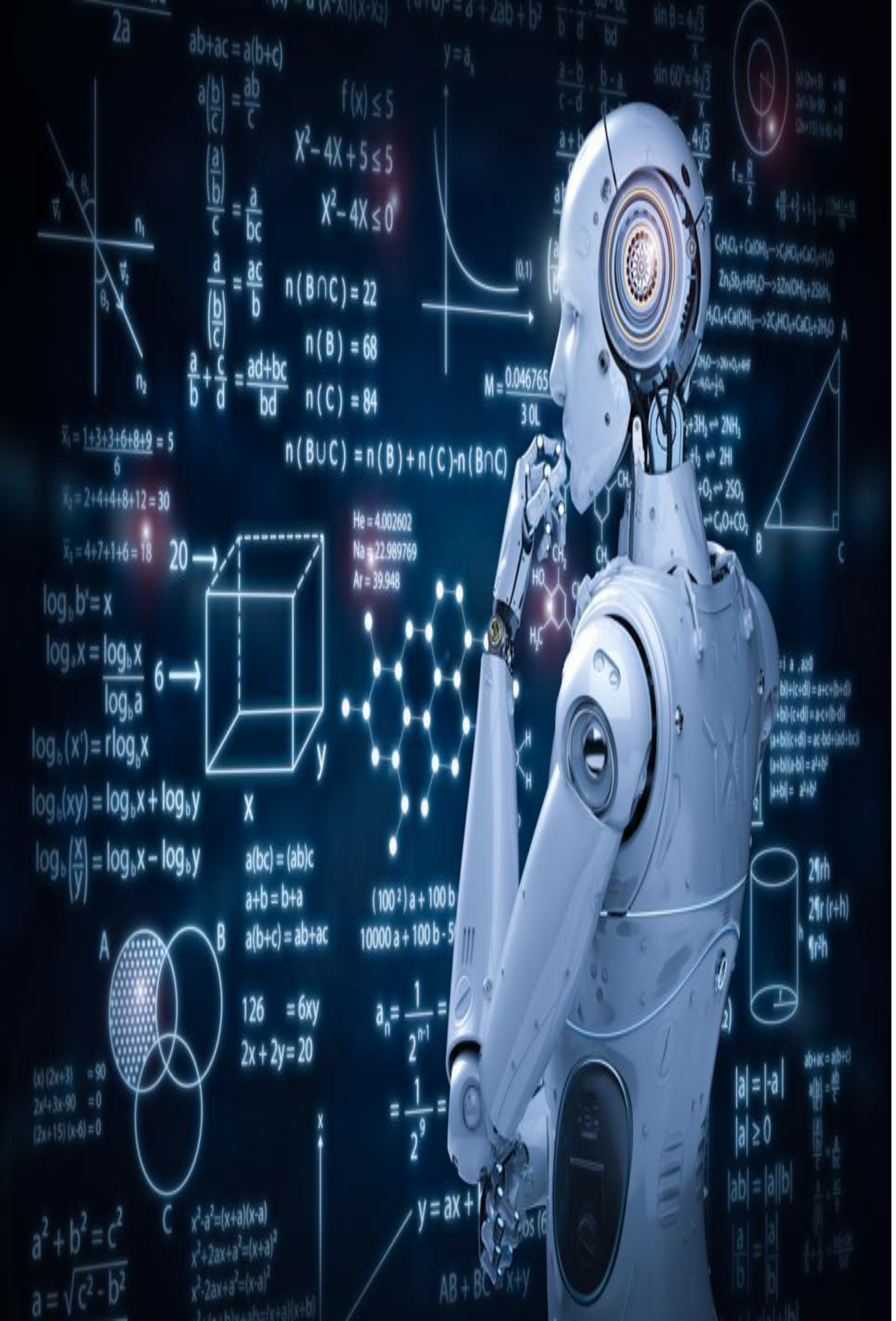


Practical Machine Learning

Day 12: Sep22 DBDA

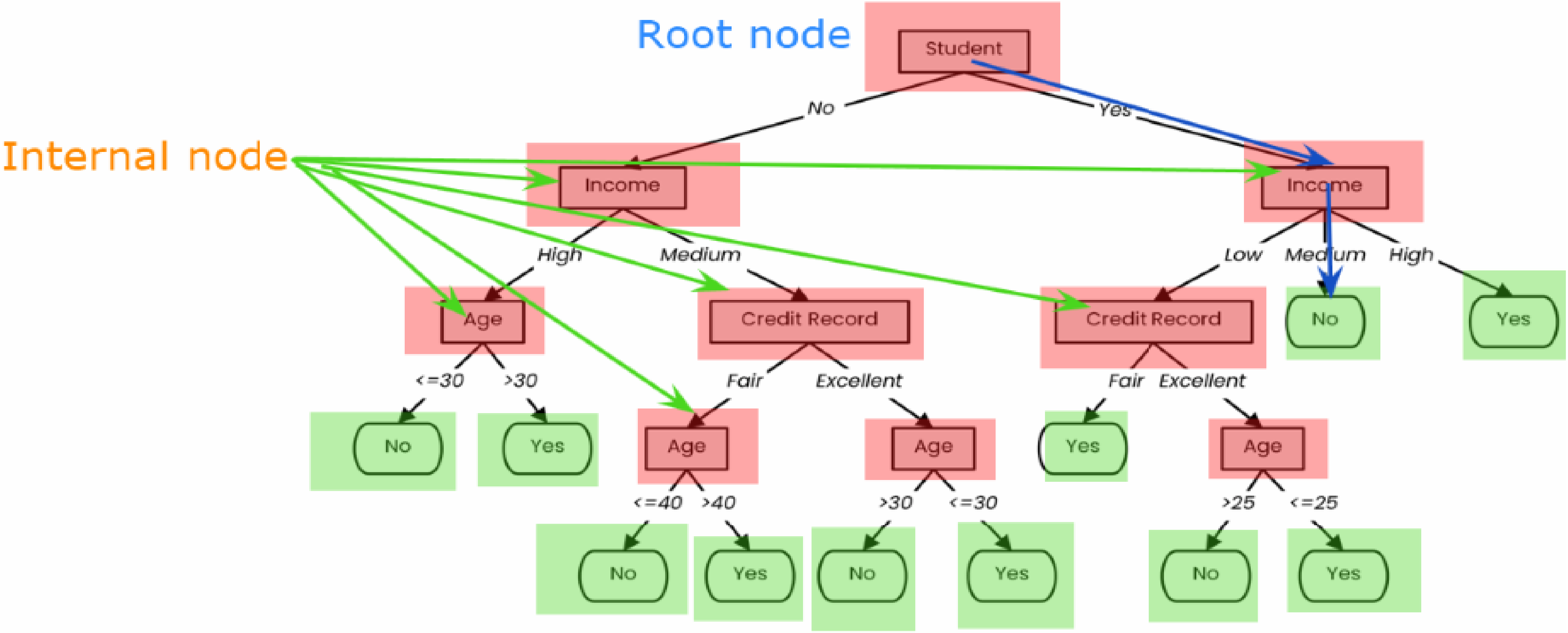
Kiran Waghmare

Agenda

- Decision Tree
- Random Forest

Definition

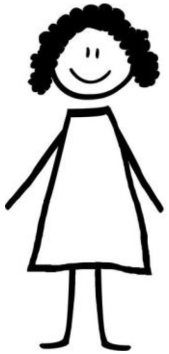
- A tree-like model that illustrates series of events leading to certain decisions
- Each node represents a test on an attribute and each branch is an outcome of that test



Definition

- A tree-like model that illustrates series of events leading to certain decisions
- Each node represents a test on an attribute and each branch is an outcome of that test

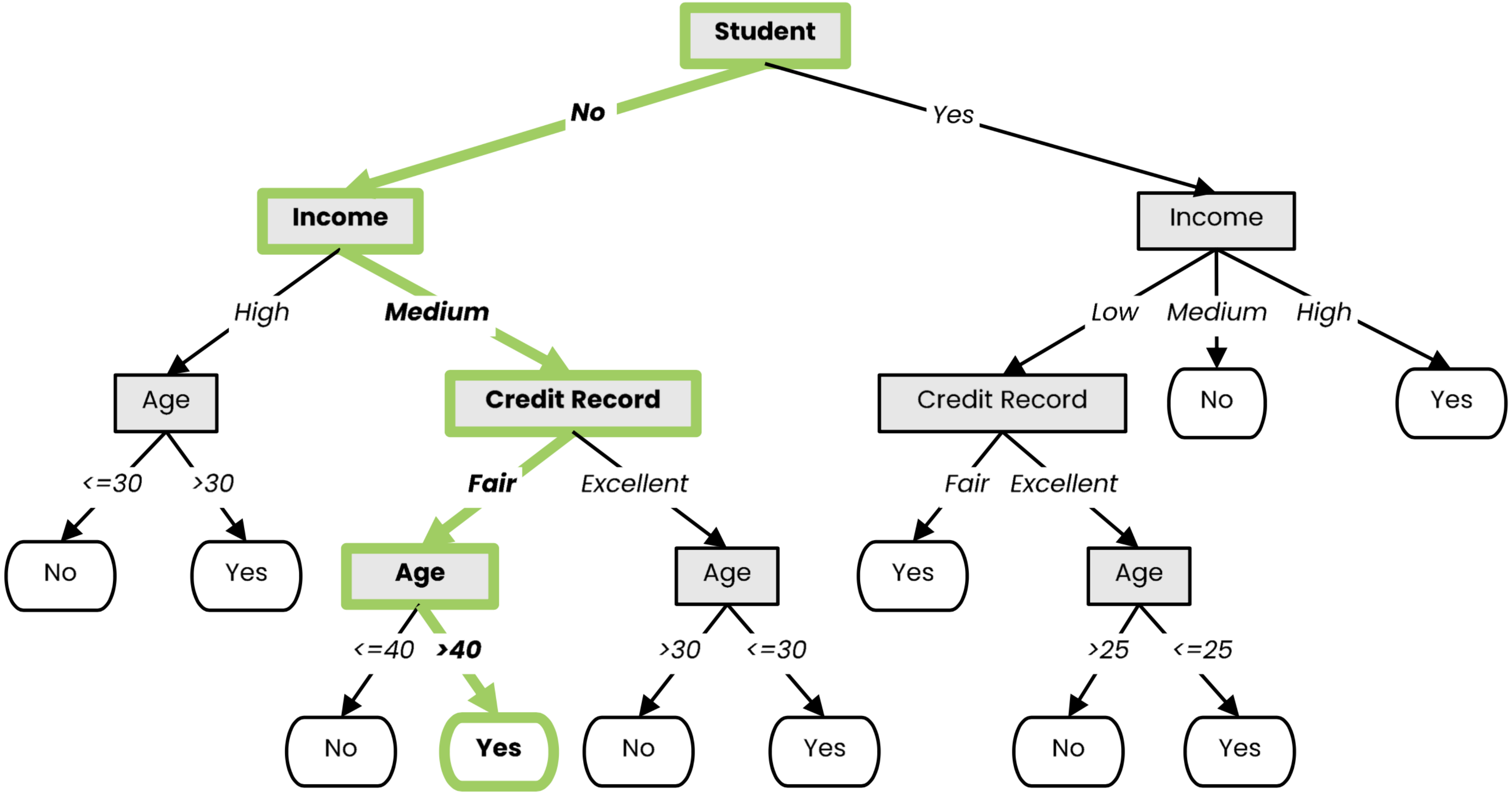
Who to loan?



- Not a student
- 45 years old
- Medium income
- Fair credit record

➢ Yes

- Student
- 27 years old
- Low income
- Excellent credit



Decision Tree Learning

- Basic step: choose an attribute and, based on its values, split the data into smaller sets
  - ➢ Recursively repeat this step until we can surely decide the label

Outlook = Sunny

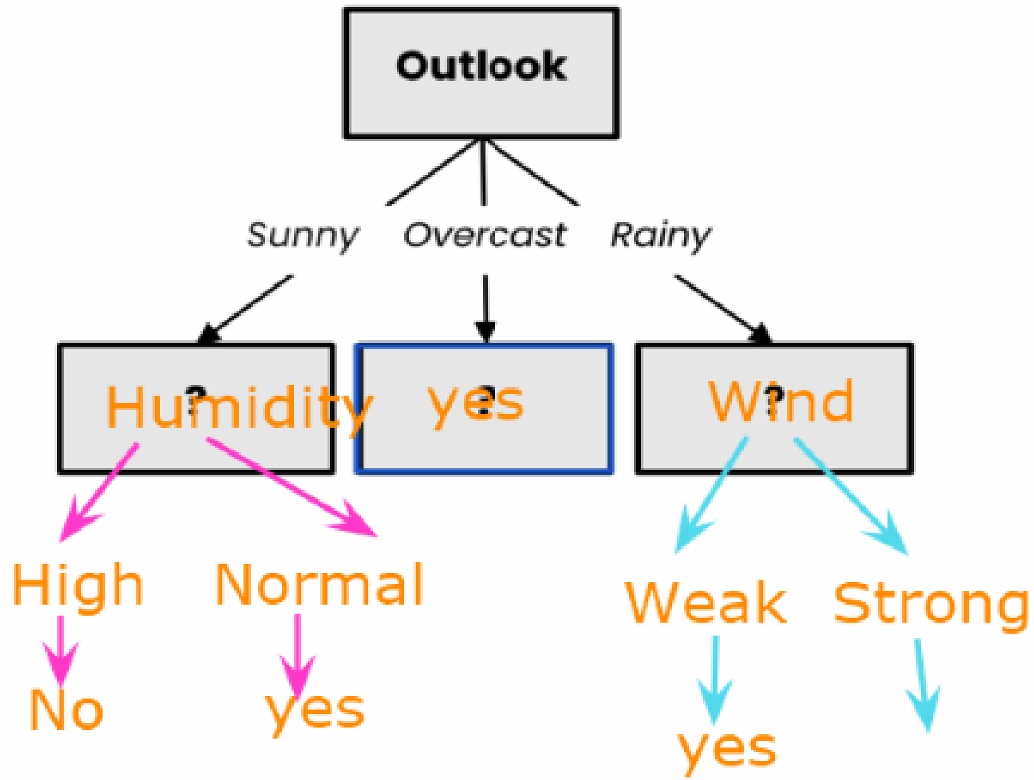
Temperature	Humidity	Wind	Play Tennis?
Hot	High	Weak	No
Hot	High	Strong	No
Mild	High	Weak	No
Cool	Normal	Weak	Yes
Mild	Normal	Strong	Yes

Outlook = Overcast

Temperature	Humidity	Wind	Play Tennis?
Hot	High	Weak	Yes
Cool	Normal	Strong	Yes
Mild	High	Strong	Yes
Hot	Normal	Weak	Yes

Outlook = Rainy

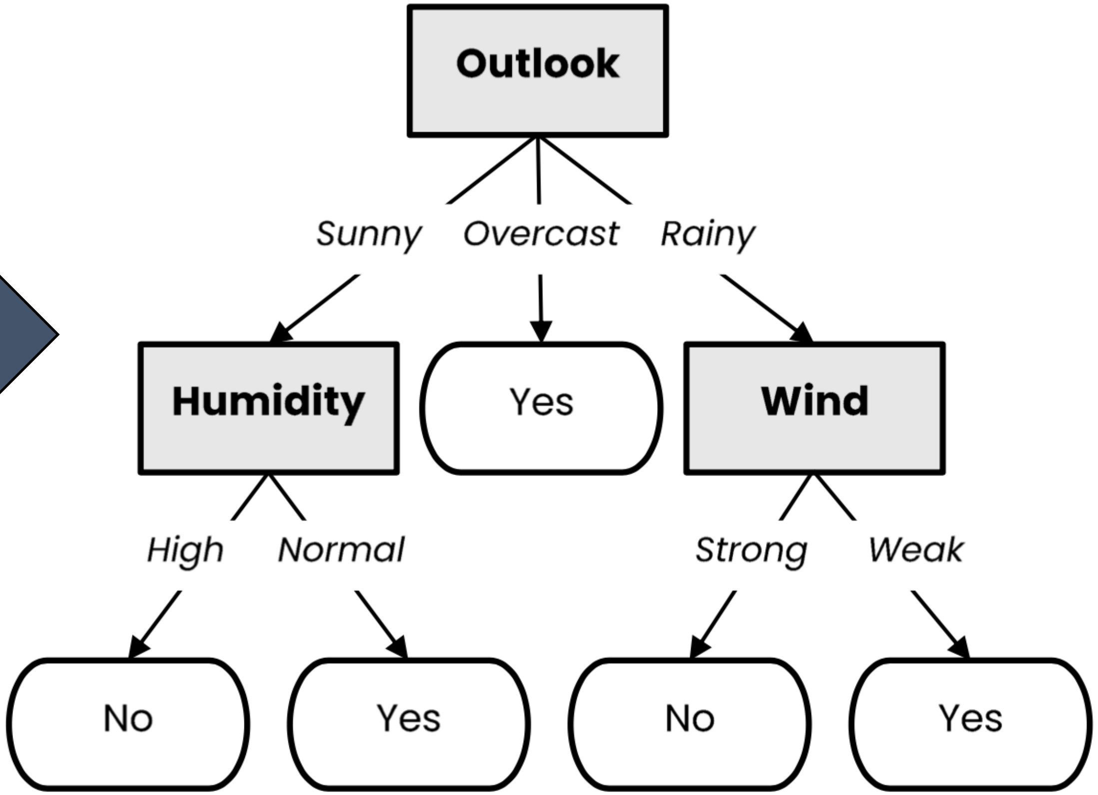
Temperature	Humidity	Wind	Play Tennis?
Mild	High	Weak	Yes
Cool	Normal	Weak	Yes
Cool	Normal	Strong	No
Mild	Normal	Weak	Yes
Mild	High	Strong	No



Decision Tree Learning

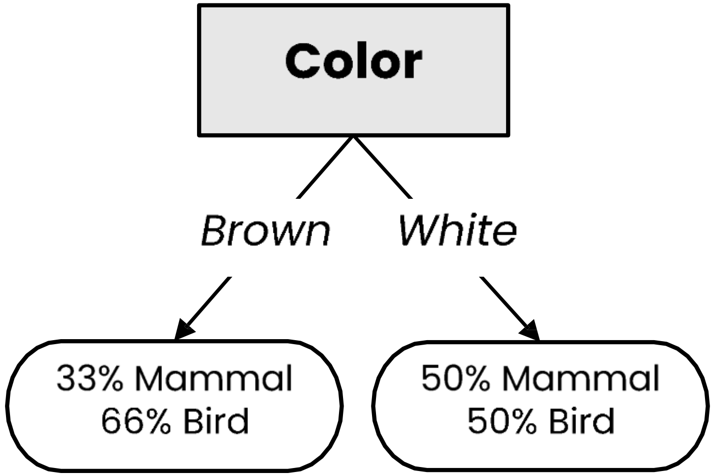
- We use labeled data to obtain a suitable decision tree for future predictions
  - ➢ We want a decision tree that works well on unseen data, while asking as few questions as possible

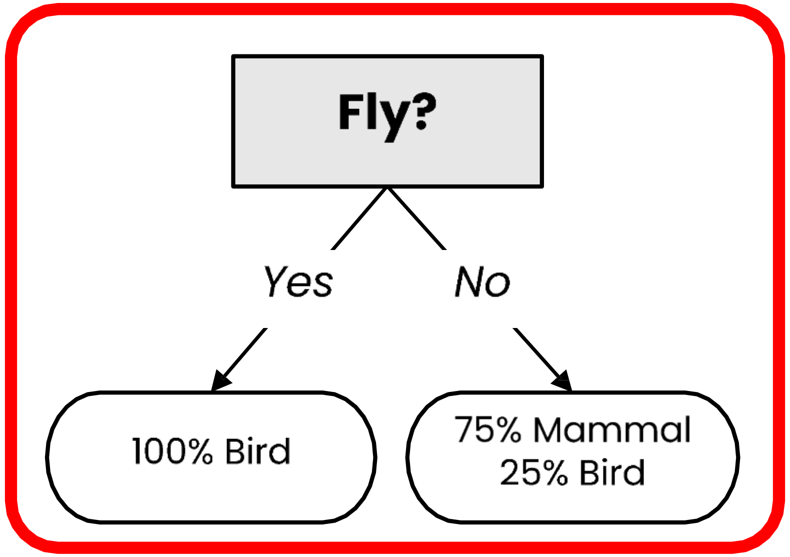
Outlook	Temperature	Humidity	Wind	Play Tennis?
Sunny	Hot	High	Weak	**No**
Sunny	Hot	High	Strong	**No**
Overcast	Hot	High	Weak	**Yes**
Rainy	Mild	High	Weak	**Yes**
Rainy	Cool	Normal	Weak	**Yes**
Rainy	Cool	Normal	Strong	**No**
Overcast	Cool	Normal	Strong	**Yes**
Sunny	Mild	High	Weak	**No**
Sunny	Cool	Normal	Weak	**Yes**
Rainy	Mild	Normal	Weak	**Yes**
Sunny	Mild	Normal	Strong	**Yes**
Overcast	Mild	High	Strong	**Yes**
Overcast	Hot	Normal	Weak	**Yes**
Rainy	Mild	High	Strong	**No**



What is a good attribute?

Does it fly?	Color	Class
No	Brown	**Mammal**
No	White	**Mammal**
Yes	Brown	**Bird**
Yes	White	**Bird**
No	White	**Mammal**
No	Brown	**Bird**
Yes	White	**Bird**





- Which attribute provides better splitting?
- Why?
  - ➢ Because the resulting subsets are more pure
  - ➢ Knowing the value of this attribute gives us more information about the label (the entropy of the subsets is lower)

Entropy

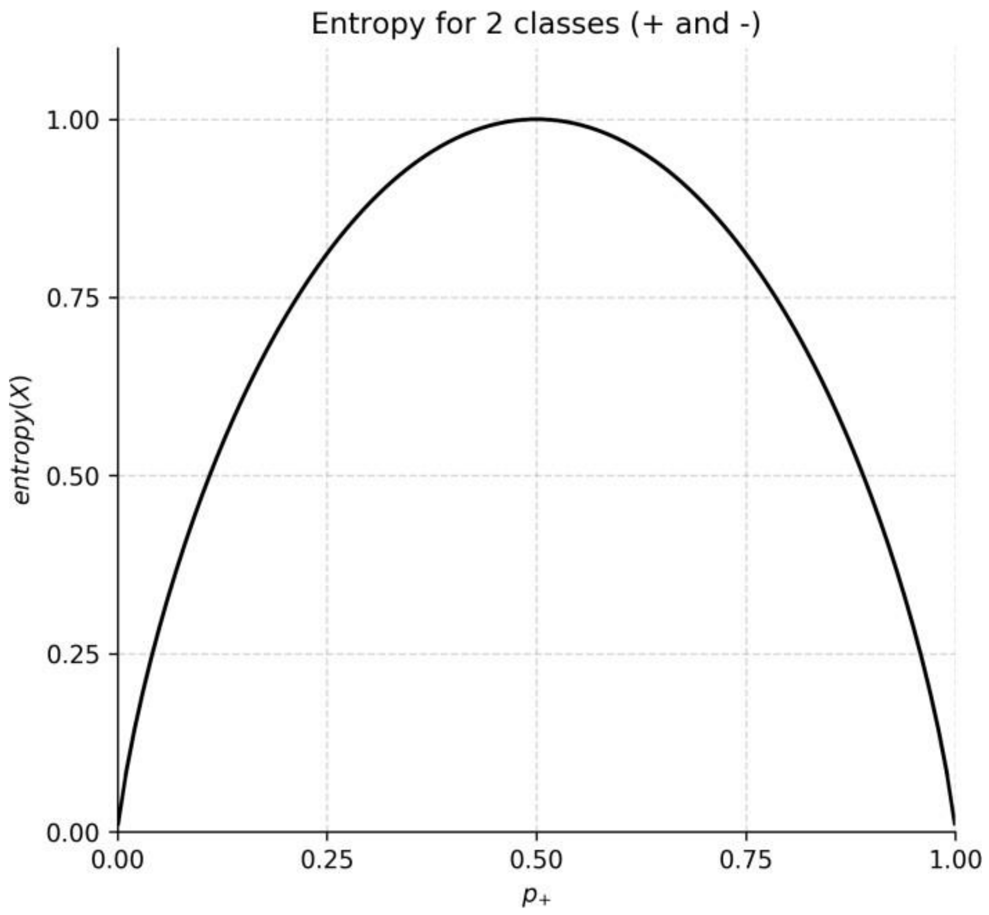
- Entropy measures the degree of randomness in data

- For a set of samples X with k classes:

$$entropy(X) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

where p_i is the proportion of elements of class i

- Lower entropy implies greater predictability!



Information Gain

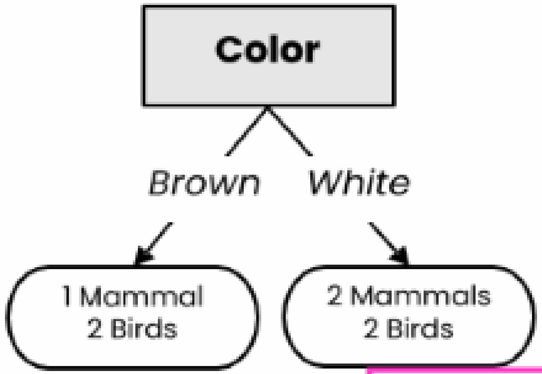
- The information gain of an attribute a is the expected reduction in entropy due to splitting on values of a:

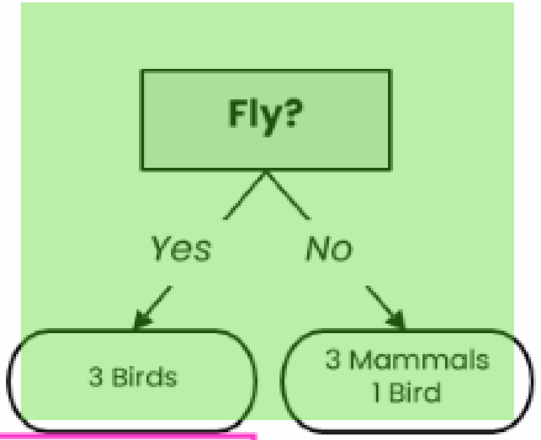
$$gain(X, a) = \overset{y}{\boxed{entropy(X)}} - \sum_{v \in Values(a)} \frac{|X_v|}{|X|} entropy(X_v)$$

where X_v is the subset of X for which $a = v$

Best attribute = highest information gain

Does it fly?	Color	Class
No	Brown	**Mammal**
No	White	**Mammal**
Yes	Brown	**Bird**
Yes	White	**Bird**
No	White	**Mammal**
No	Brown	**Bird**
Yes	White	**Bird**





$$entropy\ (X) = -p_{\text{mammal}} \log_2 p_{\text{mammal}} - p_{\text{bird}} \log_2 p_{\text{bird}} = \boxed{\boxed{-\frac{3}{7} \log_2 \frac{3}{7}} - \boxed{\frac{4}{7} \log_2 \frac{4}{7}}} \approx \boxed{0.985}$$

$$entropy\ (X_{color=brown}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \approx \boxed{0.918} \qquad entropy\ (X_{color=white}) = 1$$

$$\boldsymbol{gain\ (X, color) = 0.985 - \frac{3}{7} \cdot 0.918 - \frac{4}{7} \cdot 1 \approx 0.020}$$

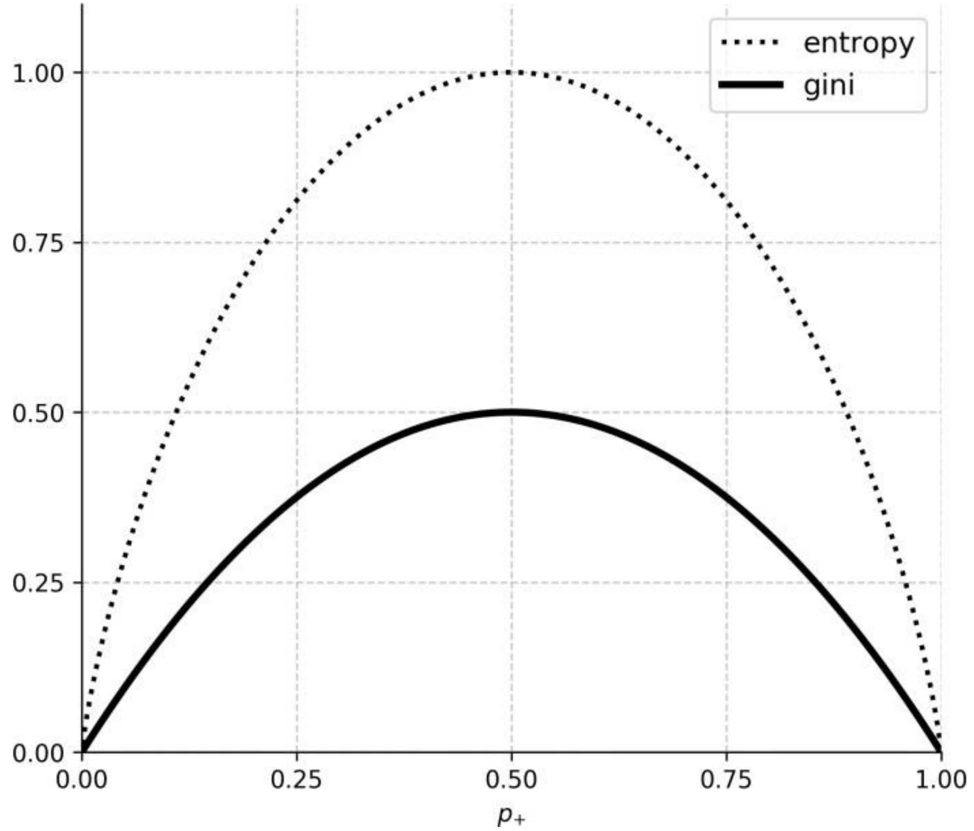
$$entropy\ (X_{fly=yes}) = 0 \qquad entropy\ (X_{fly=no}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \approx 0.811$$

$$\boldsymbol{gain\ (X, fly) = 0.985 - \frac{3}{7} \cdot 0 - \frac{4}{7} \cdot 0.811 \approx 0.521}$$

Gini Impurity

- Gini impurity measures how often a randomly chosen example would be incorrectly labeled if it was randomly labeled according to the label distribution

Error of classifying randomly picked fruit with randomly picked label



- For a set of samples X with k classes:

$$gini(X) = 1 - \sum_{i=1}^{k} p_i^2$$

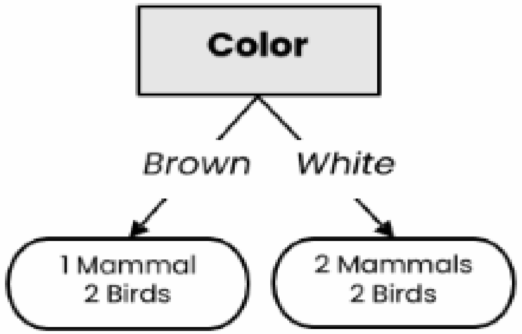
where p_i is the proportion of elements of class i

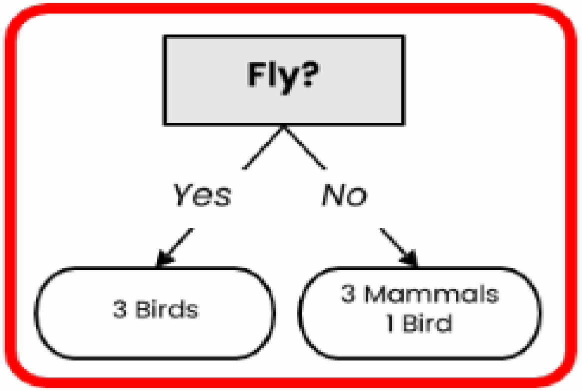
- Can be used as an alternative to entropy for selecting attributes!

Best attribute = highest impurity decrease

In practice, we compute $gini(X)$ only once!

Does it fly?	Color	Class
No	Brown	Mammal
No	White	Mammal
Yes	Brown	Bird
Yes	White	Bird
No	White	Mammal
No	Brown	Bird
Yes	White	Bird





$$gini\ (X) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 \approx 0.489$$

$$gini\ (X_{color=brown}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \approx 0.444$$

$$gini\ (X_{color=white}) = 0.5$$

$$\triangle\ \boldsymbol{gini\ (X, color)} = \boxed{\mathbf{0.489}} - \boxed{\frac{\mathbf{3}}{\mathbf{7}} \cdot \mathbf{0.444}} - \boxed{\frac{\mathbf{4}}{\mathbf{7}} \cdot \mathbf{0.5}} \approx \mathbf{0.013}$$

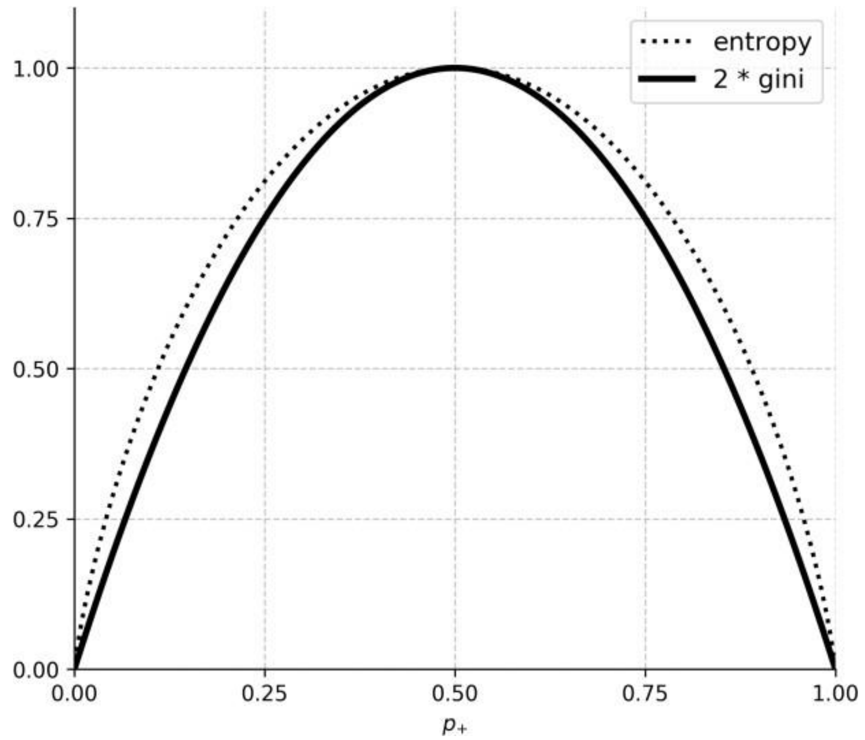
$$gini\ (X_{fly=yes}) = 0$$

$$gini\ (X_{fly=no}) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 \approx 0.375$$

$$\triangle\ \boldsymbol{gini\ (X, fly)} = \mathbf{0.489} - \frac{\mathbf{3}}{\mathbf{7}} \cdot \mathbf{0} - \frac{\mathbf{4}}{\mathbf{7}} \cdot \mathbf{0.375} \approx \boxed{\mathbf{0.274}}$$

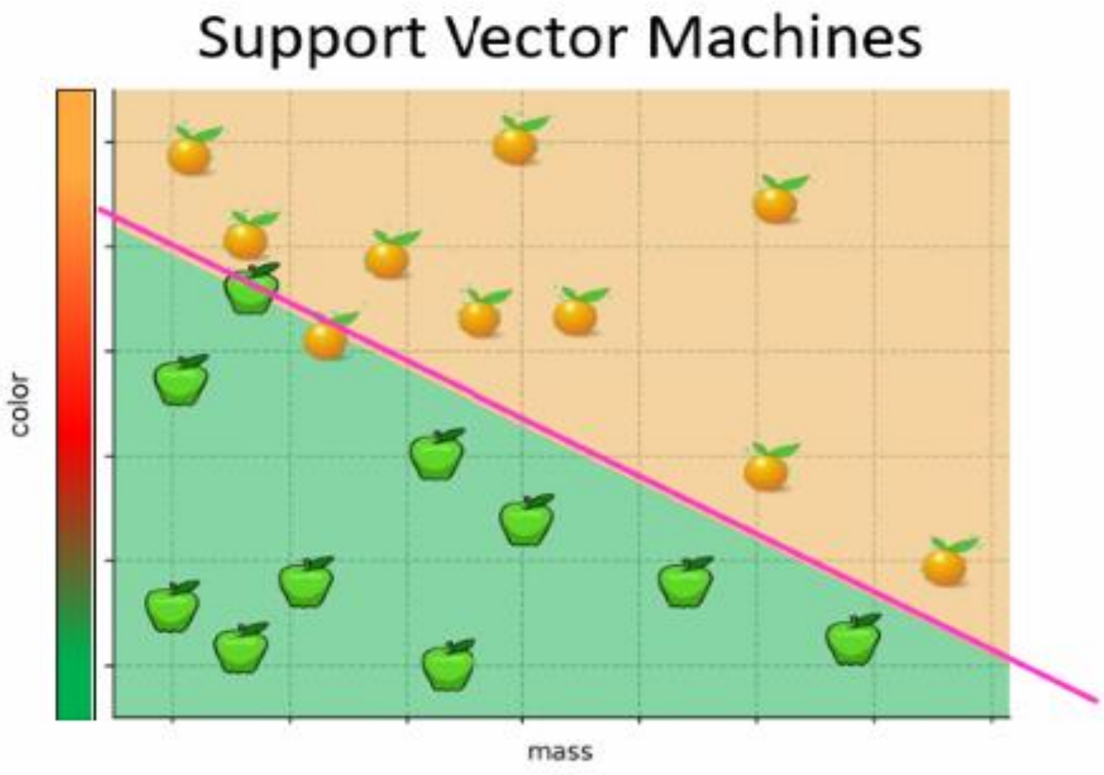
Entropy versus Gini Impurity

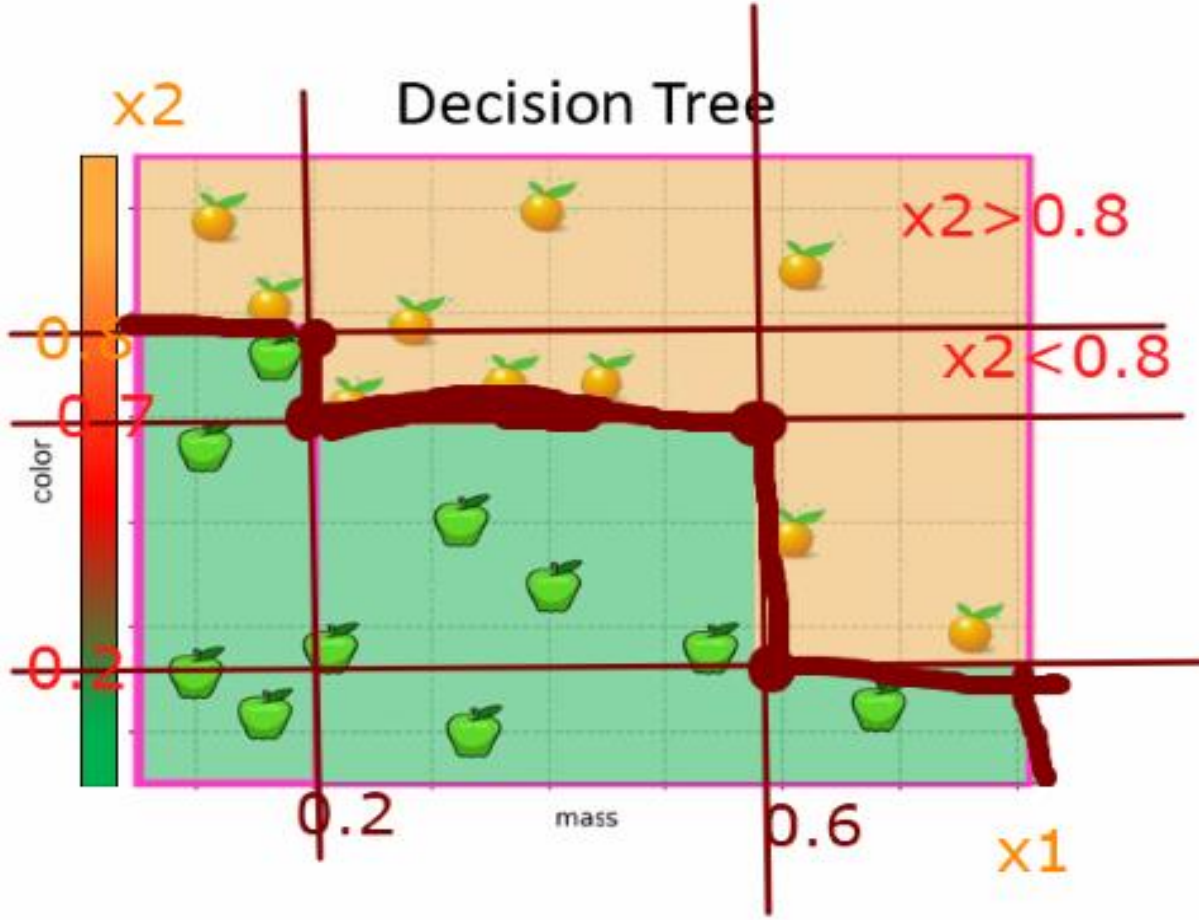
- Entropy and Gini Impurity give similar results in practice
  - ➢ They only disagree in about 2% of cases
    "Theoretical Comparison between the Gini Index and Information Gain Criteria" [Răileanu & Stoffel, AMAI 2004]
  - ➢ Entropy might be slower to compute, because of the log



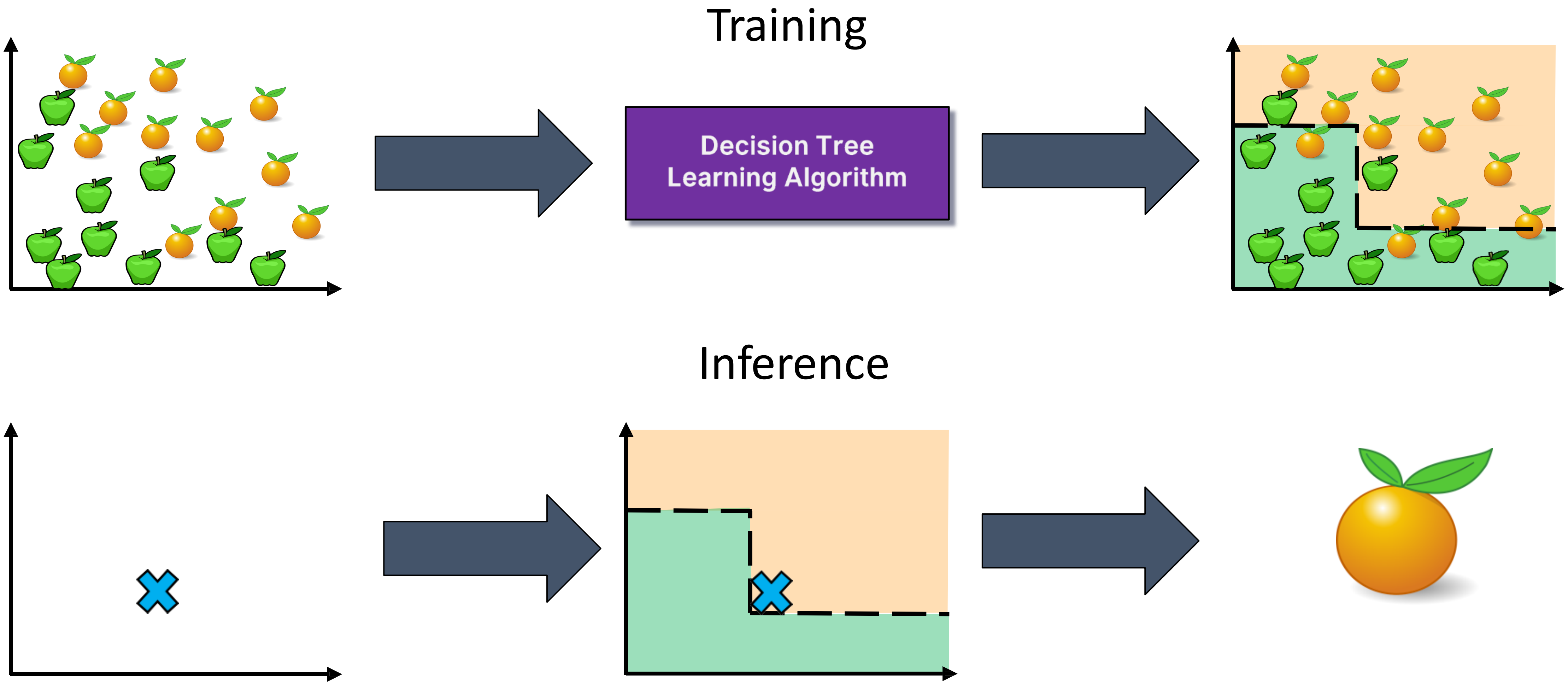
Decision Boundaries

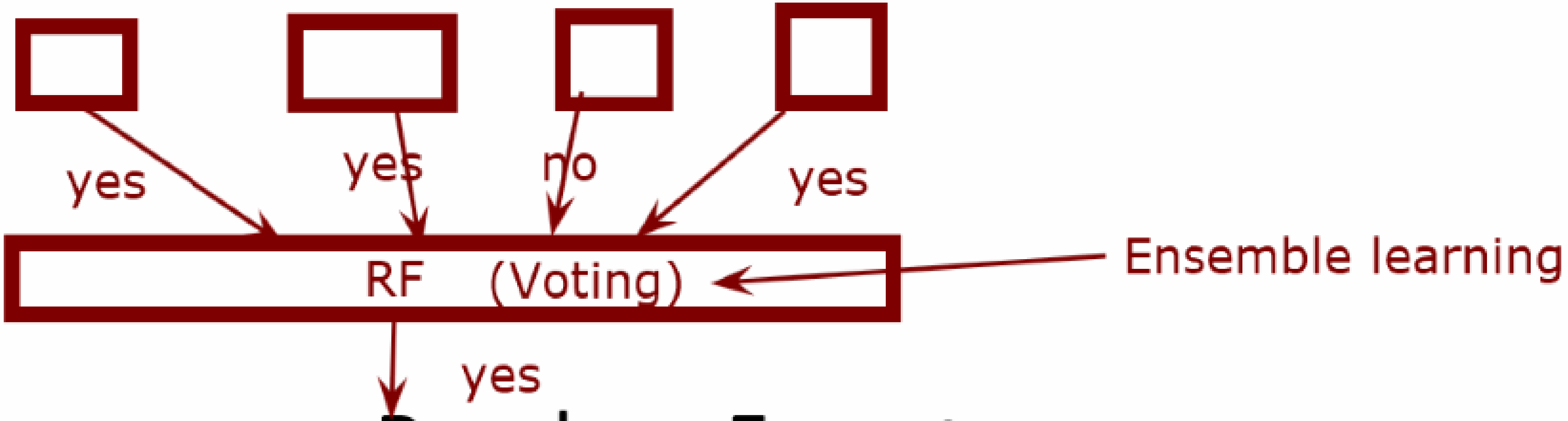
- Decision trees produce non-linear decision boundaries





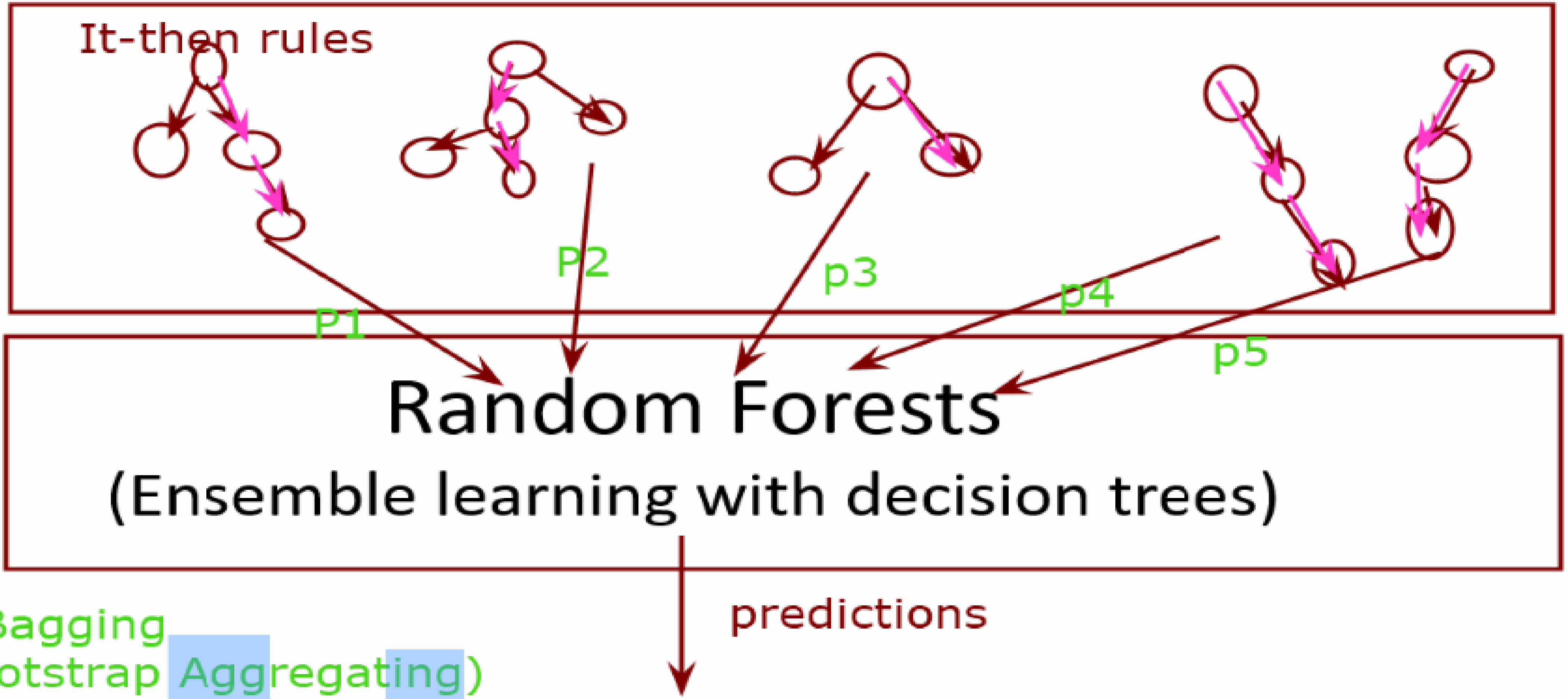
Decision Trees: Training and Inference





Random Forests

(Ensemble learning with decision trees)



1. Bagging
(Bootstrap Aggregating)

2. Random Subspace Method
(Feature Bagging)

Random Forests

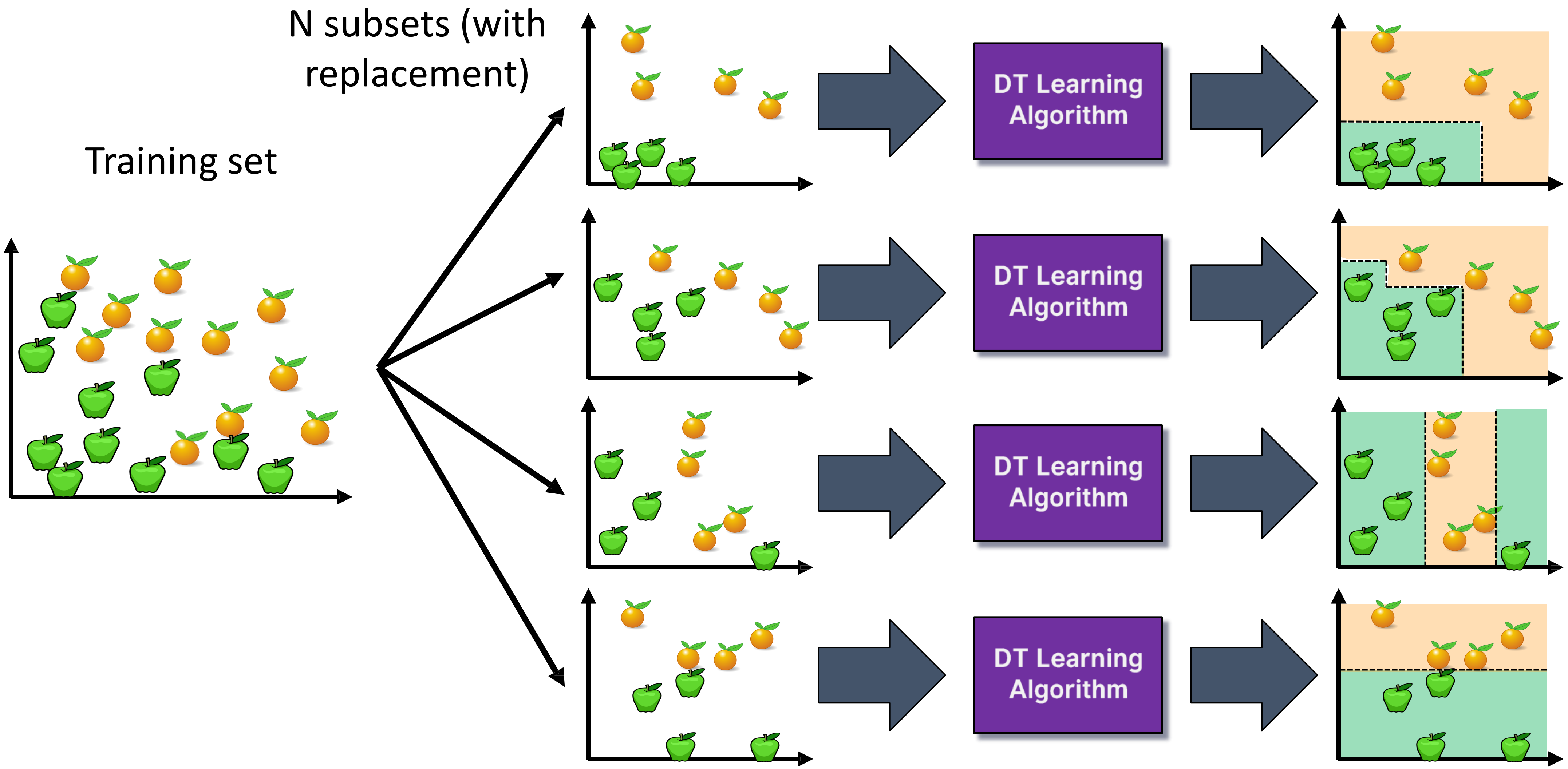
- Random Forests:
  - ➢ Instead of building a single decision tree and use it to make predictions, build many slightly different trees and combine their predictions

- We have a single data set, so how do we obtain slightly different trees?

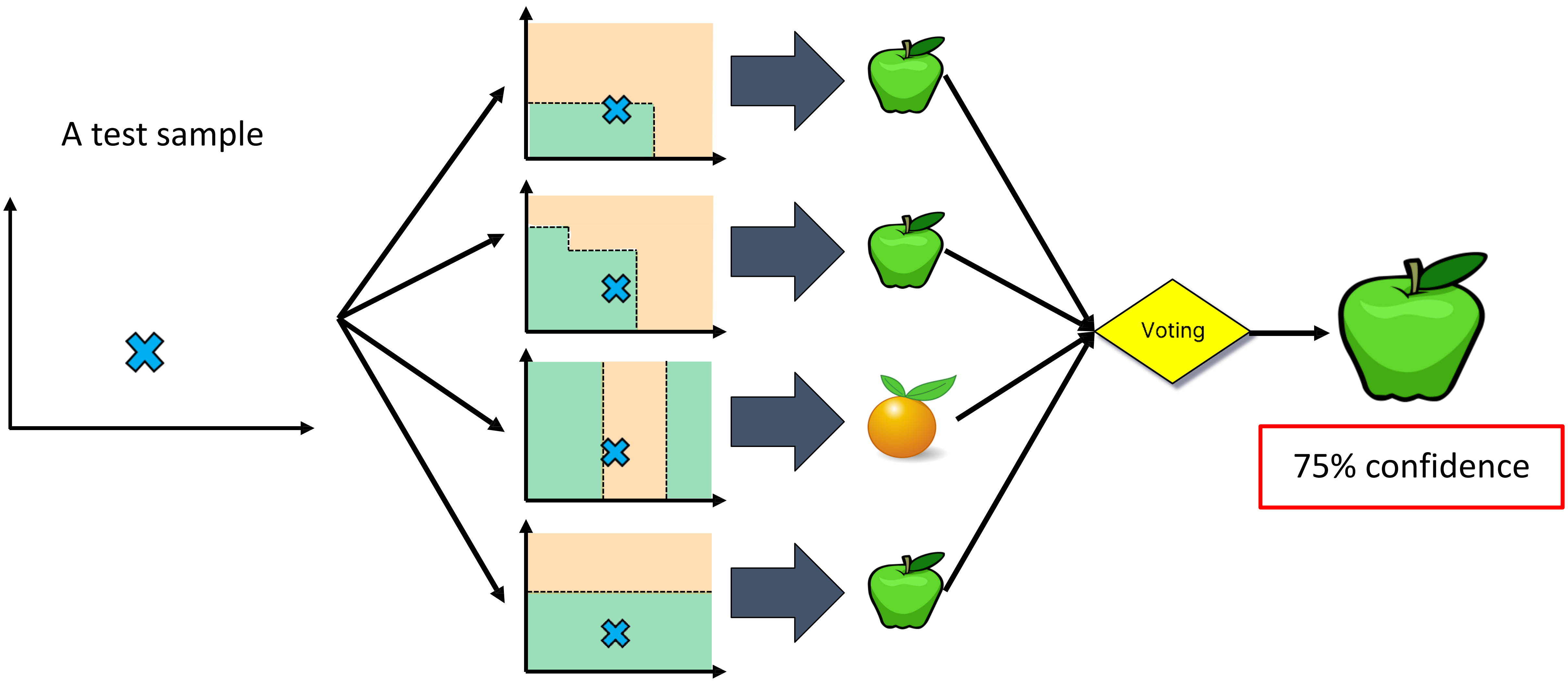
1. Bagging (**B**ootstrap **Agg**regat**ing**):
   - ➢ Take random subsets of data points from the training set to create N smaller data sets
   - ➢ Fit a decision tree on each subset

2. Random Subspace Method (also known as Feature Bagging):
   - ➢ Fit N different decision trees by constraining each one to operate on a random subset of features

Bagging at training time



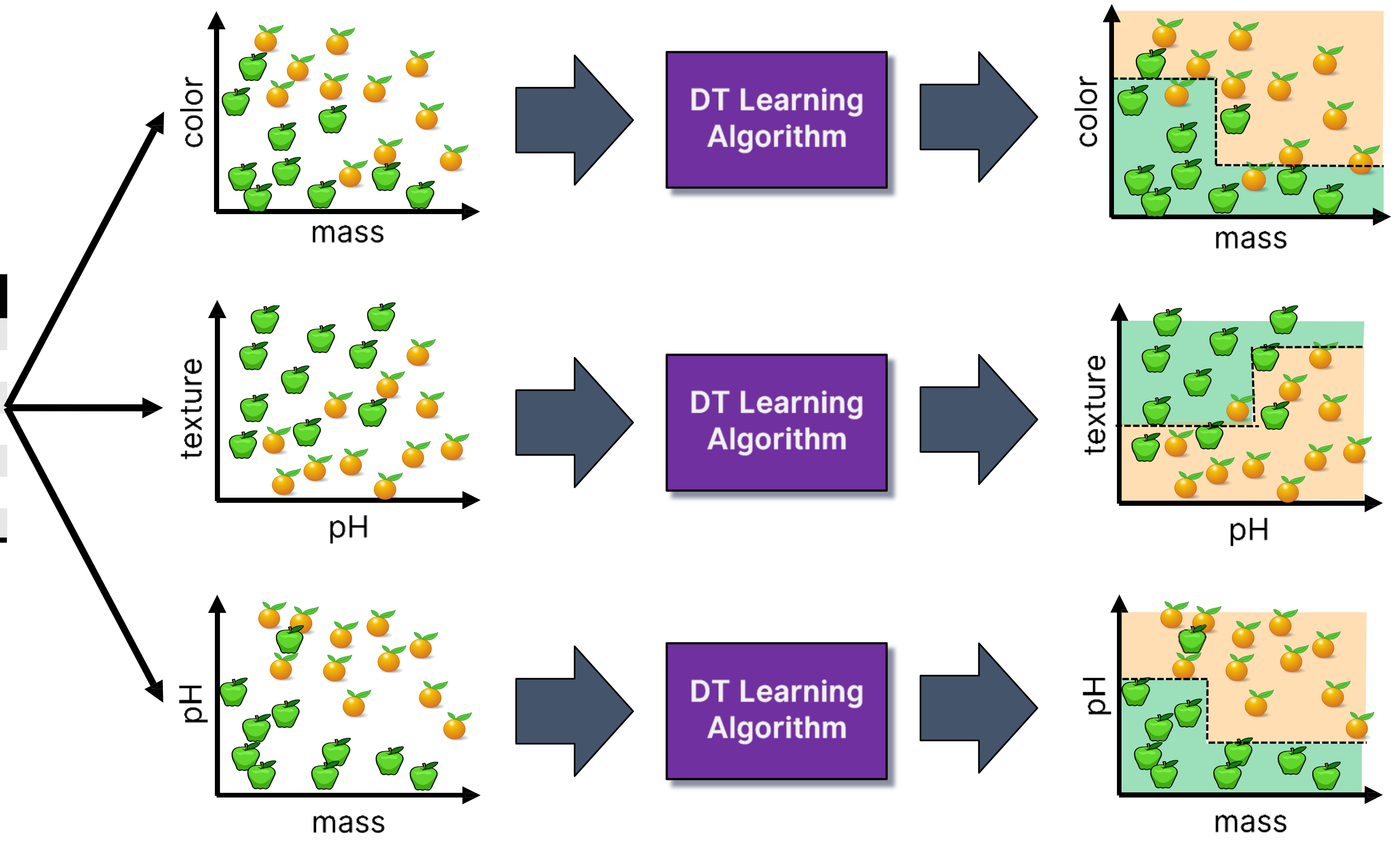
Bagging at inference time



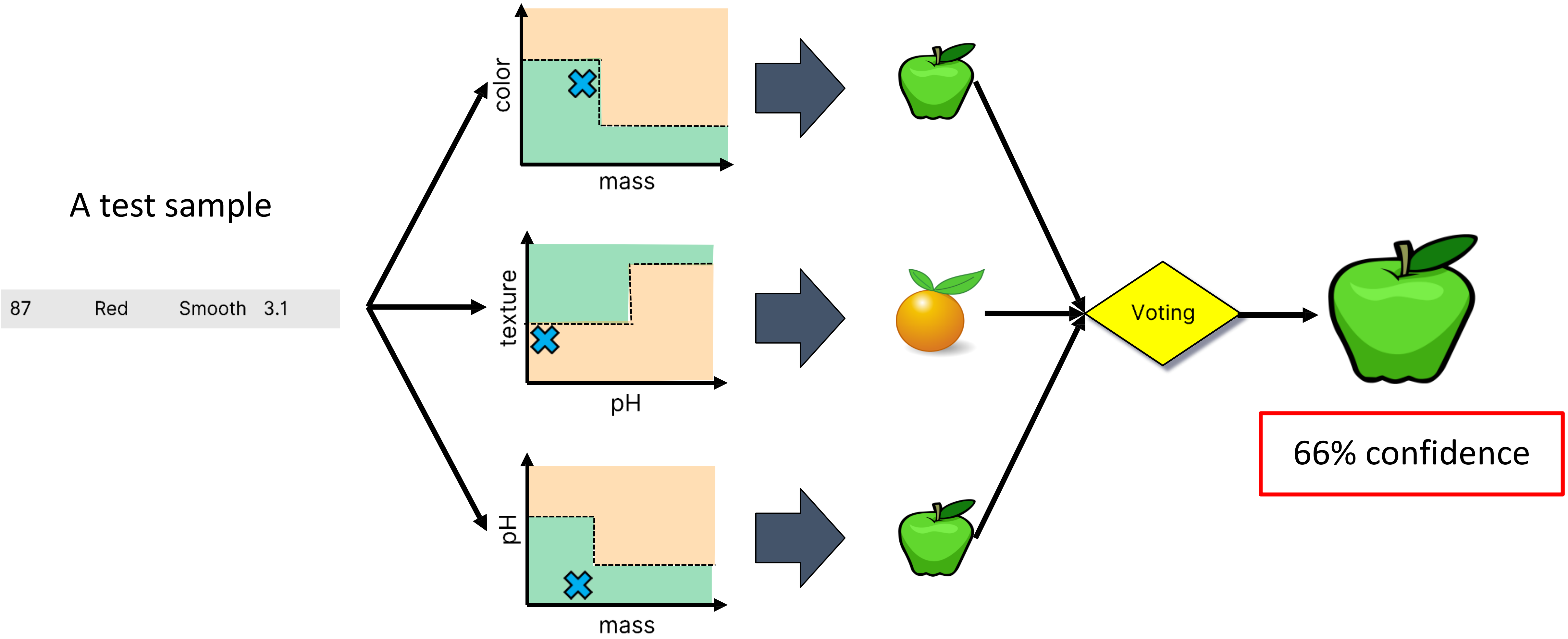
Random Subspace Method at training time

Training data

Mass (g)	Color	Texture	pH	Label
84	Green	Smooth	3.5	**Apple**
121	Orange	Rough	3.9	**Orange**
85	Red	Smooth	3.3	**Apple**
101	Orange	Smooth	3.7	**Orange**
111	Green	Rough	3.5	**Apple**
...				
117	Red	Rough	3.4	**Orange**

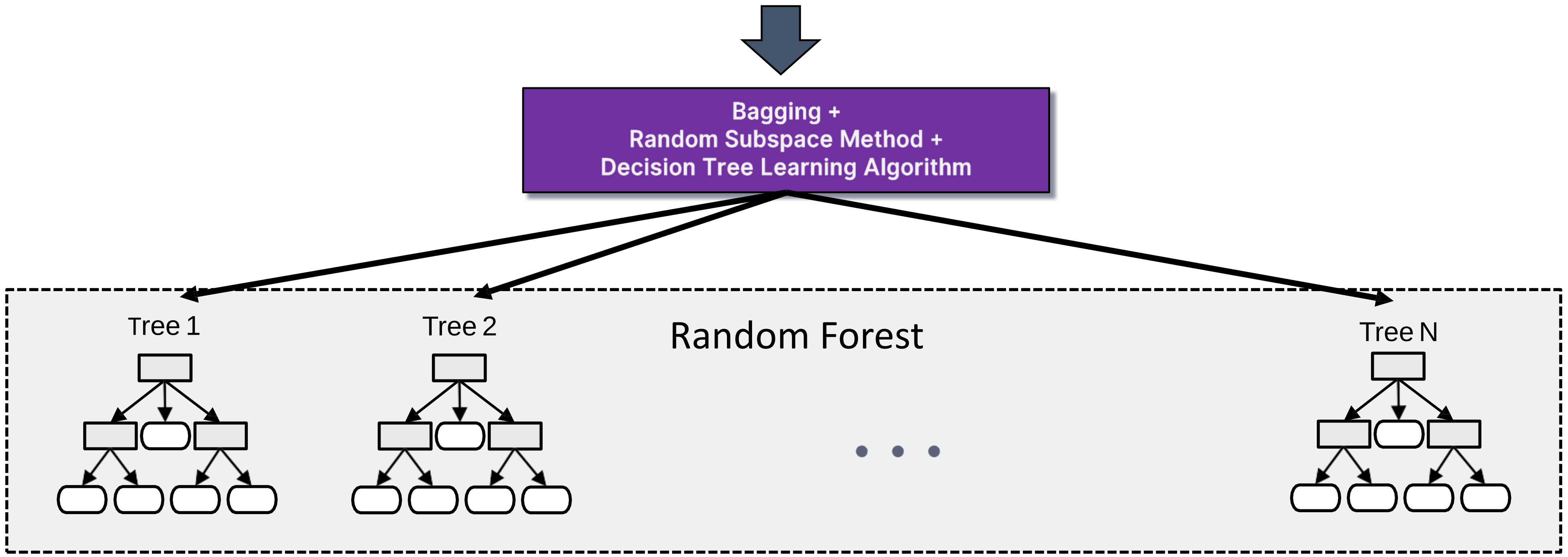


Random Subspace Method at inference time



Random Forests

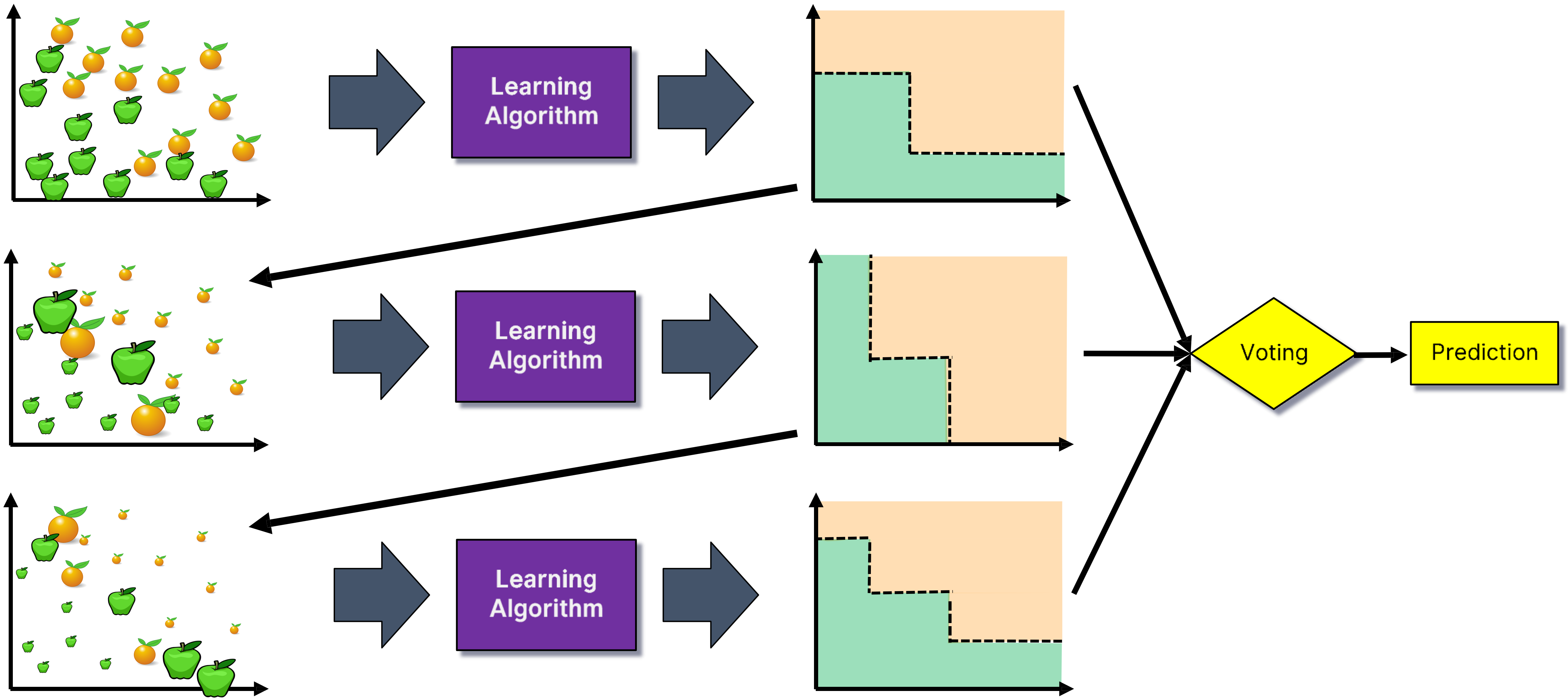
Mass (g)	Color	Texture	pH	Label
84	Green	Smooth	3.5	**Apple**
121	Orange	Rough	3.9	**Orange**
85	Red	Smooth	3.3	**Apple**
101	Orange	Smooth	3.7	**Orange**
111	Green	Rough	3.5	**Apple**
...				
117	Red	Rough	3.4	**Orange**

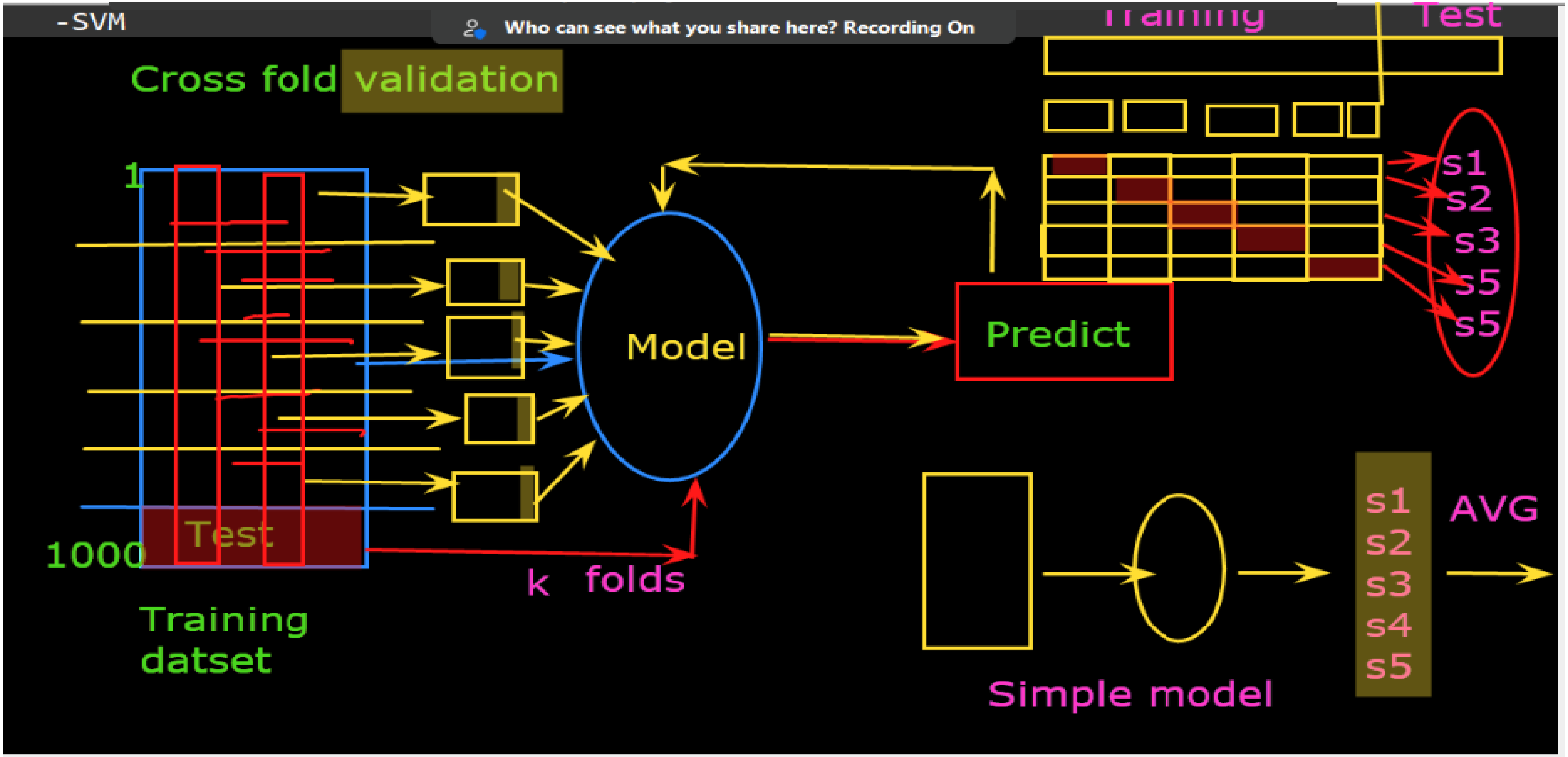


Ensemble Learning

- Ensemble Learning:
  - Method that combines multiple learning algorithms to obtain performance improvements over its components
- **Random Forests** are one of the most common examples of ensemble learning
- Other commonly-used ensemble methods:
  - Bagging: multiple models on random subsets of data samples
  - Random Subspace Method: multiple models on random subsets of features
  - Boosting: train models iteratively, while making the current model focus on the mistakes of the previous ones by increasing the weight of misclassified samples

Boosting


Learning Algorithm
Learning Algorithm
Learning Algorithm
Voting
Prediction


-SVM
Who can see what you share here? Recording On
Cross fold validation
Training
Test
1
1000
Test
Training
datset
Model
k folds
Predict
s1
s2
s3
s5
s5
s1
s2
s3
s4
s5
AVG
Simple model