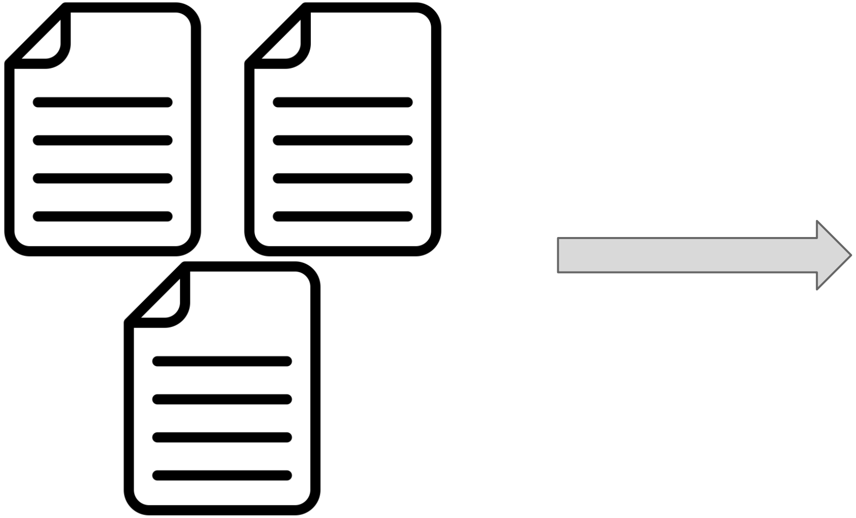
Decision Trees and Random Forests

Esteban Villalobos

A bit on Machine Learning...



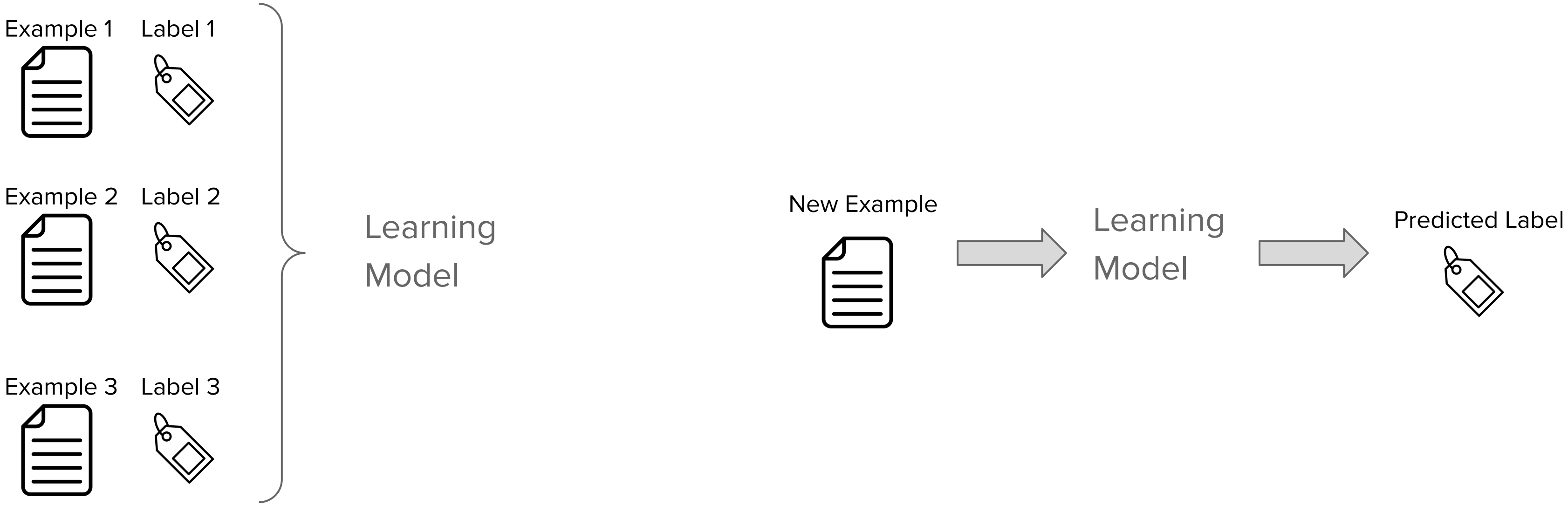
Predictions

Created by Vectorstall
from Noun Project

A bit on Machine Learning...

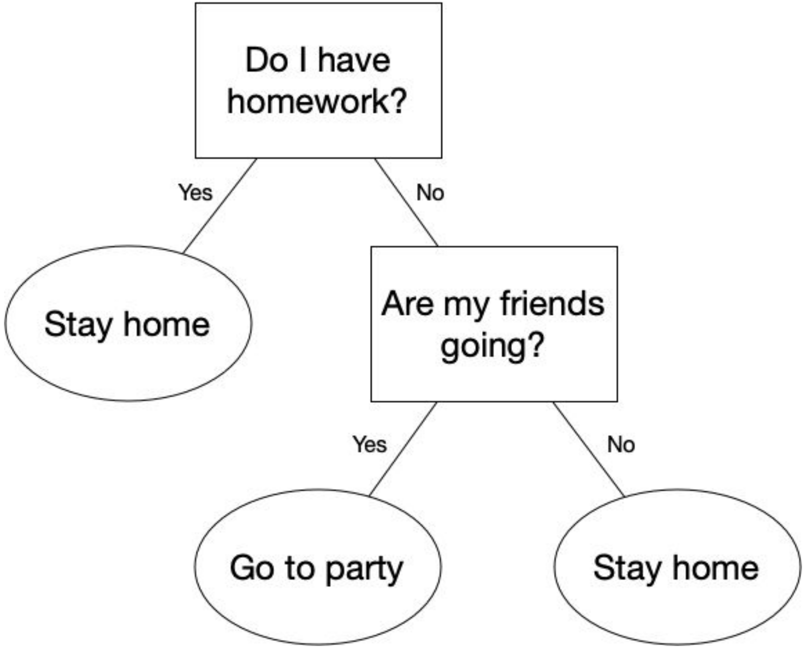
Training:

Testing:



Created by Vectorstall
from Noun Project

What are Decision Trees?



What are Decision Trees?

"A.I. is just a bunch of 'if' statements"

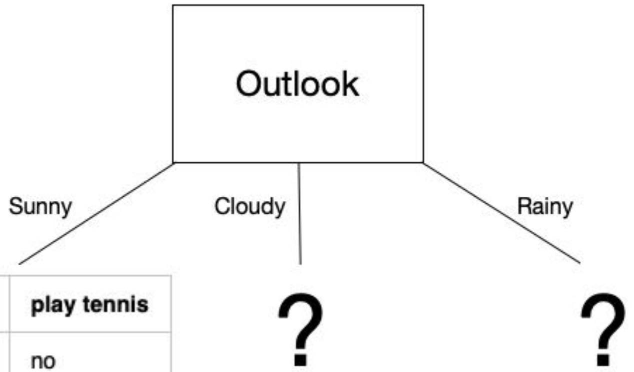
Decision Tree Learning

	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
3	cloudy	high	high	weak	yes
4	rainy	medium	high	weak	yes
5	rainy	low	normal	weak	yes
6	rainy	low	normal	strong	no
7	cloudy	low	normal	strong	yes
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
10	rainy	medium	normal	weak	yes
11	sunny	medium	normal	strong	yes
12	cloudy	medium	high	strong	yes
13	cloudy	high	normal	weak	yes
14	rainy	medium	high	strong	no

Decision Tree Learning

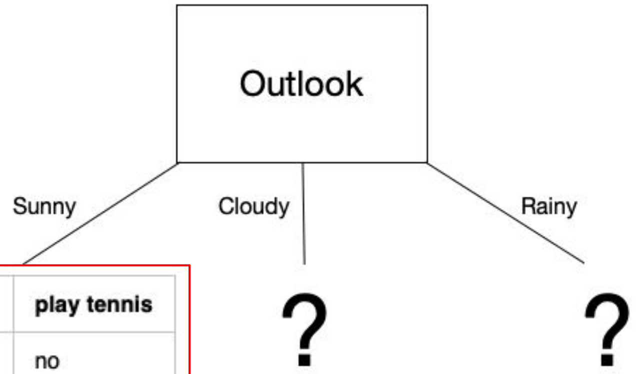
How do we make the problem easier?

Decision Tree Learning



	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
11	sunny	medium	normal	strong	yes

Decision Tree Learning



	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
11	sunny	medium	normal	strong	yes

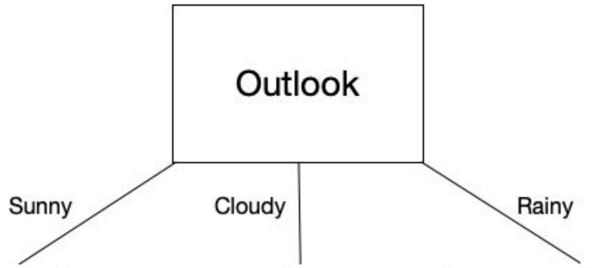
Decision Tree Learning

	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
3	cloudy	high	high	weak	yes
4	rainy	medium	high	weak	yes
5	rainy	low	normal	weak	yes
6	rainy	low	normal	strong	no
7	cloudy	low	normal	strong	yes
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
10	rainy	medium	normal	weak	yes
11	sunny	medium	normal	strong	yes
12	cloudy	medium	high	strong	yes
13	cloudy	high	normal	weak	yes
14	rainy	medium	high	strong	no

Which one looks easier?

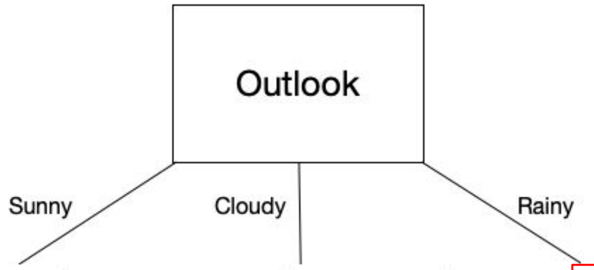
	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
11	sunny	medium	normal	strong	yes

Decision Tree Learning



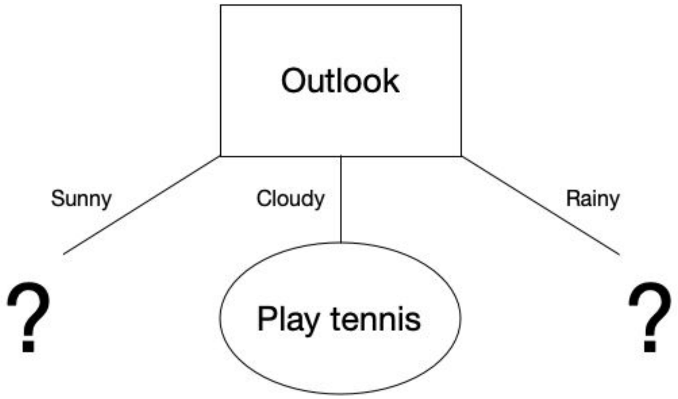
	~~outlook~~	temperature	humidity	wind	play tennis
3	cloudy	high	high	weak	yes
7	cloudy	low	normal	strong	yes
12	cloudy	medium	high	strong	yes
13	cloudy	high	normal	weak	yes

Decision Tree Learning



	~~**outlook**~~	**temperature**	**humidity**	**wind**	**play tennis**
3	cloudy	high	high	weak	yes
7	cloudy	low	normal	strong	yes
12	cloudy	medium	high	strong	yes
13	cloudy	high	normal	weak	yes

Decision Tree Learning



General Framework

1) If all examples are classified correctly, stop splitting and create a leaf node
2) If not, find the best feature to split in the dataset
3) Build the node corresponding to the split, and repeat algorithm each subset of examples.

General Framework: Best split

- We need to define an impurity measure
- This measure should be:
  - Zero when all the examples have the same label
  - Maximum when all examples have different labels
- Some examples:
  - Entropy
  - Gini Index

General Framework: Best split

- Entropy:
  - Measures amount of “uncertainty” in the labels

$$H(X) = -\sum_{x \in vals(X)} p(x) \log_2(p(x))$$

- Gini Index:
  - Measures probability of being wrong when assigning a random label to a random example

$$Gini(X) = 1 - \sum_{x \in vals(X)} p(x)^2$$

General Framework: Best split

	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
3	cloudy	high	high	weak	yes
4	rainy	medium	high	weak	yes
5	rainy	low	normal	weak	yes
6	rainy	low	normal	strong	no
7	cloudy	low	normal	strong	yes
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
10	rainy	medium	normal	weak	yes
11	sunny	medium	normal	strong	yes
12	cloudy	medium	high	strong	yes
13	cloudy	high	normal	weak	yes
14	rainy	medium	high	strong	no

$$H(play\ tennis) = -\left(p_{yes}\log_2(p_{yes}) + p_{no}\log_2(p_{no})\right)$$

$$= -\left(\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{5}{14}\log_2\left(\frac{5}{14}\right)\right)$$

$$= 0.940$$

General Framework: Best split

- We want to find which feature will reduce the impurity the most
- For this we introduce a gain measure to see how "good" a split is

$$Gain\ Measure(D, A) = H(D) - \sum_{a \in splitsFor(A)} \frac{|D_a|}{|D|} H(D_a)$$

D is the whole dataset

A is the feature we are splitting in

D_a is a subset of examples after splitting the data

General Framework: Best split

	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
11	sunny	medium	normal	strong	yes

$D_{outlook=sunny}$

$H(D_{outlook=sunny}) = 0.907$

General Framework: Best split

$$H(D_{outlook=sunny}) = 0.907$$

$$H(D_{outlook=cloudy}) = 0$$

$$H(D_{outlook=rainy}) = 0.907$$

$$GainMeasure(Play\ Tennis,\ Outlook) = 0.246$$

General Framework: Best split

$$\boxed{GainMeasure(Play\ Tennis,\ Outlook) = 0.246}$$

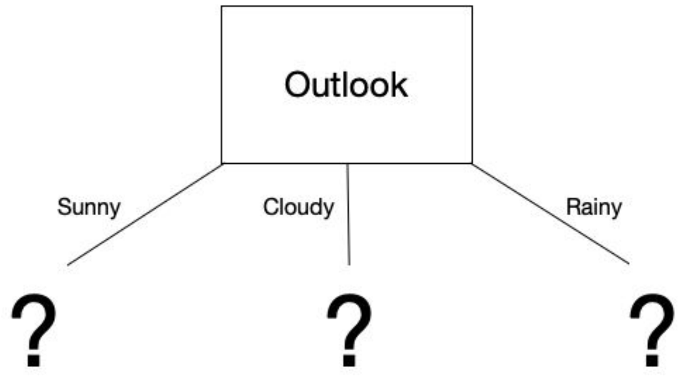
GainMeasure(Play Tennis, Temperature) = 0.029

GainMeasure(Play Tennis, Humidity) = 0.151

GainMeasure(Play Tennis, Wind) = 0.048

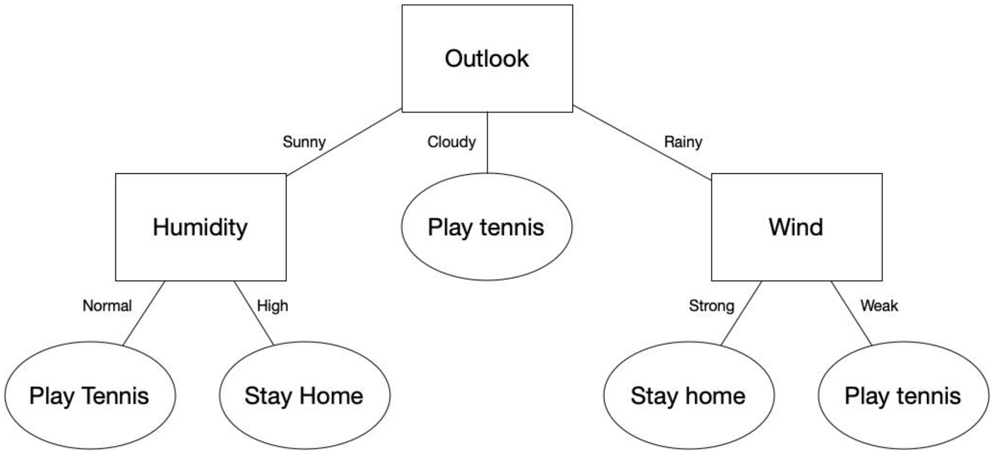
General Framework: Best split

	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
3	cloudy	high	high	weak	yes
4	rainy	medium	high	weak	yes
5	rainy	low	normal	weak	yes
6	rainy	low	normal	strong	no
7	cloudy	low	normal	strong	yes
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
10	rainy	medium	normal	weak	yes
11	sunny	medium	normal	strong	yes
12	cloudy	medium	high	strong	yes
13	cloudy	high	normal	weak	yes
14	rainy	medium	high	strong	no



General Framework: Best split

	outlook	temperature	humidity	wind	play tennis
1	sunny	high	high	weak	no
2	sunny	high	high	strong	no
3	cloudy	high	high	weak	yes
4	rainy	medium	high	weak	yes
5	rainy	low	normal	weak	yes
6	rainy	low	normal	strong	no
7	cloudy	low	normal	strong	yes
8	sunny	medium	high	weak	no
9	sunny	low	normal	weak	yes
10	rainy	medium	normal	weak	yes
11	sunny	medium	normal	strong	yes
12	cloudy	medium	high	strong	yes
13	cloudy	high	normal	weak	yes
14	rainy	medium	high	strong	no



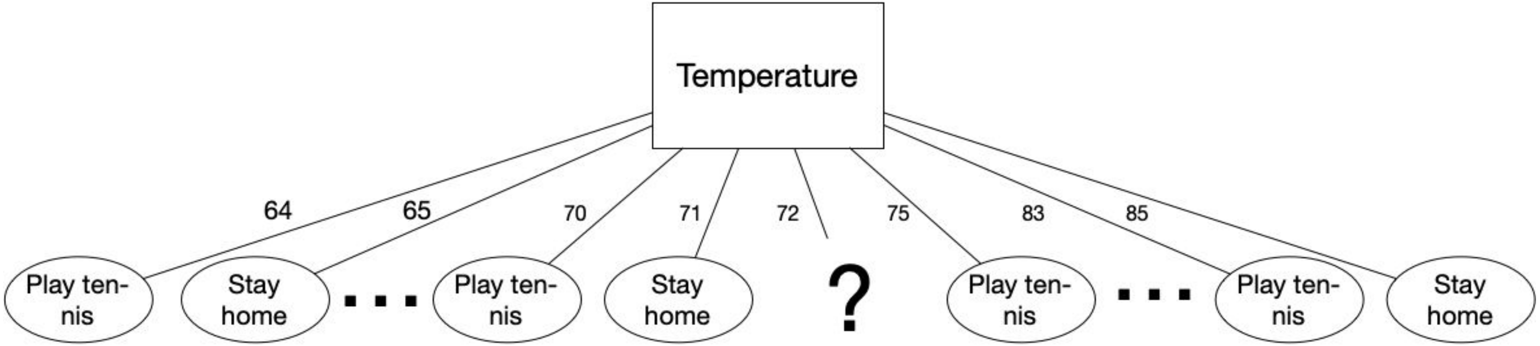
General Framework: Type of splits

- What happens if we have numerical data?

	outlook	temperature	humidity	wind	play tennis
1	sunny	85	85	weak	no
2	sunny	80	90	strong	no
3	cloudy	83	78	weak	yes
4	rainy	70	96	weak	yes
5	rainy	68	80	weak	yes
6	rainy	65	70	strong	no
7	cloudy	64	65	strong	yes
8	sunny	72	95	weak	no
9	sunny	69	70	weak	yes
10	rainy	75	80	weak	yes
11	sunny	75	70	strong	yes
12	cloudy	72	90	strong	yes
13	cloudy	81	75	weak	yes
14	rainy	71	80	strong	no

General Framework: Type of splits

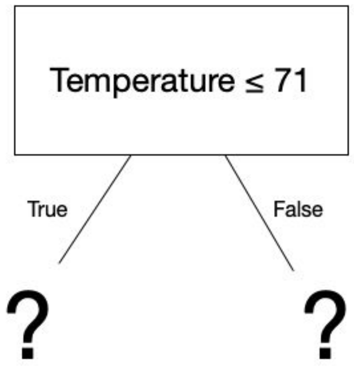
- What happens if we have numerical data?



This is not good...

General Framework: Type of splits

- What happens if we have numerical data?



This is better but...

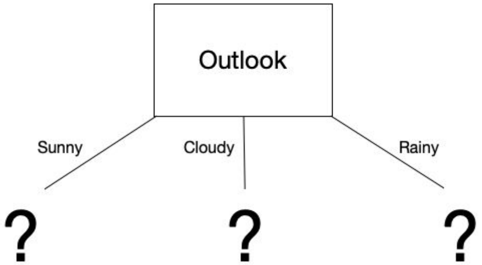
General Framework: Type of splits

- Now we also need find which is the best threshold

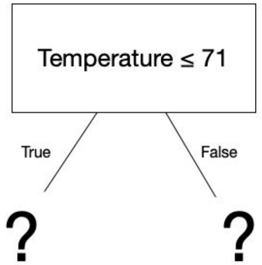
	outlook	temperature	humidity	wind	play tennis
1	sunny	85	85	weak	no
2	sunny	80	90	strong	no
3	cloudy	83	78	weak	yes
4	rainy	70	96	weak	yes
5	rainy	68	80	weak	yes
6	rainy	65	70	strong	no
7	cloudy	64	65	strong	yes
8	sunny	72	95	weak	no
9	sunny	69	70	weak	yes
10	rainy	75	80	weak	yes
11	sunny	75	70	strong	yes
12	cloudy	72	90	strong	yes
13	cloudy	81	75	weak	yes
14	rainy	71	80	strong	no

General Framework: Type of splits

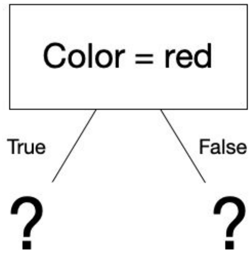
Multiway splits



Threshold splits



OneVsAll Splits



Different algorithms

	ID3	C4.5	CART
Impurity Measure	Entropy	Entropy	Gini index
Gain Measure	Information Gain	Gain Ratio	Information Gain (with Gini)
Splitting strategy: Categorical variable	Multiway Splits	Multiway Splits	One-vs-All Splits
Splitting strategy: Numerical variable	Multiway Splits	Threshold Splits	Threshold Splits

Different algorithms

Main problem:

These trees are over-specialized!

Random Forest

- Idea: Train many decision trees and then agglomerate their answers
- How does it work?

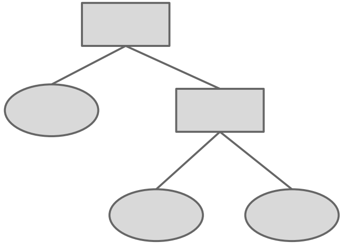
1) Select random examples

	outlook	temperature	humidity	wind	play tennis
1	sunny	85	85	weak	no
2	sunny	80	90	strong	no
3	cloudy	83	78	weak	yes
4	rainy	70	96	weak	yes
5	rainy	68	80	weak	yes
6	rainy	65	70	strong	no
7	cloudy	64	65	strong	yes
8	sunny	72	95	weak	no
9	sunny	69	70	weak	yes
10	rainy	75	80	weak	yes
11	sunny	75	70	strong	yes
12	cloudy	72	90	strong	yes
13	cloudy	81	75	weak	yes
14	rainy	71	80	strong	no

2) Select random features

	outlook	temperature	humidity	wind	play tennis
1	sunny	85	85	weak	no
2	sunny	80	90	strong	no
5	rainy	68	80	weak	yes
6	rainy	65	70	strong	no
7	cloudy	64	65	strong	yes
11	sunny	75	70	strong	yes

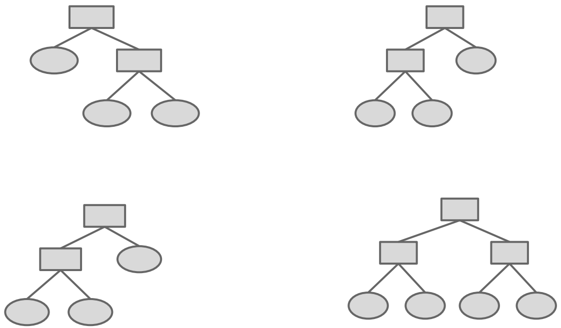
3) Train a decision tree



Random Forest

- Idea: Train many decision trees and then agglomerate their answers
- How does it work?

4) Repeat



Random Forest

- How to make a decision?