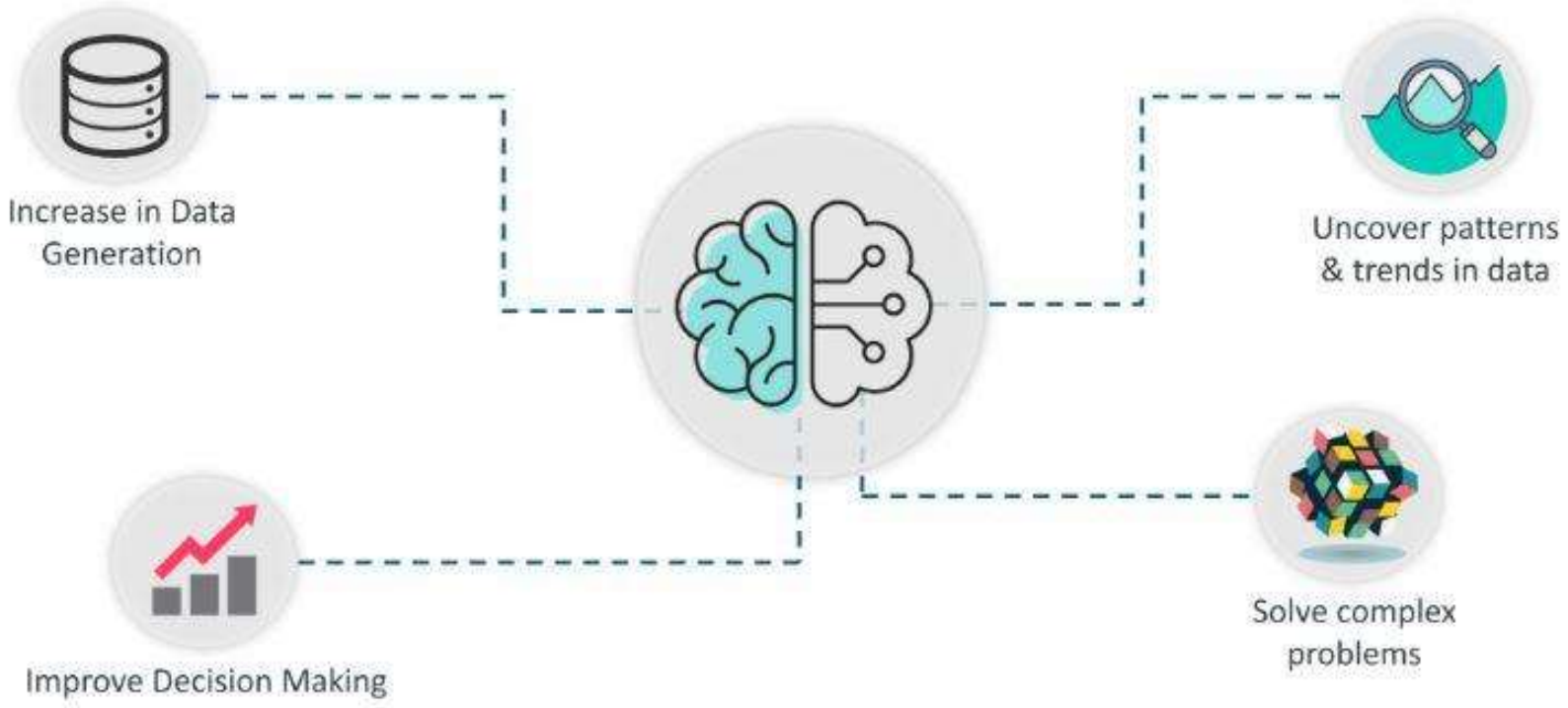
Introduction to ML

(Unit II _Part II)

Dr. Mukti Padhya

Assistant Professor, NFSU

Importance of Machine Learning

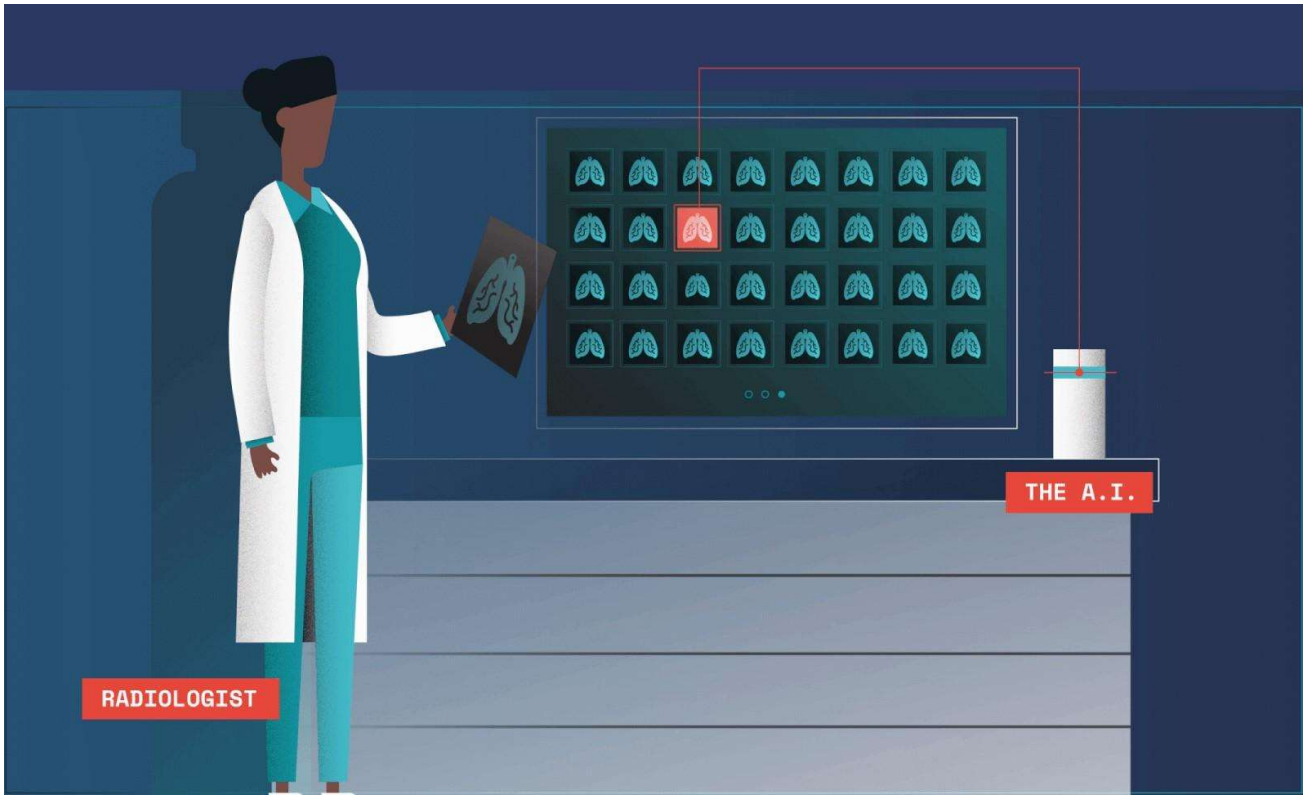


Importance of Machine Learning

- Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better.
- Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data.

Example : ML
Traffic Prediction

Example : ML
Illness Prediction



Example : ML

Smart Cart : E - Commerce

Importance of Machine Learning

- **Netflix's Recommendation Engine:**
- **Facebook's Auto-tagging feature:**
- **Amazon's Alexa:**
- **Google's Spam Filter:**

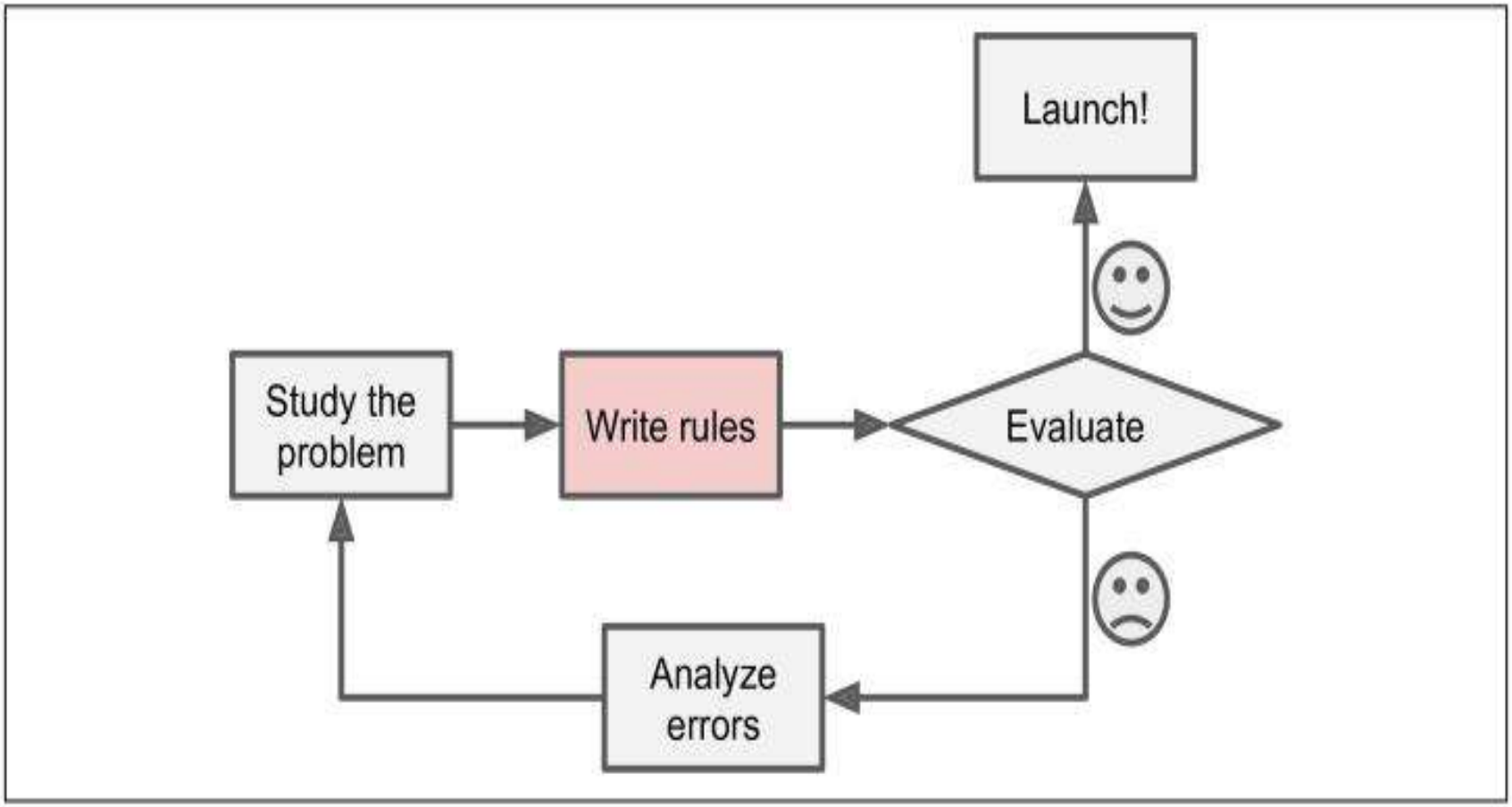
Spam Filter : Traditional Approach

- First you would look at what spam typically looks like. You might notice that some words or phrases (such as “credit card,” “free,” and “amazing offer”) tend to come up a lot in the subject. Perhaps you would also notice a few other patterns in the sender’s name, the email’s body, and so on.

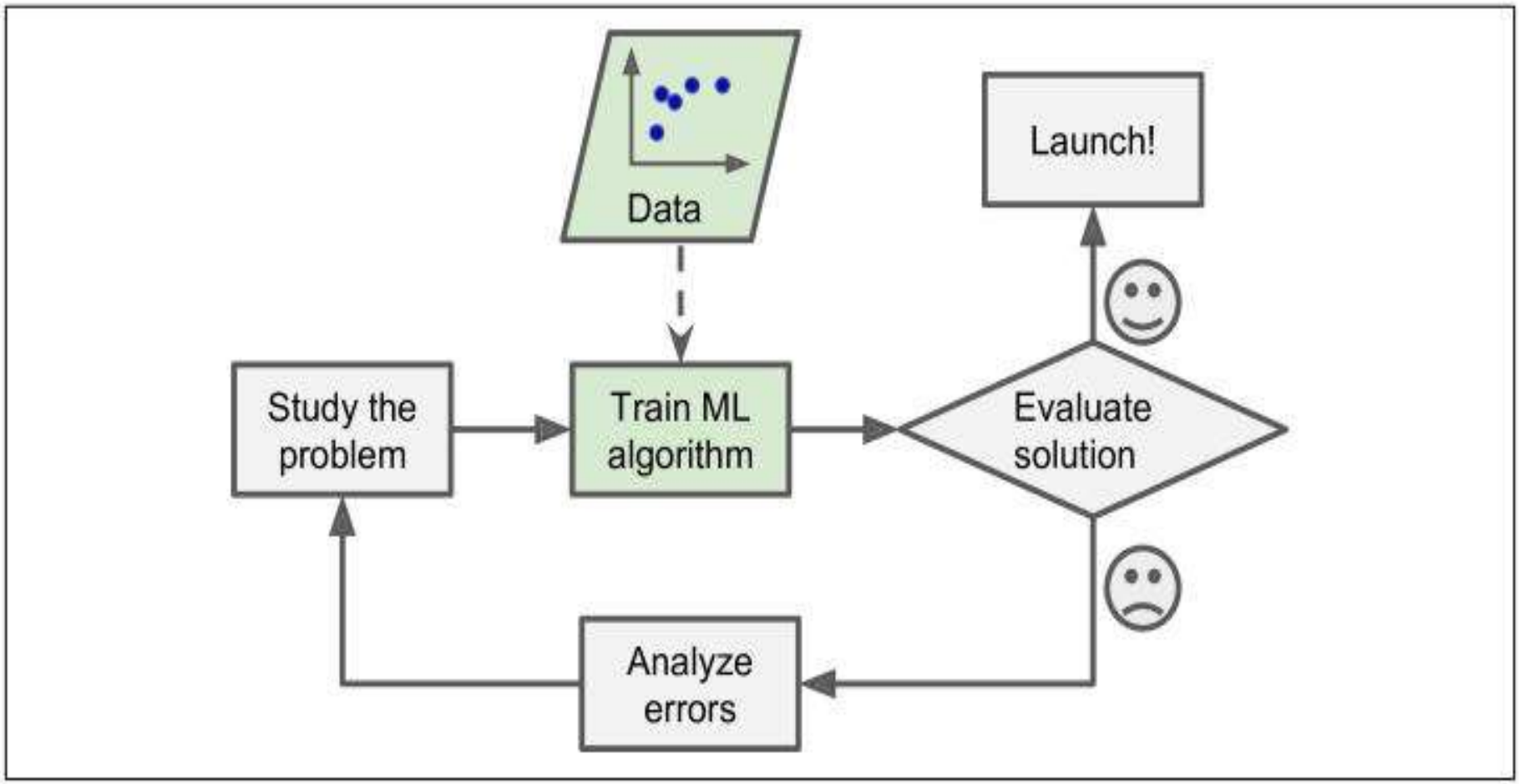
- You would write a detection algorithm for each of the patterns that you noticed, and your program would flag emails as spam if a number of these patterns are detected.

- You would test your program, and repeat steps 1 and 2 until it is good enough.

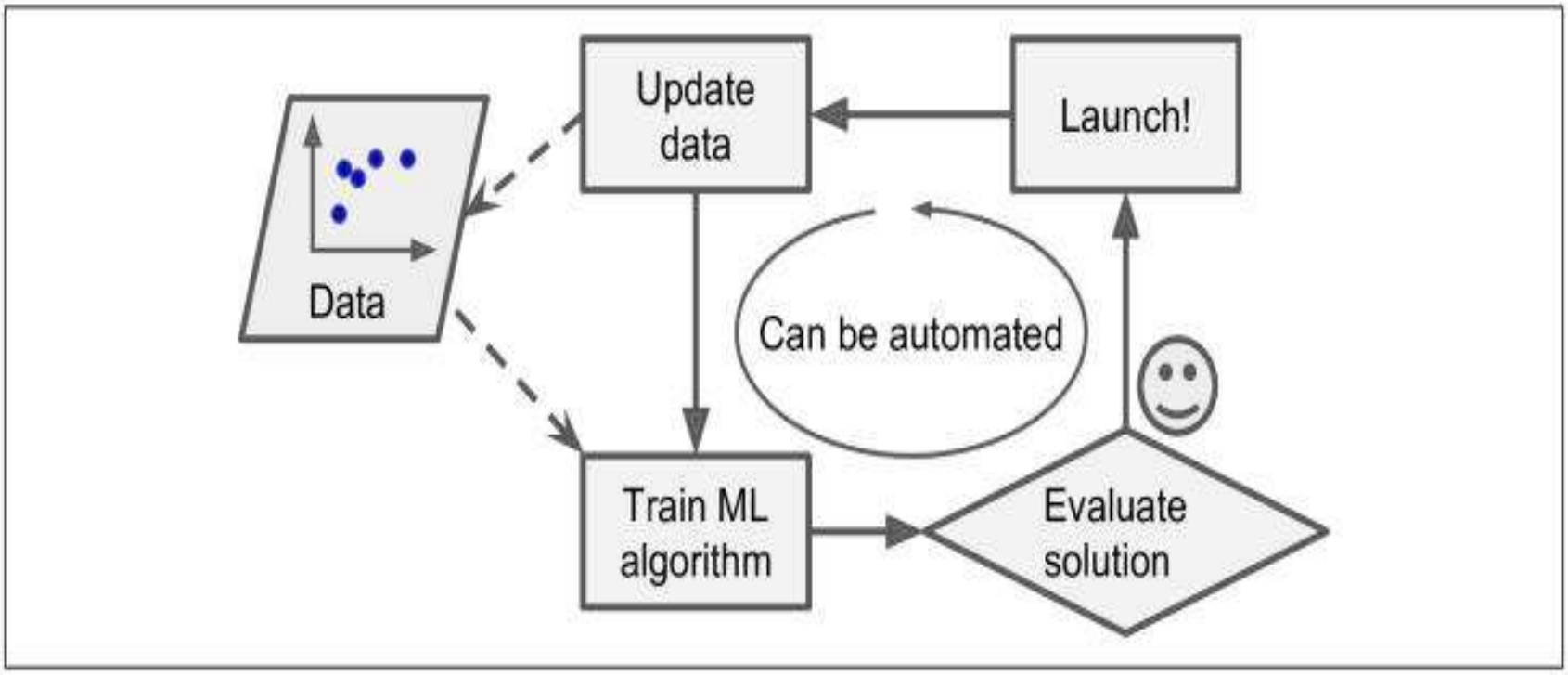
Spam Filter : Traditional Approach



Spam Filter : ML Approach



Automatically adapting change



What is Machine Learning?

- Machine Learning is the science (and art) of programming computers so they can learn from data

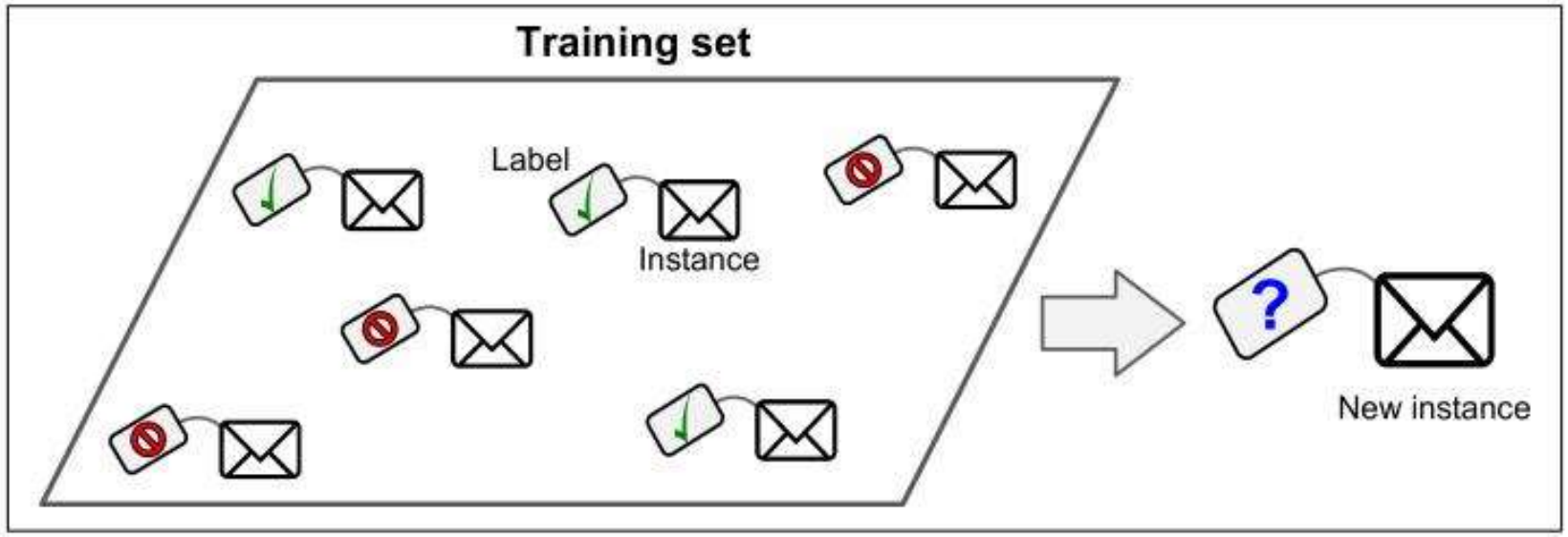
- *A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.* —Tom Mitchell, 1997

- *Machine learning is a subset of Artificial Intelligence (AI) which provides machines the ability to learn automatically & improve from experience without being explicitly programmed to do so.*

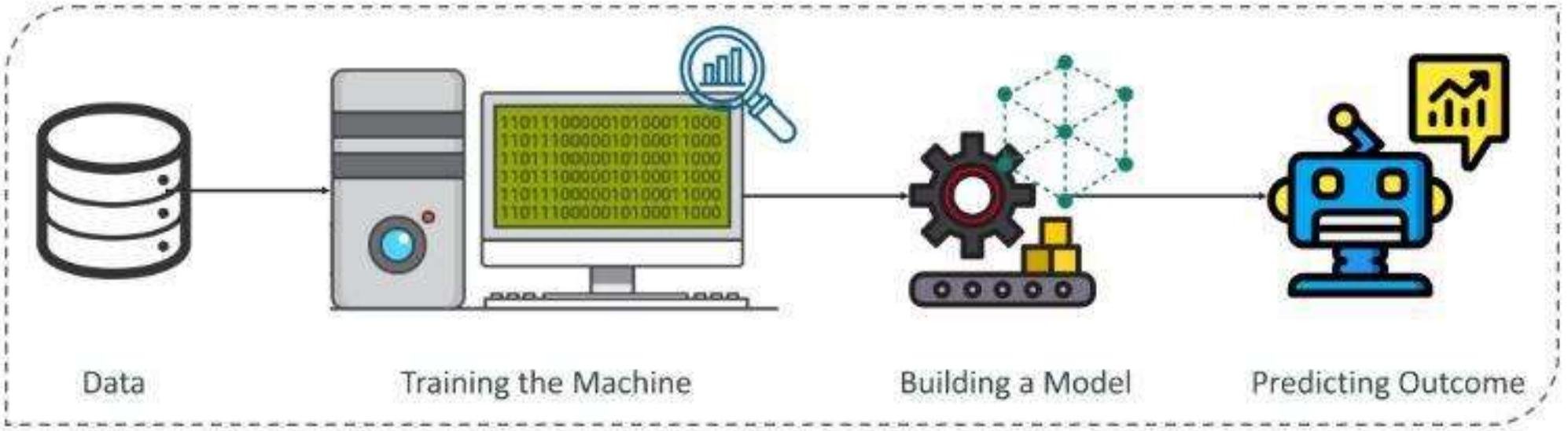
- In the sense, it is the practice of getting Machines to solve problems by gaining the ability to think.

- The first ML application was the **Spam Filter (1990)**

Example: Spam Filtering



What is Machine Learning?



Terminology of Machine Learning

- **Algorithm:** A Machine Learning algorithm is a set of rules and statistical techniques used to learn patterns from data and draw significant information from it. It is the logic behind a Machine Learning model. An example of a Machine Learning algorithm is the Linear Regression algorithm.

- **Model:** A model is trained by using a Machine Learning Algorithm. An algorithm maps all the decisions that a model is supposed to take based on the given input, in order to get the correct output.

- **Predictor Variable:** It is a feature(s) of the data that can be used to predict the output.

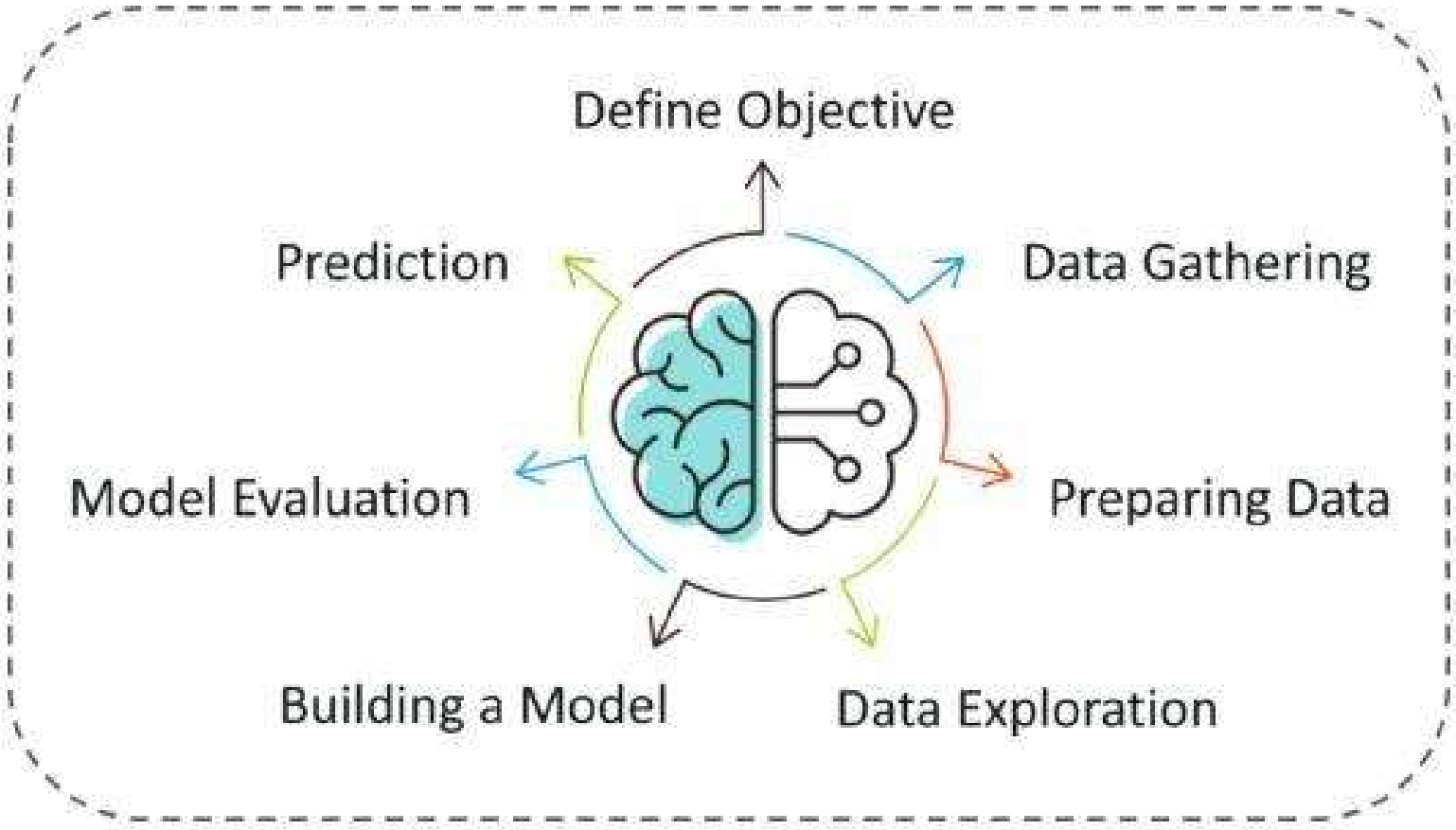
Terminology of Machine Learning

- **Response Variable:** It is the feature or the output variable that needs to be predicted by using the predictor variable(s).

- **Training Data:** The Machine Learning model is built using the training data. The training data helps the model to identify key trends and patterns essential to predict the output.

- **Testing Data:** After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.

Machine Learning Process



Challenges of Machine Learning

- **Insufficient Quantity of Training Data**
  - it takes a lot of data for most Machine Learning algorithms to work properly.
  - Even for very simple problems you typically need thousands of examples, and for complex problems such as image or speech recognition you may need millions of examples
- **Nonrepresentative Training Data**
  - In order to generalize well, it is crucial that your training data be representative of the new cases you want to generalize to.
  - if the sample is too small, you will have sampling noise (i.e., nonrepresentative data as a result of chance), but even very large samples can be nonrepresentative if the sampling method is flawed. This is called sampling bias.

Challenges of Machine Learning

- **Poor-Quality Data**
  - If your training data is full of errors, outliers, and noise (e.g., due to poor quality measurements), it will make it harder for the system to detect the underlying patterns, so your system is less likely to perform well.
  - It is often well worth the effort to spend time cleaning up your training data.

- **Slow Implementation:** Machine Learning is a Complex Process

- **Imperfections in the Algorithm When Data Grows**

Challenges of Machine Learning

- **Overfitting the Training Data**
  - The model performs well on the training data, but it does not generalize well.
  - Constraining a model to make it simpler and reduce the risk of overfitting is called regularization
- **Underfitting the Training Data**
  - It occurs when your model is too simple to learn the underlying structure of the data
  - So its predictions are bound to be inaccurate, even on the training examples
  - To reduce this, a more powerful model, with more parameters and better features must be selected

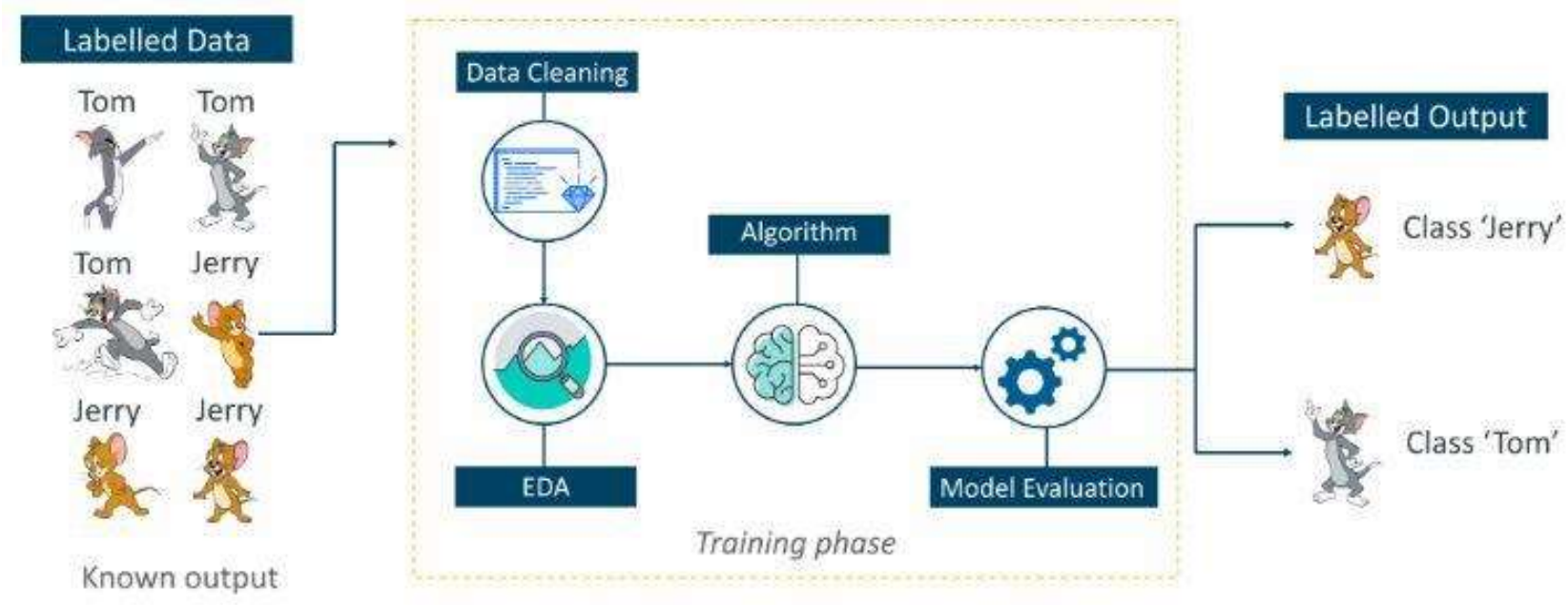
Types of ML

- Machine Learning systems can be classified according to the amount and type of supervision they get during training.

- There are four major categories:
  - Supervised learning,
  - Unsupervised learning,
  - Reinforcement Learning.

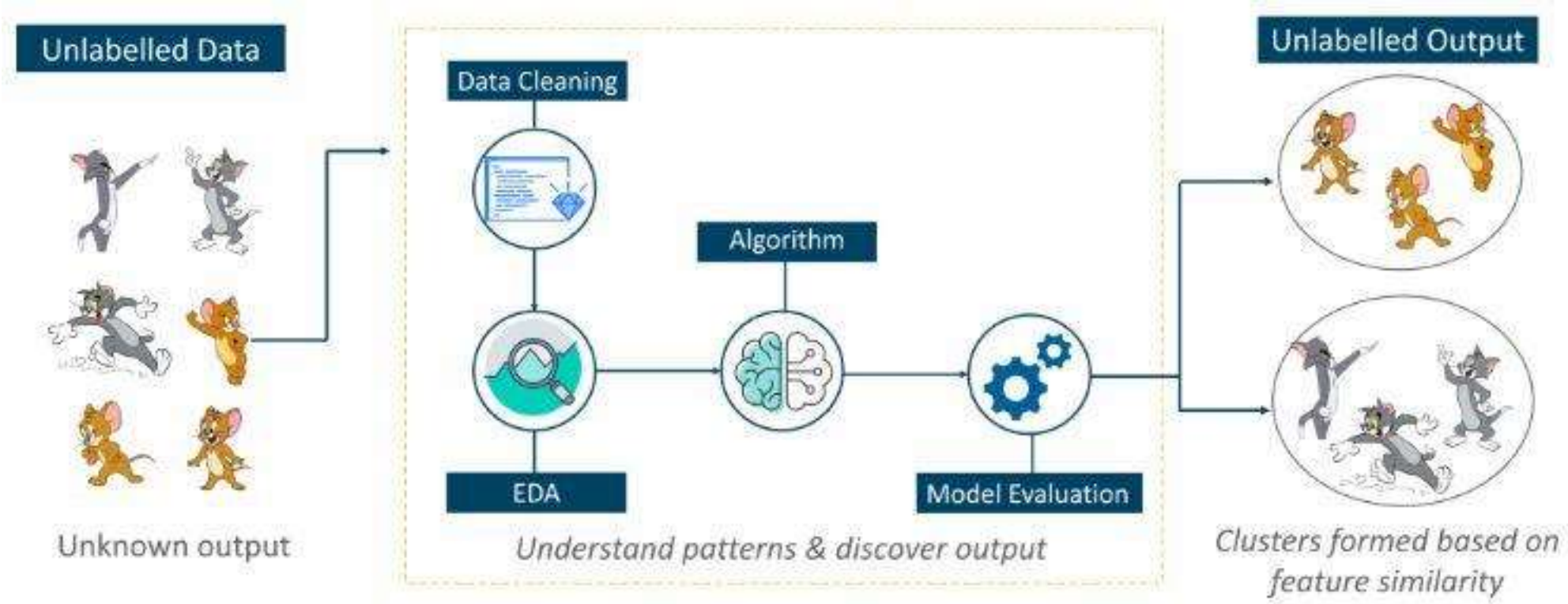
Supervised learning

"Supervised learning is a technique in which we teach or train the machine using data which is well labeled"



Unsupervised learning

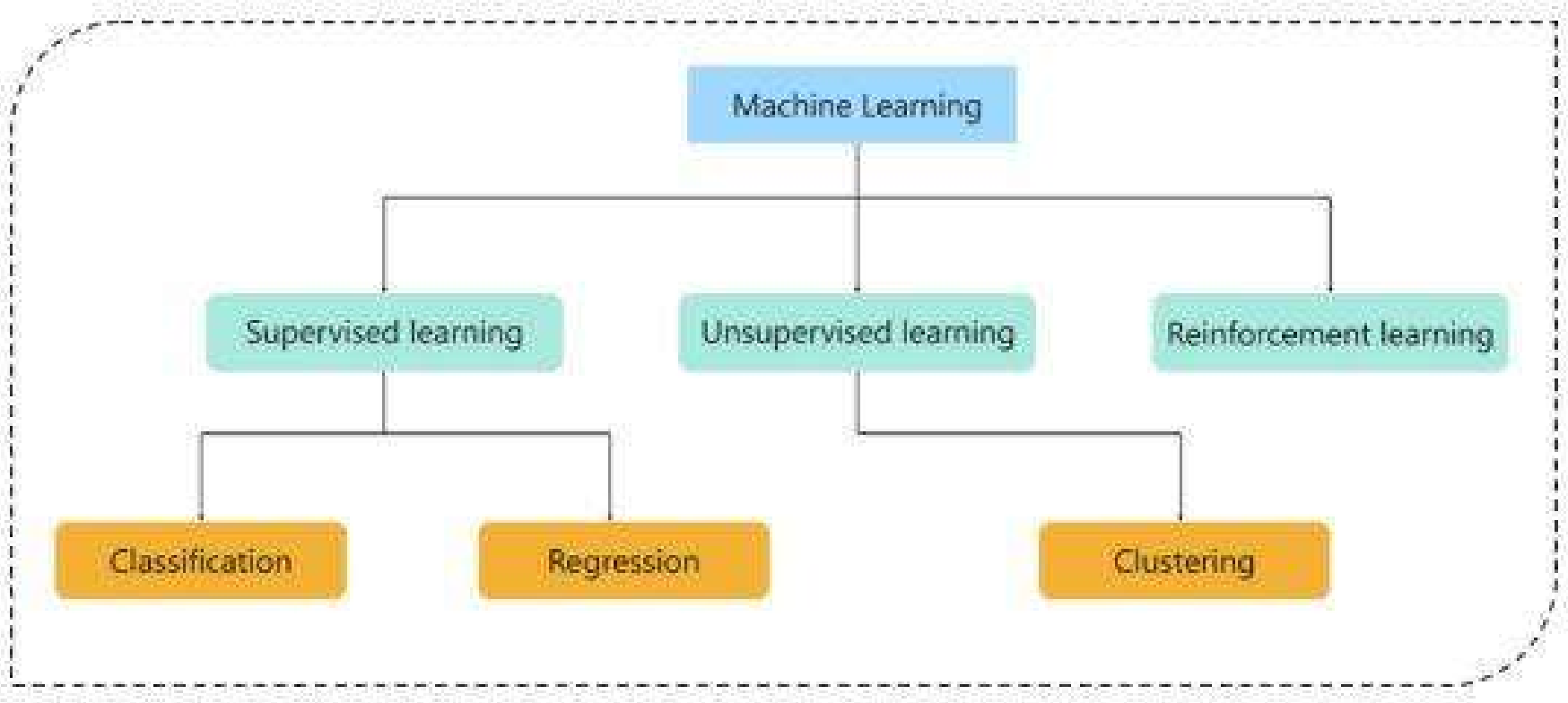
"Unsupervised learning involves training by using unlabeled data and allowing the model to act on that information without guidance."



Reinforcement Learning

"Reinforcement Learning is a part of Machine learning where an agent is put in an environment and he learns to behave in this environment by performing certain actions and observing the rewards which it gets from those actions."

Type of Problems Solved Using ML



Regression vs Classification vs Clustering

Regression

- Supervised Learning
- Output is a continuous quantity
- Main aim is to forecast or predict
- Eg: Predict stock market price
- Algorithm: Linear Regression

Classification

- Supervised Learning
- Output is a categorical quantity
- Main aim is to compute the category of the data
- Eg: Classify emails as spam or non-spam
- Algorithm: Logistic Regression

Clustering

- Unsupervised Learning
- Assigns data points into clusters
- Main aim is to group similar items clusters
- Eg: Find all transactions which are fraudulent in nature
- Algorithm: K-means

Supervised learning

- This learning is trained with human supervision with amount and type of supervision they get during training
- The training data you feed to the algorithm includes the desired solutions, called labels
  - i.e. some data is already tagged with the correct answer.
- After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

Supervised learning

- Another task of supervised learning is to predict a target numeric value, such as the price of a car, given a set of features (mileage, age, brand, etc.) called predictors.
- To train the system, you need to give it many examples of cars, including both their predictors and their labels (i.e., their prices)
- This can be achieved by regression (the way of predicting value of one variable from another where the relationship can be defined by a linear model).
- Few regression algorithms can be used for classification and probability of belonging to the same class.

Supervised learning

- **Supervised learning is classified into two categories of algorithms:**
  - **Classification:** A classification problem is when the output variable is a category, such as “Red” or “blue” or “disease” and “no disease”.
  - **Regression:** A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

- Supervised Learning Algorithms
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines (SVMs)
  - Decision Trees and Random Forests
  - Naive Bayes
  - K-Nearest Neighbors

Supervised learning

☐ **Advantages**

- Supervised learning allows collecting data and produces data output from previous experiences.
- Helps to optimize performance criteria with the help of experience.
- Supervised machine learning helps to solve various types of real-world computation problems.

☐ **Disadvantages**

- Classifying big data can be challenging.
- Training for supervised learning needs a lot of computation time. So, it requires a lot of time.

Prediction

- Prediction is like something that may go to happen in the future.
- We identify or predict the missing or unavailable data for a new observation based on the previous data that we have and based on the future assumptions.
- In prediction, the output is a continuous value.
- Machine learning model predictions allow businesses to make highly accurate guesses as to the likely outcomes of a question based on historical data

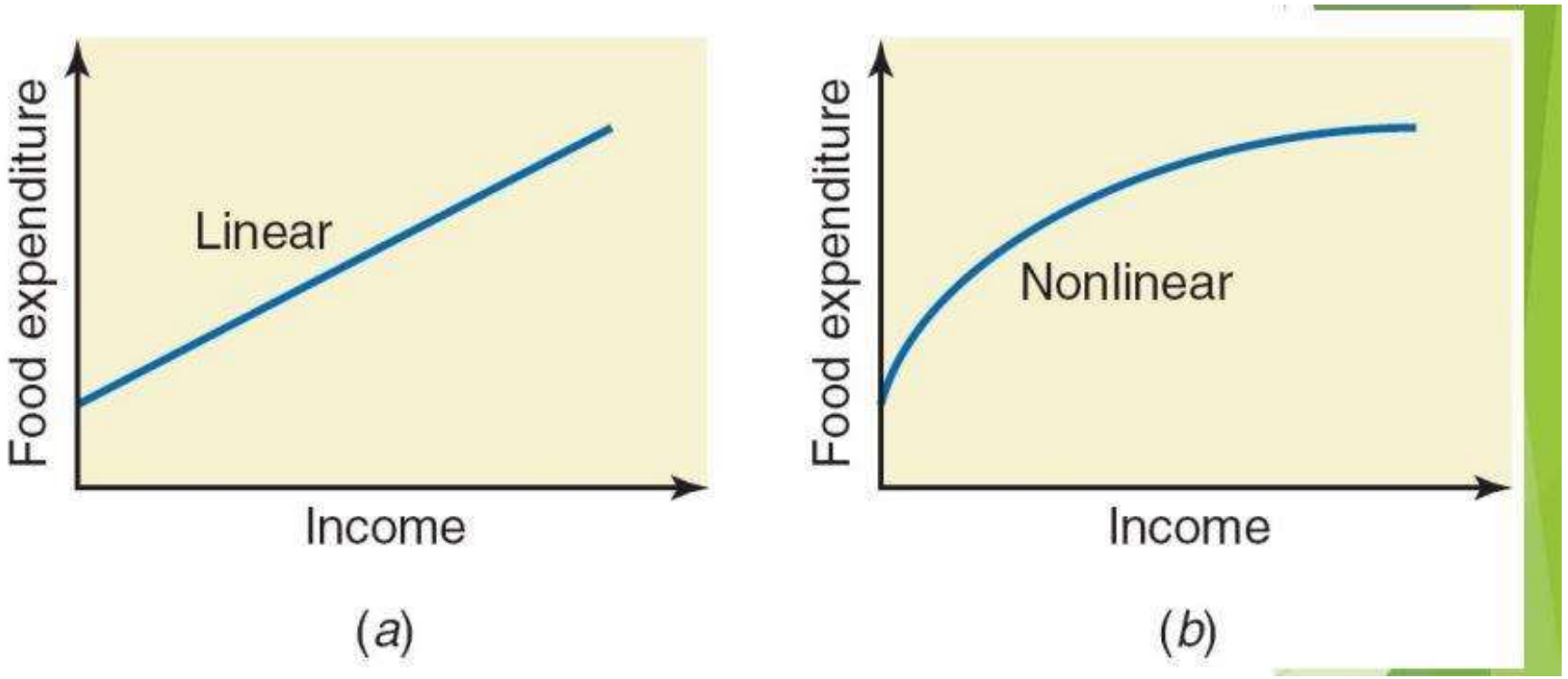
Classification

- Classification is the process of finding a good model that describes the data classes or concepts.
- The purpose of classification is to predict the class of objects whose class label is unknown.
- We can think of Classification as categorizing the incoming new data based on our current or past assumptions that we have made and the data that we already have with us.

Prediction vs. Classification

Prediction	Classification
Prediction is about predicting a missing/unknown element(continuous value) of a dataset	Classification is about determining a (categorial) class (or label) for an element in a dataset
Eg. predict the speed of a car given the distance	Eg. Classifying emails into two classes, spam and non-spam
The predictor is constructed from a training set and its accuracy refers to how well it can estimate the value of new data.	A classifier is also constructed from a training set composed of the records of databases and their corresponding class names
The model used to predict the unknown value is called a predictor.	The model used to classify the unknown value is called a classifier.

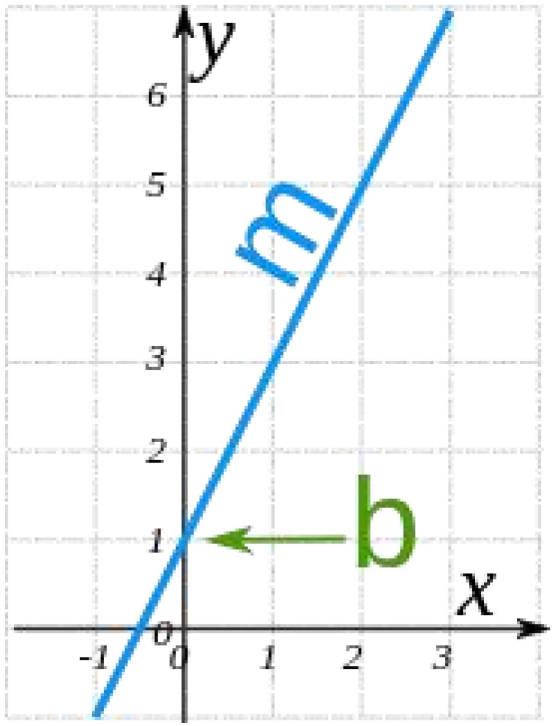
Linear Equation



Linear Equation

y=mx+b

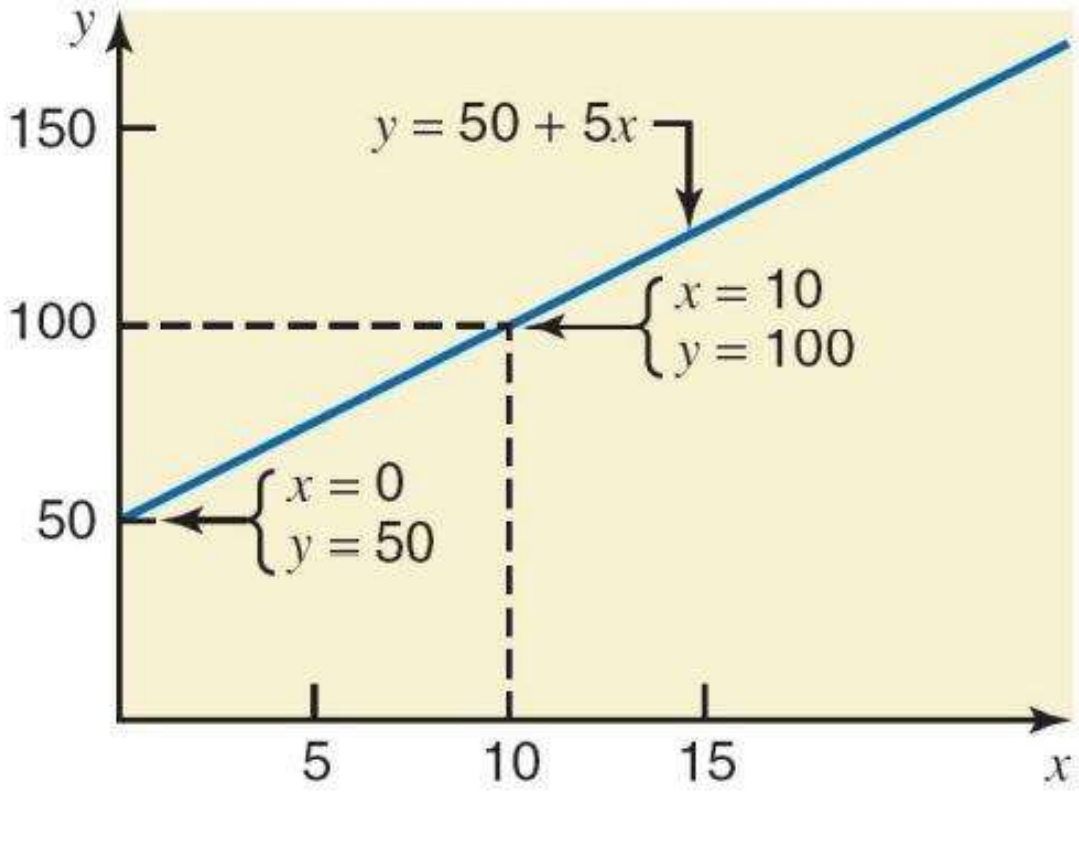
m=slope of line
b= intercept with y axis



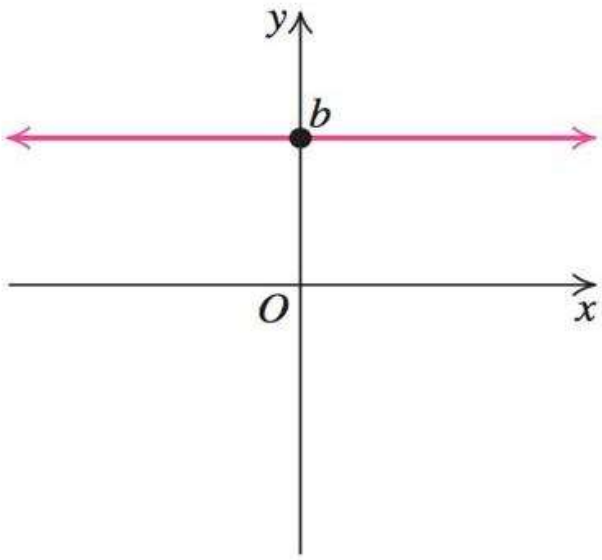
Sometimes a linear equation is written as a function, with f(x) instead of y:

f(x) = 2x − 3

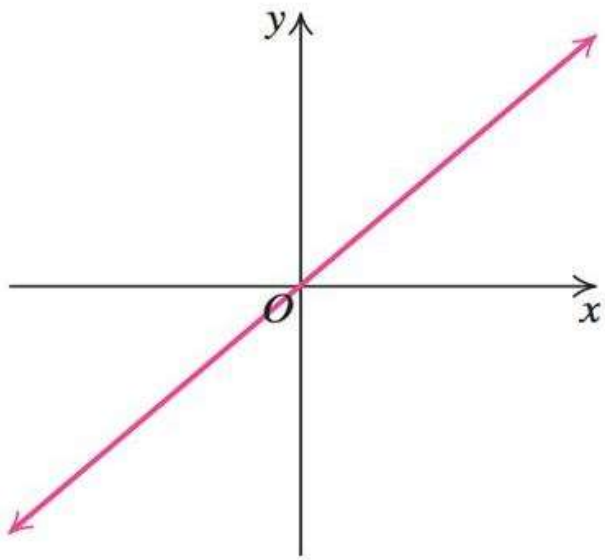
Plotting : Linear Equation



Linear Equation



(c)
$m = 0$
$f(x) = b$



(d)
$m = 1, b = 0$
$f(x) = x$

Linear Regression

- One of the most common statistical methods is linear regression.
- It's used to express the mathematical relationship between two variables or attributes if there is a linear relationship between an outcome variable and a predictor.
- A regression model that gives straight line relationship between variables is called linear regression model.

Linear Regression Model : Definition

In the regression model $y = c + mx + \varepsilon$, c is called the y-intercept or constant term, m is the slope, and ε is the random error term.The dependent and independent variables are y and x, respectively.

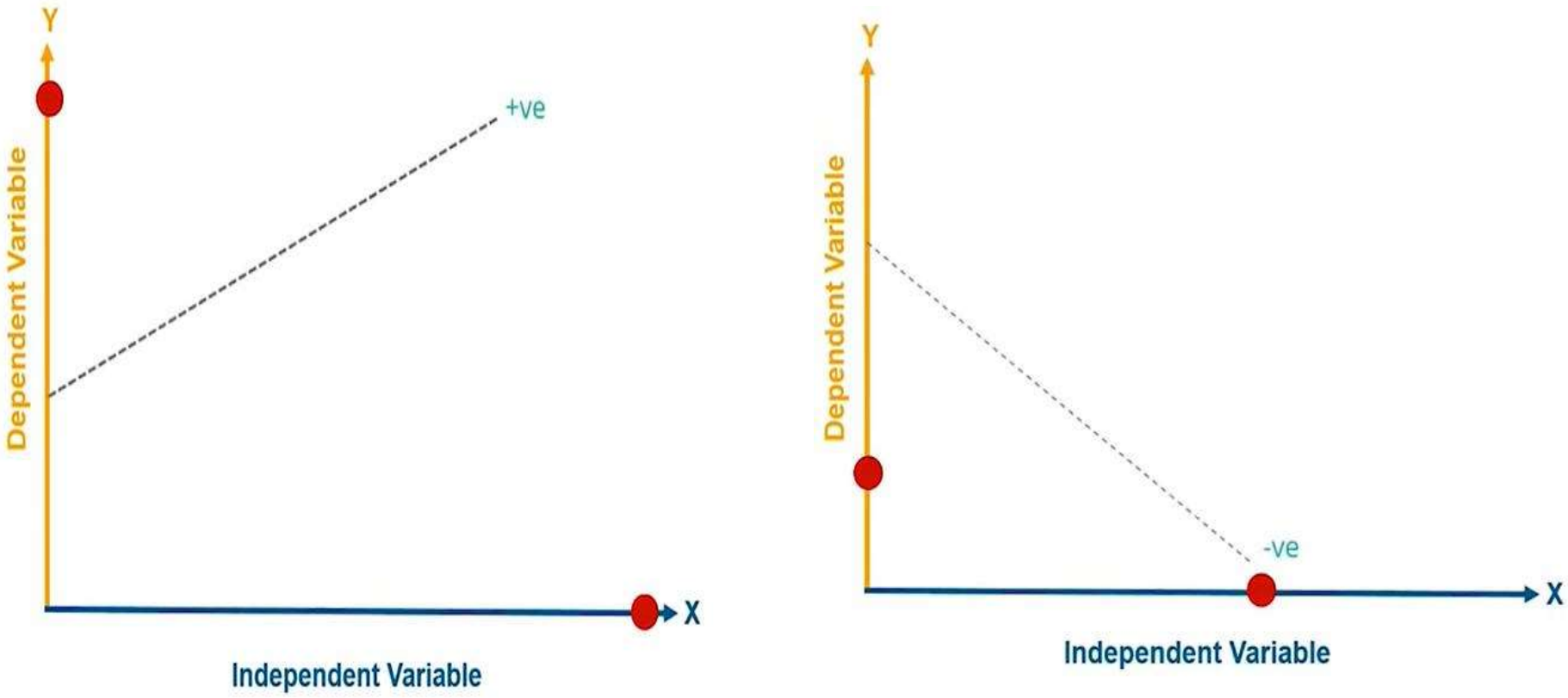
When to use Linear Regression?

- Evaluating Trends and Sales Estimation
- Analyse Impact of Price Changes
- Assessment of Risk in Financial Services
- Analyse the Advantages of Insurance Domain

Dependant and Independent Variable

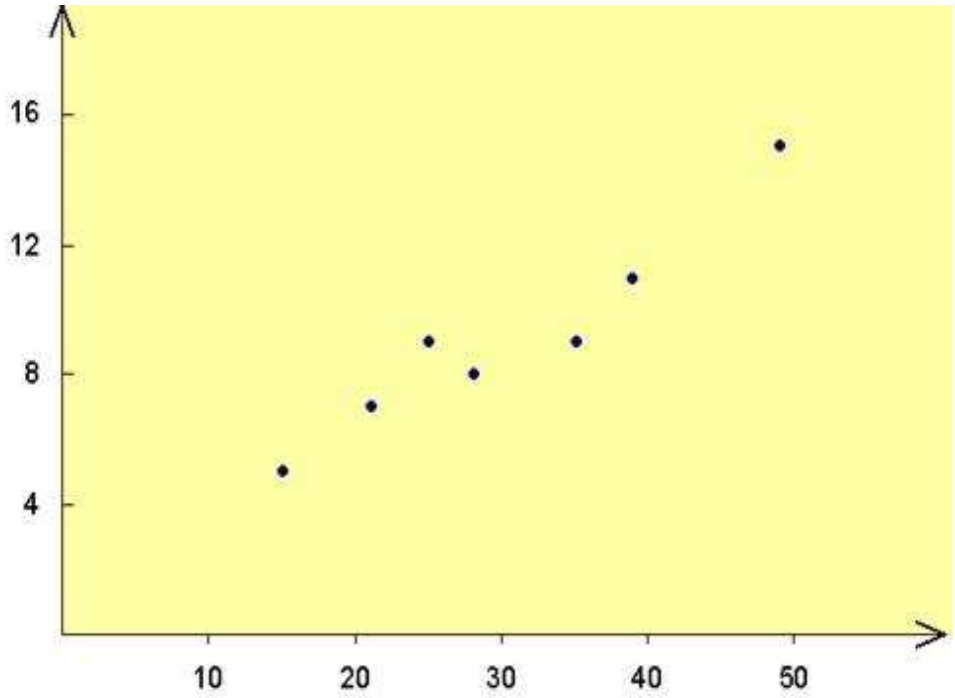
- The variable being predicted is called as **dependent variable**. (denoted by **y** in the equation)
- Variable used to predict the value of dependent variables are called as **independent variable**. (denoted by **x** in the equation)

Dependant and Independent Variable

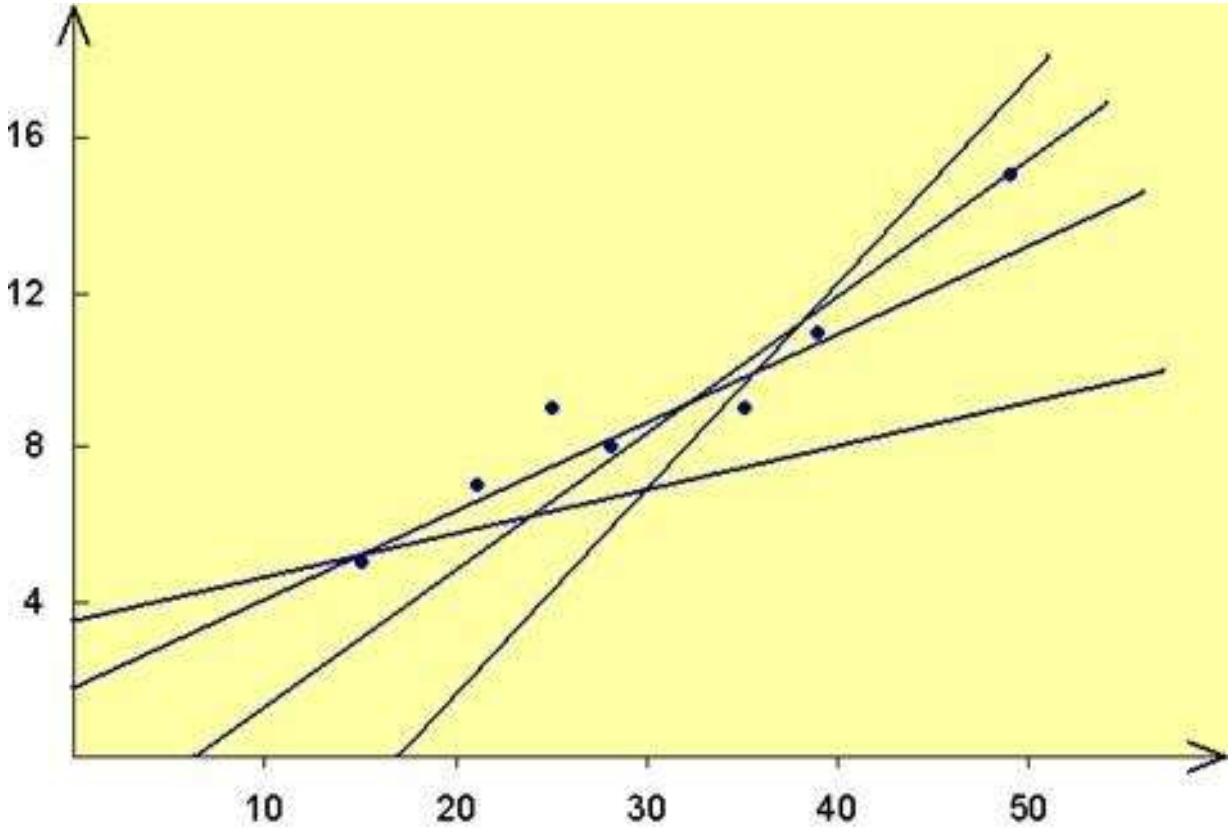


Example

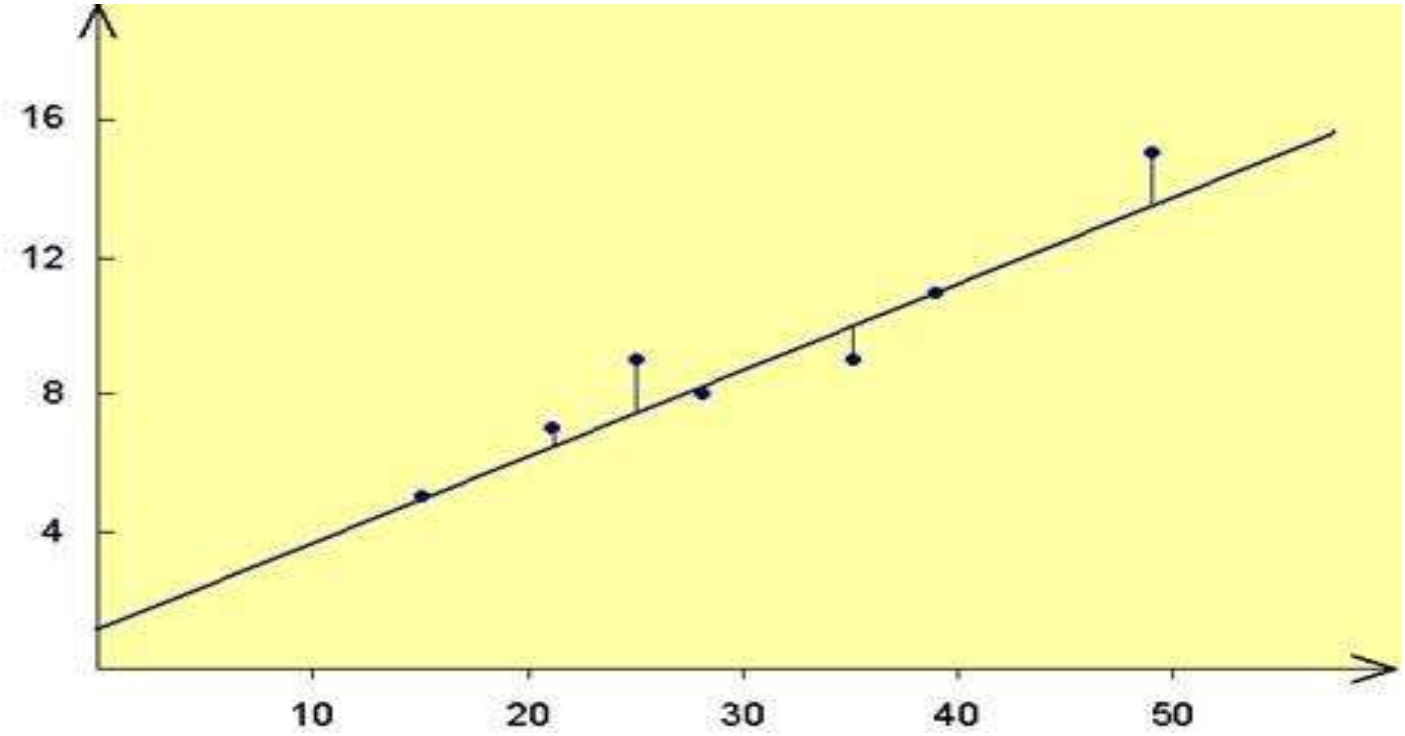
Income	Food Expenditure
35	9
49	15
21	7
39	11
15	5
28	8
25	9



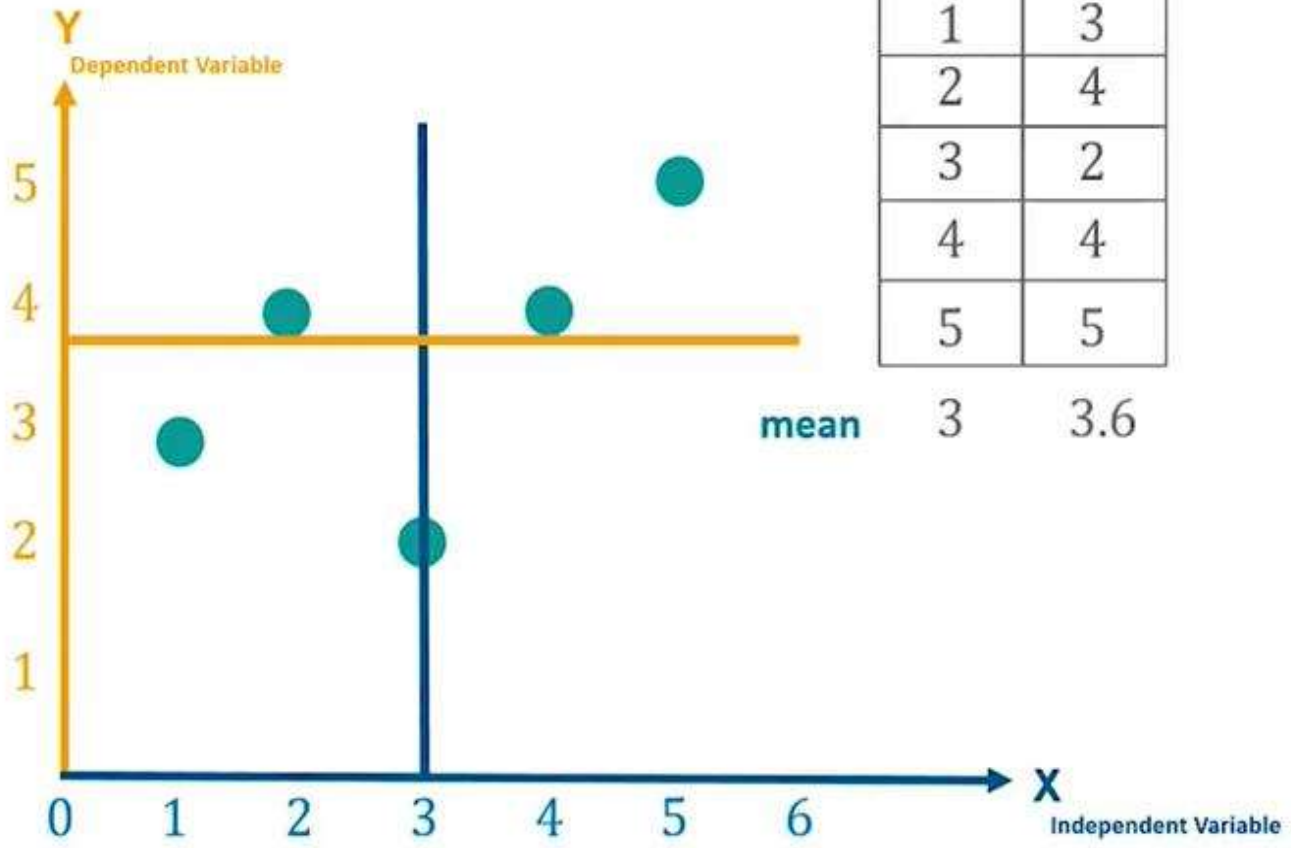
Which line to choose?



Regression line and error



Regression line : Example

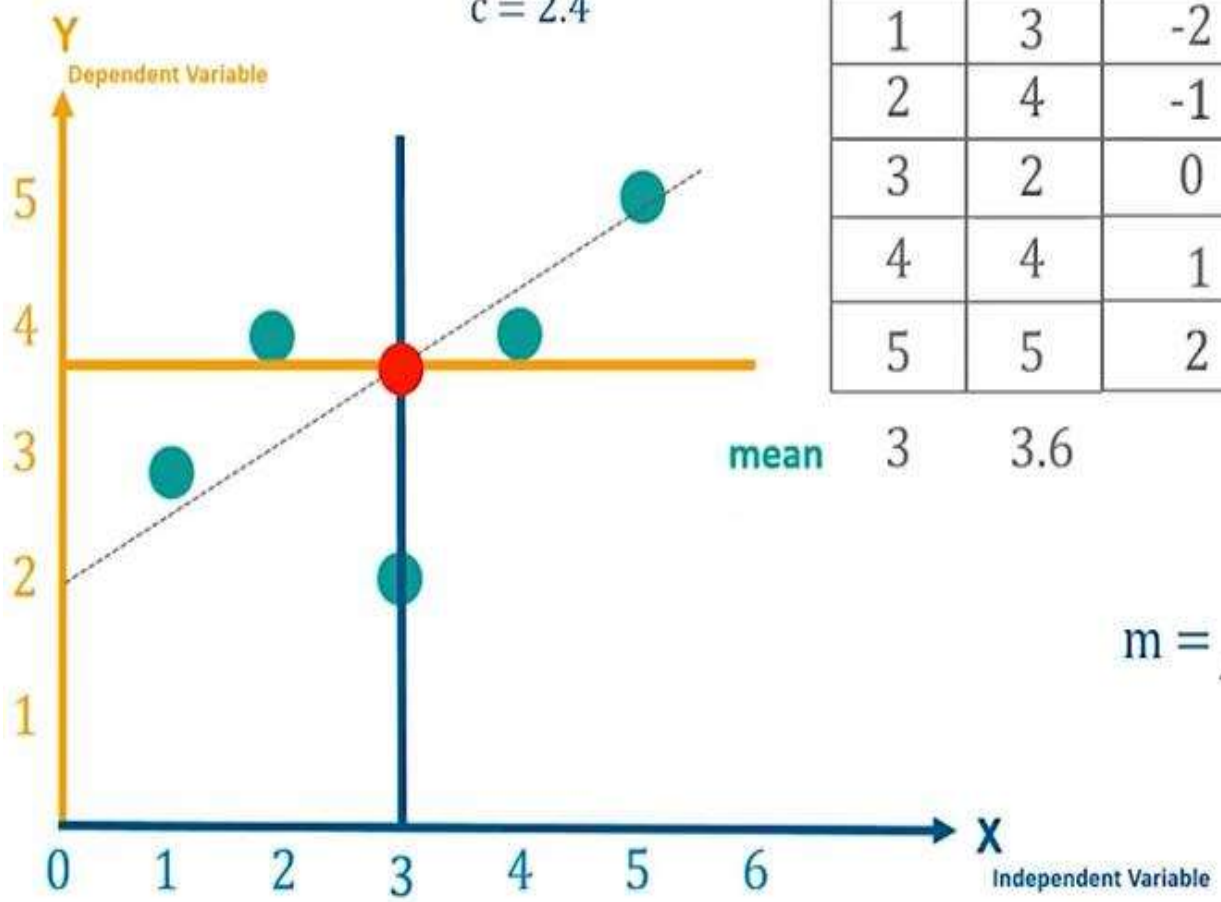


x	y
1	3
2	4
3	2
4	4
5	5
mean	3	3.6

Regression line : Example

$y = mx+c$

$c = 2.4$



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
mean	3	3.6		$\Sigma = 10$	$\Sigma = 4$

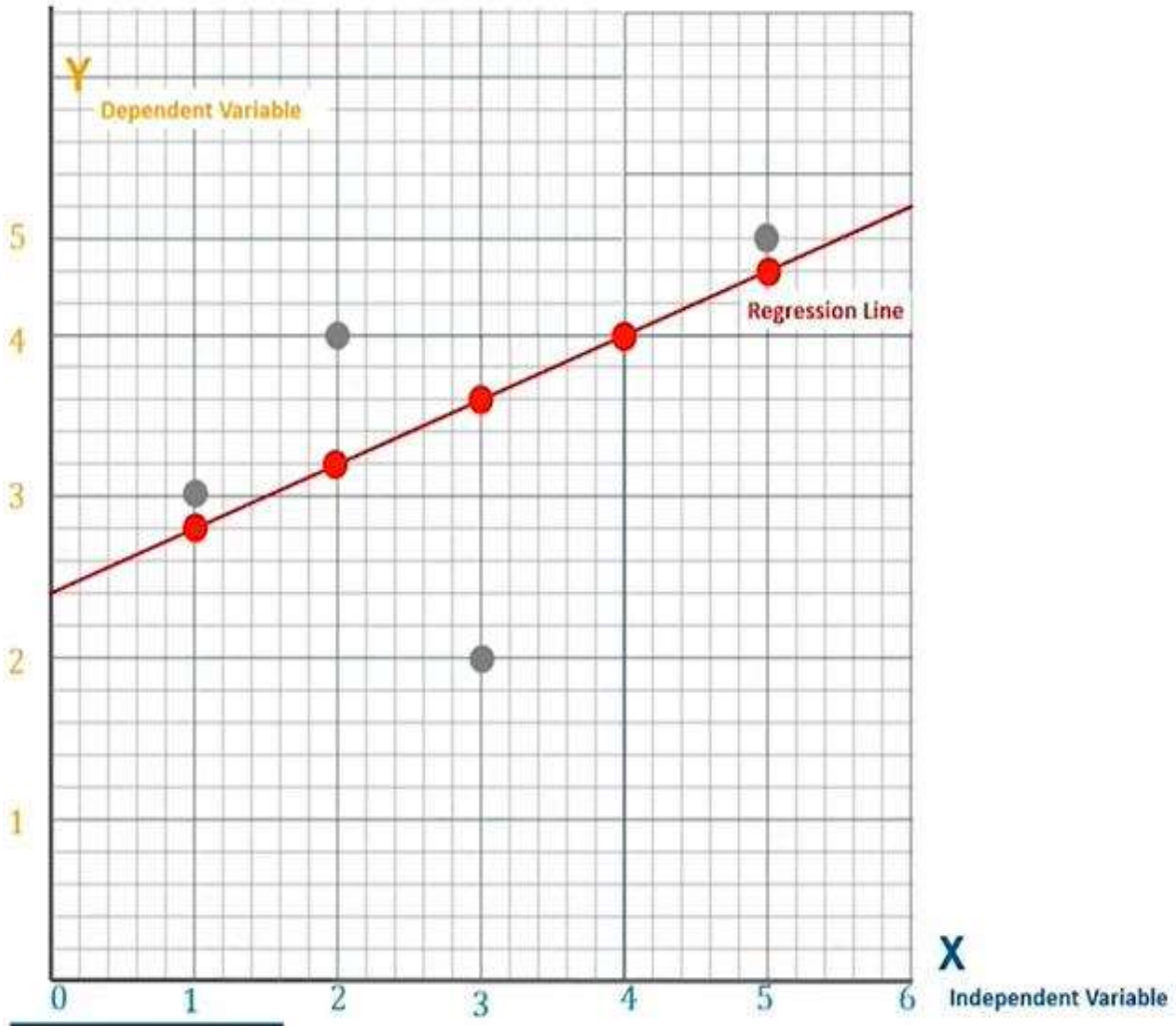
$$m = \sum \frac{(x - \bar{x})(y - \bar{y})}{(x - \bar{x})^2} = \frac{4}{10}$$

$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

Mean Square Error



$$m = 0.4$$
$$c = 2.4$$
$$y = 0.4x + 2.4$$

For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$$y = 0.4 \times 1 + 2.4 = 2.8$$
$$y = 0.4 \times 2 + 2.4 = 3.2$$
$$y = 0.4 \times 3 + 2.4 = 3.6$$
$$y = 0.4 \times 4 + 2.4 = 4.0$$
$$y = 0.4 \times 5 + 2.4 = 4.4$$

Example:

x	y
2	3
4	7
6	5
8	10

Example:

x	y
2	3
4	7
6	5
8	10

Ans:

Intercept :1.5 Slope:0.95

Y=1.5 +0.95X

Cost Function (J)

- Regression model aims to predict y value such that the error difference between predicted value and actual value is minimum.
- It is very important to update the θ1 and θ2 values(slope and coefficient), to reach the best value that minimize the error between predicted y value (pred) and actual y value (y).

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Cost function

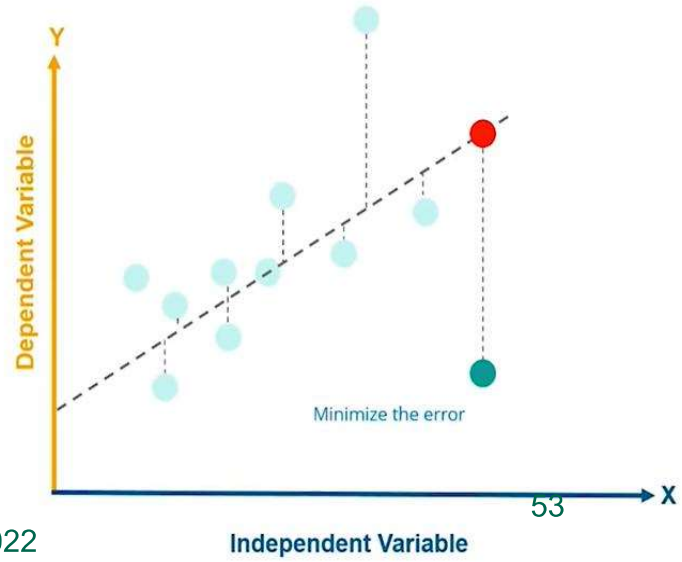
- In ML, cost functions are used to estimate how models are performing.
- ***cost function is a measure of how wrong the model is in terms of its ability to estimate the relationship between X and y.***
- This is typically expressed as a difference or distance between the predicted value and the actual value. The cost function (you may also see this referred to as *loss* or *error*.) can be estimated by iteratively running the model to compare estimated predictions against actual value— the known values of *y*.

Find error using cost function

X	Actual value (y)	Predicted value (pred) y=1.14+0.26x	Error = (Pred-y)	(pred-y)^2
35	9	10.24	1.24	1.5376
49	15	13.88	-1.12	1.2544
21	7	6.6	-0.4	0.16
39	11	11.28	0.28	0.0784
15	5	5.04	0.04	0.0016
28	8	8.42	0.42	0.1764
25	9	7.64	-1.36	1.8496

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

J=1/7*(5.058)
= 0.72

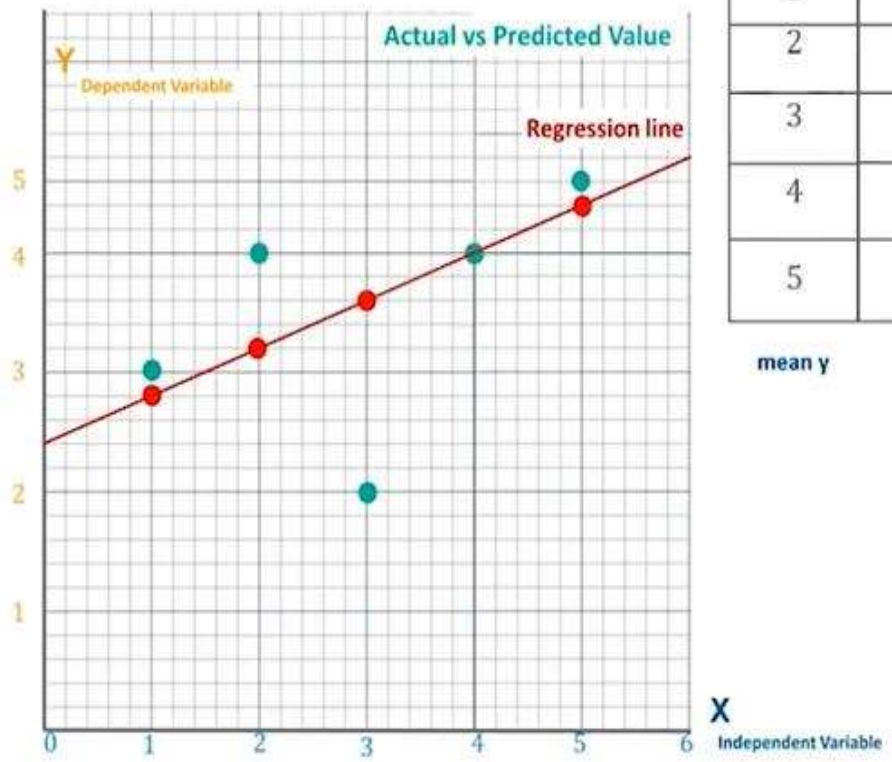


Value of R2

- R- squared value is the measure of how close the data are to the fitted regression line
- It is also known as coefficient of determination.
- Actual vs Predicted Values
- Distance Actual Mean vs Distance Predicted Mean

This is nothing but $R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$

Example



x	y	$y-y$	$(y-y)^2$	y_p	(y_p-y)	$(y_p-y)^2$
1	3	- 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	-1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		$\sum$ 5.2			$\sum$ 1.6

$$R^2 = \frac{1.6}{5.2} = \frac{\sum \left((y_p - \bar{y}\right)^2}{\sum (y-\dot{y})^2}$$

R^2 is Approximately 0.3

What is Classification?

- The goal of data classification is to organize and categorize data in distinct classes.
  - A model is first created based on the data distribution.
  - The model is then used to classify new data.
  - Given the model, a class can be predicted for new data.
- Classification = prediction for discrete and nominal values

Classification Example

- Example training database
  - Two predictor attributes: Age and Car-type (Sport, Minivan and Truck)
  - Age is ordered, Car-type is categorical attribute
  - Class label indicates whether person bought product
  - Dependent attribute is *categorical*

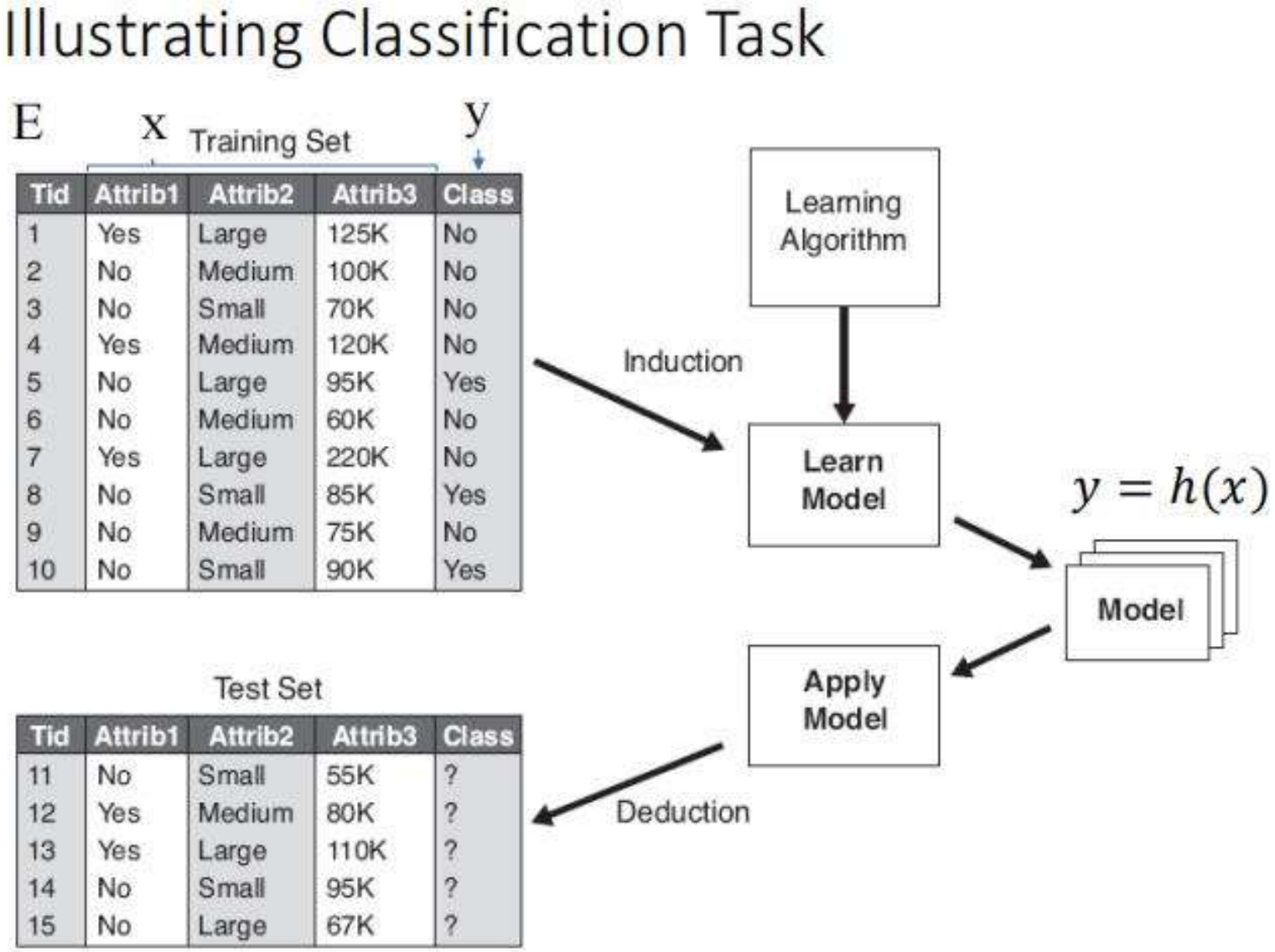
Age	Car	Class
20	M	Yes
30	M	Yes
25	T	No
30	S	Yes
40	S	Yes
20	T	No
30	M	Yes
25	M	Yes
40	M	Yes
20	S	No

Regression (Prediction) Example

- Example training database
  - Two predictor attributes: Age and Car-type (Sport, Minivan and Truck)
  - Spent indicates how much person spent during a recent visit to the web site
  - Dependent attribute is *numerical*

Age	Car	Spent
20	M	$200
30	M	$150
25	T	$300
30	S	$220
40	S	$400
20	T	$80
30	M	$100
25	M	$125
40	M	$500
20	S	$420

Supervised learning process: 3 steps



Classification is a 3-step process

- **1. Model construction (Learning):**
  - Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the **class label**.
  - The set of all tuples used for construction of the model is called **training set**.

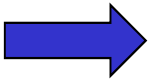
  – The model is represented in the following forms:
    - Classification rules, (IF-THEN statements),
    - Decision tree
    - Mathematical formulae

1. Classification Process (Learning)

Name	Income	Age	Credit rating
Samir	Low	<30	bad
Ahmed	Medium	[30...40]	good
Salah	High	<30	good
Ali	Medium	>40	good
Sami	Low	[30..40]	good
Emad	Medium	<30	bad

Training Data

class



Classification Method

Classification Model

IF Income = 'High'
OR Age > 30
THEN Class = 'Good

OR

Decision Tree

OR

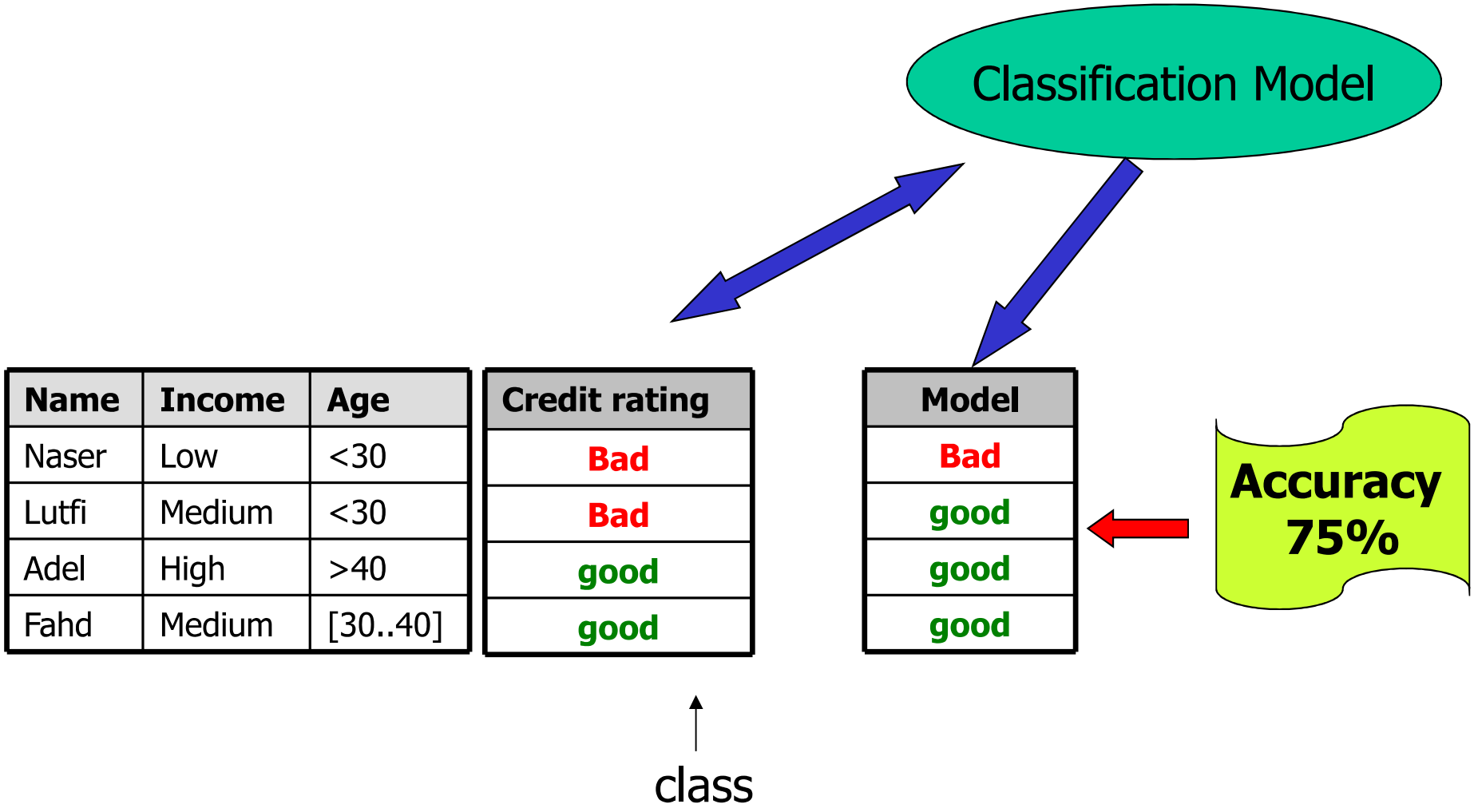
Mathematical For

Classification is a 3-step process

2. Model Evaluation (Accuracy):

- Estimate accuracy rate of the model based on a **test set**.
- The known label of test sample is compared with the classified result from the model.
- Accuracy rate is the percentage of test set samples that are correctly classified by the model.
- Test set is independent of training set otherwise over-fitting will occur

2. Classification Process (Accuracy Evaluation)

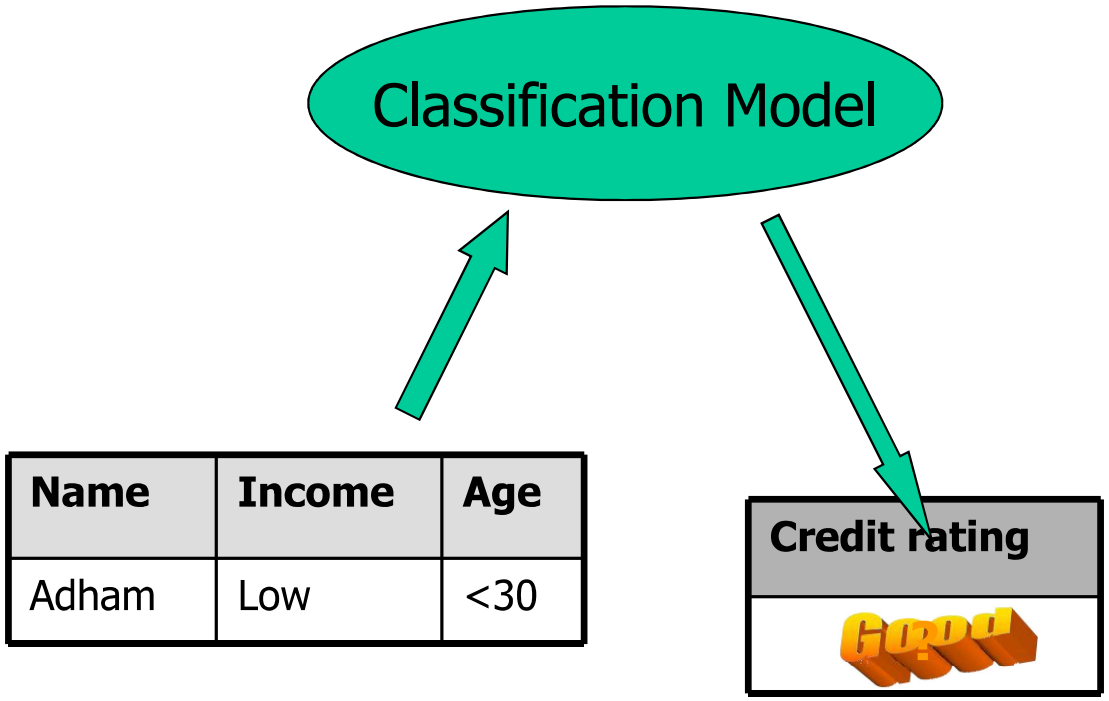


Classification is a 3-step process

3. Model Use (Classification):

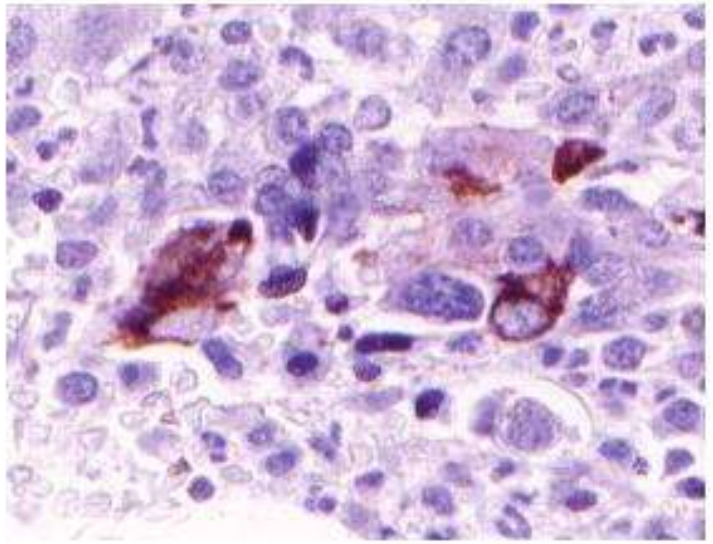
- The model is used to classify unseen objects.
  - Give a class label to a new tuple
  - Predict the value of an actual attribute

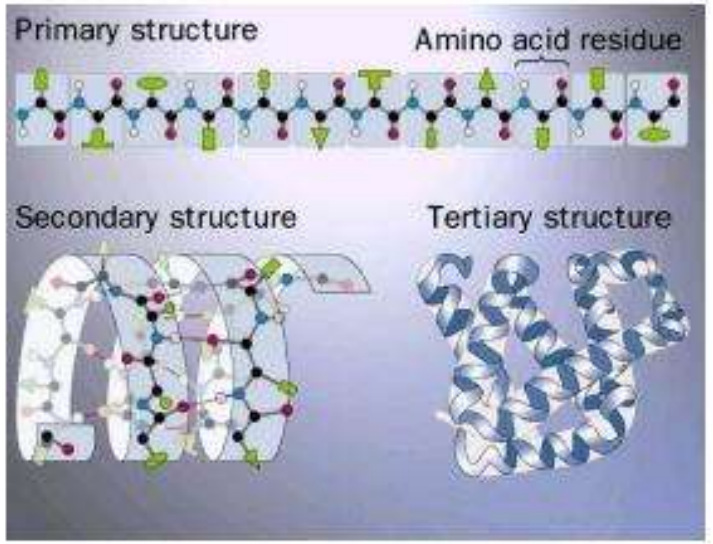
3. Classification Process (Use)



Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc





Classification Techniques


Decision Tree based Methods


Rule-based Methods


Memory based reasoning


Neural Networks / Deep Learning


Naïve Bayes and Bayesian Belief Networks

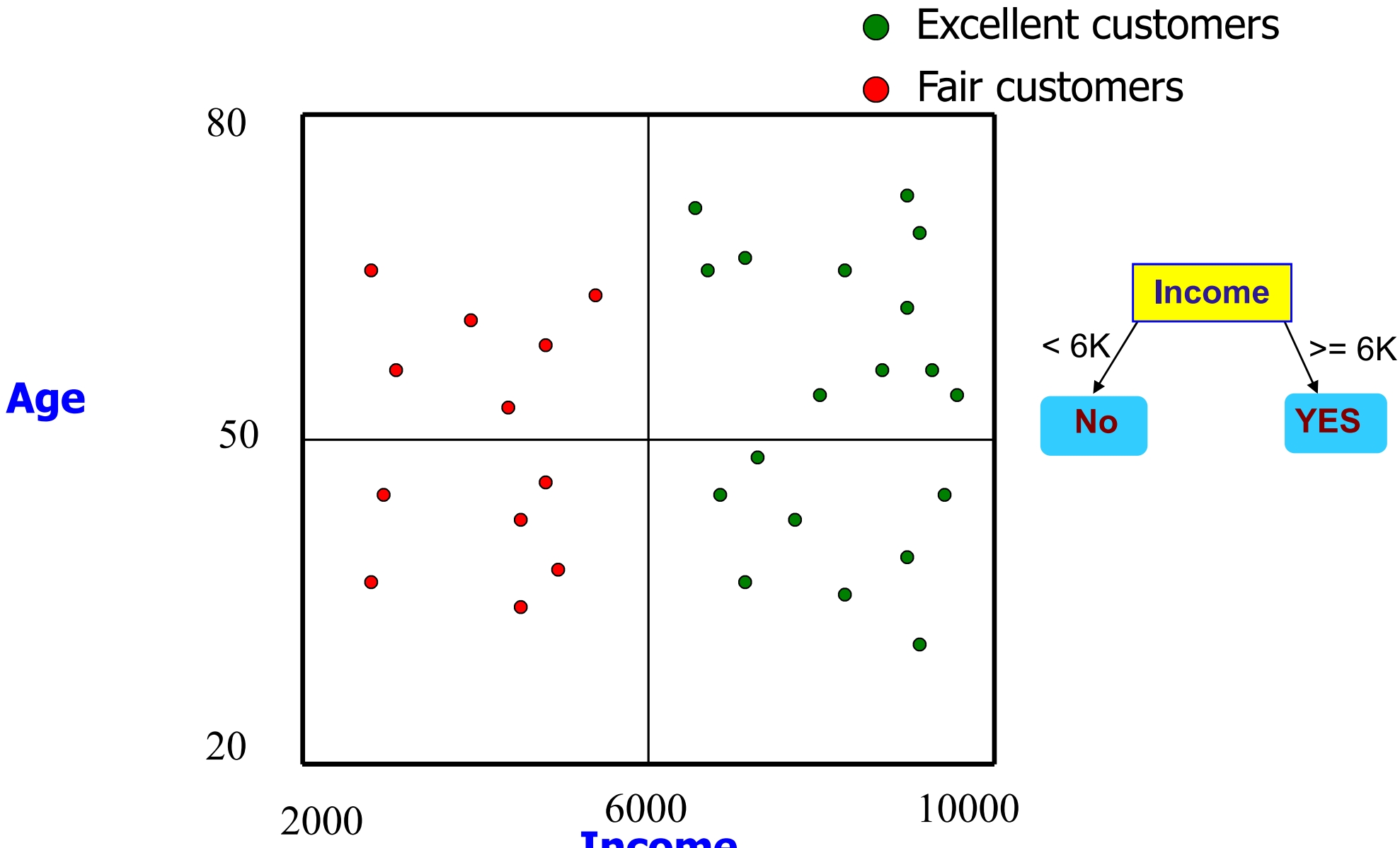

Support Vector Machines

Decision Tree

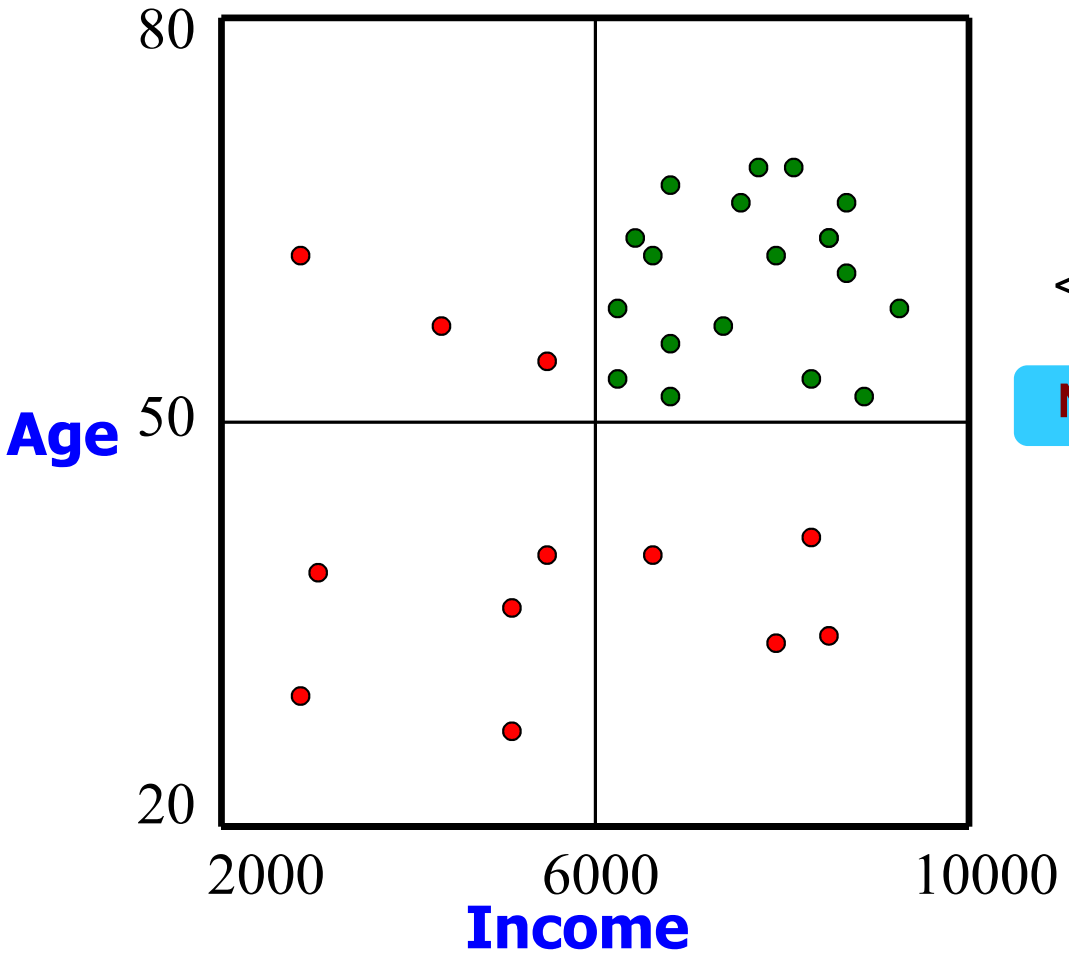
What is a Decision Tree?

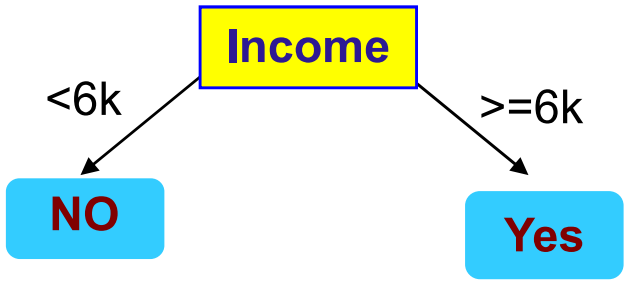
- A decision tree is a flow-chart-like tree structure.
  - Internal node denotes a test on an attribute
  - Branch represents an outcome of the test
- Leaf node represents class label

Sample Decision Tree

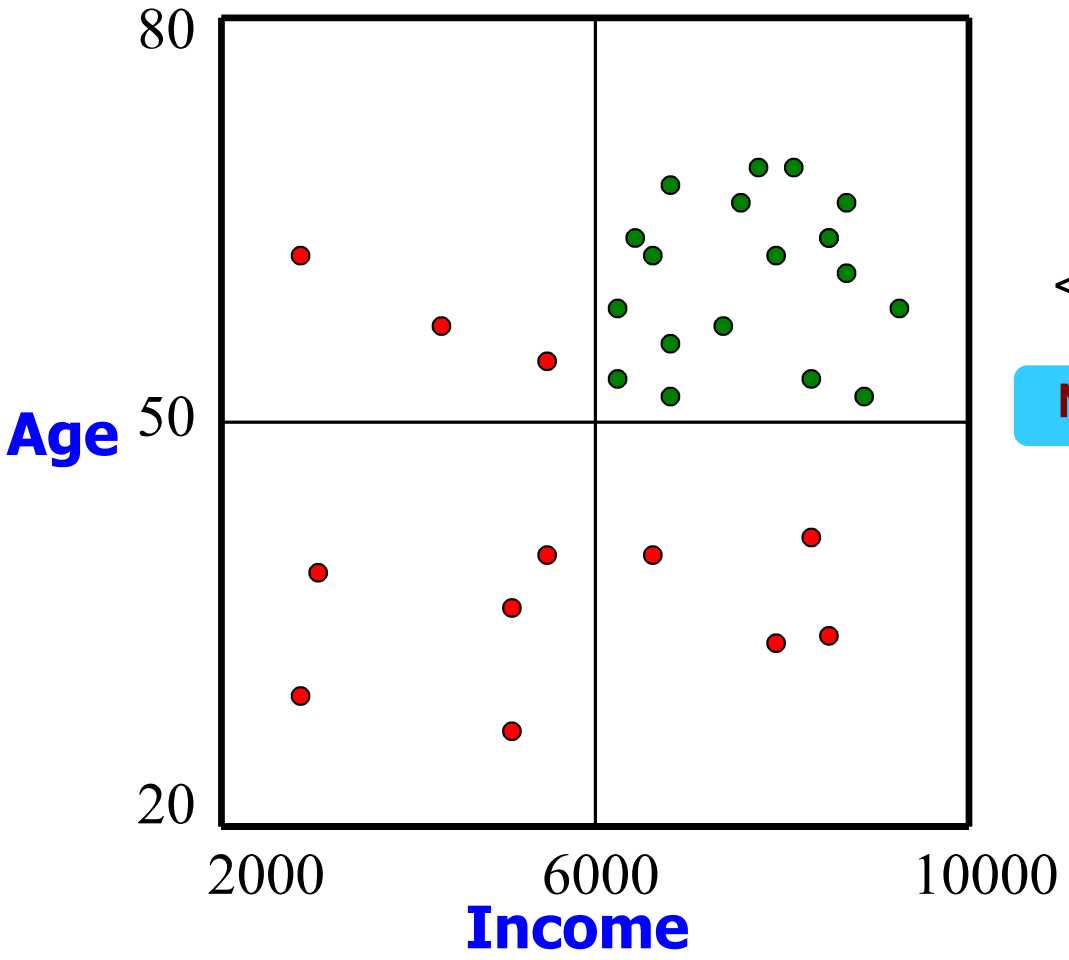


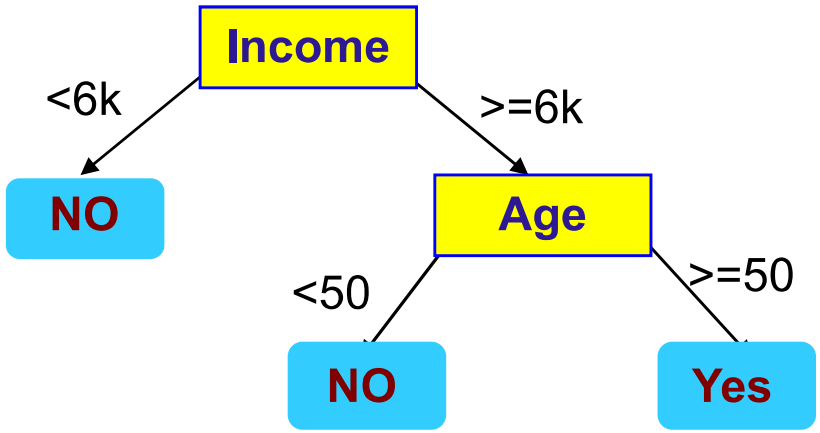
Sample Decision Tree





Sample Decision Tree

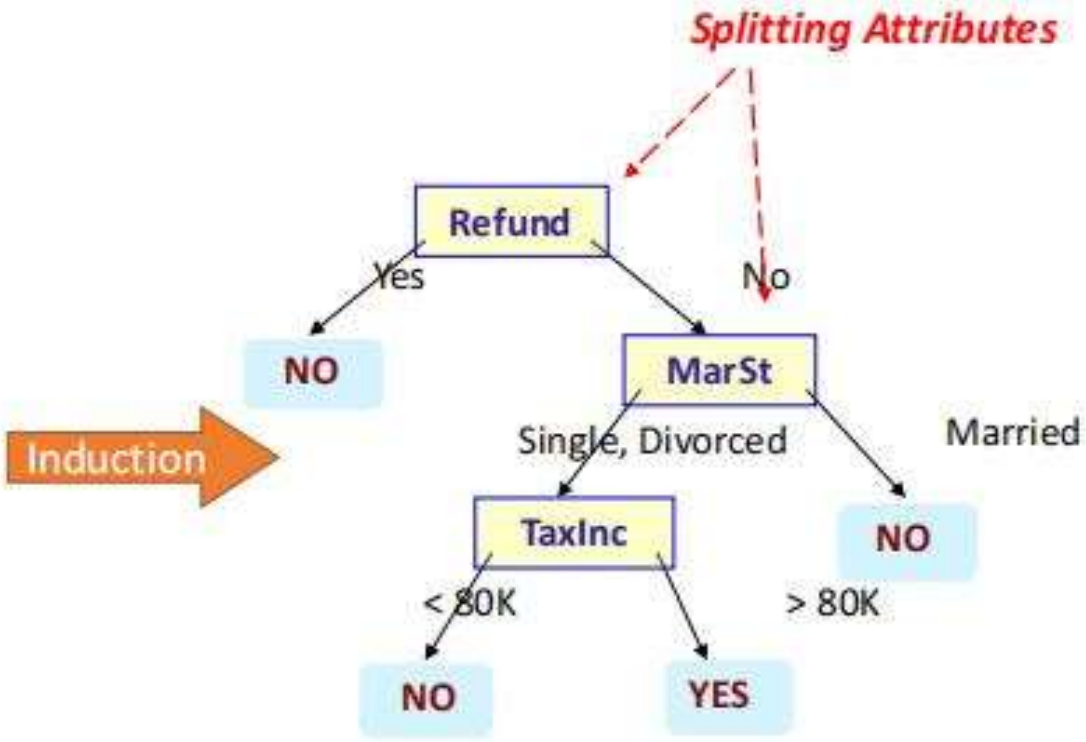




Example of a Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data

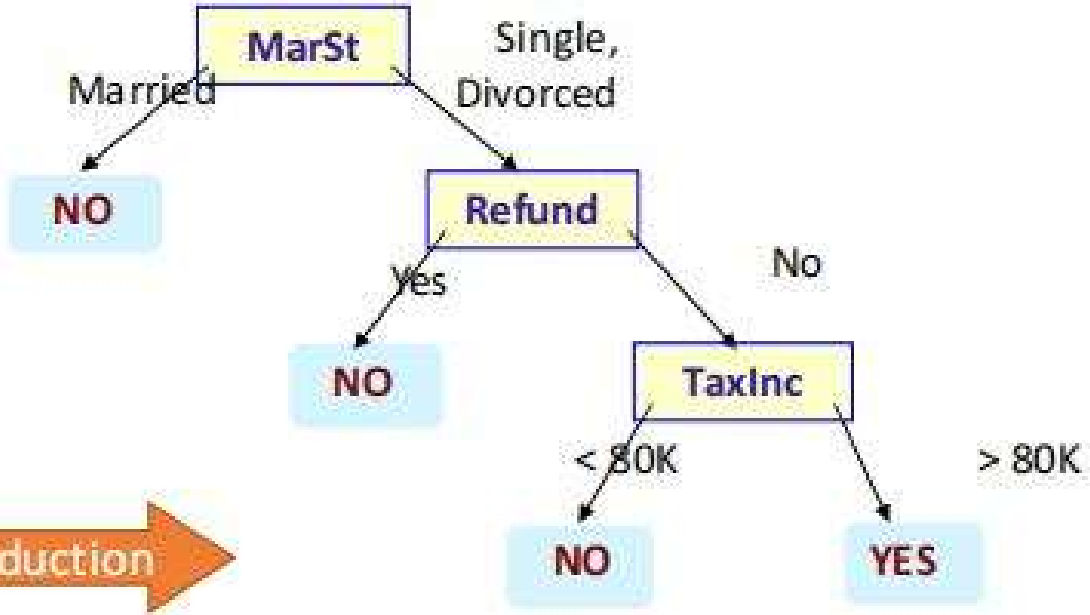


Model: Decision Tree

Another Example of Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Induction

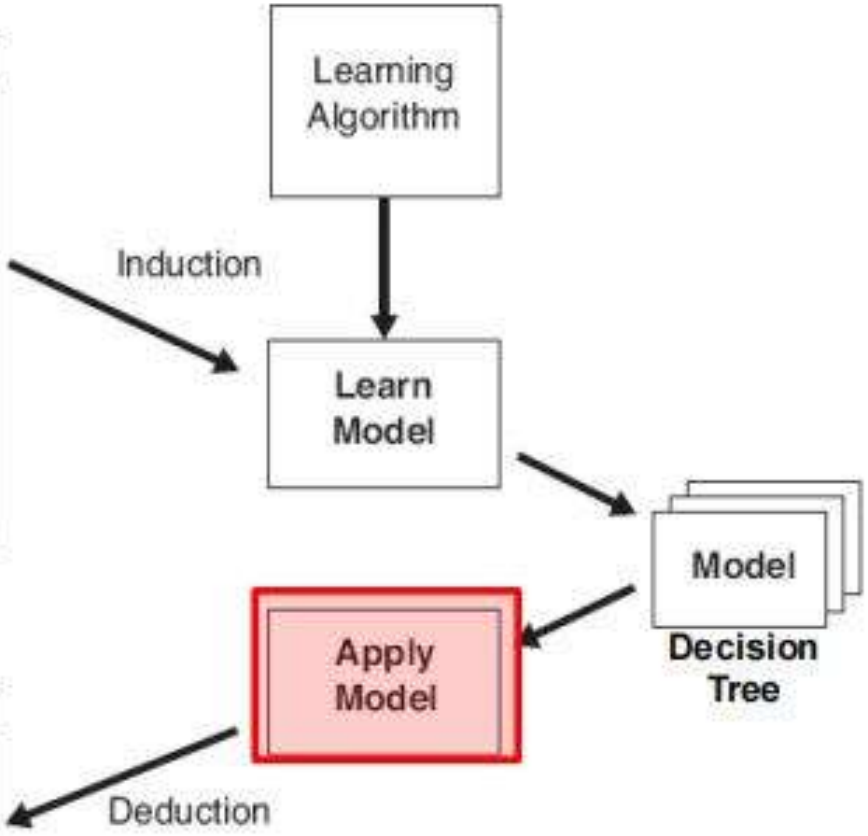


There could be more than one tree that fits the same data!

Decision Tree: Deduction

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

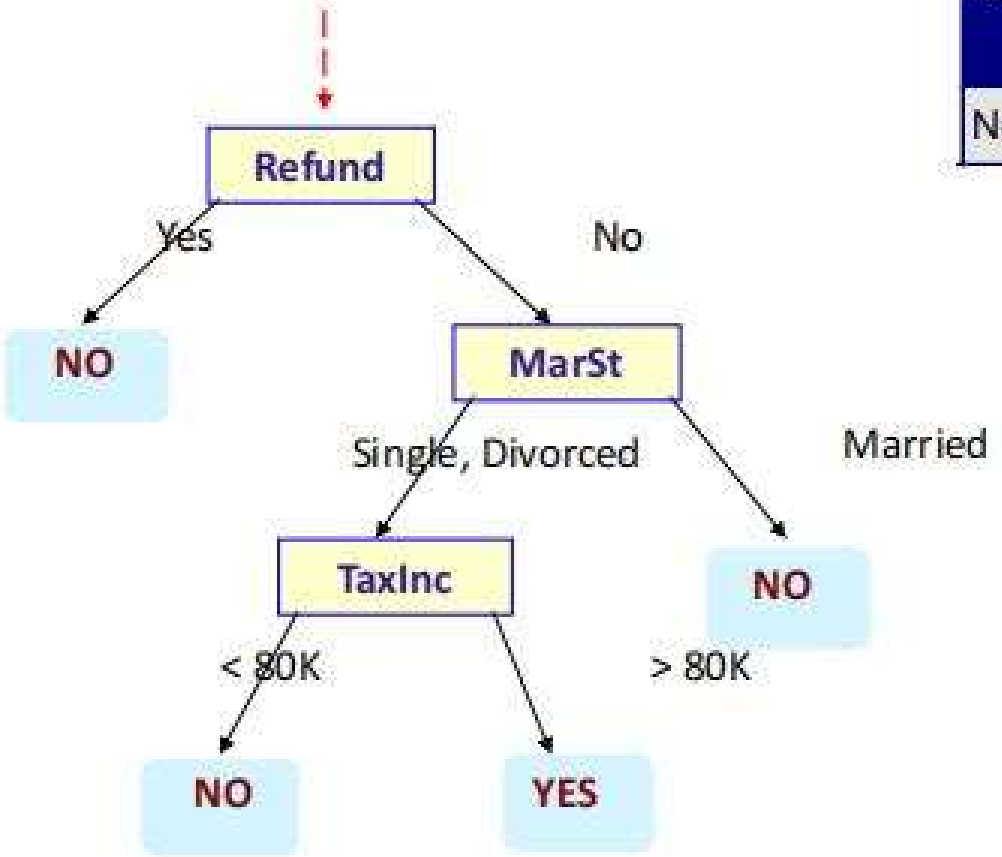


Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Apply Model to Test Data

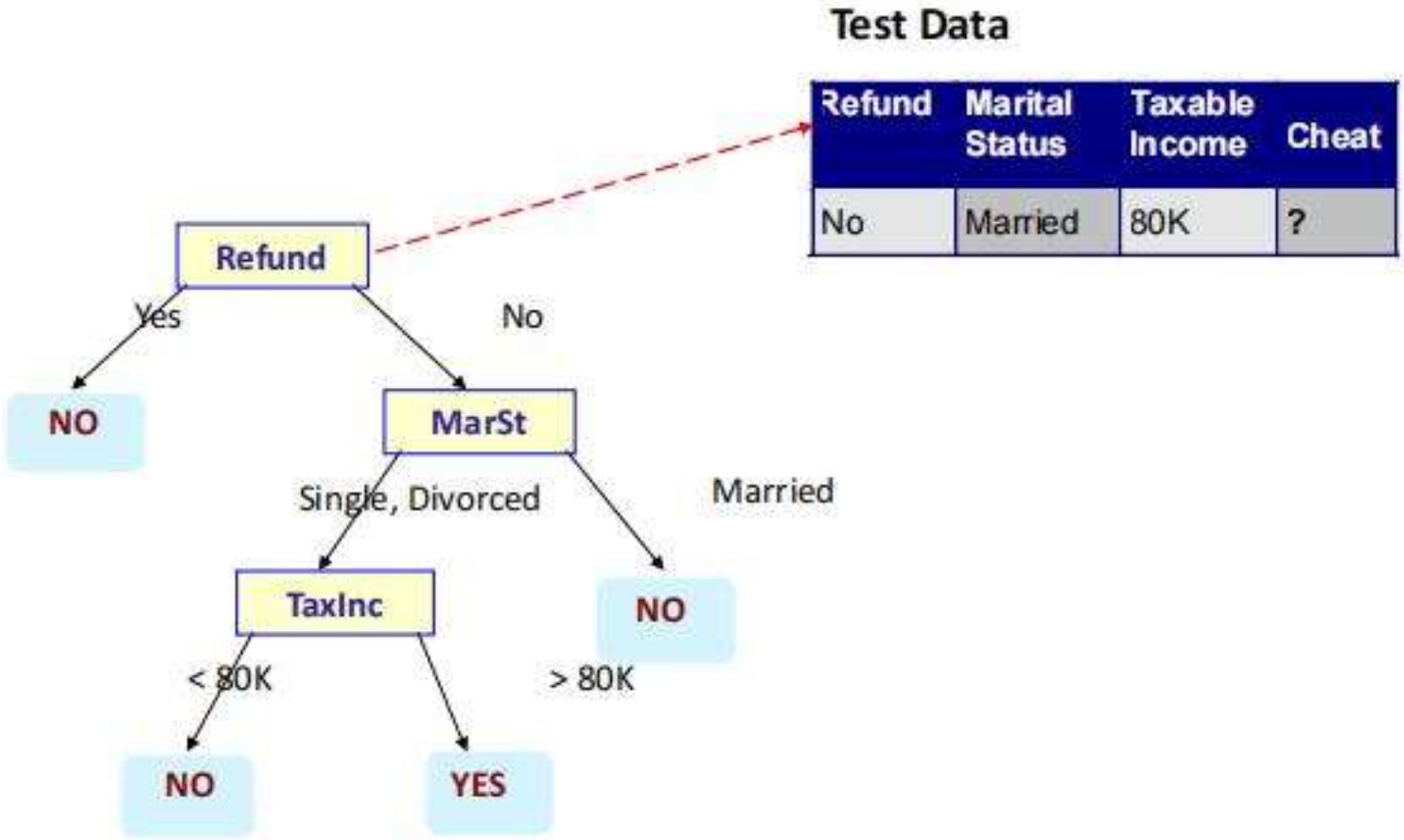
Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

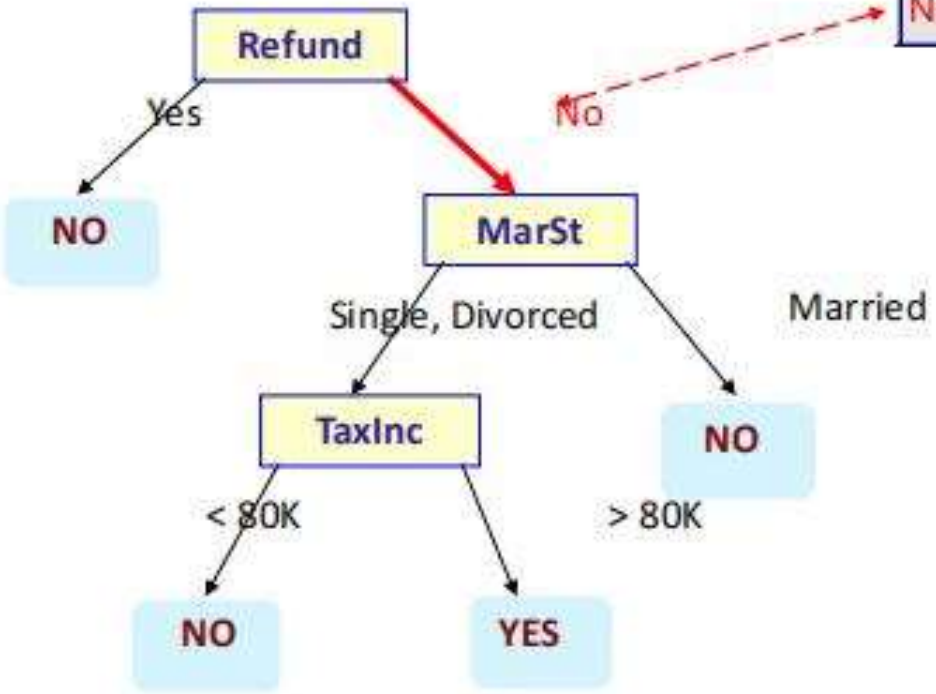
Apply Model to Test Data



Apply Model to Test Data

Test Data

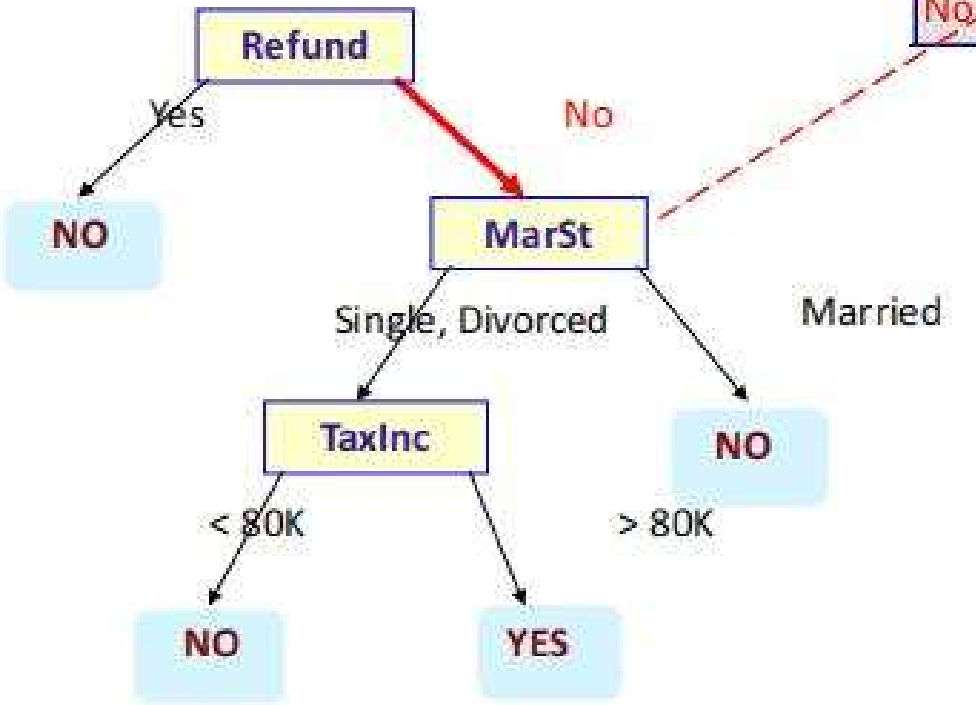
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

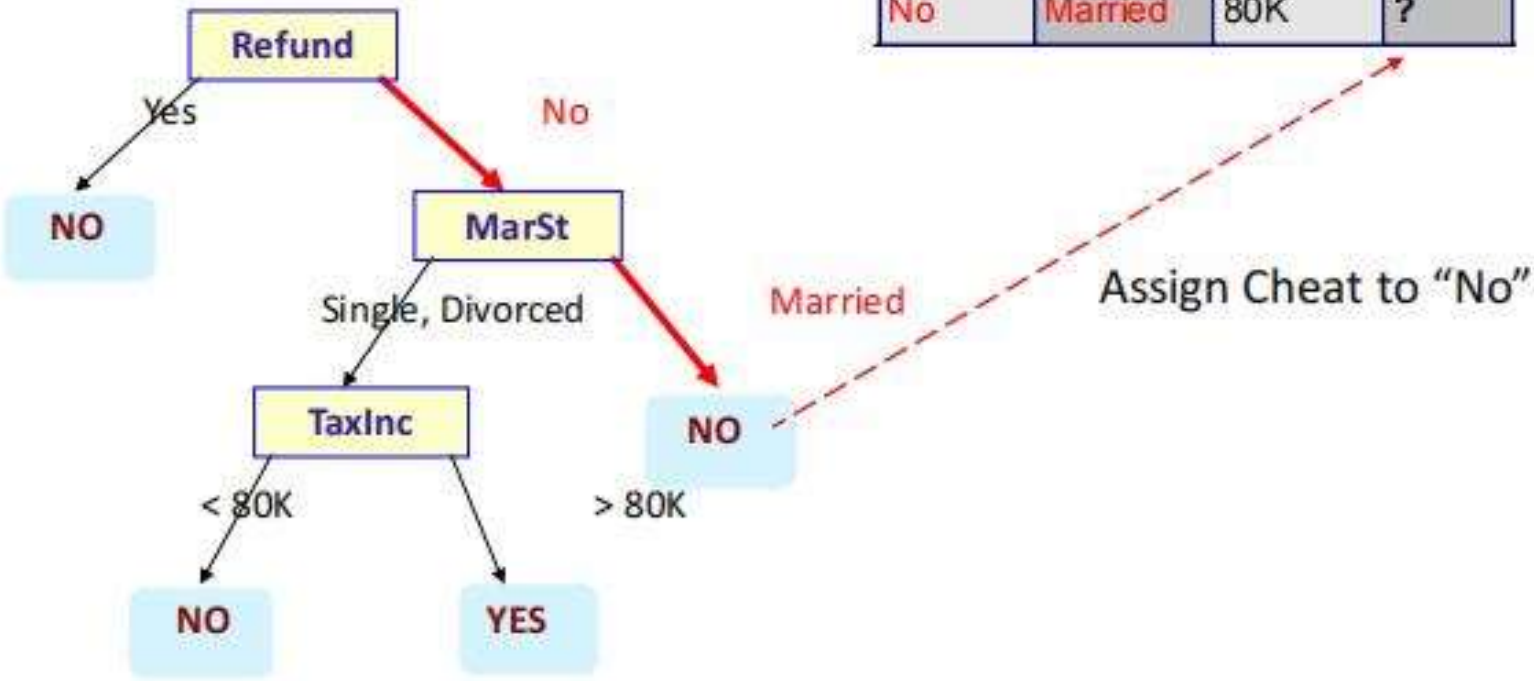
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



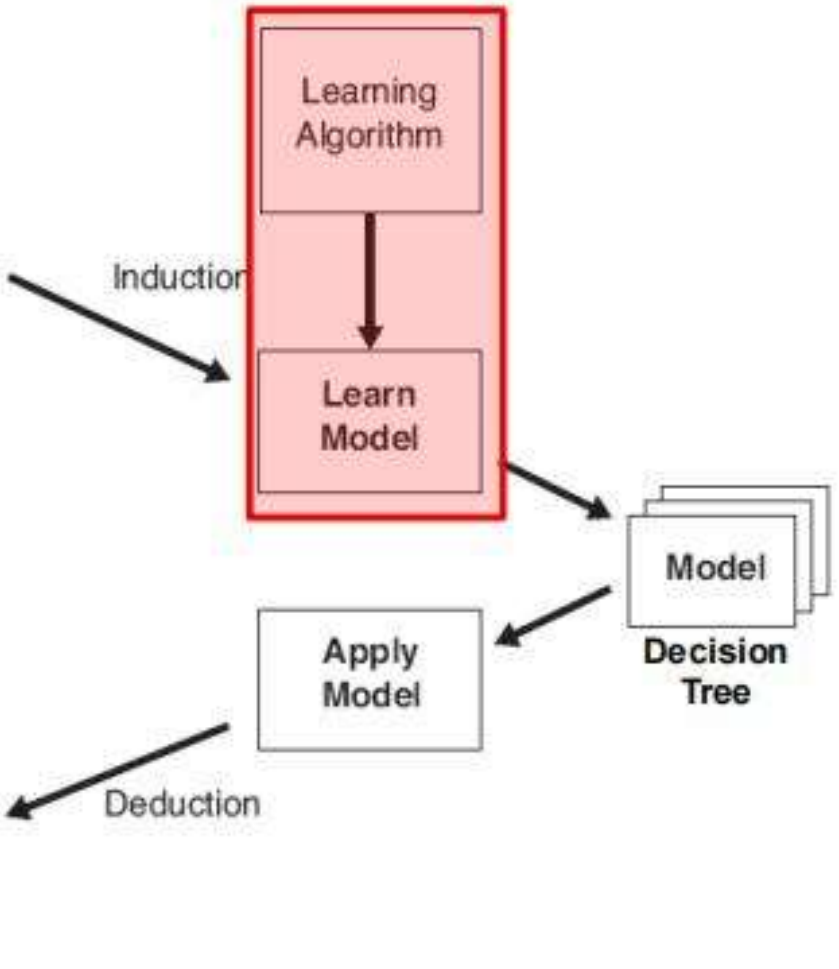
Decision Tree: Induction

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?



Decision-Tree Classification Methods

- The basic top-down decision tree generation approach usually consists of two phases:

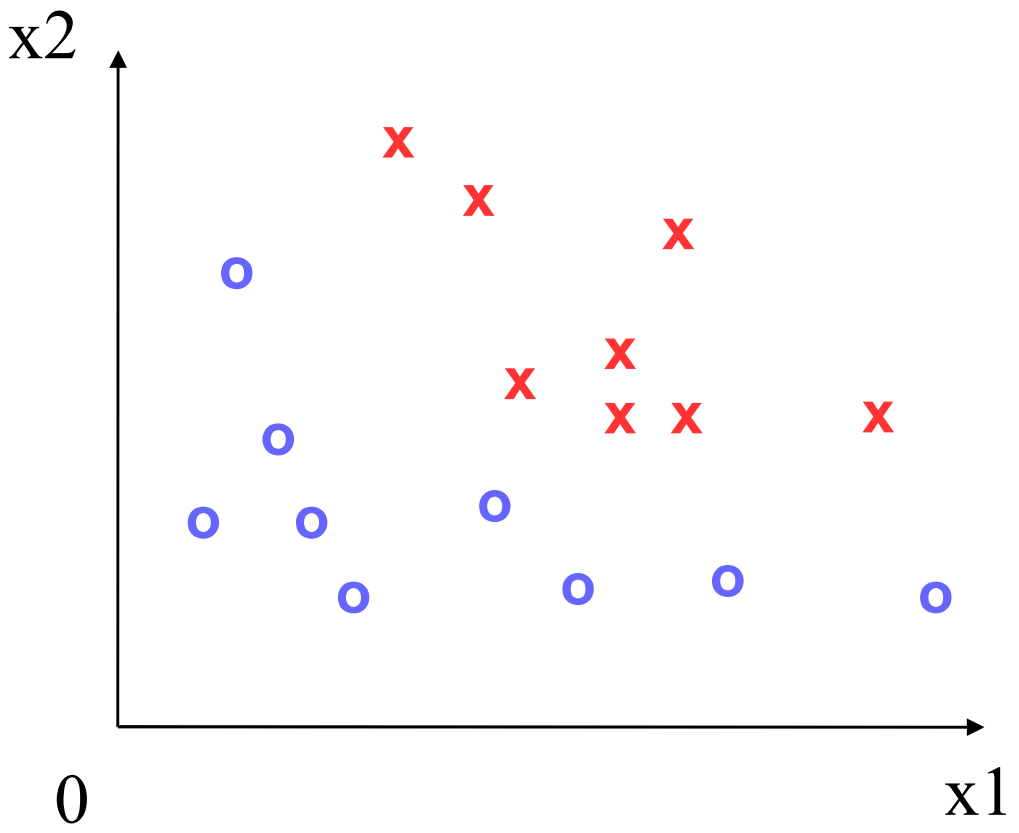
1. **Tree construction**
   - At the start, all the training examples are at the root.
   - Partition examples are recursively based on selected attributes.
2. **Tree pruning**
   - Aiming at removing tree branches that may reflect noise in the training data and lead to errors when classifying test data → improve classification accuracy

Decision Tree Induction

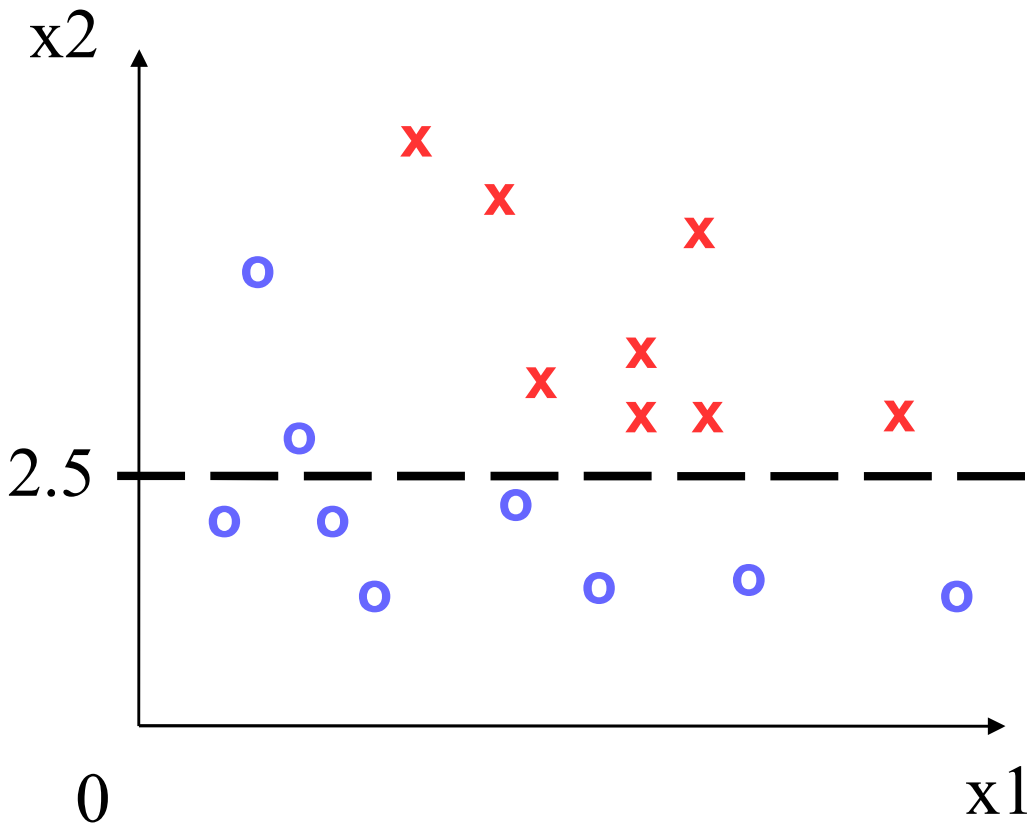
Many Algorithms:

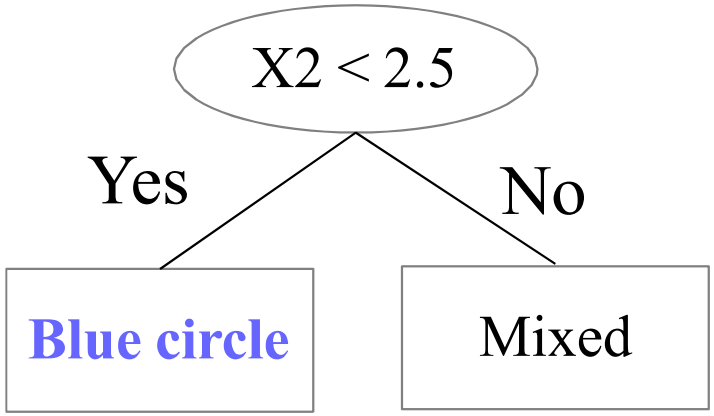
- Hunt's Algorithm (one of the earliest)
- CART (Classification And Regression Tree)
- ID3, C4.5, C5.0 (by Ross Quinlan, information gain)
- CHAID (CHi-squared Automatic Interaction Detection)
- MARS (Improvement for numerical features)
- SLIQ, SPRINT
- Conditional Inference Trees (recursive partitioning using statistical tests)

Example : Creating a Decision Tree

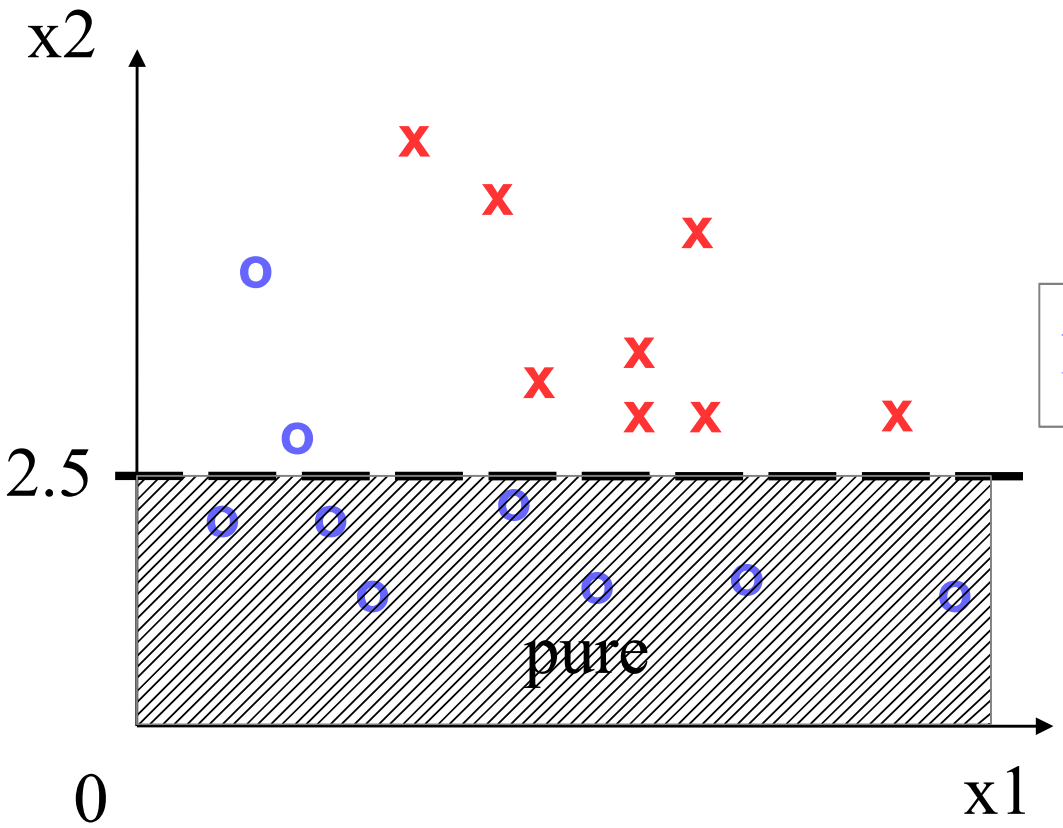


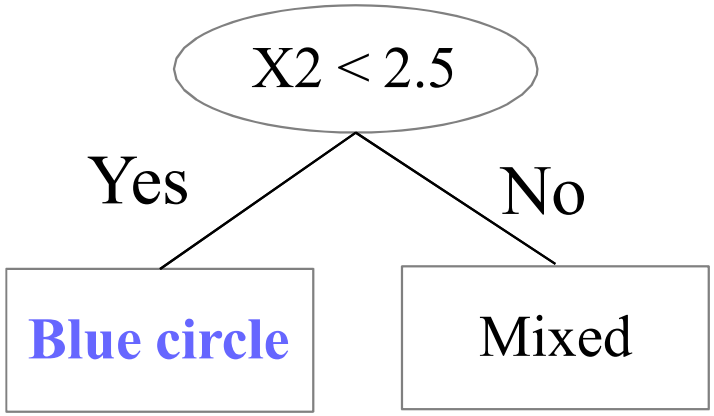
Example : Creating a Decision Tree



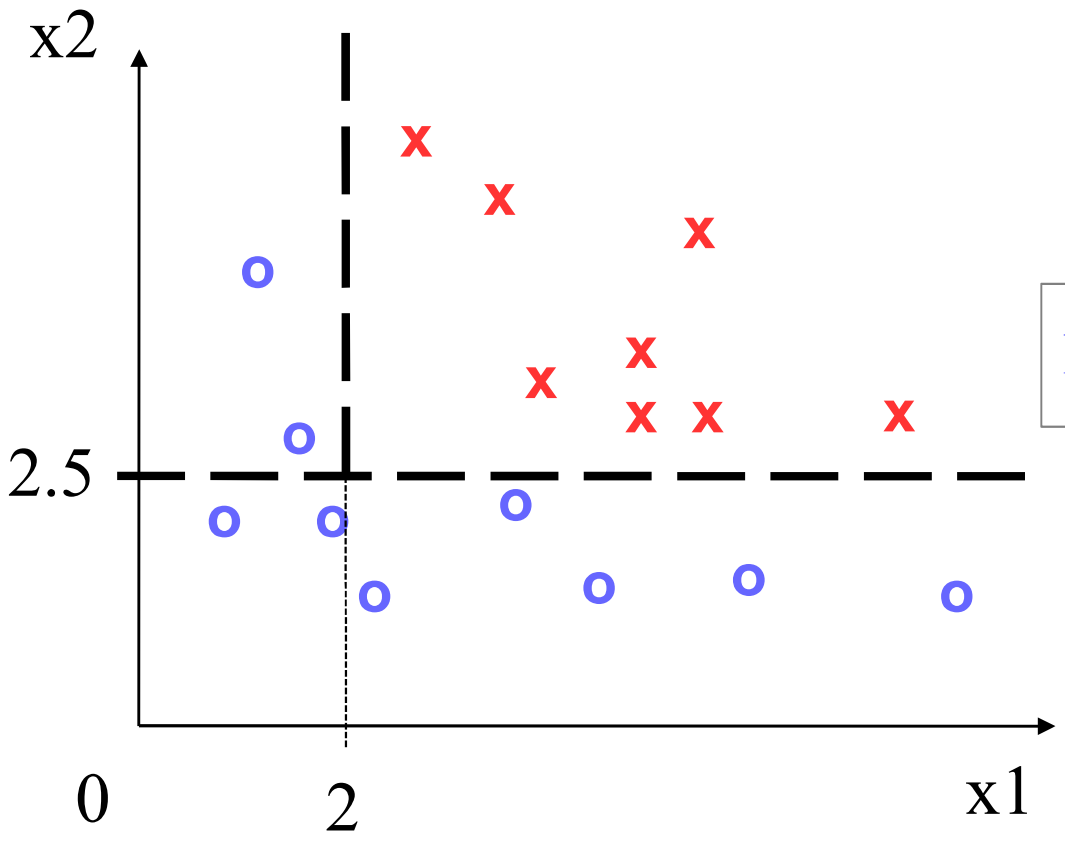


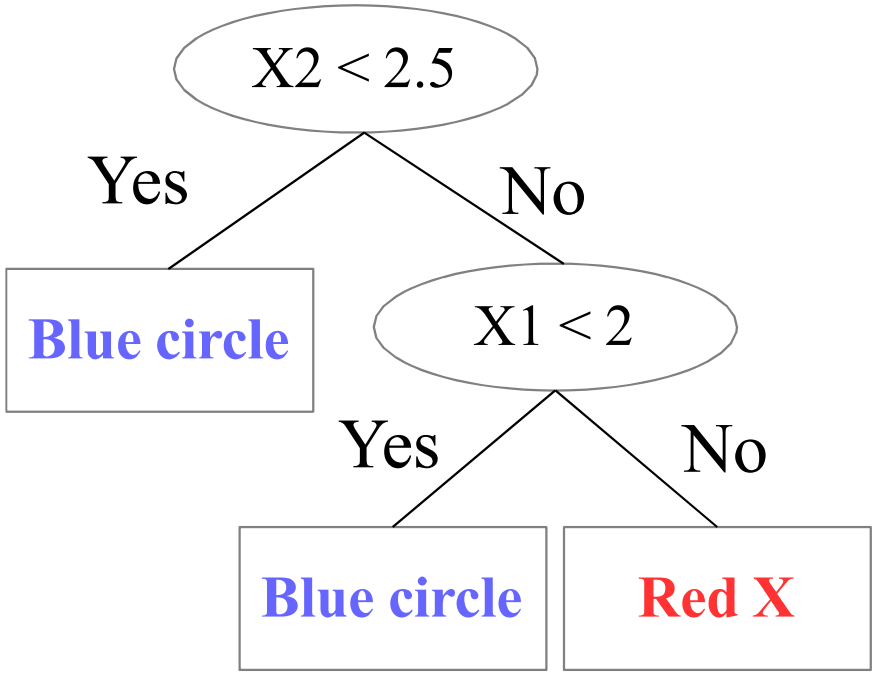
Example : Creating a Decision Tree





Example : Creating a Decision Tree

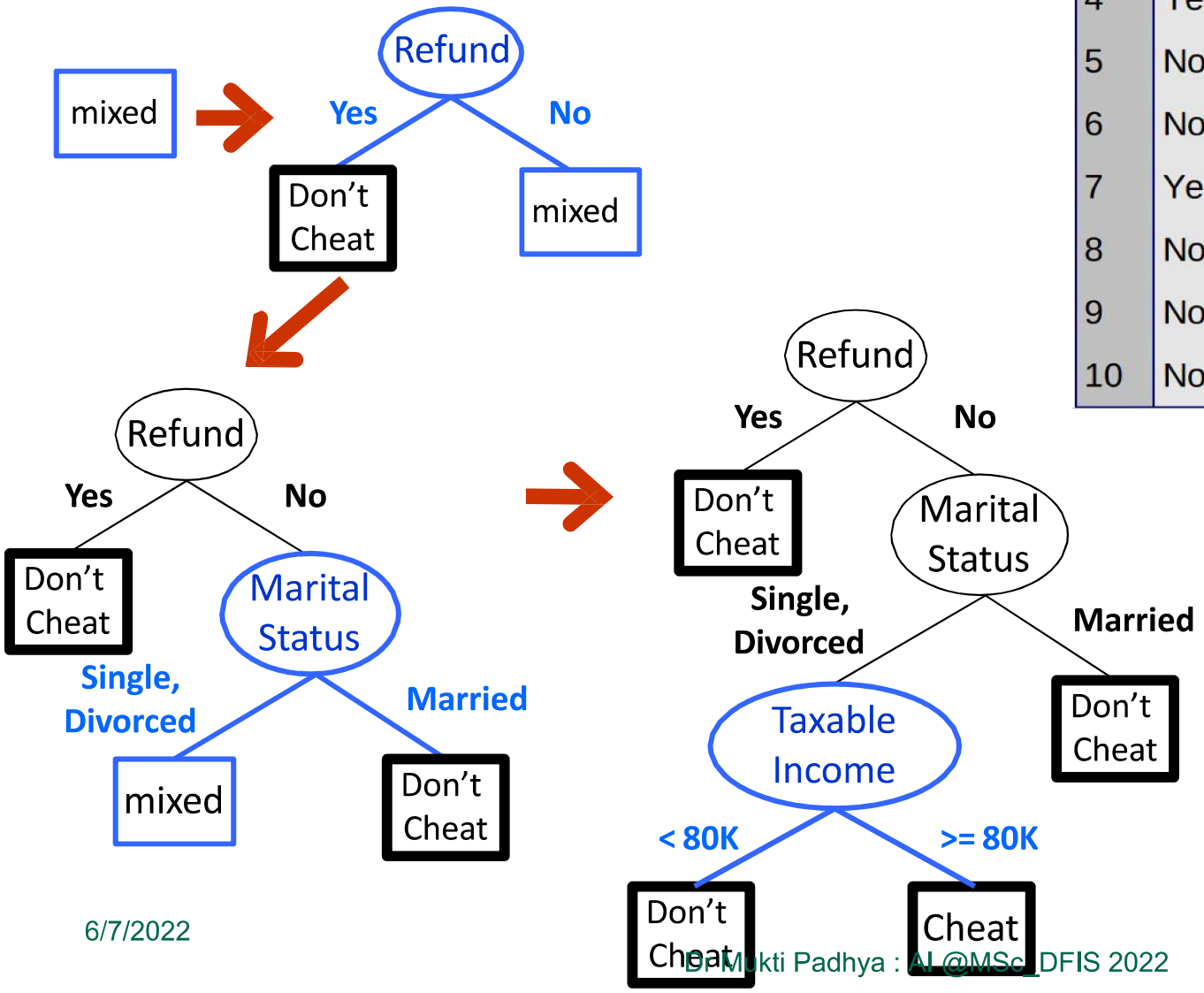




Hunt's Algorithm

"Use attributes to split the data recursively, till each split contains only a single class."

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Tree Induction

- Greedy strategy
  - —Split the records based on an attribute test that optimizes a certain criterion.

- Issues
  - —Determine how to split the record using different attribute types.
  - —How to determine the best split?
  - —Determine when to stop splitting

Tree Induction

- Greedy strategy
  - —Split the records based on an attribute test that optimizes a certain criterion.

- Issues
  - —**Determine how to split the record using different attribute types.**
  - —How to determine the best split?
  - —Determine when to stop splitting

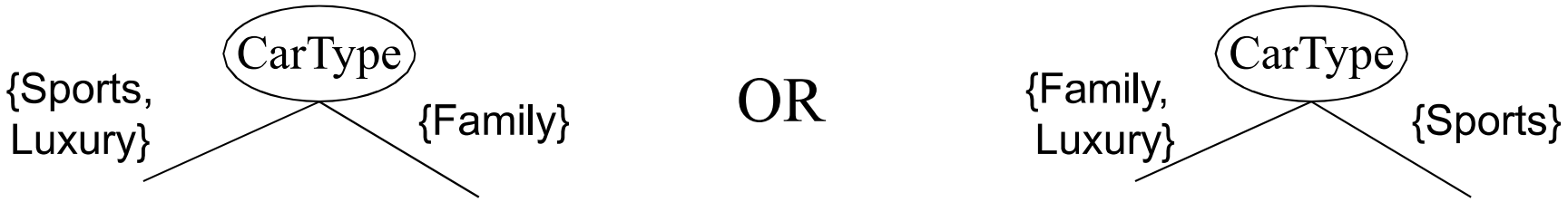
How to Specify Test Condition?

- Depends on attribute types
  - Nominal
  - Ordinal
  - Continuous

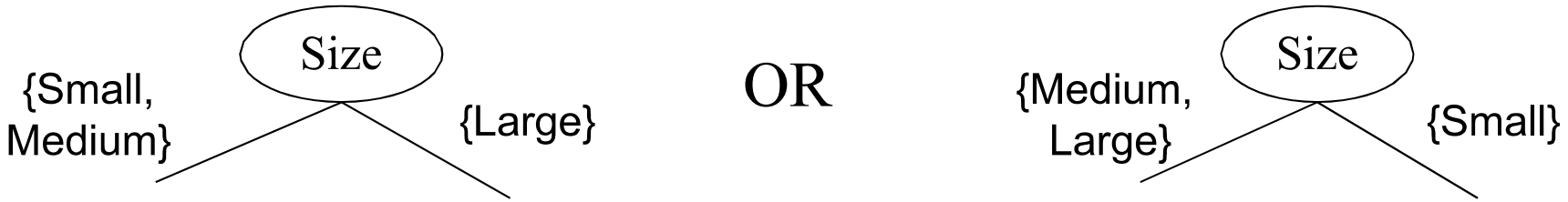
- Depends on number of ways to split
  - 2-way split
  - Multi-way split

Splitting Based on Nominal Attributes

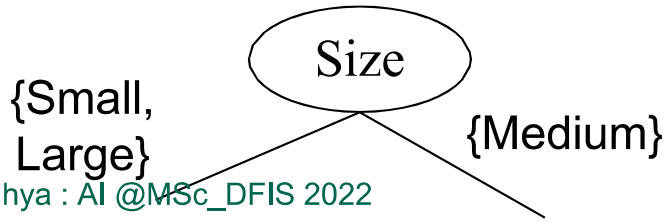
- Nominal Attribute: Divides values into two subsets. Need to find optimal partitioning.



- Ordinal Atribute: Divides values into two subsets. Need to find optimal partitioning.



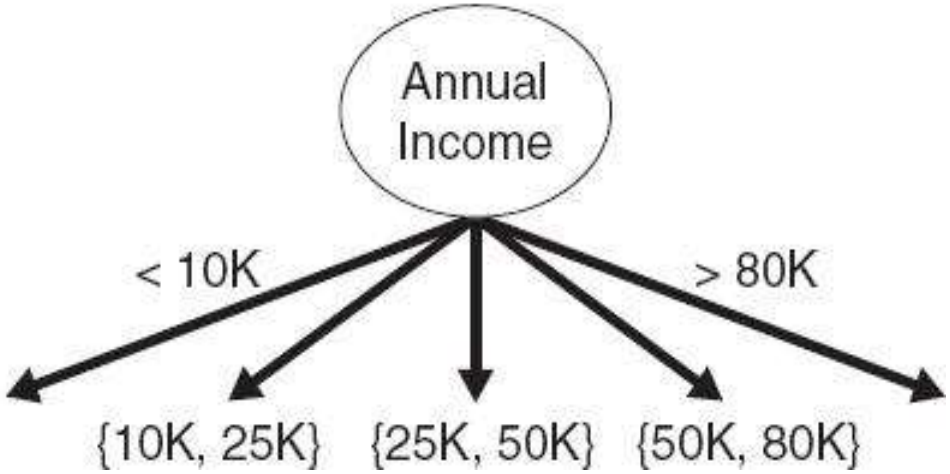
  - What about this split?



Splitting Based on ContinuousAttributes

Binary split

Multi-way split



Discretization to form an ordinal categorical attribute:

- **Static** – discretize the data set once at the beginning (equal interval, equal frequency, etc.).
- **Dynamic** – discretize during the tree construction.
  - Example: For a binary decision ($A < v$) or ($A \geq v$) consider all possible splits and finds the best cut. This can be done efficiently.

Tree Induction

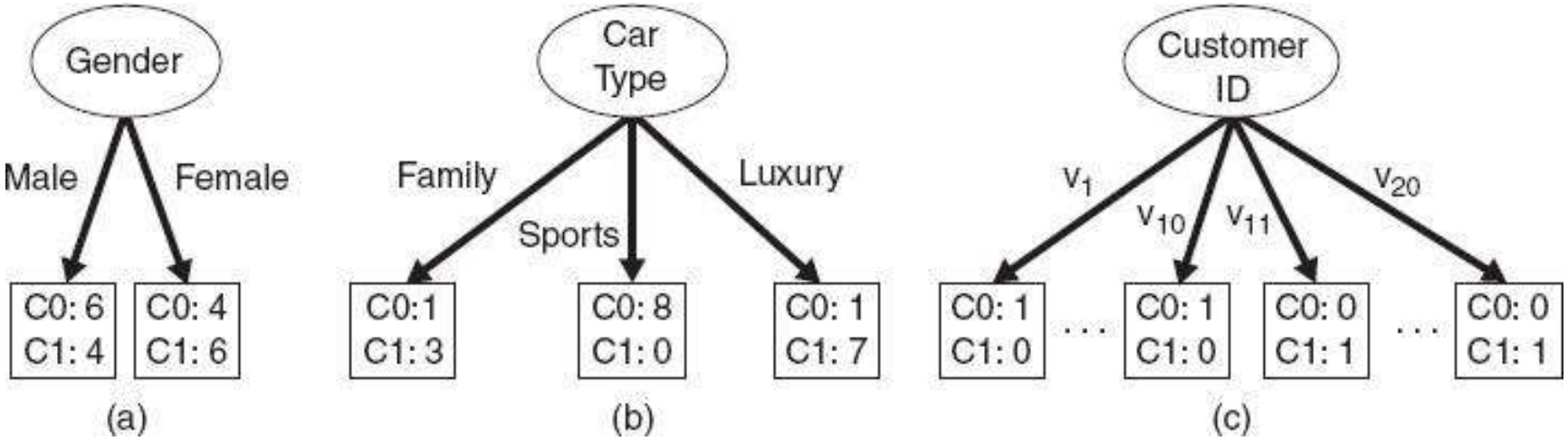
- Greedy strategy
  - —Split the records based on an attribute test that optimizes a certain criterion.

- Issues
  - —Determine how to split the record using different attribute types.
  - —**How to determine the best split?**
  - —Determine when to stop splitting

How to determine the Best Split

Before Splitting: 10 records of class 0,
10 records of class 1

C0: 10
C1: 10



Which test condition is the best?

How to determine the Best Split

- Greedy approach:
  - —Nodes with homogeneous class distribution are preferred
- Need a measure of node impurity:

C0:	**5**
C1:	**5**

Non-homogeneous,

High degree of impurity

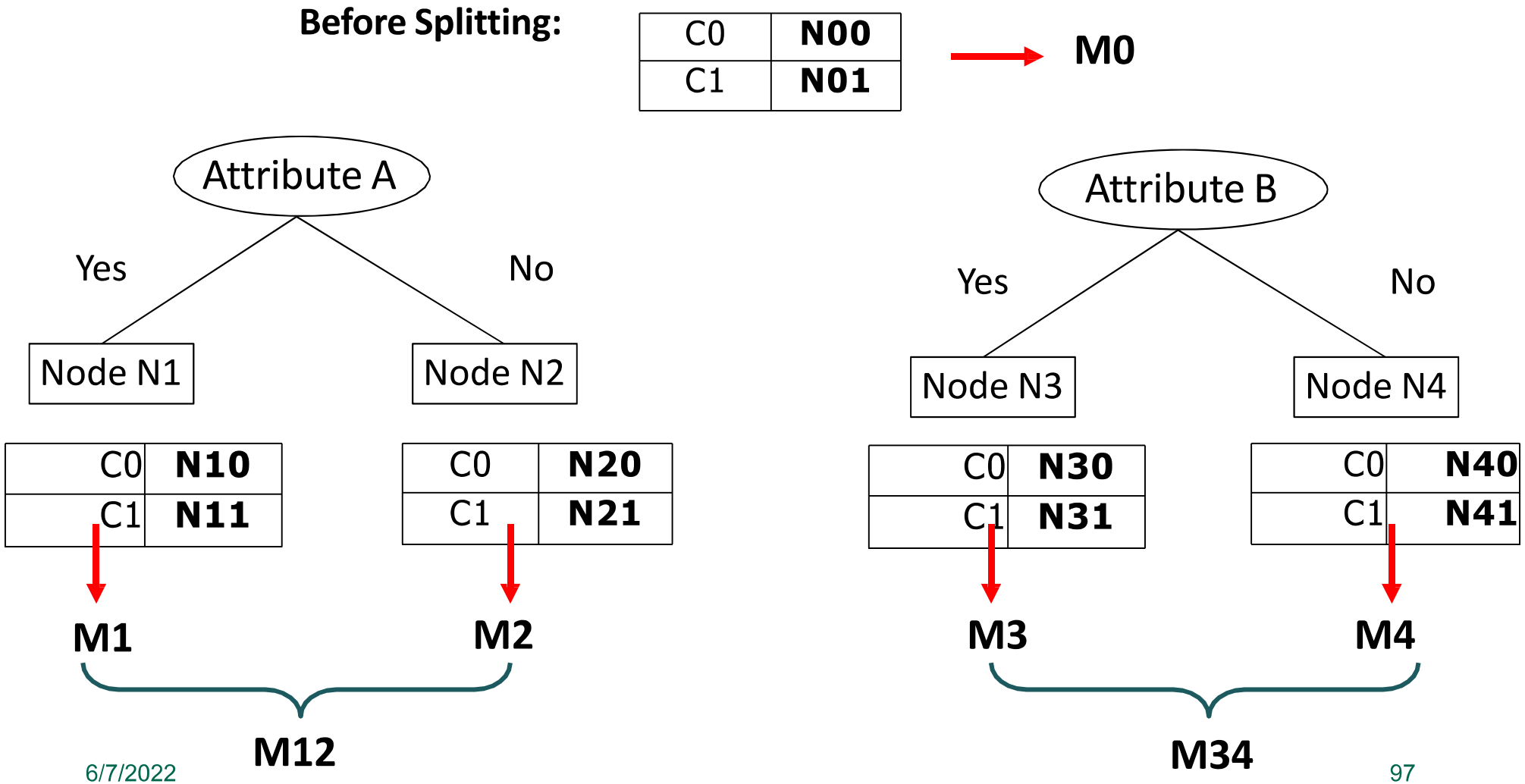
C0:	**9**
C1:	**1**

Homogeneous,

Low degree of impurity

Find the Best Split -General Framework

Assume we have a measure **M** that tells us how "pure" a node is.



Gain = M0 – M12 vs M0 – M34 → Choose best split

Tree Induction

- Greedy strategy
  - —Split the records based on an attribute test that optimizes a certain criterion.

- Issues
  - —Determine how to split the record using different attribute types.
  - —How to determine the best split?
  - —**Determine when to stop splitting**

Stopping Criteria for Tree Induction

- Stop expanding a node when **all the records belong to the same class**. Happens guaranteed when there is only one observation left in the node (e.g., Hunt's algorithm).
- Stop expanding a node when all the records in the node have the **same attribute values**. Splitting becomes impossible.
- **Early termination criterion** (to be discussed later with tree pruning)

Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Classification Algorithms

- ID3
  - Uses information gain
- C4.5
  - Uses Gain Ratio
- CART
  - Uses Gini

Measures of Node Impurity

Gini Index

Entropy

Classification error

Measure of Impurity: Entropy

- Entropy at a given node t:

$$\text{Entropy(t)} = -\sum_j p(j \mid t) \log(p(j \mid t))$$

$p(j \mid t)$ is the relative frequency of class j at node t;
$0 \log(0) = 0$ is used!

- Measures homogeneity of a node (originally a measure of uncertainty of a random variable or information content of a message).
- Maximum: $\log(n_c)$ when records are equally distributed among all classes = maximal impurity.
- Minimum: 0 when all records belong to one class = maximal purity.

Examples for computing Entropy

$$\text{Entropy(t)} = -\sum_j p(j \mid t) \log(p(j \mid t))$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

C1	**3**
C2	**3**

P(C1) = 3/6 P(C2) = 3/6

Entropy = – (3/6) $\log_2$ (3/6) – (3/6) $\log_2$ (3/6) = 1

Information Gain

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Parent Node, p is split into k partitions;
n_i is number of records in partition i

- Measures reduction in Entropy achieved because of the split. Choose the split that achieves most reduction (maximizes GAIN)
- Used in ID3, C4.5 and C5.0
- Disadvantage: Tends to prefer splits that result in large number of partitions, each being small but pure.

Decision Tree Induction: Training Dataset

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Attribute Selection: Information Gain

- In training data set, The class level attribute, *buys_computer,* has two distinct values (namely, {yes,no}); therefore, there are two distinct classes (m=2).

- **Let class P correspond to *yes* and N correspond to *no.***
- There are 9 samples of class yes and 5 samples of class no.

Attribute Selection: Information Gain

- Class P: buys_computer = “yes”
- Class N: buys_computer = “no”

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

Attribute Selection: Information Gain

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes's and 3 no's.

Attribute Selection: Information Gain

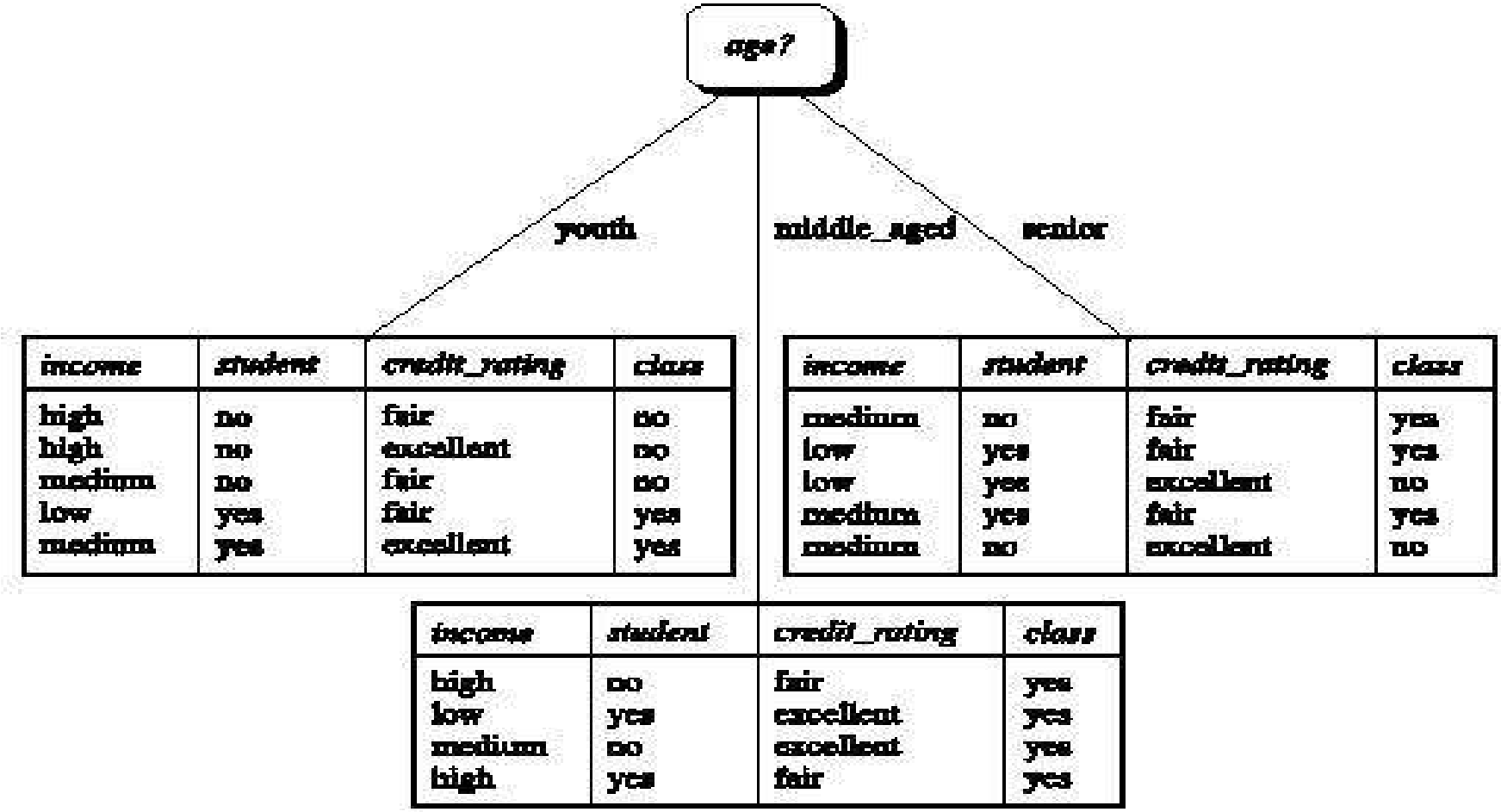
$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Attribute Selection: Information Gain

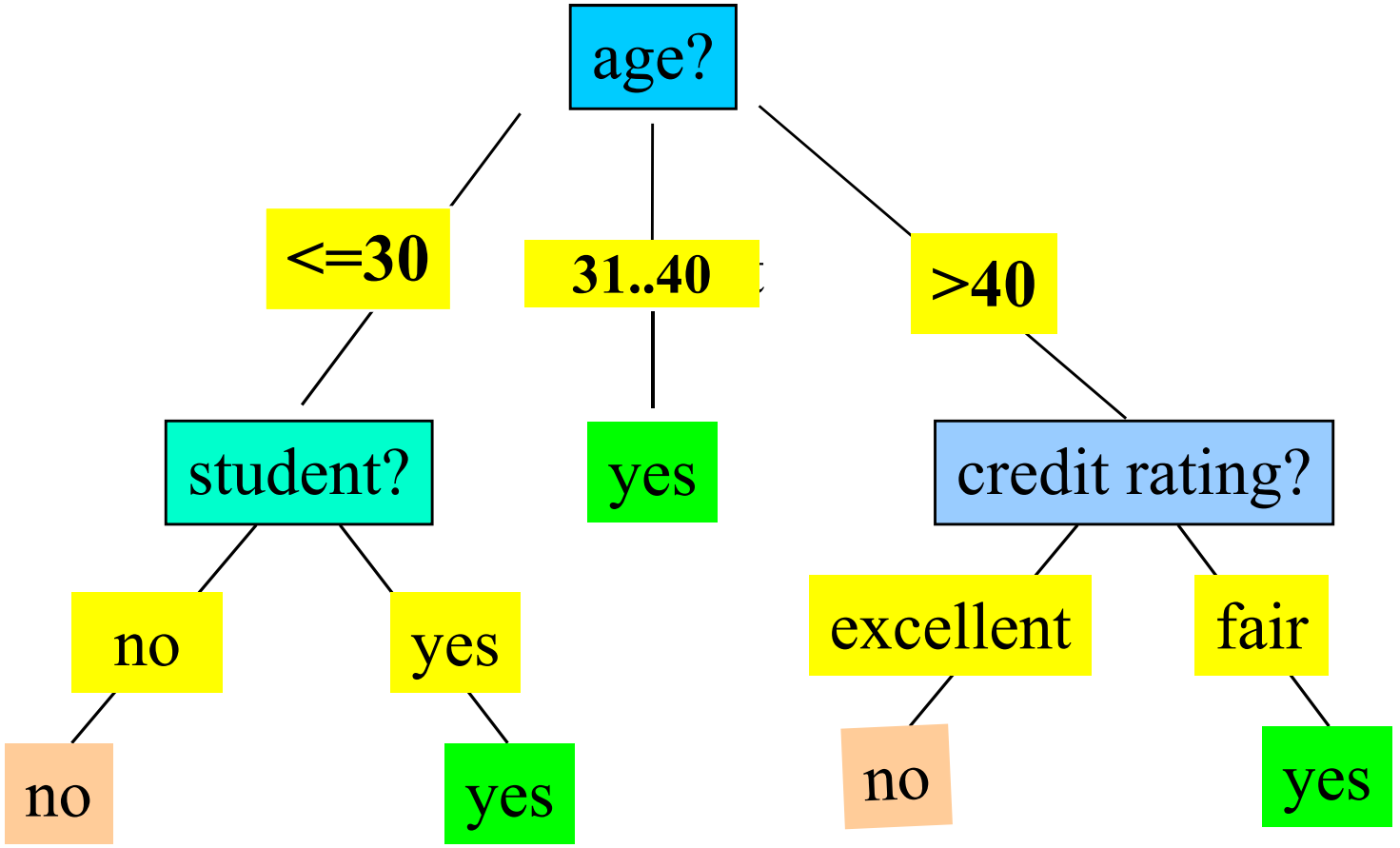


income	student	credit_rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

Output: A Decision Tree for "*buys_computer*"



Measure of Impurity: GINI

- Gini Index for a given node t :

$$GINI(t) = \sum_j p(j\,|t)(1 - p(j\,|t)) = 1 - \sum_j p(j\,|t)^2$$

$p(j \mid t)$ is estimated as the relative frequency of class j at node t

- Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
- Maximum of $1 - 1/n_c$ (number of classes) when records are equally distributed among all classes = maximal impurity.
- Minimum of 0 when all records belong to one class = complete purity.
- Examples:

C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

C1	3
C2	3
Gini=0.500	

Examples for computing GINI

$$GINI(t) = 1 - \sum_j p(j\,|t\,)^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = **0**

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)² – (5/6)² = **0.278**

C1	**2**
C2	**4**

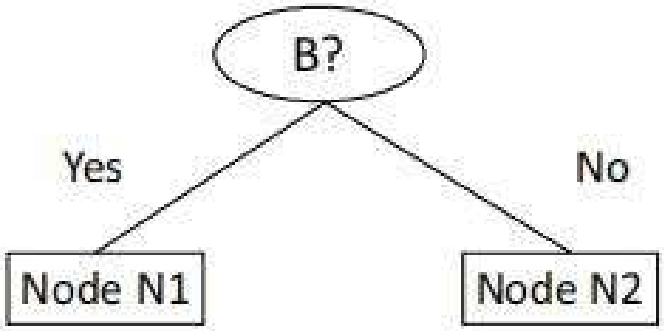
P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)² – (4/6)² = **0.444**

Maximal impurity here is ½ = .5

Binary Attributes: Computing GINI Index

- Splits into two partitions
- Effect of weighing partitions: Larger and purer partitions are sought for.



	Parent
C1	6
C2	6
Gini = 0.500	

Gini(N1)
$= 1 - (5/8)^2 - (3/8)^2$
$= 0.469$

Gini(N2)
$= 1 - (1/4)^2 - (3/4)^2$
$= 0.375$

	N1	N2
C1	5	1
C2	3	3
Gini=0.438		

Gini(Children)
$= 8/12 * 0.469 +$
$4/12 * 0.375$
$= 0.438$

GINI improves!

Advantages of Decision Tree Based Classification

INEXPENSIVE TO CONSTRUCT

EXTREMELY FAST AT CLASSIFYING UNKNOWN RECORDS

EASY TO INTERPRET FOR SMALL-SIZED TREES

ACCURACY IS COMPARABLE TO OTHER CLASSIFICATION TECHNIQUES FOR MANY SIMPLE DATA SETS

Why are decision tree classifiers so popular?

- The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery.
- Decision trees can handle high dimensional data.
- Representation of acquired knowledge in tree form is generally easy to assimilate by humans.
- The learning and classification steps of decision tree induction are simple and fast.
- In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand.
- Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, and molecular biology.
- Decision trees are the basis of several commercial rule induction systems.