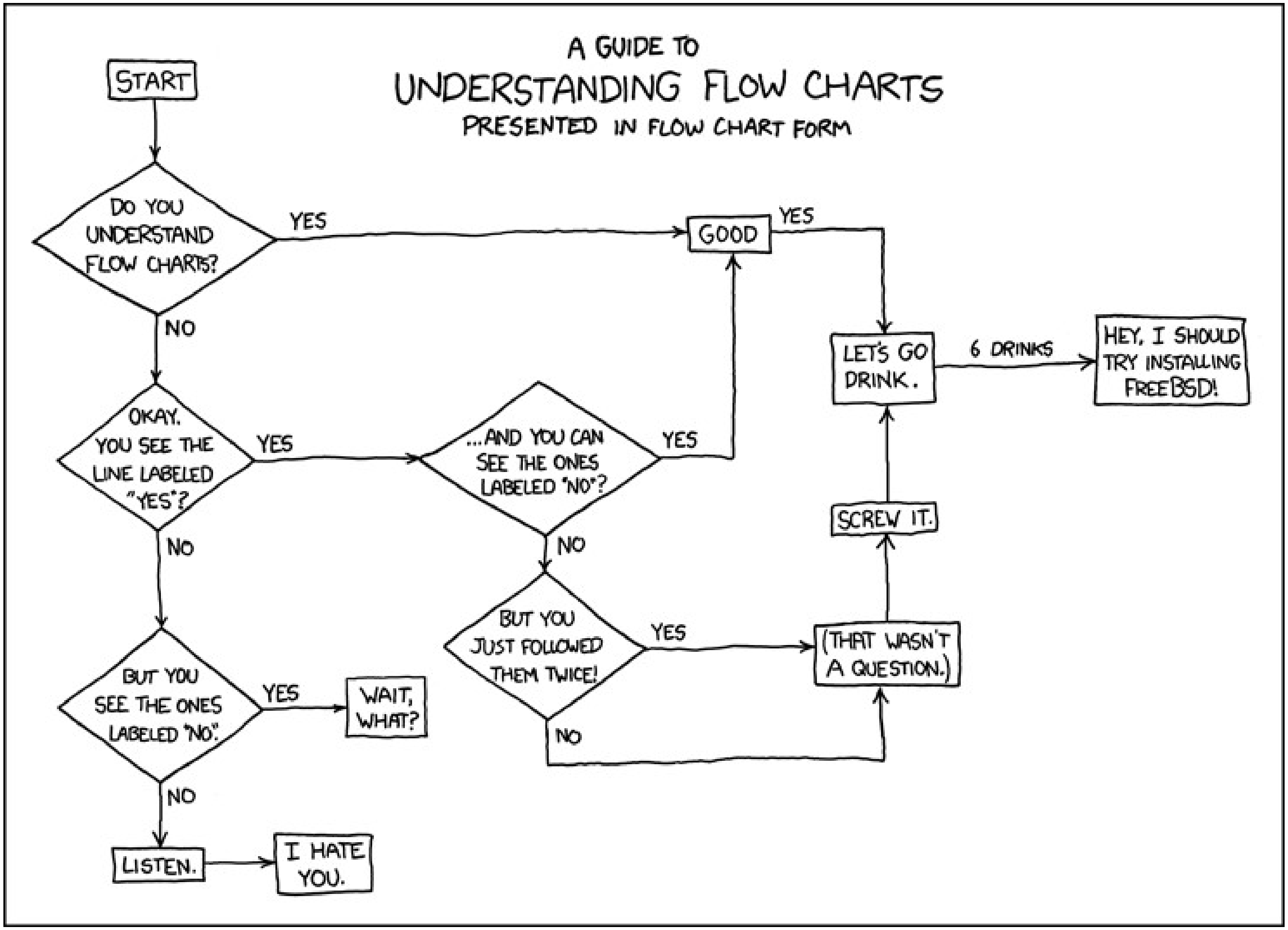
ICS Summer Academy Session II
Topic 4: Decision Trees

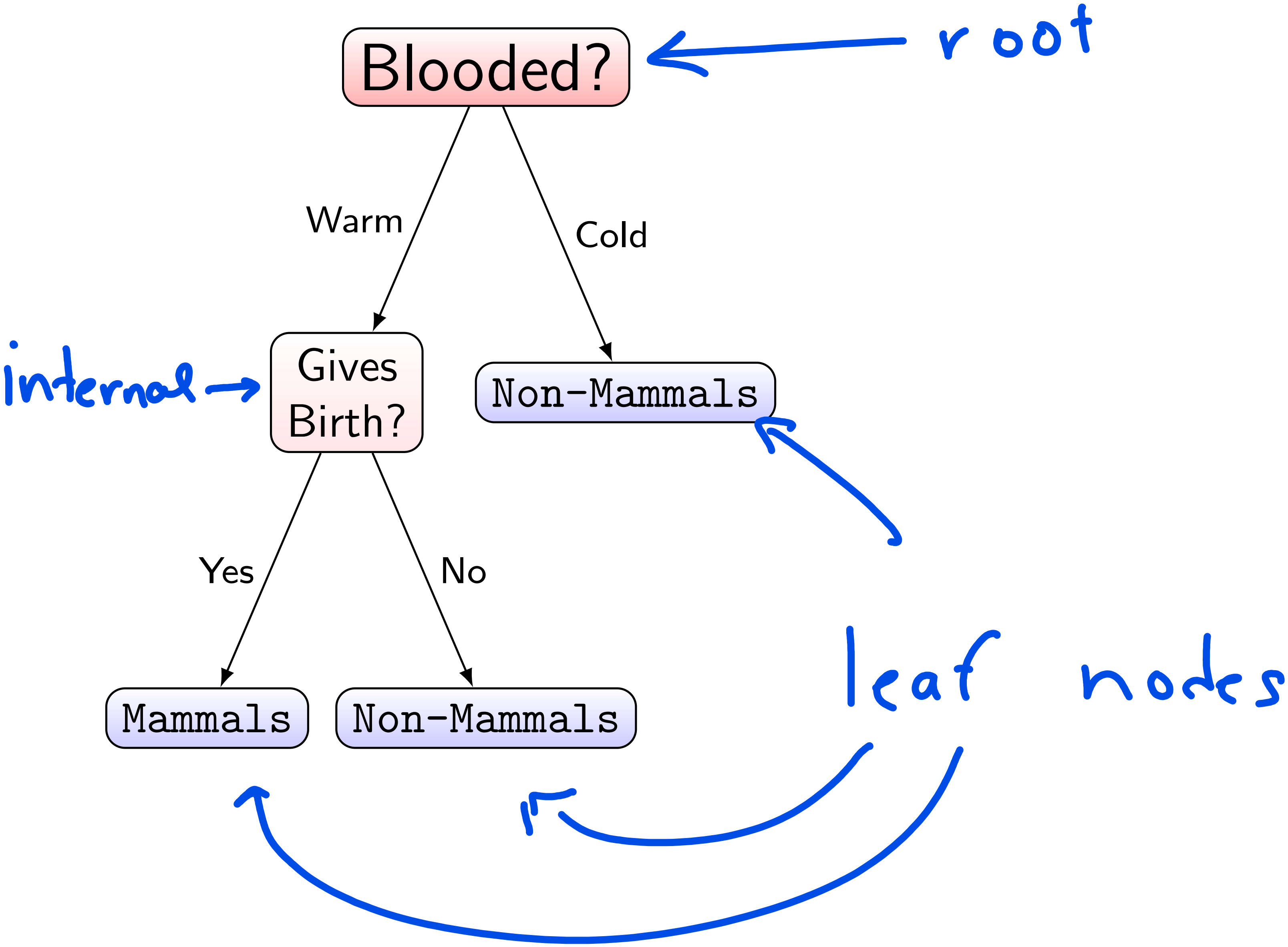
Michael Shindler

An aside: Flowcharts

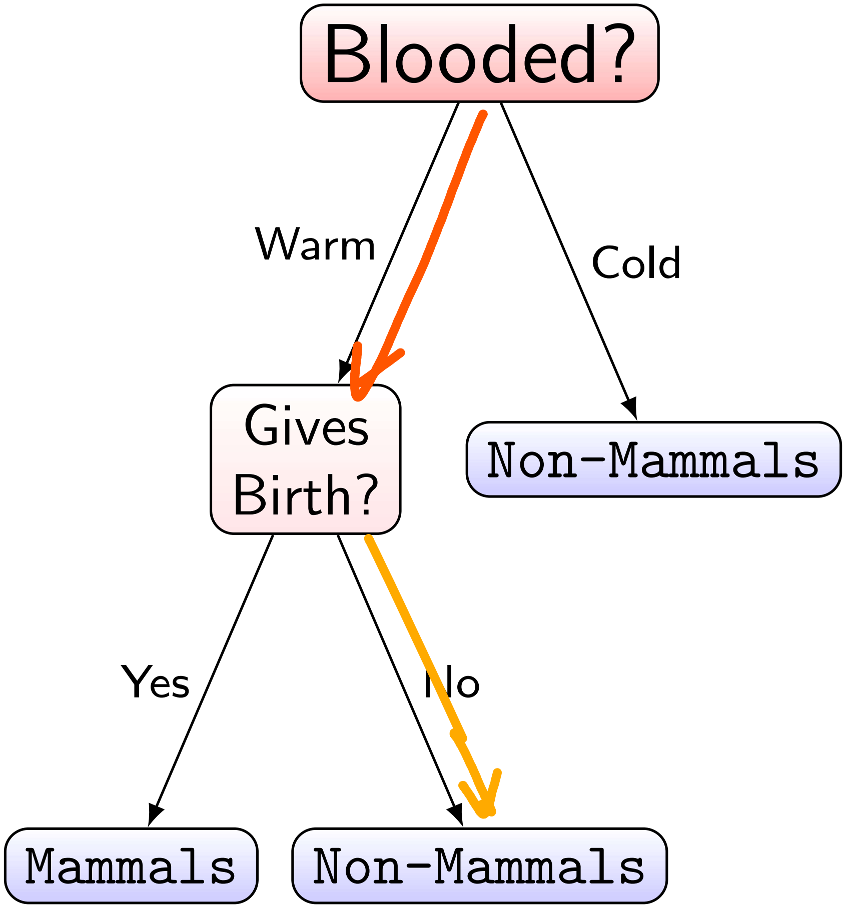


- XKCD # 518: Flow Charts. At 8 drinks, you switch the torrent from freeBSD to Microsoft Bob. C'mon, it'll be fun!
- Your professor is not endorsing underage drinking

What is a decision tree?



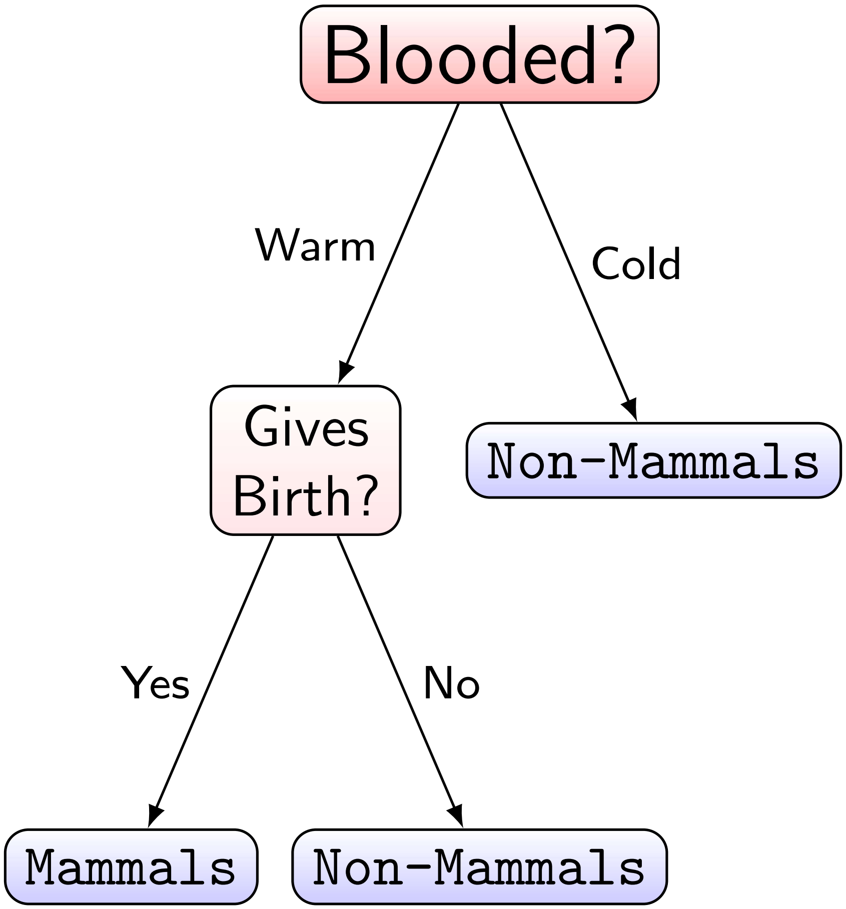
What is a decision tree?



Is a Flamingo a mammal or not?

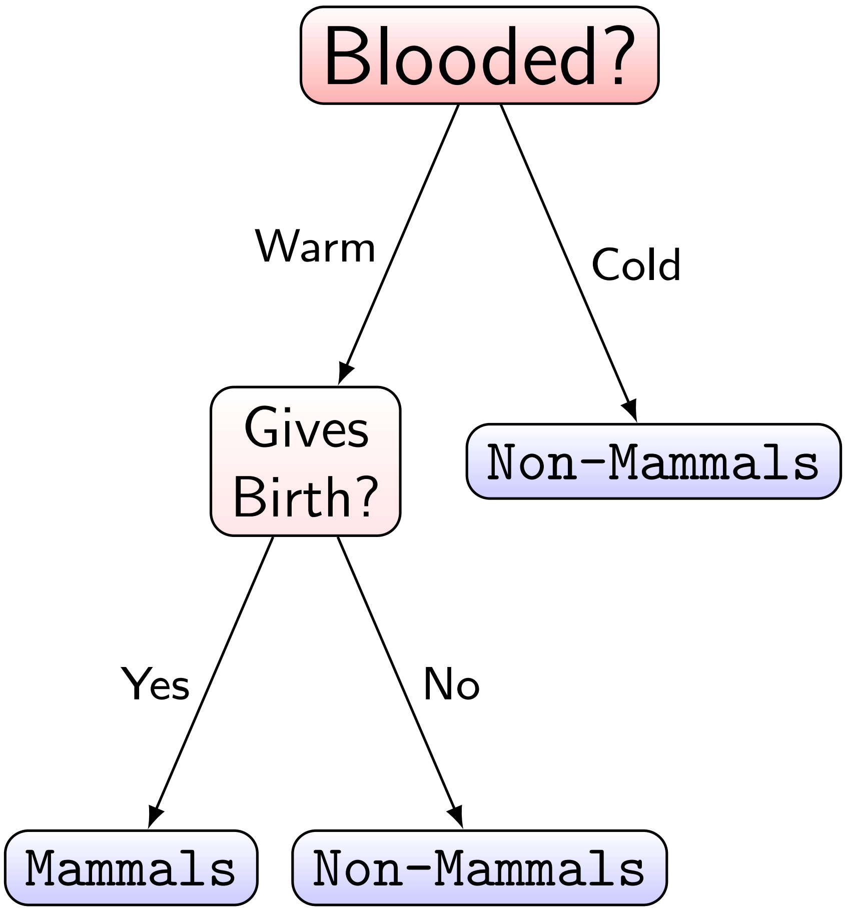
It is warm-blooded and does not give birth.

What is a decision tree?



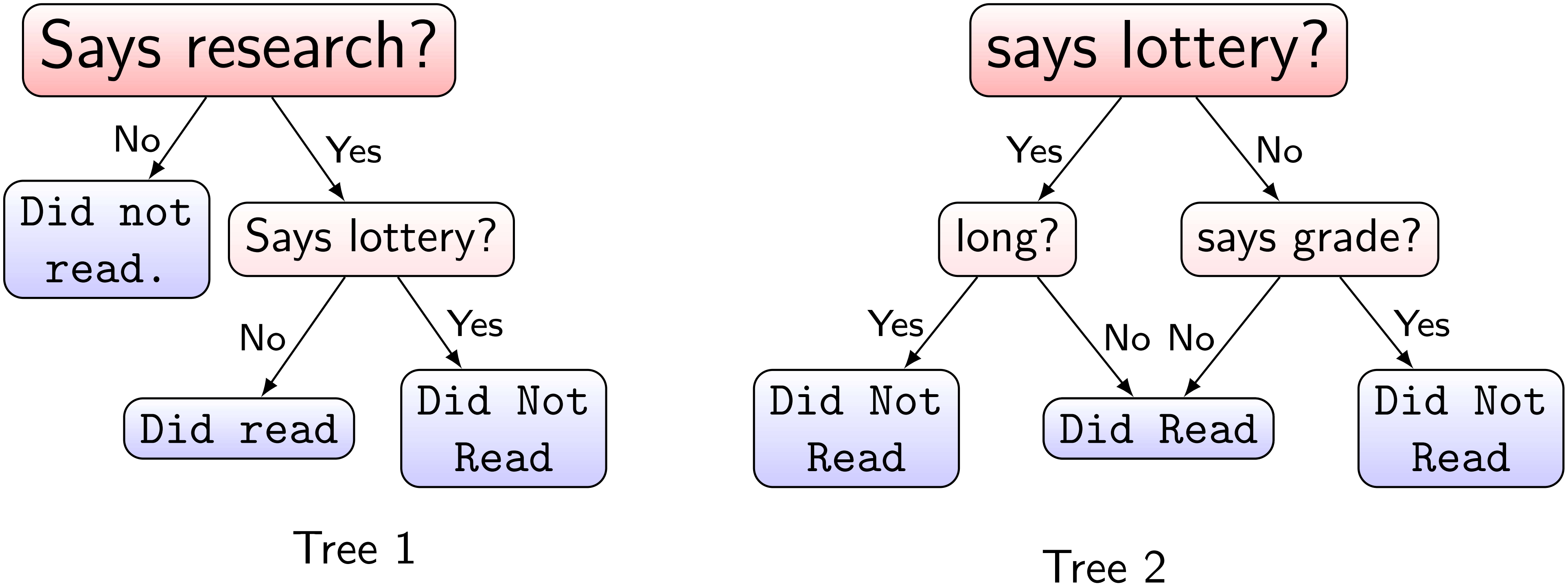
A decision tree represents a disjunction of conjunctions.
What does this represent for *Mammals?*

What is a decision tree?



A decision tree represents a disjunction of conjunctions.
What does this represent for *Non-Mammals?*

More Trees to Explain



- For each tree, write the description for both yes and no.

What problems are good for decision trees?

- Attributes have discrete values
- Output is discrete
- Disjunctive descriptions
- Training data may contain errors
- Training data may contain missing attribute values

The Basic Idea

- Look at all the data.
  - If it is one category, we are done.
  - If it is not, choose an attribute to split on.
  - Split on that attribute and construct a tree for each.

A Dataset

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Temp	Y	N
H	2	2
M	4	2
C	3	1

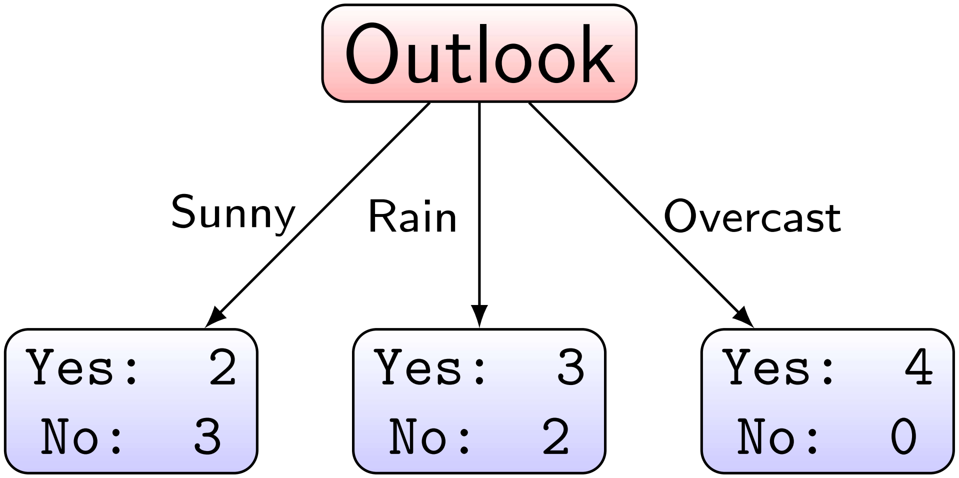
Informally, which attribute *looks good* as a root?

Deciding on a Root

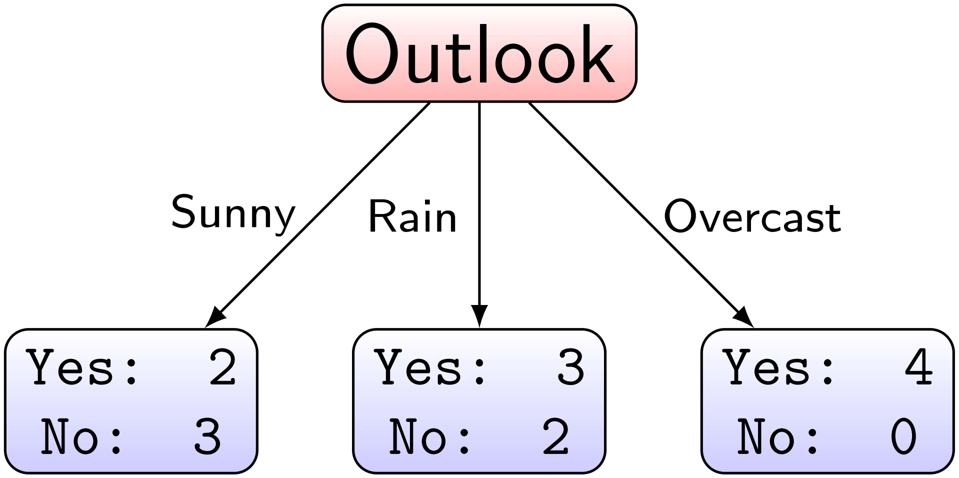
For each attribute, what is the yes/no split?

Attribute	# Yes	# No
Outlook = Sunny	2	3
Outlook = Overcast	4	0
Outlook = Rain	3	2
Temperature = Hot	2	2
Temperature = Mild	4	2
Temperature = Cool	3	1
Humidity = High	3	4
Humidity = Normal	6	1
Wind = Weak	6	2
Wind = Strong	3	3

Splitting on "Outlook"



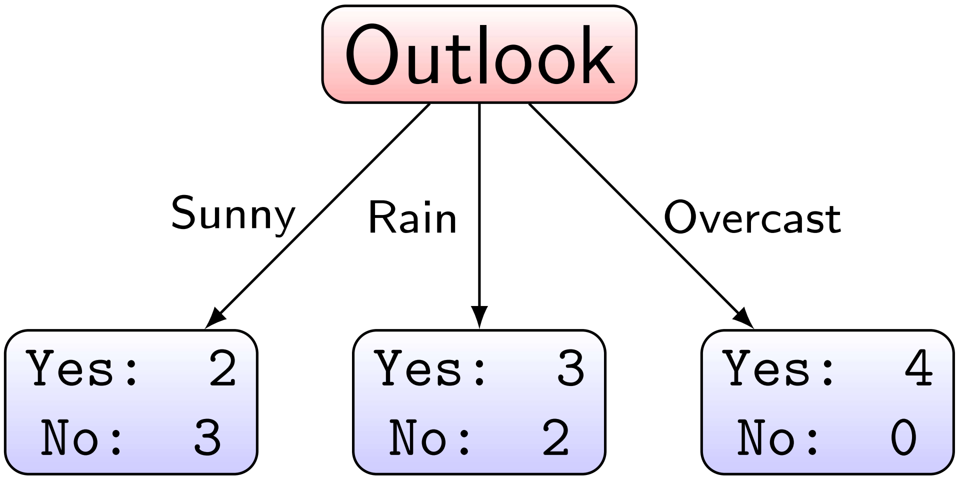
Splitting on “Outlook”



We now need to split the five examples with a Sunny outlook:

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

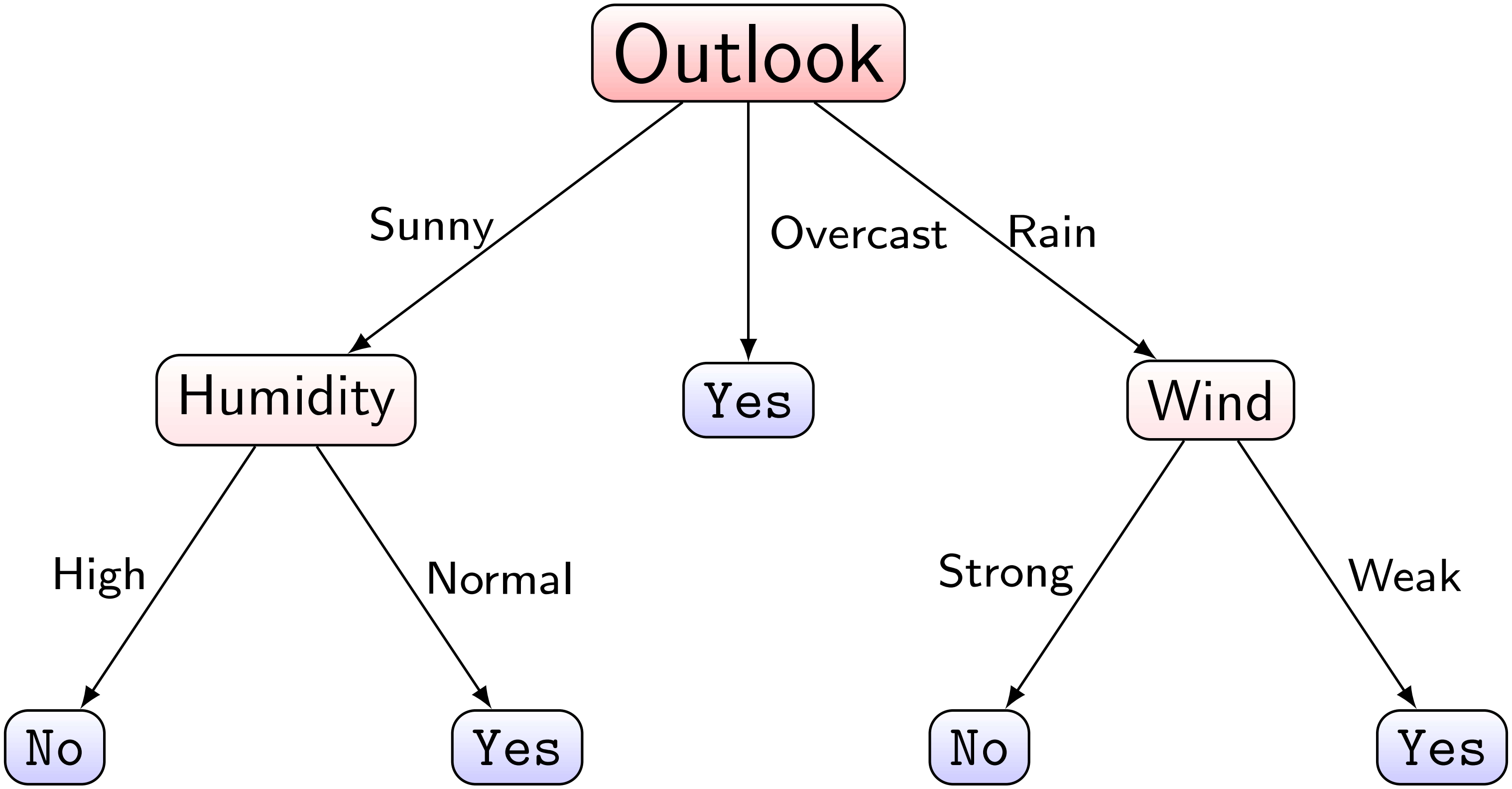
Splitting on "Outlook"



We now need to split the five examples with a Rainy outlook:

Outlook	Temperature	Humidity	Wind	PlayTennis
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Rain	Mild	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Resulting Tree



But a computer can't "eyeball" a selection

Let's look at the definition of *entropy*; you can think of it as a measure of uncertainty.

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

where p_i is the proportion of S belonging to class i.

Example: suppose S has 14 examples, 9 positive and 5 negative. Calculate the entropy

$$\frac{-9}{14} \cdot \log_2 \frac{9}{14} + \frac{-5}{14} \log_2 \frac{5}{14} \approx .940$$

But a computer can't "eyeball" a selection

Let's look at the definition of *entropy*; you can think of it as a measure of uncertainty.

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

where p_i is the proportion of S belonging to class i.

What should entropy be if everything is one category?

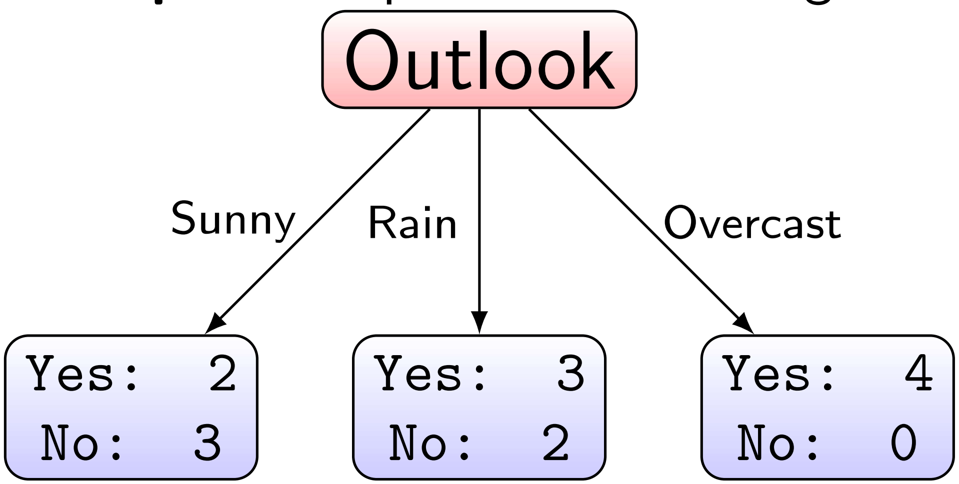
$$-\log 1 + \underbrace{-0 \log 0}_{\text{treat as } 0}$$

How to select based on entropy?

The **information gain** for selecting attribute A to split set S:

$$\text{Gain}(S, A) = \underline{\text{Entropy}(S)} - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Example: Compute information gain from Outlook:



also $E(S_{Rain})$

$$E(S_{sunny}) = \frac{-2}{5}\log_2(2/5) + \frac{-3}{5}\log_2(3/5) \approx .970$$

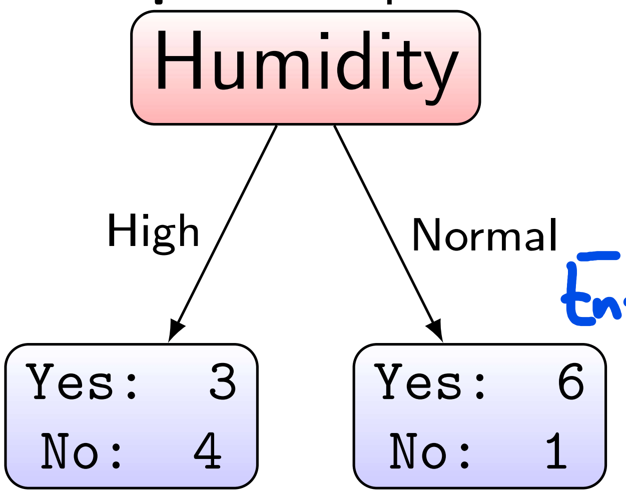
$$\text{Info Gain} = .940 - \frac{5}{14}(.970) - \frac{5}{14}(.970) - \frac{4}{14}(0) \approx .246$$

How to select based on entropy?

The **information gain** for selecting attribute A to split set S:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Example: Compute information gain from Humidity:



$$\text{Gain}: \ .940 - \frac{7}{14}(.985) - \frac{7}{14}(.592)$$

$$\text{Entropy}(S_{High}) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} \approx .985$$

$$\text{Entropy}(S_{normal}) = -\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7} \approx .592$$

Why did we select Outlook?

Information Gain for root choices:

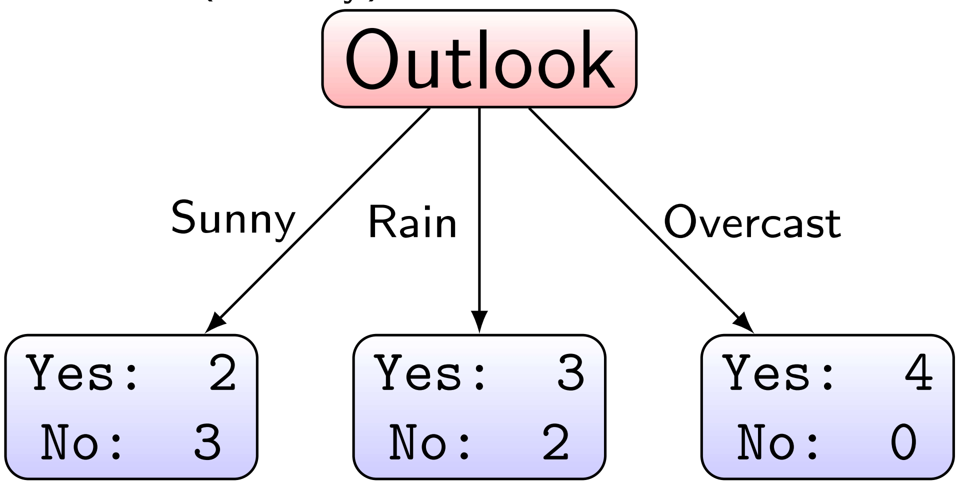
Attribute	Information Gain
Outlook	.246
Humidity	.151
Wind	.048
Temperature	.029

You can calculate the remaining ones if you want to.

Finishing the tree

What is the information gain for splitting Sunny Outlook category on attribute Humidity?

Entropy(S_{Sunny}) $= .970$



Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

Sunburn Dataset

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

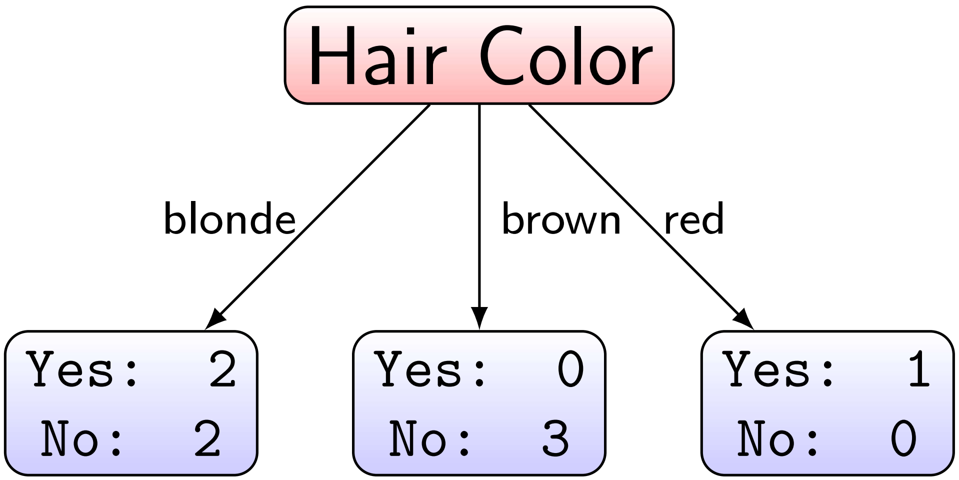
$$\text{Entropy}(S) = \frac{-3}{8}\log\left(3/8\right) + \frac{-5}{8}\log 5/8 .$$

$$\approx .954$$

Sunburn data gather

Attribute	# Yes	# No
hair=blonde	2	2
hair=brown	0	3
hair=red	1	0
height = short	1	2
height = tall	0	2
height = average	2	1
weight = light	1	1
weight = average	1	2
weight = heavy	1	2
lotion = no	3	2
lotion = yes	0	3

Hair color as the root



Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
blonde	short	average	no	sunburned
blonde	short	light	yes	none

- Which attribute to split?

Wrapping up ID3

- What hypothesis space does ID3 search?

- How many hypotheses does ID3 maintain?
  Can we use it to get a set of consistent hypotheses?

Contrasting ID3 to Candidate-Elimination

- CANDIDATE-ELIMINATION does not search the complete hypothesis space
- ID3 searches a *complete space*:
  all finite discrete-valued functions are possible to find.

Contrasting ID3 to Candidate-Elimination

- What types of trees does ID3 produce?

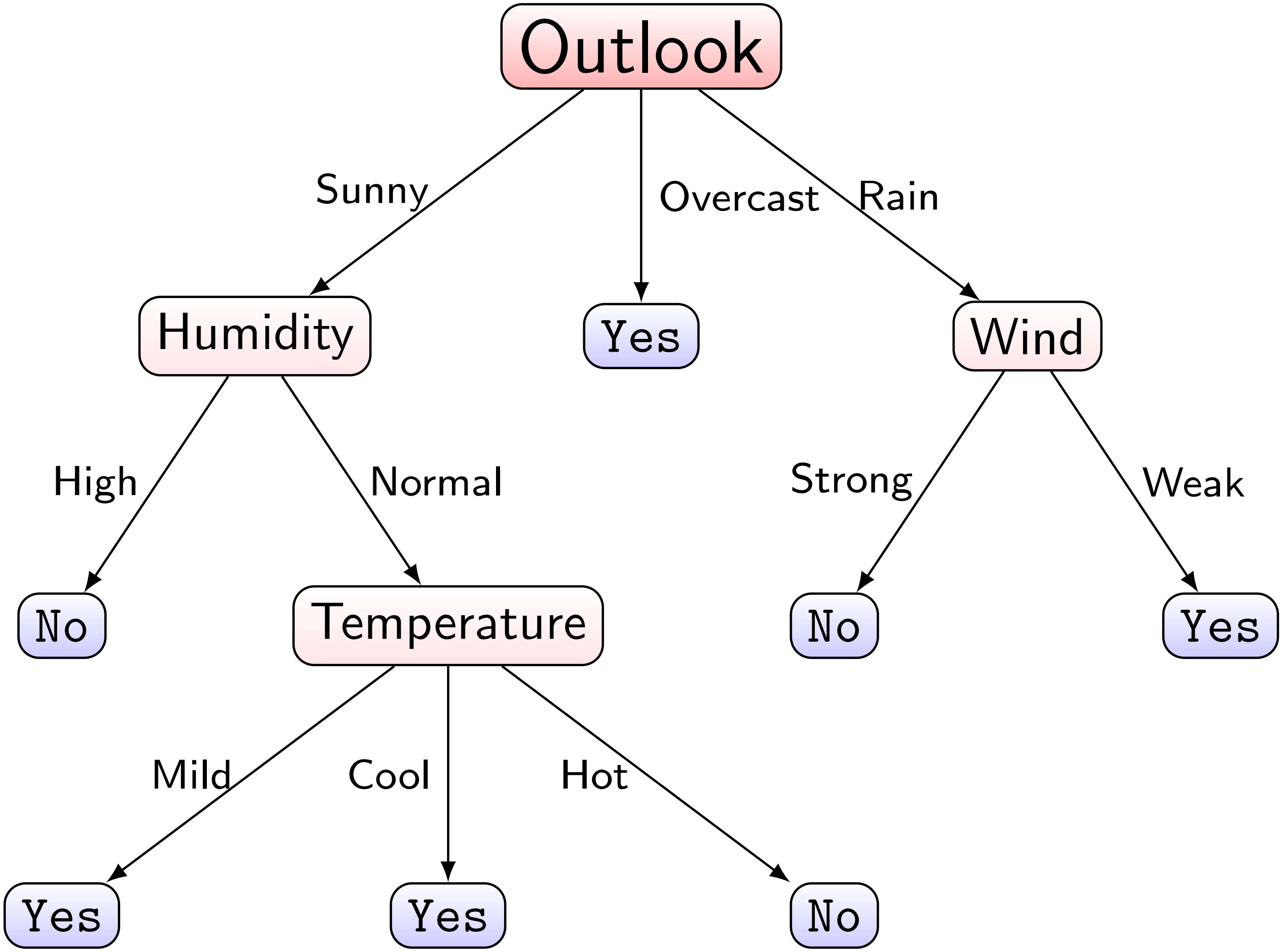
- How do the preferences of ID3 and CANDIDATE-ELIMINATION differ?

Overfitting

- We saw ID3 can achieve 100% training accuracy.
  Is this always desirable?

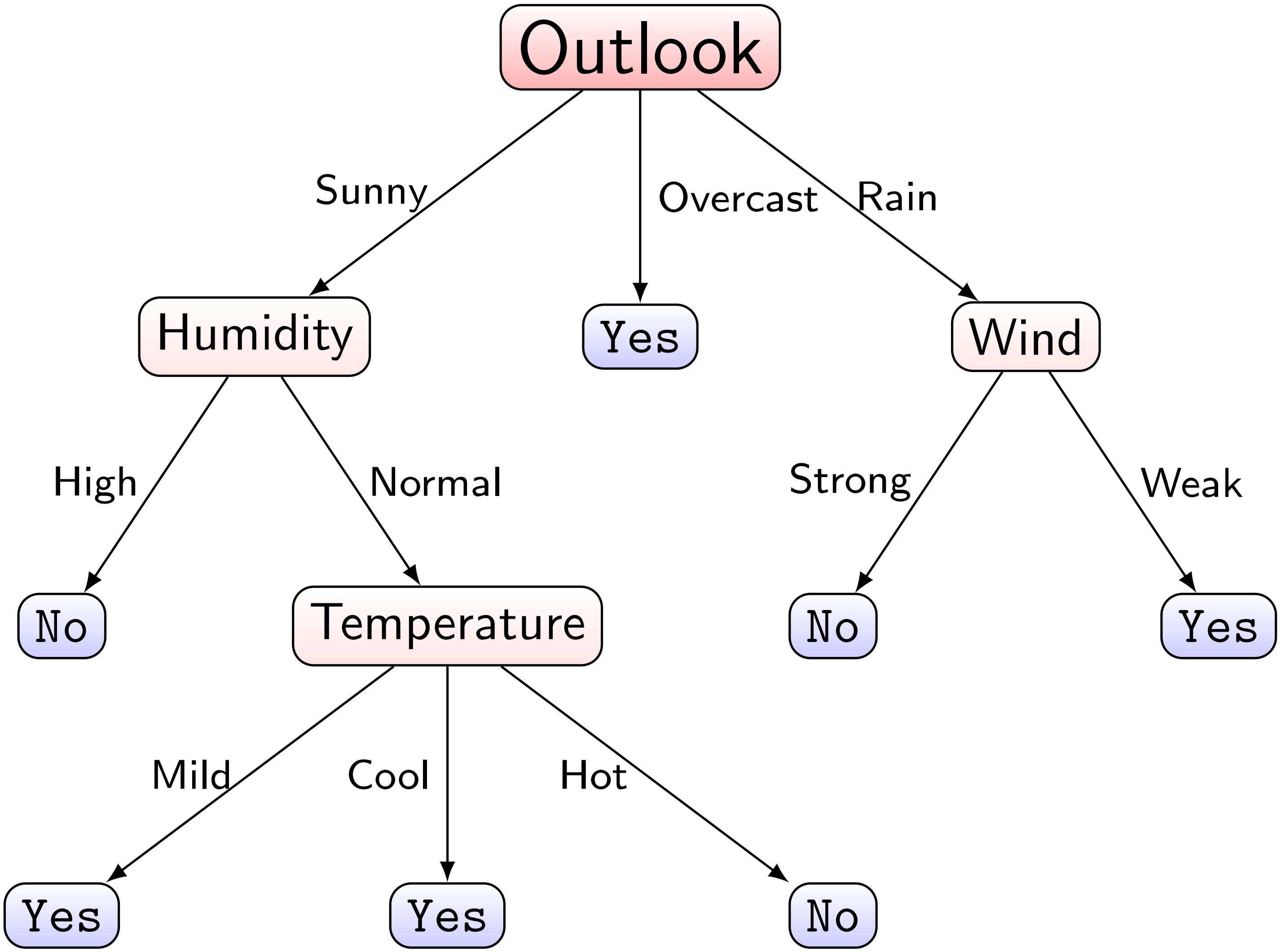
- What should cause us to suspect overfitting in a decision tree?

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	Normal	Strong	No



We saw that larger trees are suspected of overfitting.
What can we change about ID3 to combat this?

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	Normal	Strong	No



How can we decide branches as candidates for pruning?

(end 7/26: will finish tmw)