Clarify from last time:
- If number Positive = number Negative => do not break ties. Partition again with another feature.
- Keep handout example as is.

Play Tennis example

Suppose chose features: Outlook, Temp, Humidity, Wind.

At every level, we chose same feature, may not be most efficient.
Notice, at every level we chose the same feature, this may not be the most efficient, i.e. may not yield the smallest tree.
Want a small tree: depends on order of attribute selection.

Which is the "best" attribute to choose?

(Compare Type + Patrons)

Greedy Decision Tree Construction - at each node choose the feature that divides the examples into sets that are almost all positive or almost all negative.

How to find this feature?

Feature that:
- gives most information
- has "certainty"
- has least entropy

Expected/Average entropy of a random variable Y:
- estimates the uncertainty of a random variable
- measures the amount of info needed to find value of Y.
  \[ H(Y) = -\sum p(y) \log p(y) \]
- lower entropy = higher certainty = less info needed (better)
- value \( \in [0,1] \)
  
  0: no entropy (completely certain/little info needed)
  1: complete entropy (no certainty/lots of info needed)

ex. random variable is flip of:
- biased coin that is always heads: entropy = 0
- fair "": entropy = 1
Ultimate goal: choose feature with most information gain (use entropy to find this).

Formulas:

$$\text{Gain}(F, P) = \text{info gain of node } F \text{ with previous value } P.$$  

$$= \text{expected entropy of } P - \text{expected remaining entropy after choosing } F.$$  

But this is all probability-based, so we need expected:

$$= H(P) - R(F).$$

What about root? Has no previous value.

(Discuss later)

If F is root:

$$\text{Gain}(F, \text{Initial}) = \text{info node of root } F.$$. 