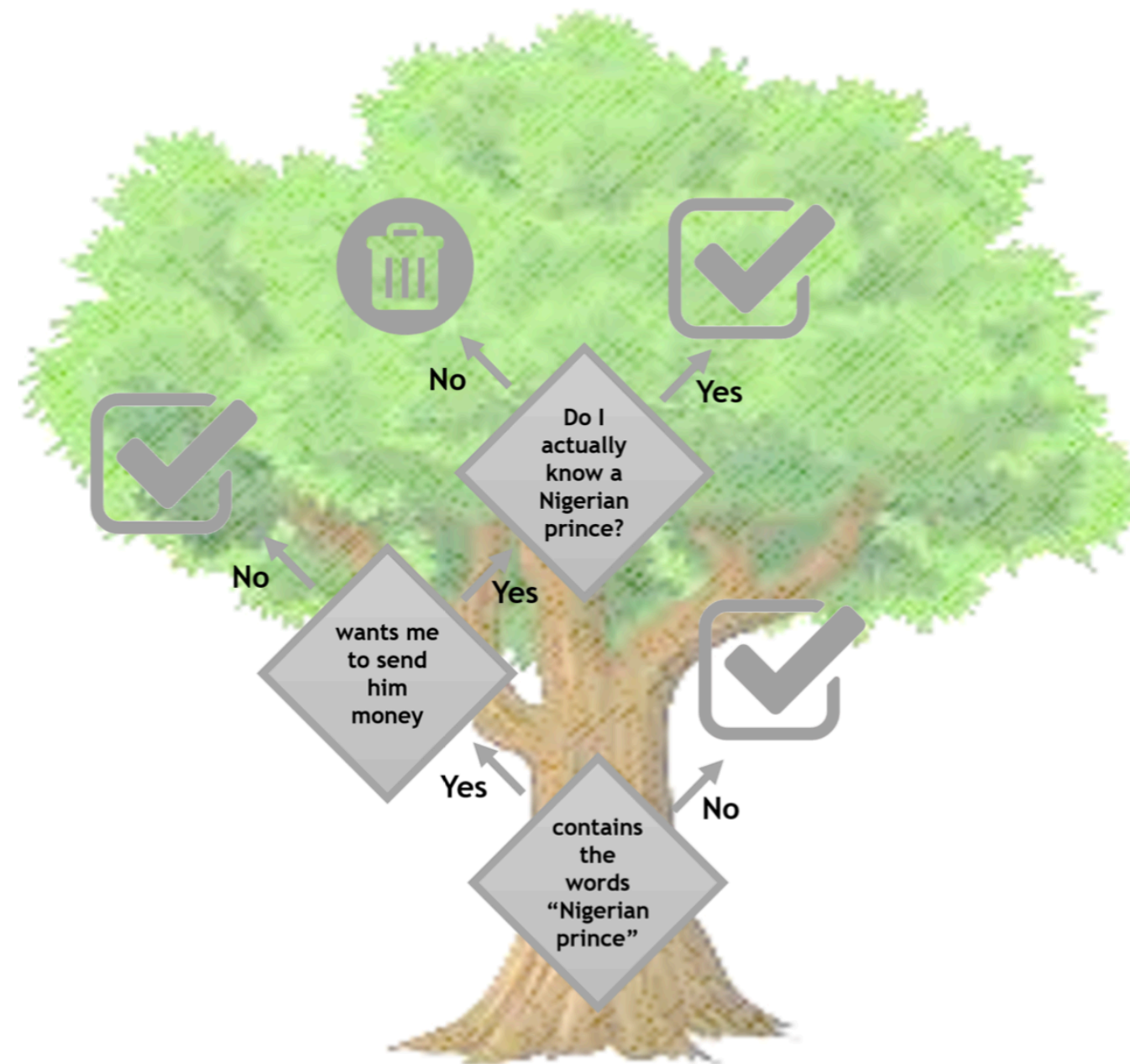
- Our features are all qualitative factors.
- How could we make rules out of quantitative data? Example: 2hr plasma glucose (mg/dL) $x = [148, 85, 183, 89, 137, 116, \dots]$
 1. Define threshold. $x > 100$: $[T, F, T, F, T, T]$
 2. Divide into ranges (e.g. < 80 , $80 \leq x < 120$, ≥ 120). Affect on tree?

Diabetes diagnosis with a decision tree



Data: Pima Diabetic data. Source: Barale & Shirke (2016)

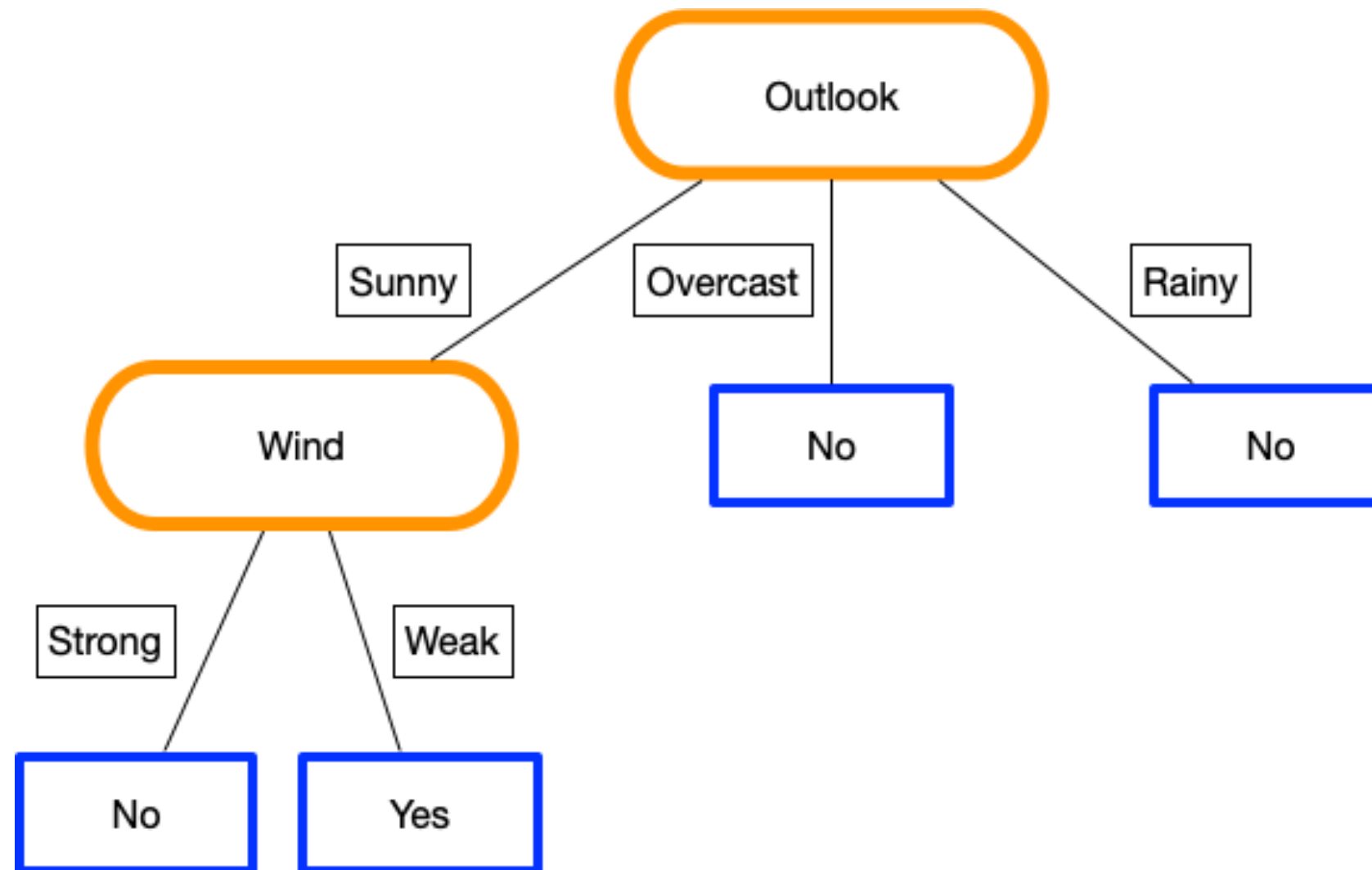
Another application of decision trees: spam filtering



Decision trees and conditionals

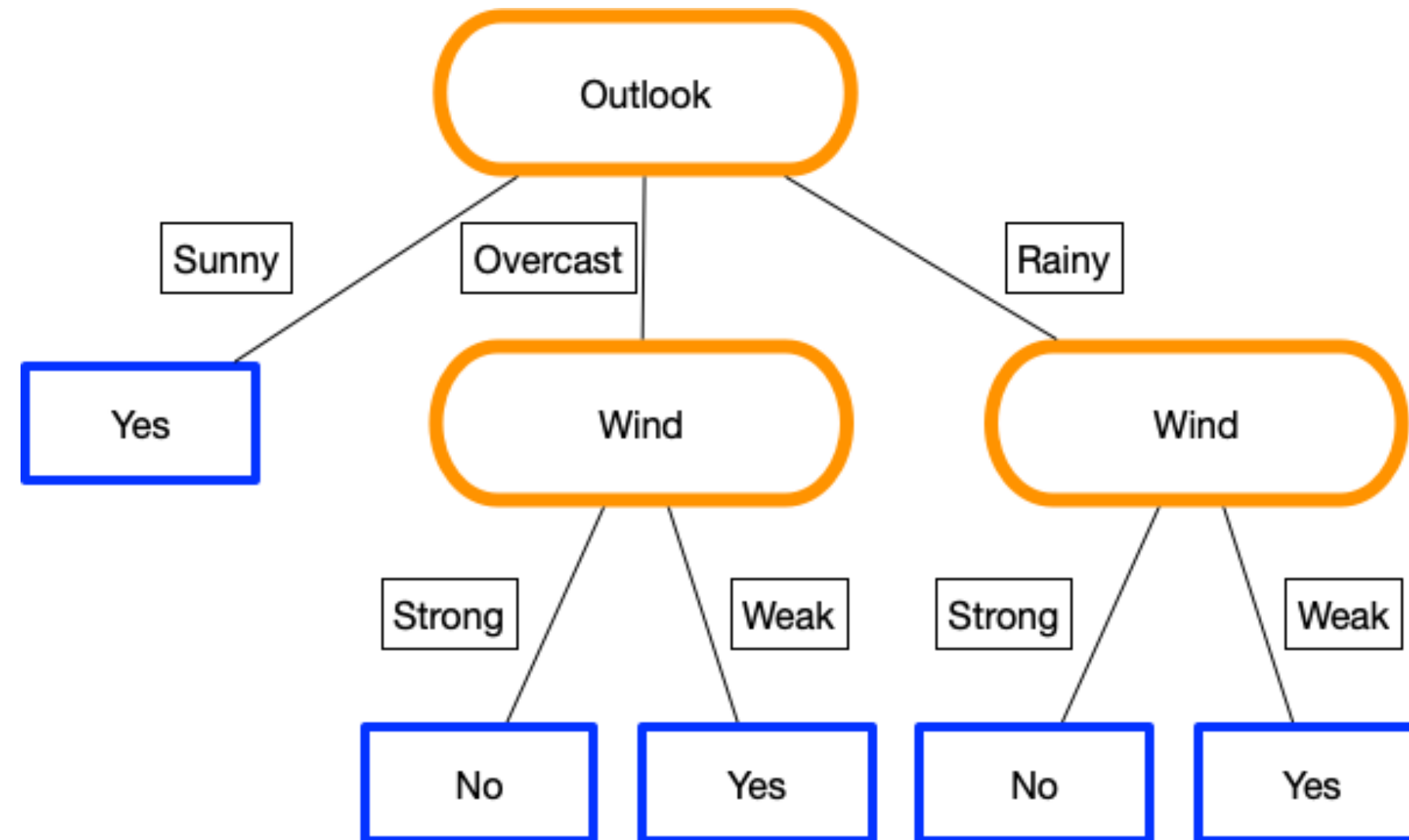
- Every path from the root to a leaf node can be represented by a long if statement (e.g. if blood pressure low and glucose low, you don't have diabetes).
- Trees themselves can encode boolean logic on the node features.

Decision tree for playing tennis (1/3)



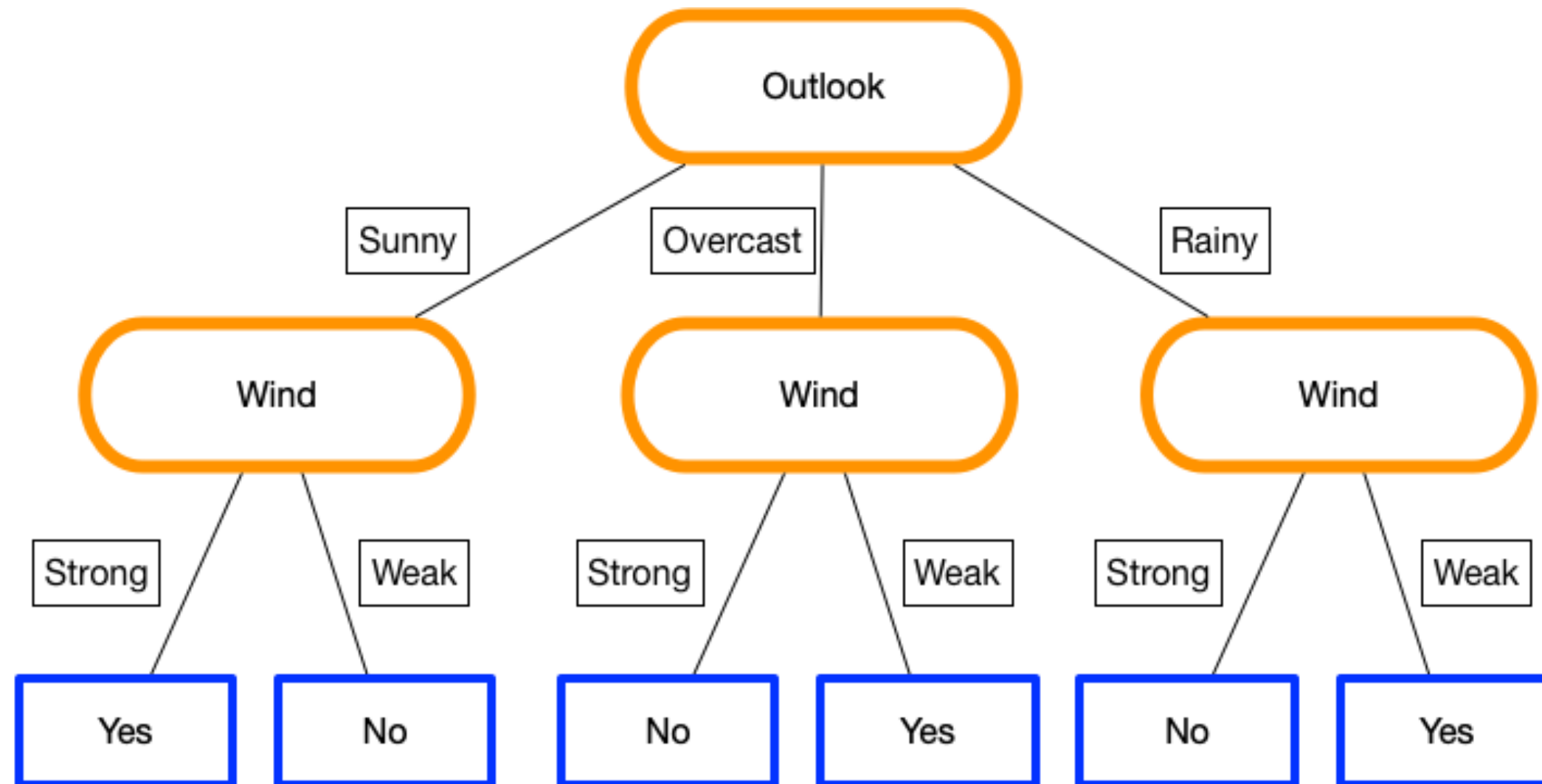
- Tree represents the rule: $(Outlook == Sunny) \implies (Wind == Weak)$

Decision tree for playing tennis (2/3)



- Tree represents the rule: $(Outlook == Sunny) \implies (Wind == Weak)$

Decision tree for playing tennis (3/3)



- Tree represents the rule: $(Outlook == Sunny) \implies (Wind == Weak)$

How do we determine the best question to ask at each node?

1. We want each question (node) to maximize the information we learn about how to classify — can this question tell us a lot about the output class, regardless of yes/no answer?
 - Best case scenario: toss out 1/2 dataset/available options each time.
2. Probability that we correctly guess the output class increases as we descend down the tree.

Decision nodes and information theory

- A question (decision node) that classifies data $\approx 100\%$ into a single class is called **pure** (low level of **impurity**).
- A **pure node** gives us little uncertainty about the likely output class ($N = 10$: [10 class A, 0 class B]).
- An **impure node** gives lots of uncertainty about the likely output class ($N = 10$: [5 class A, 5 class B]).
- Information theory quantifies uncertainty as **entropy**. Let's draw the entropy curve.