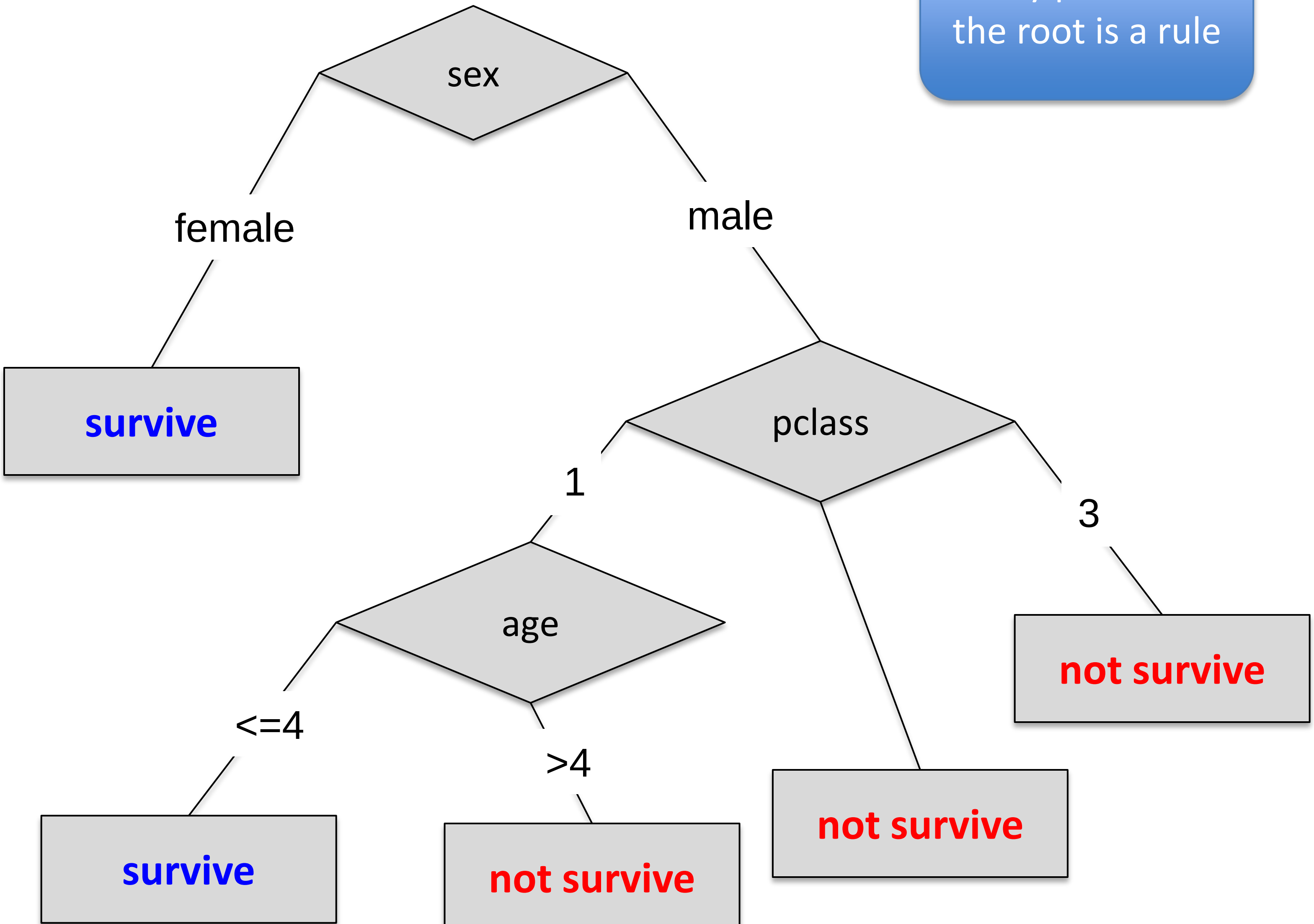
How far can we go?

We might consider grouping redundant conditions

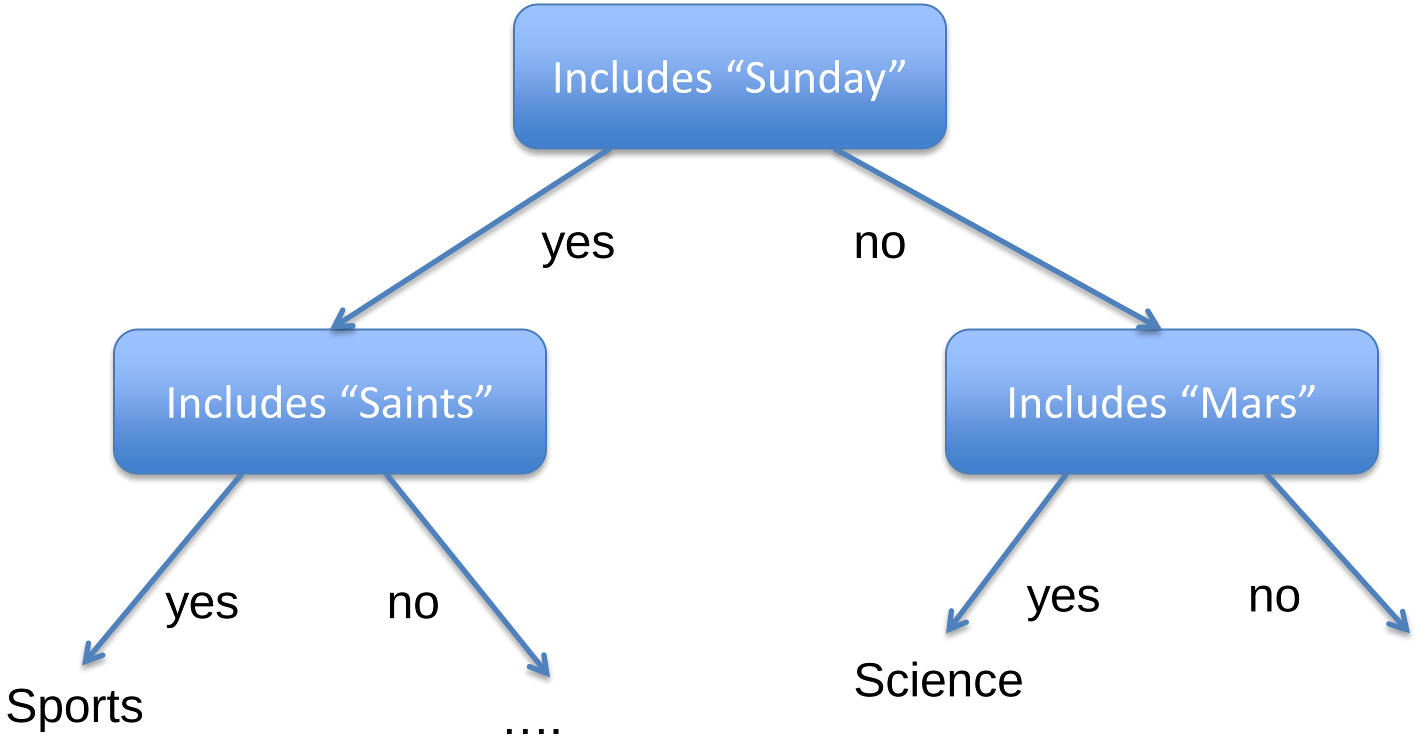
```
IF pclass='1' THEN
    IF sex='female' THEN survive=yes
    IF sex='male' AND age < 5 THEN survive=yes
IF pclass='2'
    IF sex='female' THEN survive=yes
    IF sex='male' THEN survive=no
IF pclass='3'
    IF sex='male' THEN survive=no
    IF sex='female'
        IF age < 4 THEN survive=yes
        IF age >= 4 THEN survive=no
```

A decision tree

Every path from the root is a rule



Document Classification Example



Aside on Entropy

Consider two sequences of coin flips

HHTHTTHHHHTTHTHTHTTTT….
TTHHTTHTHTTTTHHHTHTTT….

How much information do we get after flipping each coin once?

We want some function "Information" that satisfies:

$\text{Information}_{1\&2}(p_1 p_2) = \text{Information}_1(p_1) + \text{Information}_2(p_2)$

$$I(X) = \log_2 p_x$$

Expected Information = "Entropy"

$$H(X) = E(I(X)) = \sum_x p_x I(x) = -\sum_x p_x \log_2 p_x$$

Example: Flipping a Coin

$$\text{Entropy} = -\sum_i p_x \log_2 p_x$$

$$= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5)$$

$$= 1$$

Example: Rolling a die

$$p_1 = \frac{1}{6},\ p_2 = \frac{1}{6},\ p_3 = \frac{1}{6},\ \ldots$$

$$\begin{aligned} \text{Entropy} &= -\sum_i p_i \log_2 p_i \\ &= -6 \times \left(\frac{1}{6} \log_2 \frac{1}{6}\right) \\ &\approx 2.58 \end{aligned}$$

Example: Rolling a weighted die

$$p_1 = 0.1,\ p_2 = 0.1,\ p_3 = 0.1,\ \ldots\ p_6 = 0.5$$

$$\text{Entropy} = -\sum_i p_x \log_2 p_x$$

$$= -5 \times (0.1 \log_2 0.1) - 0.5 \log_2 0.5$$

$$= 2.16$$

The weighted die is more unpredictable than a fair die

How unpredictable is your data?

- 342/891 survivors in titanic training set

$$-\left(\frac{342}{891}\log_2\frac{342}{891}+\frac{549}{891}\log_2\frac{549}{891}\right)=0.96$$

- Say there were only 50 survivors

$$-\left(\frac{50}{891}\log_2\frac{50}{891}+\frac{841}{891}\log_2\frac{841}{891}\right)=0.31$$

Back to decision trees

- Which attribute do we choose at each level?
- The one with the highest information gain
  - The one that reduces the unpredictability the most