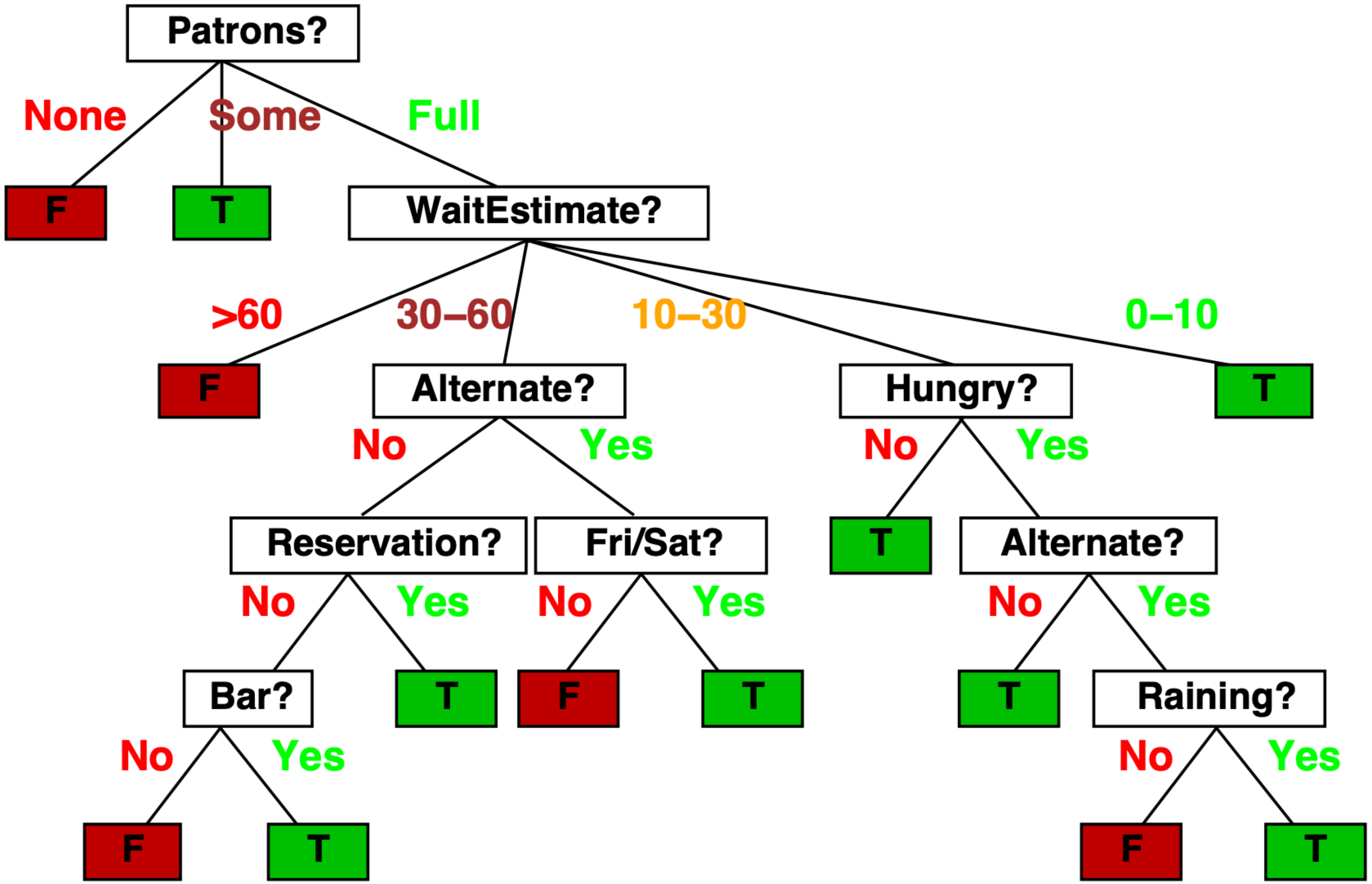
# Classification Tree

# Classification vs Regression

| Example | Attributes | | | | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Alt* | *Bar* | *Fri* | *Hun* | *Pat* | *Price* | *Rain* | *Res* | *Type* | *Est* | *WillWait* |
| $X_1$ | *T* | *F* | *F* | *T* | *Some* | *$$$* | *F* | *T* | *French* | *0–10* | *T* |
| $X_2$ | *T* | *F* | *F* | *T* | *Full* | *$* | *F* | *F* | *Thai* | *30–60* | *F* |
| $X_3$ | *F* | *T* | *F* | *F* | *Some* | *$* | *F* | *F* | *Burger* | *0–10* | *T* |
| $X_4$ | *T* | *F* | *T* | *T* | *Full* | *$* | *F* | *F* | *Thai* | *10–30* | *T* |
| $X_5$ | *T* | *F* | *T* | *F* | *Full* | *$$$* | *F* | *T* | *French* | *>60* | *F* |
| $X_6$ | *F* | *T* | *F* | *T* | *Some* | *$$* | *T* | *T* | *Italian* | *0–10* | *T* |
| $X_7$ | *F* | *T* | *F* | *F* | *None* | *$* | *T* | *F* | *Burger* | *0–10* | *F* |
| $X_8$ | *F* | *F* | *F* | *T* | *Some* | *$$* | *T* | *T* | *Thai* | *0–10* | *T* |
| $X_9$ | *F* | *T* | *T* | *F* | *Full* | *$* | *T* | *F* | *Burger* | *>60* | *F* |
| $X_{10}$ | *T* | *T* | *T* | *T* | *Full* | *$$$* | *F* | *T* | *Italian* | *10–30* | *F* |
| $X_{11}$ | *F* | *F* | *F* | *F* | *None* | *$* | *F* | *F* | *Thai* | *0–10* | *F* |
| $X_{12}$ | *T* | *T* | *T* | *T* | *Full* | *$* | *F* | *F* | *Burger* | *30–60* | *T* |

| Dosage | Age | Sex | Etc. | Drug Effect. |
|---|---|---|---|---|
| 10 | 25 | Female | ... | 98 |
| 20 | 73 | Male | ... | 0 |
| 35 | 54 | Female | ... | 100 |
| 5 | 12 | Male | ... | 44 |
| etc... | etc... | etc... | etc... | etc... |

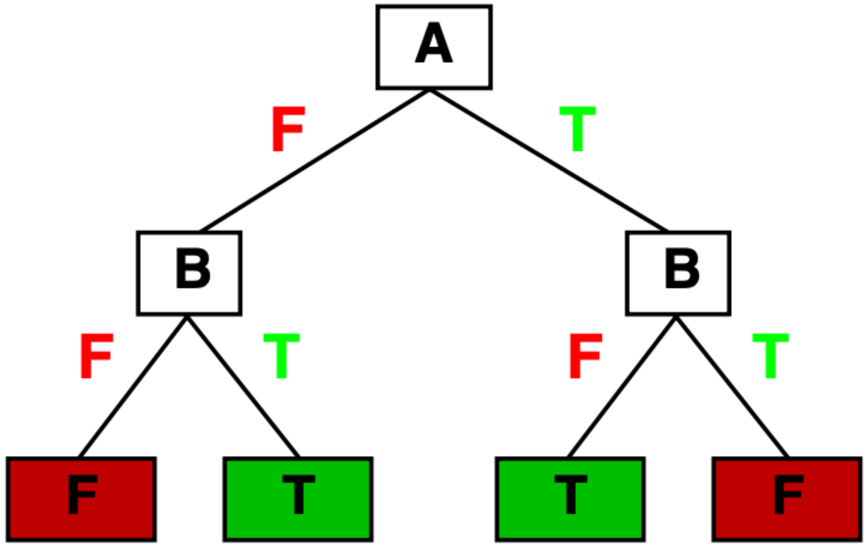
# Decision Tree



# Expressiveness

Decision trees can express any function of the input attributes.
E.g., for Boolean functions, truth table row $\longrightarrow$ path to leaf:

| A | B | A xor B |
|---|---|---|
| F | F | F |
| F | T | T |
| T | F | T |
| T | T | F |

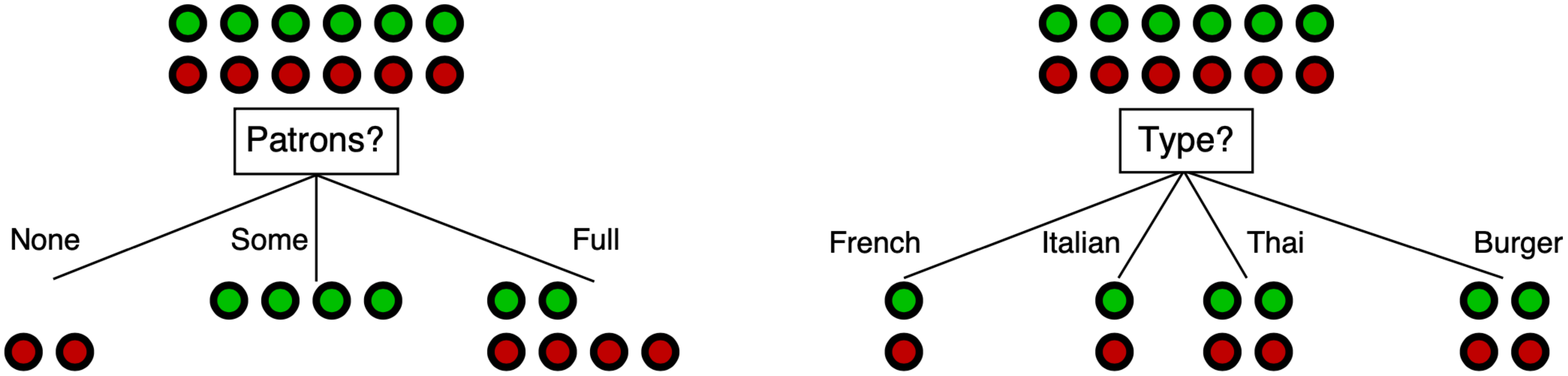


# Hypothesis Space

- The number of possible decision trees with $n$ attributes is equal to the number of possible Boolean functions with $n$ attibutes
- Given n attributes we have a truth table with $2^n$ rows
- Given a truth table we have $2^{2^n}$ possible Boolean functions

# Selecting a feature

Idea: a good attribute splits the examples into subsets that are (ideally) “all positive” or “all negative”




*Patrons?* is a better choice—gives **information** about the classification

how can we measure the discriminative power of a feature

# Entropy = Informativeness

description of entropy in information theory

Given n events we need $\log_2 n$ bits to represent them

$$\log_2 n = -\log_2 \frac{1}{n}$$

$-\log_2 \frac{1}{n}$ is the information needed to represent the events

$$-\log_2 \frac{1}{n} = -\sum_{(n\ \text{events})} \frac{1}{n} \cdot \log_2 \frac{1}{n}$$

0.5*something+0.5*something=something

$\frac{1}{n}$ is the probability of the event in case we have a uniform distribution

# Entropy = Informativeness

In the most general case, with a generic probability $p_i$ for each event $i$

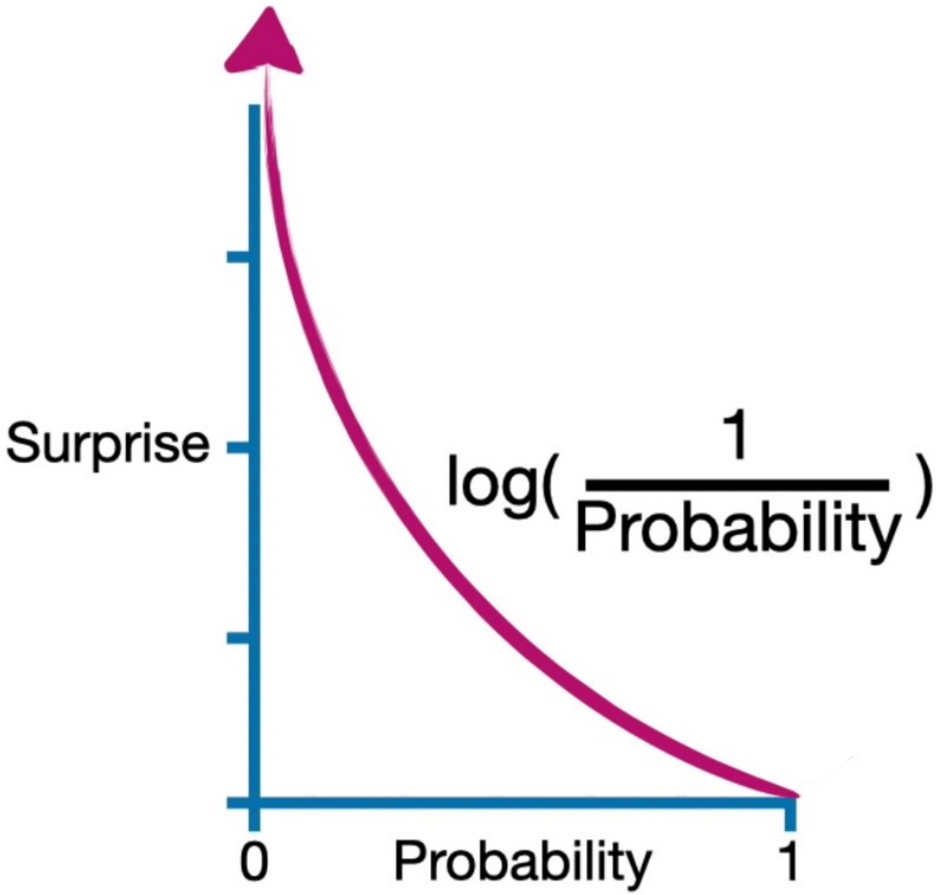
$$-\sum_{(n\ \text{events})} \frac{1}{n} \cdot \log_2 \frac{1}{n} \curvearrowright -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

move to uniformly distributed to the general case

the information needed to represent a set of stochastic prosess

entropy is a wat to measure the informativeness

# Entropy and Surprise




Entropy = Expected Surprise

# Information Gain (1/2)

compute the entropy associated to the  traget feature

$$H\left(\frac{t}{t+f}, \frac{f}{t+f}\right) = -\frac{t}{t+f} \cdot \log_2 \frac{t}{t+f} - \frac{f}{t+f} \cdot \log_2 \frac{f}{t+f}$$

how many times i have a true value over all the value, probability of true

probability of false

how informative is this feature i

| Example | Attributes | | | | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Alt* | *Bar* | *Fri* | *Hun* | *Pat* | *Price* | *Rain* | *Res* | *Type* | *Est* | *WillWait* |
| $X_1$ | T | F | F | T | Some | $$$ | F | T | French | 0–10 | T |
| $X_2$ | T | F | F | T | Full | $ | F | F | Thai | 30–60 | F |
| $X_3$ | F | T | F | F | Some | $ | F | F | Burger | 0–10 | T |
| $X_4$ | T | F | T | T | Full | $ | F | F | Thai | 10–30 | T |
| $X_5$ | T | F | T | F | Full | $$$ | F | T | French | >60 | F |
| $X_6$ | F | T | F | T | Some | $$ | T | T | Italian | 0–10 | T |
| $X_7$ | F | T | F | F | None | $ | T | F | Burger | 0–10 | F |
| $X_8$ | F | F | F | T | Some | $$ | T | T | Thai | 0–10 | T |
| $X_9$ | F | T | T | F | Full | $ | T | F | Burger | >60 | F |
| $X_{10}$ | T | T | T | T | Full | $$$ | F | T | Italian | 10–30 | F |
| $X_{11}$ | F | F | F | F | None | $ | F | F | Thai | 0–10 | F |
| $X_{12}$ | T | T | T | T | Full | $ | F | F | Burger | 30–60 | T |

# Information Gain (2/2)

$$R(Some) = \frac{t_i + f_i}{t + f} \cdot I\left(\frac{t_i}{t_i + f_i}, \frac{f_i}{t_i + f_i}\right)$$

$$R(Full) = \frac{t_i + f_i}{t + f} \cdot I\left(\frac{t_i}{t_i + f_i}, \frac{f_i}{t_i + f_i}\right)$$

$$R(None) = \frac{t_i + f_i}{t + f} \cdot I\left(\frac{t_i}{t_i + f_i}, \frac{f_i}{t_i + f_i}\right)$$

$$R(A) = \sum_{i=1}^{v} \frac{t_i + f_i}{t + f} \cdot I\left(\frac{t_i}{t_i + f_i}, \frac{f_i}{t_i + f_i}\right)$$

$$IG(A) = H\left(\frac{t}{t + f}, \frac{f}{t + f}\right) - \mathrm{R(A)}$$

| Example | Attributes | | | | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Alt$ | $Bar$ | $Fri$ | $Hun$ | $Pat$ | $Price$ | $Rain$ | $Res$ | $Type$ | $Est$ | *WillWait* |
| $X_1$ | T | F | F | T | Some | $$$ | F | T | French | 0–10 | T |
| $X_2$ | T | F | F | T | Full | $ | F | F | Thai | 30–60 | F |
| $X_3$ | F | T | F | F | Some | $ | F | F | Burger | 0–10 | T |
| $X_4$ | T | F | T | T | Full | $ | F | F | Thai | 10–30 | T |
| $X_5$ | T | F | T | F | Full | $$$ | F | T | French | >60 | F |
| $X_6$ | F | T | F | T | Some | $$ | T | T | Italian | 0–10 | T |
| $X_7$ | F | T | F | F | None | $ | T | F | Burger | 0–10 | F |
| $X_8$ | F | F | F | T | Some | $$ | T | T | Thai | 0–10 | T |
| $X_9$ | F | T | T | F | Full | $ | T | F | Burger | >60 | F |
| $X_{10}$ | T | T | T | T | Full | $$$ | F | T | Italian | 10–30 | F |
| $X_{11}$ | F | F | F | F | None | $ | F | F | Thai | 0–10 | F |
| $X_{12}$ | T | T | T | T | Full | $ | F | F | Burger | 30–60 | T |

# Building a Classification Tree

$$H\left(\frac{t}{t+f},\frac{f}{t+f}\right) = -\frac{6}{12}\cdot\log_2\frac{6}{12} - \frac{6}{12}\cdot\log_2\frac{6}{12} = 1$$

$$IG(Patrons) = 1 - \left[\frac{2}{12}\cdot H(0,1) + \frac{4}{12}\cdot H(1,0) + \frac{6}{12}\cdot H\left(\frac{2}{6},\frac{4}{6}\right)\right] \approx 0{,}541$$

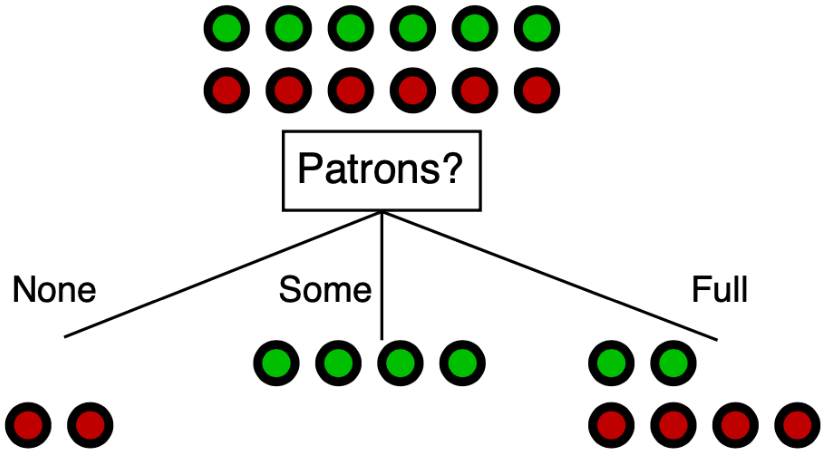
$$IG(Type) = 1 - \left[\frac{2}{12}\cdot H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{2}{12}\cdot H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{4}{12}\cdot H\left(\frac{2}{4},\frac{2}{4}\right) + \frac{4}{12}\cdot H\left(\frac{2}{4},\frac{2}{4}\right)\right] = 0$$

Select the attribute with the highest gain

in this case IG(Patrons) become the root node of our classification tree

| Example | Attributes | | | | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Alt* | *Bar* | *Fri* | *Hun* | *Pat* | *Price* | *Rain* | *Res* | *Type* | *Est* | *WillWait* |
| $X_1$ | T | F | F | T | Some | $$$ | F | T | French | 0–10 | T |
| $X_2$ | T | F | F | T | Full | $ | F | F | Thai | 30–60 | F |
| $X_3$ | F | T | F | F | Some | $ | F | F | Burger | 0–10 | T |
| $X_4$ | T | F | T | T | Full | $ | F | F | Thai | 10–30 | T |
| $X_5$ | T | F | T | F | Full | $$$ | F | T | French | >60 | F |
| $X_6$ | F | T | F | T | Some | $$ | T | T | Italian | 0–10 | T |
| $X_7$ | F | T | F | F | None | $ | T | F | Burger | 0–10 | F |
| $X_8$ | F | F | F | T | Some | $$ | T | T | Thai | 0–10 | T |
| $X_9$ | F | T | T | F | Full | $ | T | F | Burger | >60 | F |
| $X_{10}$ | T | T | T | T | Full | $$$ | F | T | Italian | 10–30 | F |
| $X_{11}$ | F | F | F | F | None | $ | F | F | Thai | 0–10 | F |
| $X_{12}$ | T | T | T | T | Full | $ | F | F | Burger | 30–60 | T |

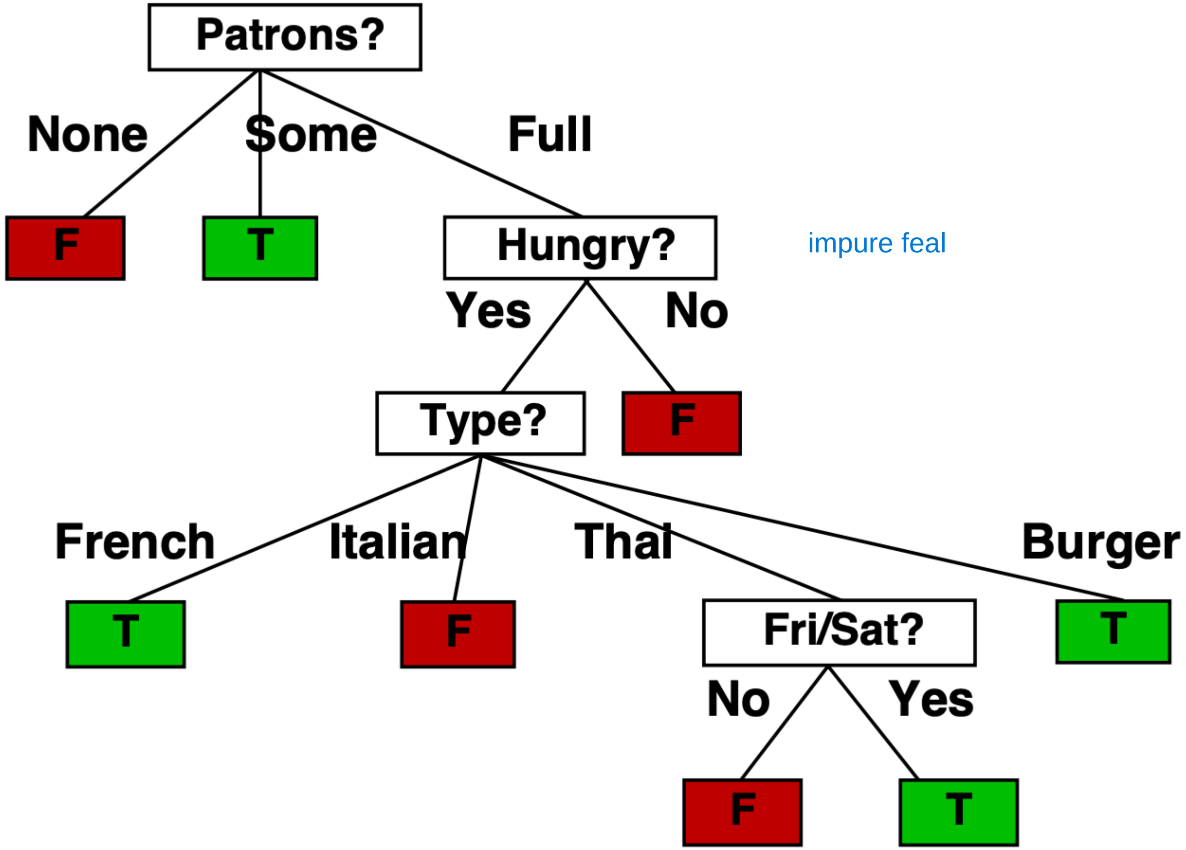
# Building a Classification Tree




| Example | Attributes | | | | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Alt* | *Bar* | *Fri* | *Hun* | *Pat* | *Price* | *Rain* | *Res* | *Type* | *Est* | *WillWait* |
| $X_1$ | ~~T~~ | ~~F~~ | ~~F~~ | ~~T~~ | ~~Some~~ | ~~$$$~~ | ~~F~~ | ~~T~~ | ~~French~~ | ~~0–10~~ | ~~T~~ |
| $X_2$ | T | F | F | T | Full | $ | F | F | Thai | 30–60 | F |
| $X_3$ | ~~F~~ | ~~T~~ | ~~F~~ | ~~F~~ | ~~Some~~ | ~~$~~ | ~~F~~ | ~~F~~ | ~~Burger~~ | ~~0–10~~ | ~~T~~ |
| $X_4$ | T | F | T | T | Full | $ | F | F | Thai | 10–30 | T |
| $X_5$ | T | F | T | F | Full | $$$ | F | T | French | >60 | F |
| $X_6$ | ~~F~~ | ~~T~~ | ~~F~~ | ~~T~~ | ~~Some~~ | ~~$$~~ | ~~T~~ | ~~T~~ | ~~Italian~~ | ~~0–10~~ | ~~T~~ |
| $X_7$ | ~~F~~ | ~~T~~ | ~~F~~ | ~~F~~ | ~~None~~ | ~~$~~ | ~~T~~ | ~~F~~ | ~~Burger~~ | ~~0–10~~ | ~~F~~ |
| $X_8$ | ~~F~~ | ~~F~~ | ~~F~~ | ~~T~~ | ~~Some~~ | ~~$$~~ | ~~T~~ | ~~T~~ | ~~Thai~~ | ~~0–10~~ | ~~T~~ |
| $X_9$ | F | T | T | F | Full | $ | T | F | Burger | >60 | F |
| $X_{10}$ | T | T | T | T | Full | $$$ | F | T | Italian | 10–30 | F |
| $X_{11}$ | ~~F~~ | ~~F~~ | ~~F~~ | ~~F~~ | ~~None~~ | ~~$~~ | ~~F~~ | ~~F~~ | ~~Thai~~ | ~~0–10~~ | ~~F~~ |
| $X_{12}$ | T | T | T | T | Full | $ | F | F | Burger | 30–60 | T |

# Building a Classification Tree

very common in exams




impure feal