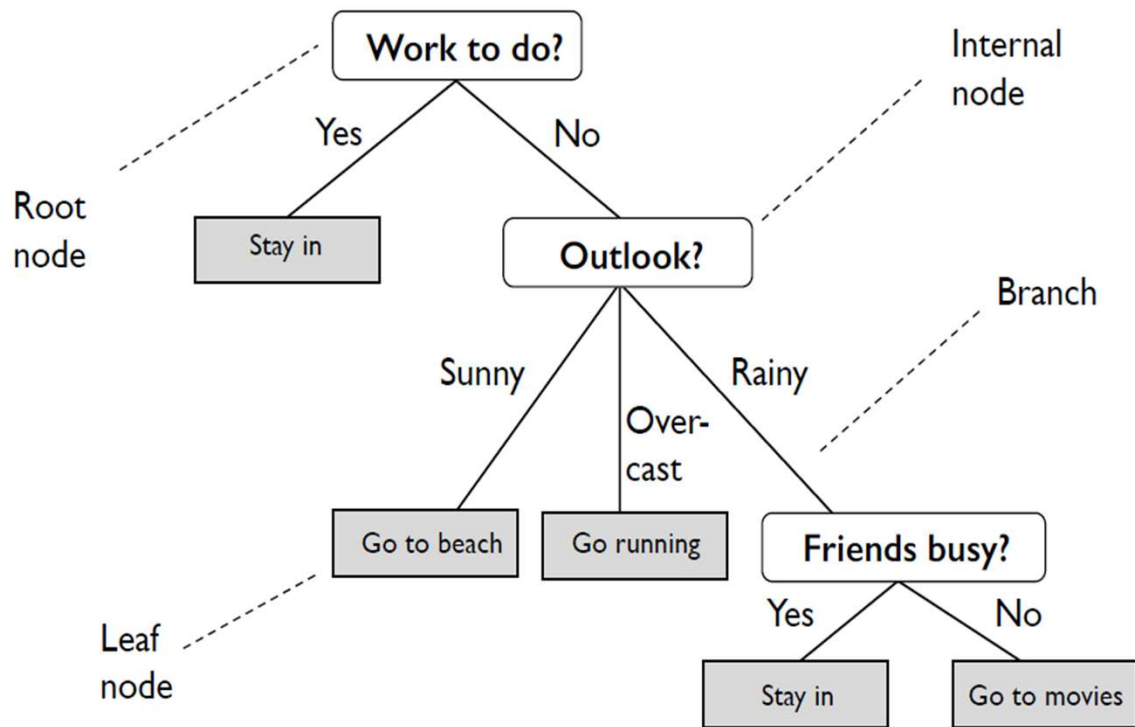


# Decision Tree

Fatemeh Mansoori

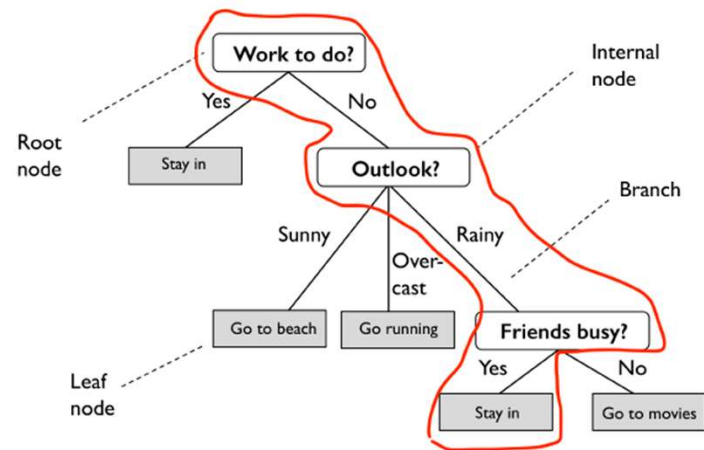
# Decision Tree Terminology



**Decision Tree** algorithms can be considered as an iterative, top-down construction method for either classifier or regressor.

Considering only binary (or Boolean) features, at each node, how many potential splits to be evaluated given that the dataset has  $m$  features?

# Decision Trees as Rulesets



IF

\_\_\_\_\_

\_\_\_\_\_

THEN

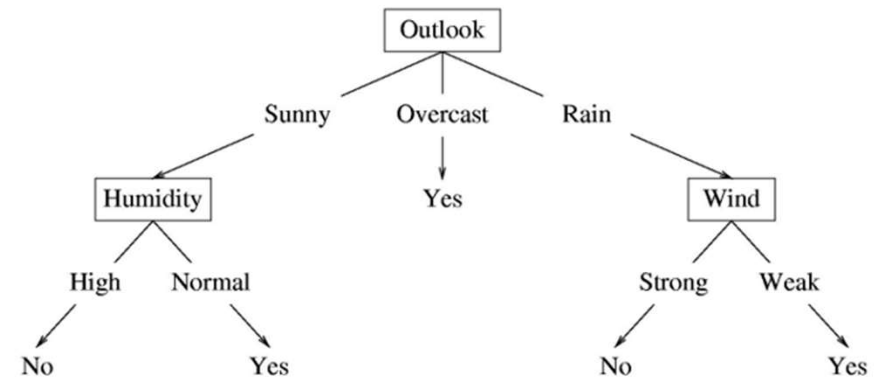
\_\_\_\_\_

# Sample Dataset

- Columns denote features  $X_i$
- Rows denote labeled instances  $\langle x_i, y_i \rangle$
- Class label denotes whether a tennis game was played

$\langle x_i, y_i \rangle$

Predictors				Response
Outlook	Temperature	Humidity	Wind	Class
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



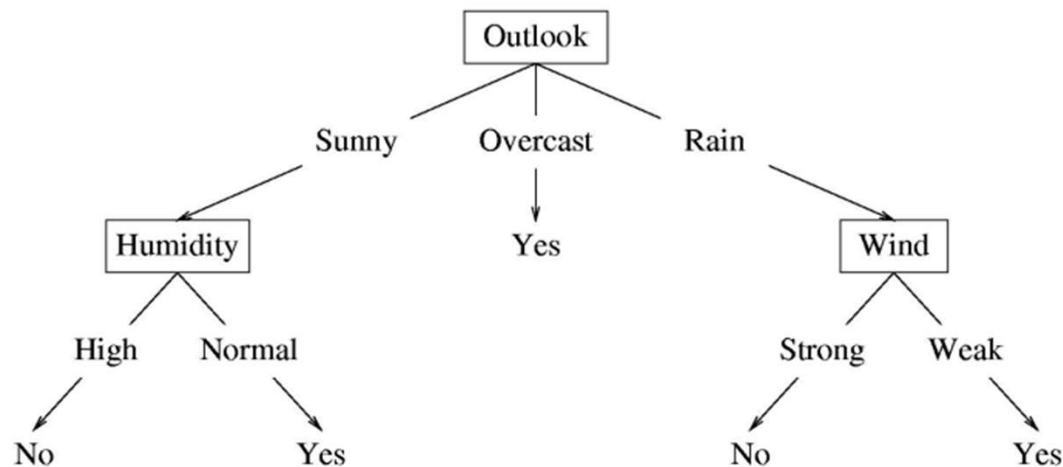
Each internal node: test one attribute  $X_i$

Each branch from a node: selects one value for  $X_i$

Each leaf node: predict  $Y$  (or  $p(Y | x \in \text{leaf})$ )

# Decision Tree

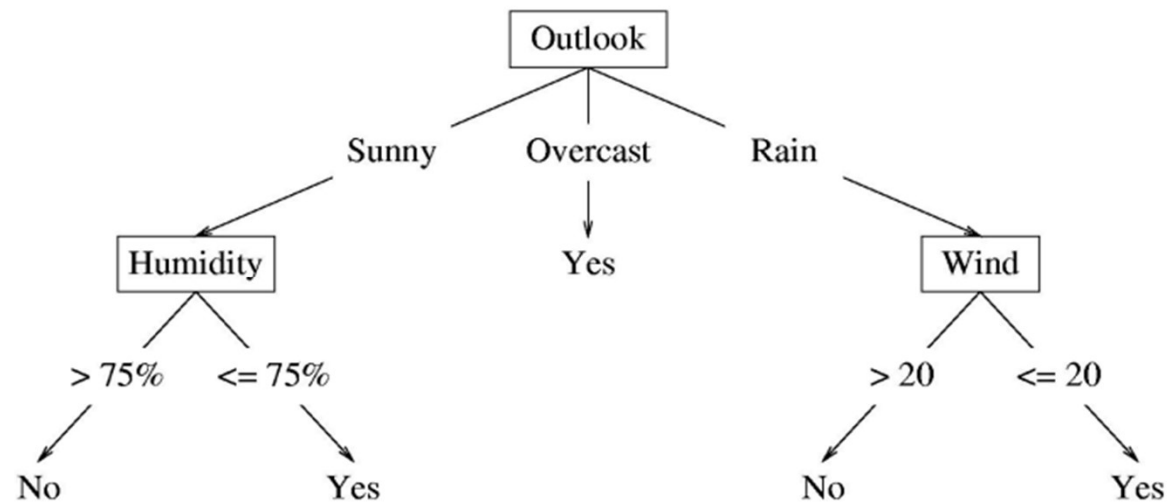
- A possible decision tree for the data:



- What prediction would we make for  
<outlook=sunny, temperature=hot, humidity=high, wind=weak> ?

# Decision Tree

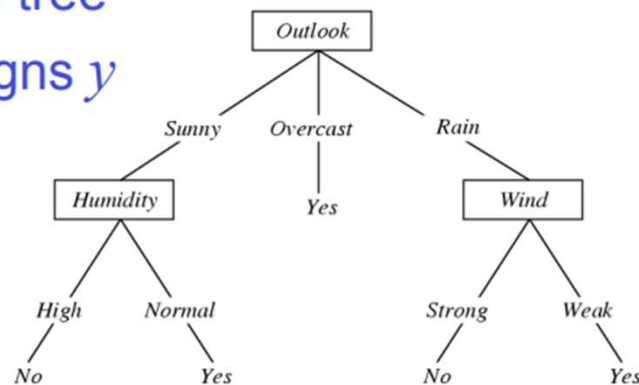
- If features are continuous, internal nodes can test the value of a feature against a threshold



# Decision Tree Learning

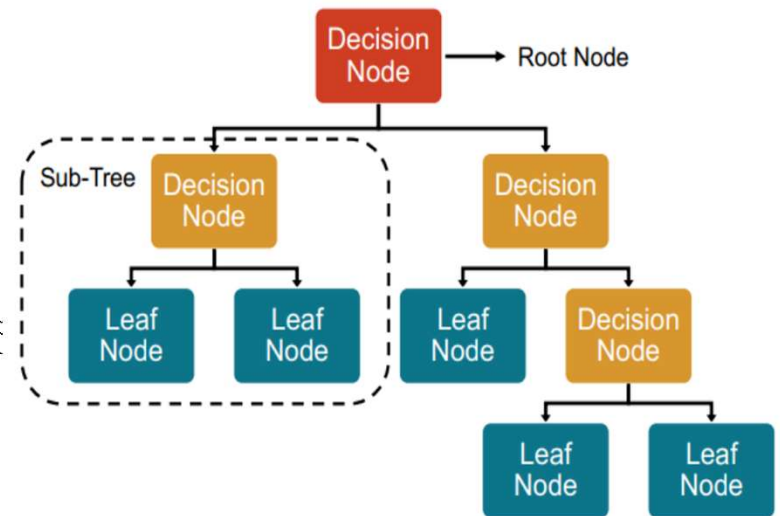
## Problem Setting:

- Set of possible instances  $X$ 
  - each instance  $x$  in  $X$  is a feature vector
  - e.g.,  $\langle \text{Humidity}=\text{low}, \text{Wind}=\text{weak}, \text{Outlook}=\text{rain}, \text{Temp}=\text{hot} \rangle$
- Unknown target function  $f: X \rightarrow Y$ 
  - $Y$  is discrete valued
- Set of function hypotheses  $H = \{ h \mid h: X \rightarrow Y \}$ 
  - each hypothesis  $h$  is a decision tree
  - trees sorts  $x$  to leaf, which assigns  $y$



# Terminology

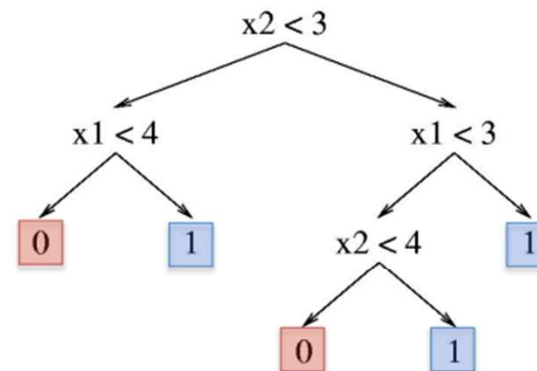
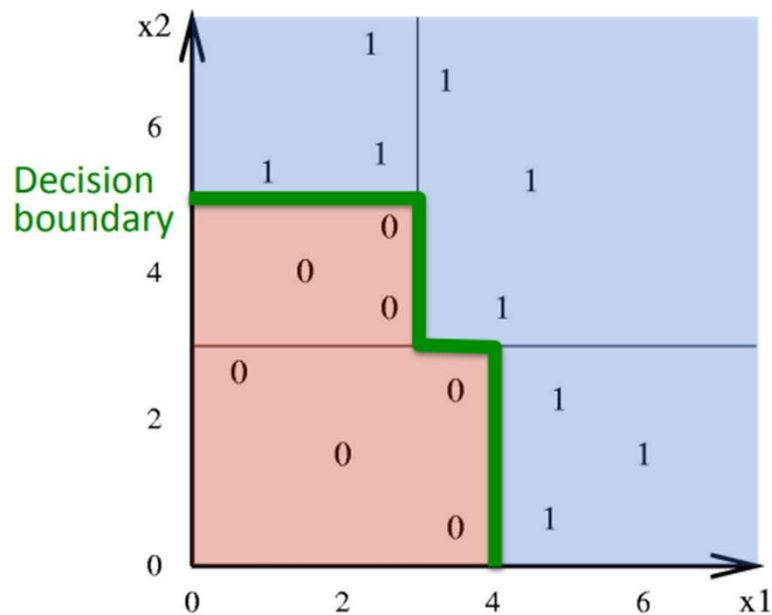
- **Root Node**: No incoming edge, zero, or more outgoing edges.
- **Internal Node**: One incoming edge, two (or more) outgoing edges.
- **Leaf Node**: Each leaf node is assigned a class label if nodes are pure; otherwise, the class label is determined by majority vote.
- **Parent & Child Node**: If a node is split, we refer to that given node as the parent node, and the resulting nodes are called child nodes.





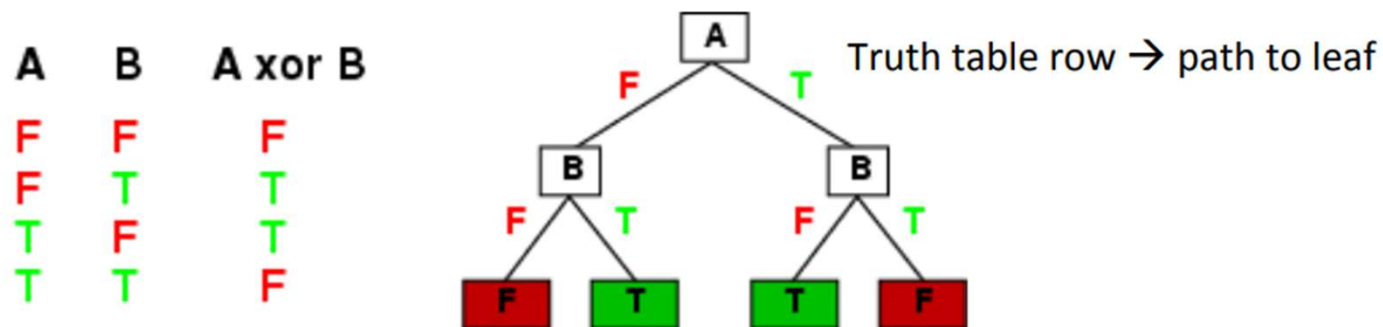
# Decision Tree – Decision Boundary

- Decision trees divide the feature space into axis-parallel (hyper-)rectangles
- Each rectangular region is labeled with one label
  - or a probability distribution over labels



# Expressiveness

- Decision trees can represent any boolean function of the input attributes



- In the worst case, the tree will require exponentially many nodes

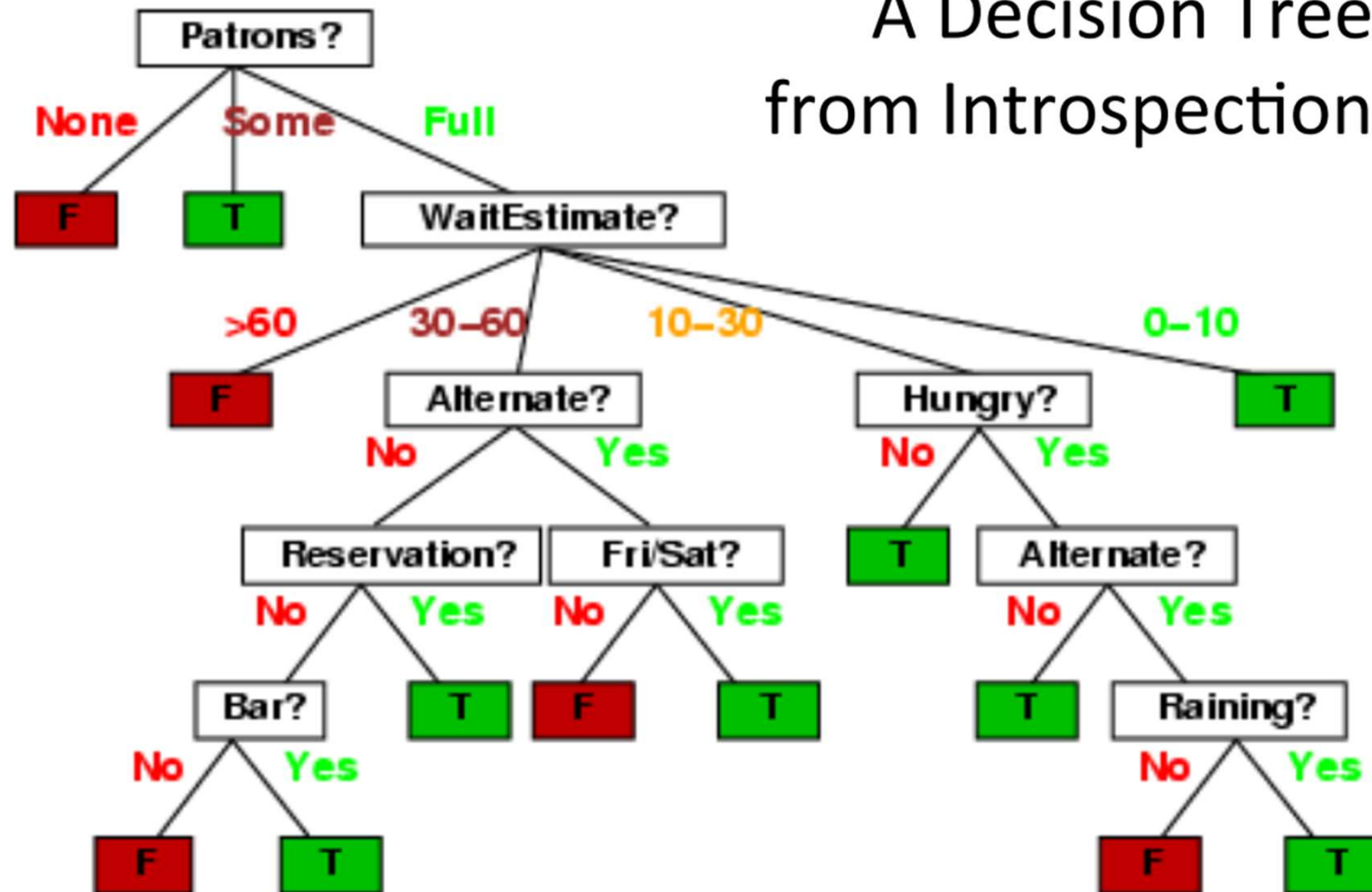
## Another Example: Restaurant Domain (Russell & Norvig)

Model a patron's decision of whether to wait for a table at a restaurant

Example	Attributes										Target
	<i>Alt</i>	<i>Bar</i>	<i>Fri</i>	<i>Hun</i>	<i>Pat</i>	<i>Price</i>	<i>Rain</i>	<i>Res</i>	<i>Type</i>	<i>Est</i>	<i>Wait</i>
$X_1$	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

~7,000 possible cases

## A Decision Tree from Introspection



Is this the best decision tree?

# Which Tree is the best?

- Smallest decision tree that correctly classifies all of the training examples is best
- Finding the provably smallest decision tree is NP-hard
- Instead of constructing the absolute smallest tree consistent with the training examples, construct one that is pretty small

# Basic Algorithm for Top-Down Induction of Decision Trees

[ID3, C4.5 by Quinlan]

*node* = root of decision tree

Main loop:

1.  $A \leftarrow$  the “best” decision attribute for the next node.
2. Assign  $A$  as decision attribute for *node*.
3. For each value of  $A$ , create a new descendant of *node*.
4. Sort training examples to leaf nodes.
5. If training examples are perfectly classified, stop.  
Else, recurse over new leaf nodes.

How do we choose which attribute is best?

# Choosing the Best Attribute

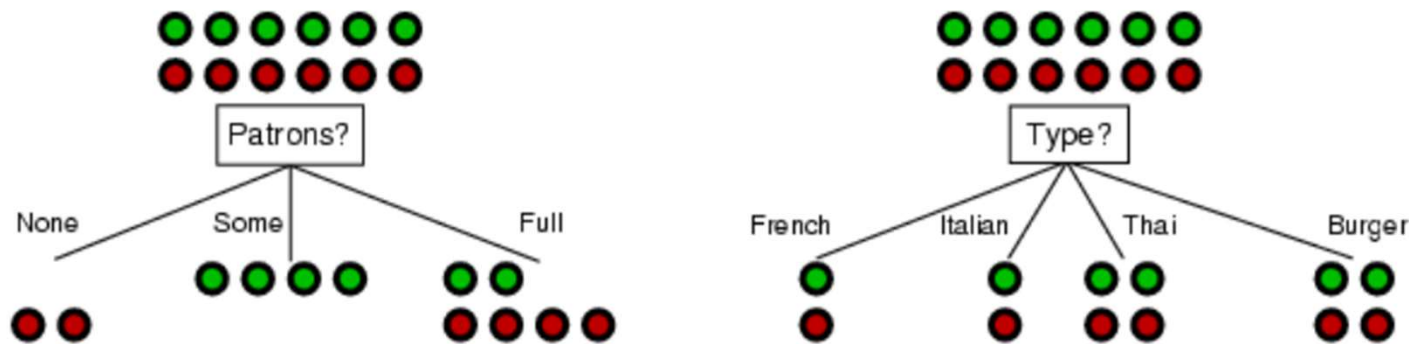
**Key problem:** choosing which attribute to split a given set of examples

- Some possibilities are:
  - **Random:** Select any attribute at random
  - **Least-Values:** Choose the attribute with the smallest number of possible values
  - **Most-Values:** Choose the attribute with the largest number of possible values
  - **Max-Gain:** Choose the attribute that has the largest expected *information gain*
    - i.e., attribute that results in smallest expected size of subtrees rooted at its children
- The ID3 algorithm uses the Max-Gain method of selecting the best attribute



# Choosing an Attribute

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



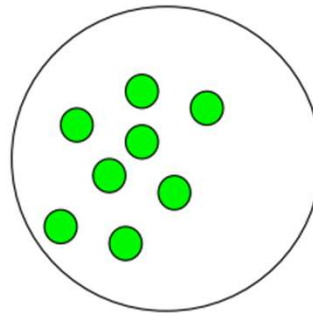
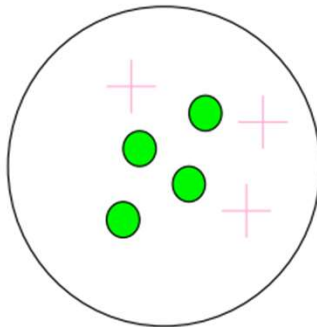
Which split is more informative: *Patrons?* or *Type?*



# Information Gain

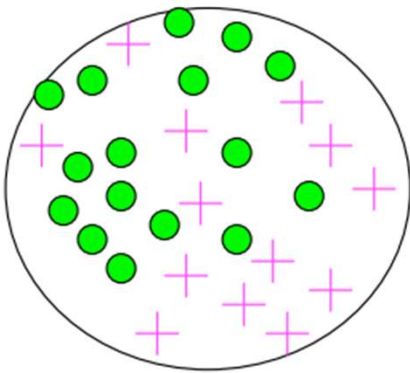
## Impurity/Entropy (informal)

- Measures the level of **impurity** in a group of examples

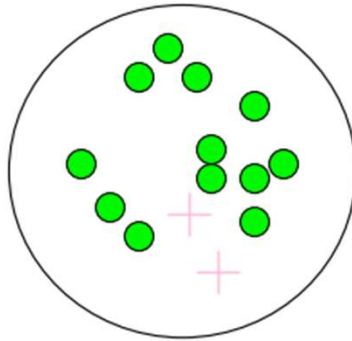


# Impurity

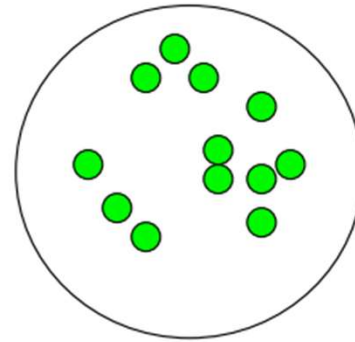
**Very impure group**



**Less impure**



**Minimum  
impurity**



# Entropy: a common way to measure impurity

Entropy  $H(X)$  of a random variable  $X$

# of possible  
values for  $X$

$$H(X) = - \sum_{i=1}^n P(X = i) \log_2 P(X = i)$$

$H(X)$  is the expected number of bits needed to encode a randomly drawn value of  $X$  (under most efficient code)

Why? Information theory:

- Most efficient code assigns  $-\log_2 P(X=i)$  bits to encode the message  $X=i$
- So, expected number of bits to code one random  $X$  is:

$$\sum_{i=1}^n P(X = i)(-\log_2 P(X = i))$$

## 2-Class Cases:

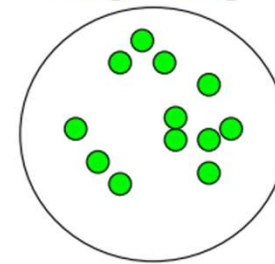
$$\text{Entropy } H(x) = - \sum_{i=1}^n P(x = i) \log_2 P(x = i)$$

- What is the entropy of a group in which all examples belong to the same class?

– entropy =  $-1 \log_2 1 = 0$

not a good training set for learning

**Minimum  
impurity**

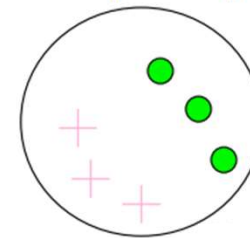


- What is the entropy of a group with 50% in either class?

– entropy =  $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

good training set for learning

**Maximum  
impurity**



# Information Gain

- We want to determine **which attribute** in a given set of training feature vectors is **most useful** for discriminating between the classes to be learned.
- **Information gain** tells us how important a given attribute of the feature vectors is.
- We will use it to decide the ordering of attributes in the nodes of a decision tree.

# From Entropy to Information Gain

Entropy  $H(X)$  of a random variable  $X$

$$H(X) = - \sum_{i=1}^n P(X = i) \log_2 P(X = i)$$

Specific conditional entropy  $H(X|Y=v)$  of  $X$  given  $Y=v$  :

$$H(X|Y = v) = - \sum_{i=1}^n P(X = i|Y = v) \log_2 P(X = i|Y = v)$$

Conditional entropy  $H(X|Y)$  of  $X$  given  $Y$  :

$$H(X|Y) = \sum_{v \in \text{values}(Y)} P(Y = v) H(X|Y = v)$$

Mutual information (aka Information Gain) of  $X$  and  $Y$  :

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

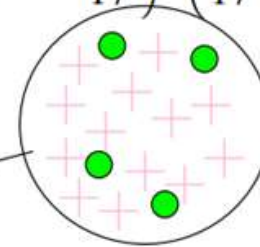
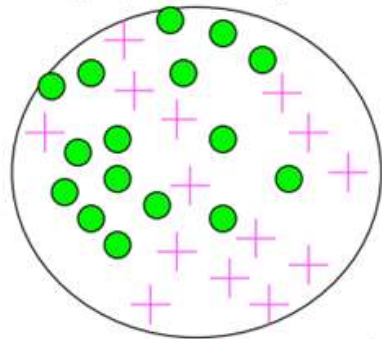


## Calculating Information Gain

**Information Gain** = entropy(parent) – [average entropy(children)]

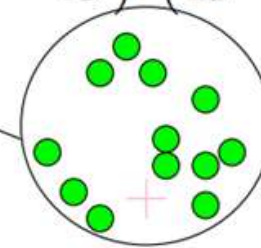
$$\text{child entropy} = -\left(\frac{13}{17} \cdot \log_2 \frac{13}{17}\right) - \left(\frac{4}{17} \cdot \log_2 \frac{4}{17}\right) = 0.787$$

Entire population (30 instances)



17 instances

$$\text{child entropy} = -\left(\frac{1}{13} \cdot \log_2 \frac{1}{13}\right) - \left(\frac{12}{13} \cdot \log_2 \frac{12}{13}\right) = 0.391$$



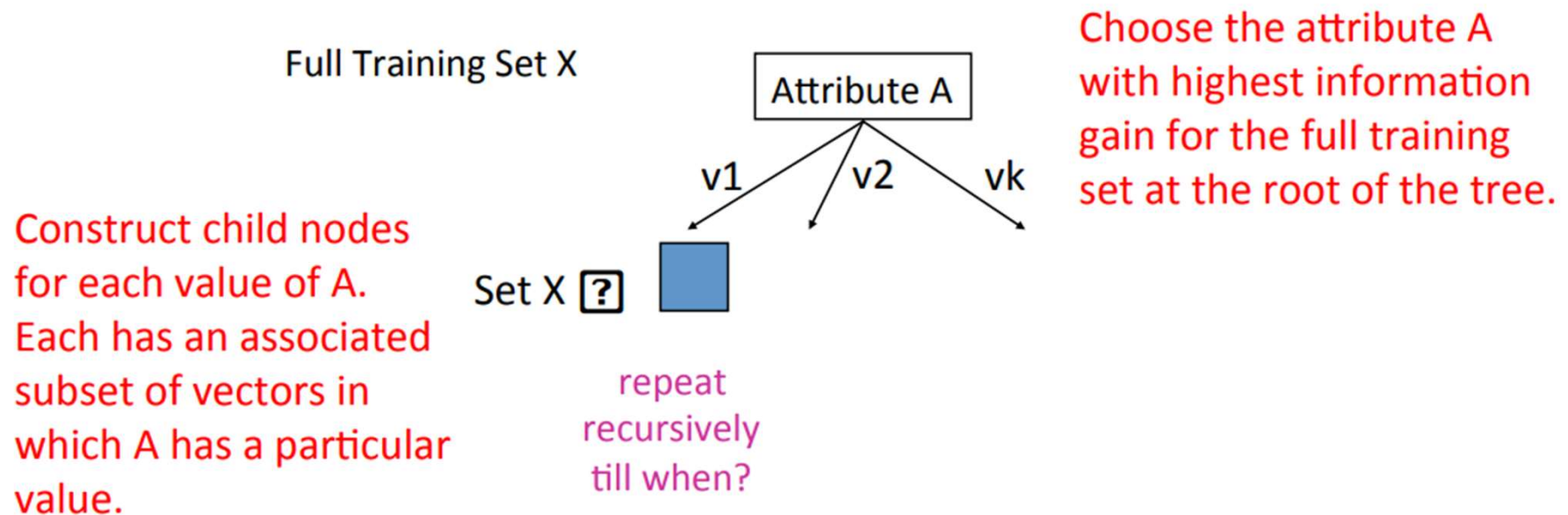
13 instances

$$\text{parent entropy} = -\left(\frac{14}{30} \cdot \log_2 \frac{14}{30}\right) - \left(\frac{16}{30} \cdot \log_2 \frac{16}{30}\right) = 0.996$$

$$\text{(Weighted) Average Entropy of Children} = \left(\frac{17}{30} \cdot 0.787\right) + \left(\frac{13}{30} \cdot 0.391\right) = 0.615$$

$$\text{Information Gain} = 0.996 - 0.615 = 0.38$$

# Using Information Gain to Construct a Decision Tree





## Training Examples

---

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

# Gini Index

- **Gini Impurity**: Let  $X$  be a random variable over a discrete space  $\Omega$  with probability mass function  $P$ . Then the Gini coefficient of  $X$  is:

$$G(X) = \sum_{x \in X} \mathbb{P}(x)(1 - \mathbb{P}(x)) = 1 - \sum_{x \in X} \mathbb{P}(x)^2$$

# Misclassifications Error

Misclassification error seems to be another reasonable choice, where:

$$ERR(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^n L(\hat{y}^{[i]}, y^{[i]})$$

with the 0-1 Loss,

$$L(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y, \\ 1 & \text{if } \hat{y} \neq y \end{cases}$$

This, in case of the training set, is equal to

$$ERR(p) = 1 - \max_i (p(i \mid x_j))$$

for a given node if we use majority voting at this node.

Can we apply this measure as the criteria for splitting?

# Algorithm

- ID3: Iterative Dichotomizer 3
  - It can only be used for discrete features, and cannot handle numeric features!
  - It supports multi-category splits.
  - It does not apply any pruning, and is prone to overfitting.
  - It results in short and wide trees.
  - It maximizes information gain/minimizes entropy.
- C4.5
  - It supports Continuous and discrete features (continuous feature splitting is very expensive because must consider all possible ranges).
  - It can handles missing attributes (ignores them in information gain computation)
  - It performs post-pruning

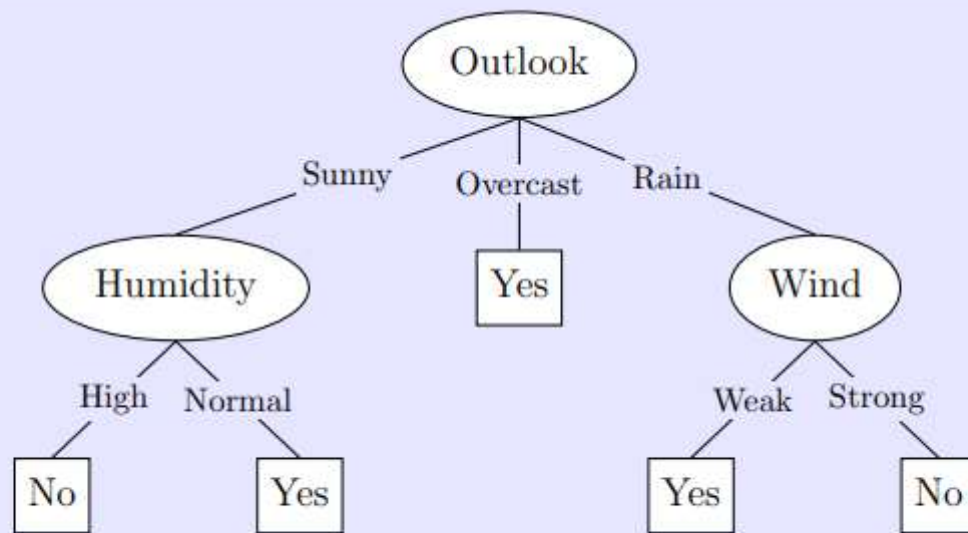
# Decision Tree for Regression

- It is the same as classification with a few subtle modifications:
- The main difference is that instead of predicting a class in each node, it predicts a value.
- The prediction for each node is simply the average target value of the all training instances associated to this leaf node.
- Instead of trying to split the training set in a way that minimizes impurity, CART algorithm now tries to split the training set in a way that minimizes the **MSE**.
  - The MSE at a given node is also often referred to as **intra-node variance**, and the splitting criterion is thus called **variance reduction**.
  - Likewise to classification tasks, decision trees are prone to overfitting when dealing with regression tasks.

# Overfitting(example)

**Example:** The Jeeves data set has 14 data points, this is an extremely small data set; we can still learn a reasonable decision tree for this data set.

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

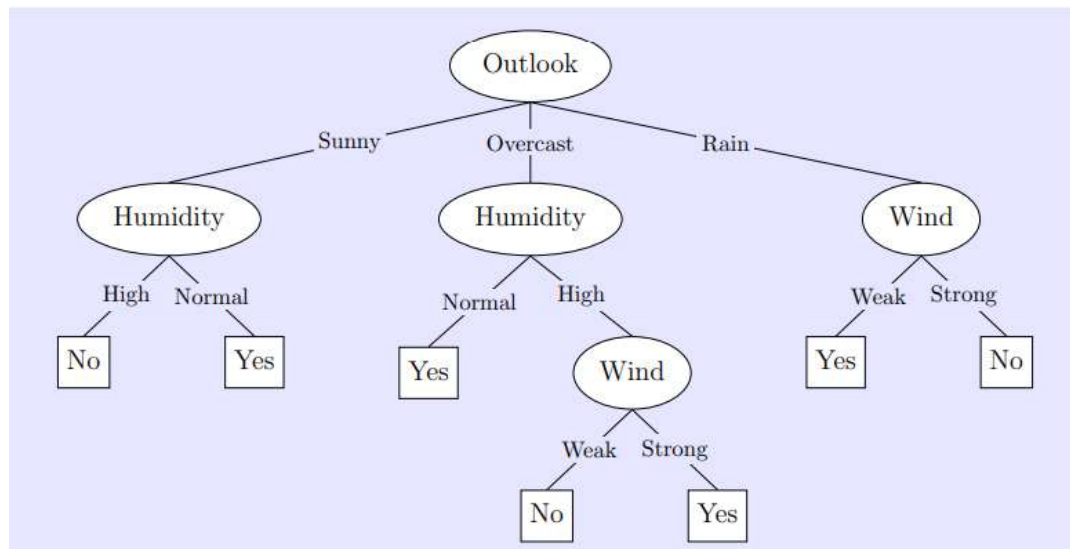


Test error is 0 out of 14.

- Suppose that you sent the training set and the test set to your friend. For some reason the data set gets corrupted during the transmission. Your friend gets a corrupted training set where only one data point is different. For this third data, point, it changed from the label "yes" to the label "no".

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	<b>No</b>
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No





We grew an entire middle sub-tree to accommodate one corrupted data point, and this small corruption caused a dramatic change to the tree. This new tree is more complicated and it likely won't generalize well to unseen data.

# Overfitting

- Decision Trees make very few assumptions about the training data!
- If left unconstrained, their structure will adapt itself to the training data, fitting it very closely, and most likely overfitting it!
- When tree is too complex and capture noise in the training samples the overfitting is occurred
- It would be better to grow a smaller and shallower tree.
- Smaller and shallower tree may not predict all of the training data points perfectly but it may generalize to test data better.
- have two options to prevent over-fitting when learning a decision tree:
  - Pre-pruning: stop growing the tree early.
  - Post-pruning: grow a full tree first and then trim it afterwards

# Pre pruning

- decide not to split the examples at a node and stop growing the tree
- use the majority label as the decision for that leaf node
  - **Maximum depth**: not to split the examples if the depth of that node has reached a maximum value that we decided beforehand.
  - **Minimum number of examples at the leaf node**: We can decide not to split the examples if the number of examples remaining at that node is less than a predefined threshold value.
  - **Minimum information gain**: We can decide not to split the examples if the benefit of splitting at that node is not large enough. We can measure the benefit by calculating the expected information gain
  - **Reduction in training error**: We can decide not to split the examples at a node if the reduction in training error is less than a predefined threshold value.

# Post Pruning

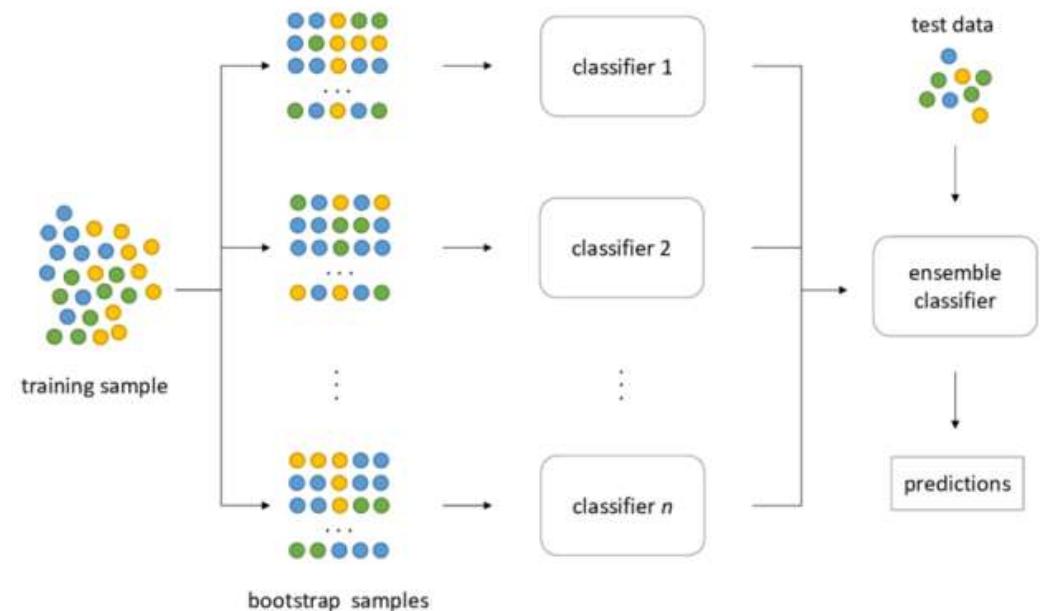
- Post-pruning: grow a full tree first and then trim it afterwards
- When we may use Post Pruning?

# Idea of Ensemble Classifiers

- Using a single decision tree could be problematic
  - highly sensitive to small changes in data
- One solution :
  - instead of just using a single decision tree, train multiple decision trees to make it more robust (Bagging)
- Each decision tree is a different way of classifying the objects.
- The collection of the decision trees is called tree ensemble.

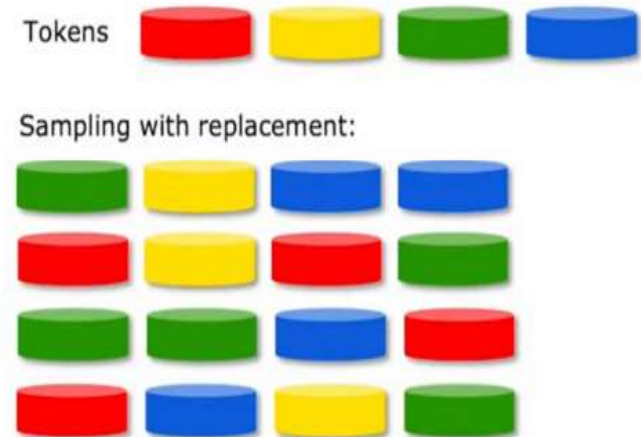
# Bagging (bootstrap aggregation)

- learning method utilizing several different models
  - built by training on different sample subset
- The average, aggregation, or majority of the predictions from different models is calculated for prediction
- In the context of Decision Tree, tree ensemble and majority decision is used for bagging.



# Sampling with replacement

- to build tree ensemble, used sampling with replacement technique
- Sampling with replacement :
  - randomly select a sample from the dataset
  - after recording the sample that was picked, put in back to the dataset
  - pick another random sample from the dataset record it, and put it back again



# Random Forest

- used for classification and regression
- Several decision trees are used in this ensemble learning technique to increase accuracy and decrease overfitting.
- construct a huge number of decision trees, each of which is trained using a unique part of the training data and a random sample of features
- Each tree in the forest makes a forecast for an output given an input
- final prediction is formed by collecting the majority vote of all the trees in the forest
- Random feature selection and data sampling contribute to reducing overfitting.



# Random Forest

*RandomForest = DecisionTree + Bagging + Bootstrapping + RandomSplits*

Given training set of size  $m$

For  $b = 1$  to  $B$ :

    Use sampling with replacement to create a new training set of size  $m$

    Train a decision tree on the new dataset

$B$  stands for bagging and denotes the number of decision trees we are going to train and is often recommended to be picked as 64 or 128.

- The key idea is that even with sampling with replacement, you always ended up with same root nodes and similar splits.
- Random Forest improve this by randomizing the feature choice.
- At each node, when choosing a feature to use to split, if  $n$  features are available, pick a random subset of  $k < n$  features and allow the algorithm to only choose from that subset of features



**Random forests, adaptive boosting, gradient boosting**