Introduction to Machine Learning with R

By: Muhammad Apriandito Arya Saputra

About Me

Muhammad Apriandito Arya Saputra

- Junior Researcher Digital Business Ecosystem Research Center, Telkom University.
- Product Designer Telkom University School of Data Science.
- Social Media Strategist, Asosiasi Ilmuwan Data Indonesia

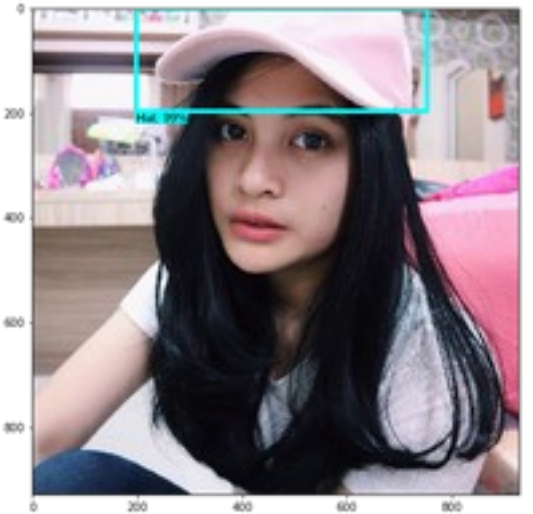
Research Area: Big Data Analytic, Computer Vision, Machine Learning, Natural Language Processing

ASOSIASI ILMUWAN DATA INDONESIA

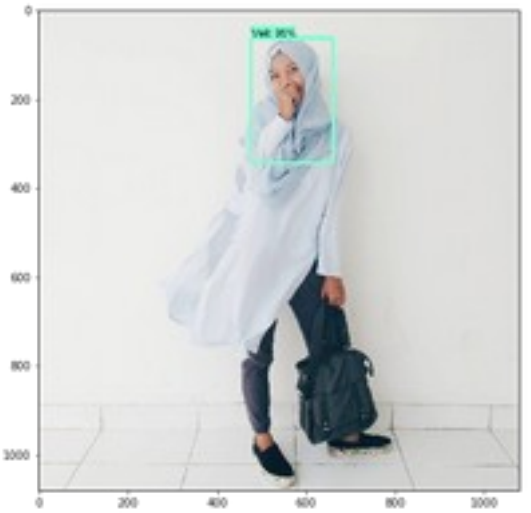
DIGITAL BUSINESS ECOSYSTEM RESEARCH CENTER

Research and Publication

- Object Detection using Convolutional Neural Network to Identify Popular Fashion Product. *(Journal of Physics: Conference Series Volume 1192, Number 1)*
- A Comparative Study of Hollywood Movie Successfulness Prediction Model.
- Forecasting Portfolio Optimization using Artificial Neural Network and Genetic Algorithm.
- Analysis of Customer Chat using Text Mining for Customer Relationship Management.




Eyeglasses: 99%




Eyeglasses: 99%

Telkom University School of Data Science

- **Data Science:**
  - R Academy
  - Python Academy
  - Social Media Analytic Academy
  - International Big Data Certification
  - Big Data for High School

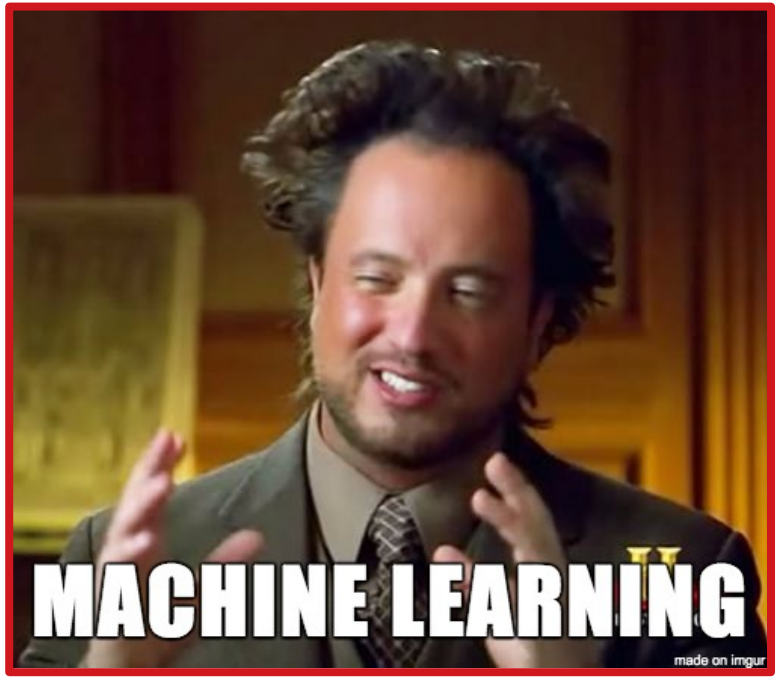
- **Outside Data Science:**
  - International Block Chain Certification

Outline

- **Day 1**
  - Introduction to Machine Learning.
  - Machine Learning Process.
- **Day 2**
  - Clustering and Classification Model.
  - Regression Model.
- **Day 3**
  - Capstone Project.

Objective

- After the training program, the participant's should be able to:
  - Understand how Machine Learning algorithm work.
  - Create Machine Learning Model using R.
  - Use Machine Learning to solve specific problem.

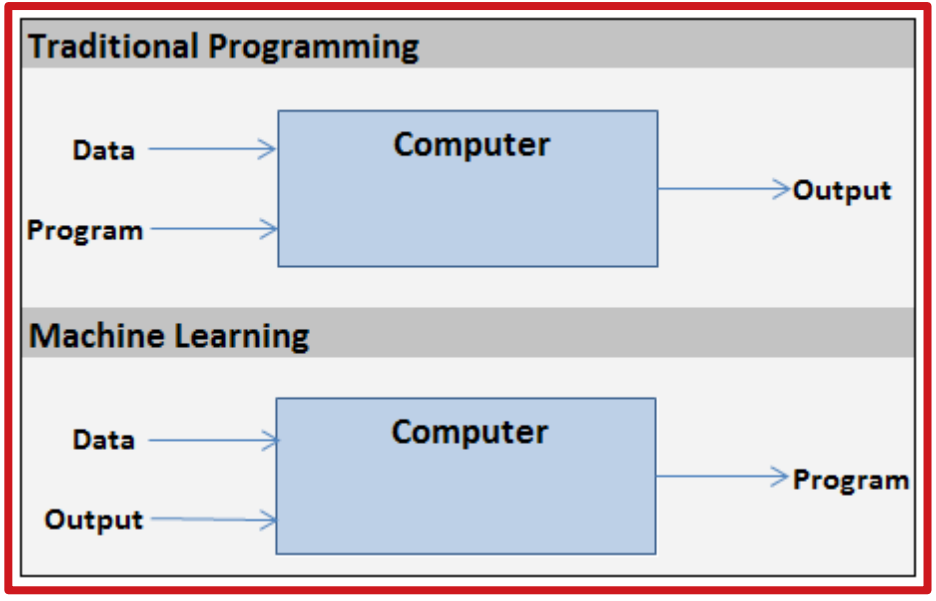


Machine Learning?

What's in Your head, the first time you heard about Machine Learning?

Machine Learning

- Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given.
- Machine Learning aim to uncover hidden patters and unknown correlation, and to find useful information from data.



Story Behind Machine Learning

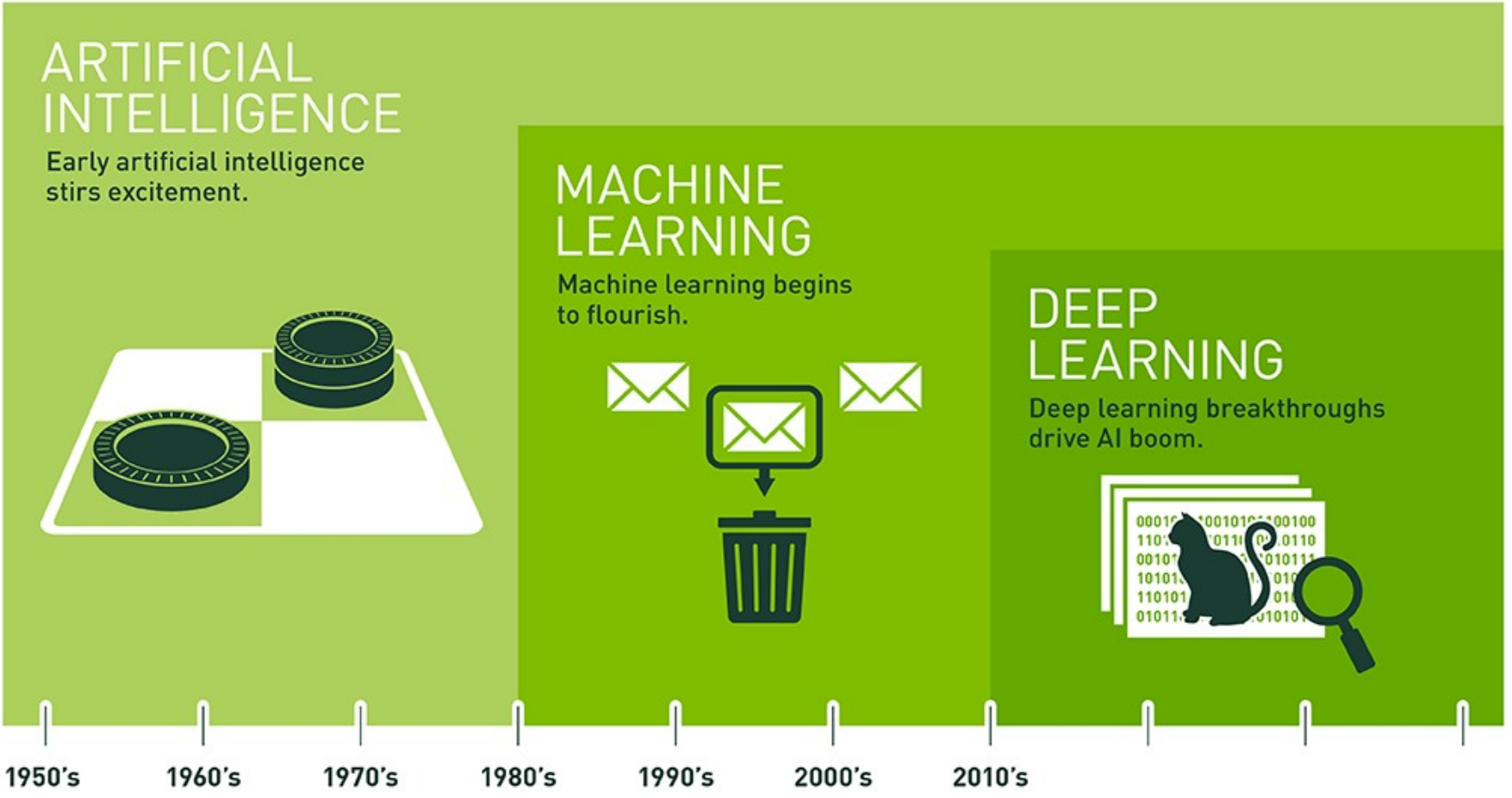
Experience	Salary (Rupees)
2	3,00,000
4	6,00,000
6	9,00,000
10	15,00,000
12	24,00,000
14	28,00,000

```
if (experience < = 10)
{ salary = experience * 1.5 * 100000}
else if(experience >10)
{ salary = experience * 2 * 100000}
```

Experience	Job Level	Rare Skill?	Salary (Rupees)
2	3	Yes	4,50,000
4	3	No	6,00,000
6	4	No	7,50,000
10	5	Yes	18,00,000
12	5	No	15,00,000
14	6	No	18,00,000

Experience	Job Level	Rare Skill?	Salary (Rupees)
2	3	Yes	4,50,000
4	3	No	6,00,000
6	4	No	7,50,000
10	5	Yes	18,00,000
12	5	No	15,00,000
14	6	No	18,00,000

```
Salary = Experience * Magic_Number_1 + JobLevel * Magic_Number_2 +
Skill * Magic_Number_3 + Magic_Number_4
```



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Why Now

- There is an extraordinary convergence of large volumes of Big Data, unprecedented computing power, and sophisticated self-learning algorithms taking place.
- The affordability, viability, and feasibility of these three technologies are the driving forces behind why machine learning is becoming more and more prevalent today.

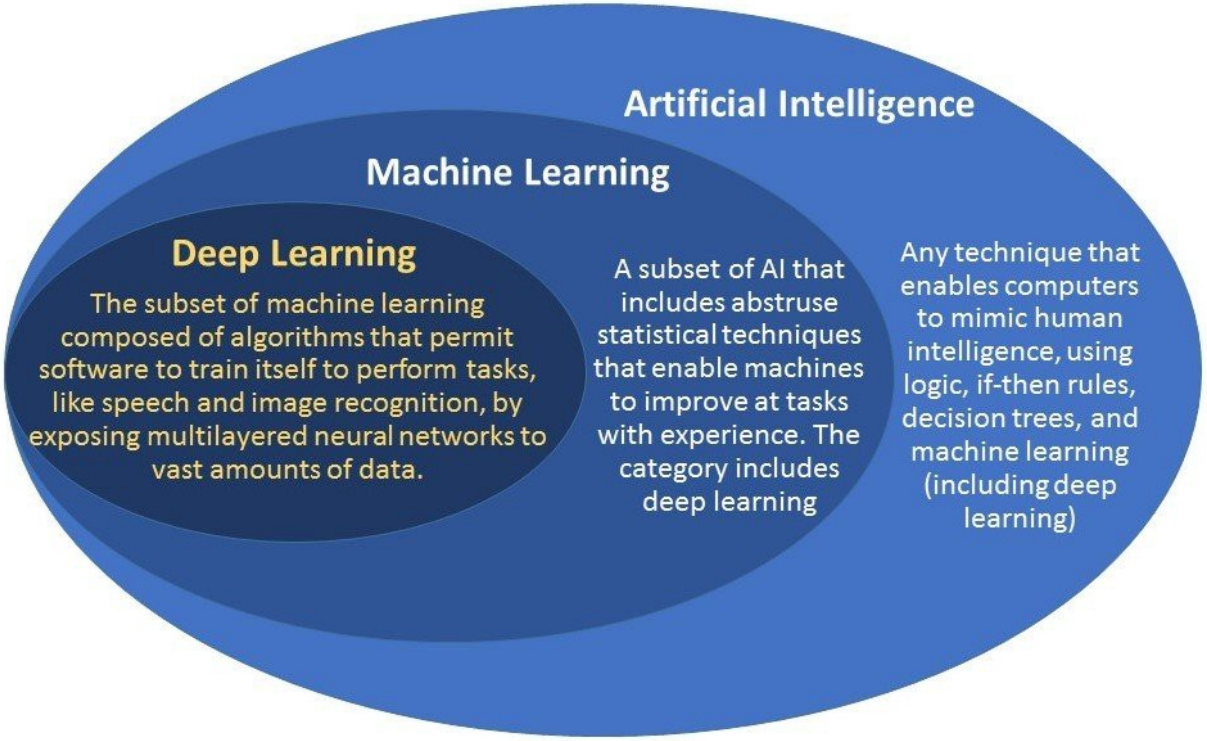
Machine Learning Application

- **Image Recognition**
  - For Face/Object Recognition
  - For Character Recognition
- **Spech Recognition**
- **Financial Services**
- **Sentiment Analysis**
- **News Classification**

Machine Learning Application (Cont.)

- **Services of Social Media**
- **Recommendation for Products and Services**
- **Online Customer Supports**
- **Virtual Personal Assistant**

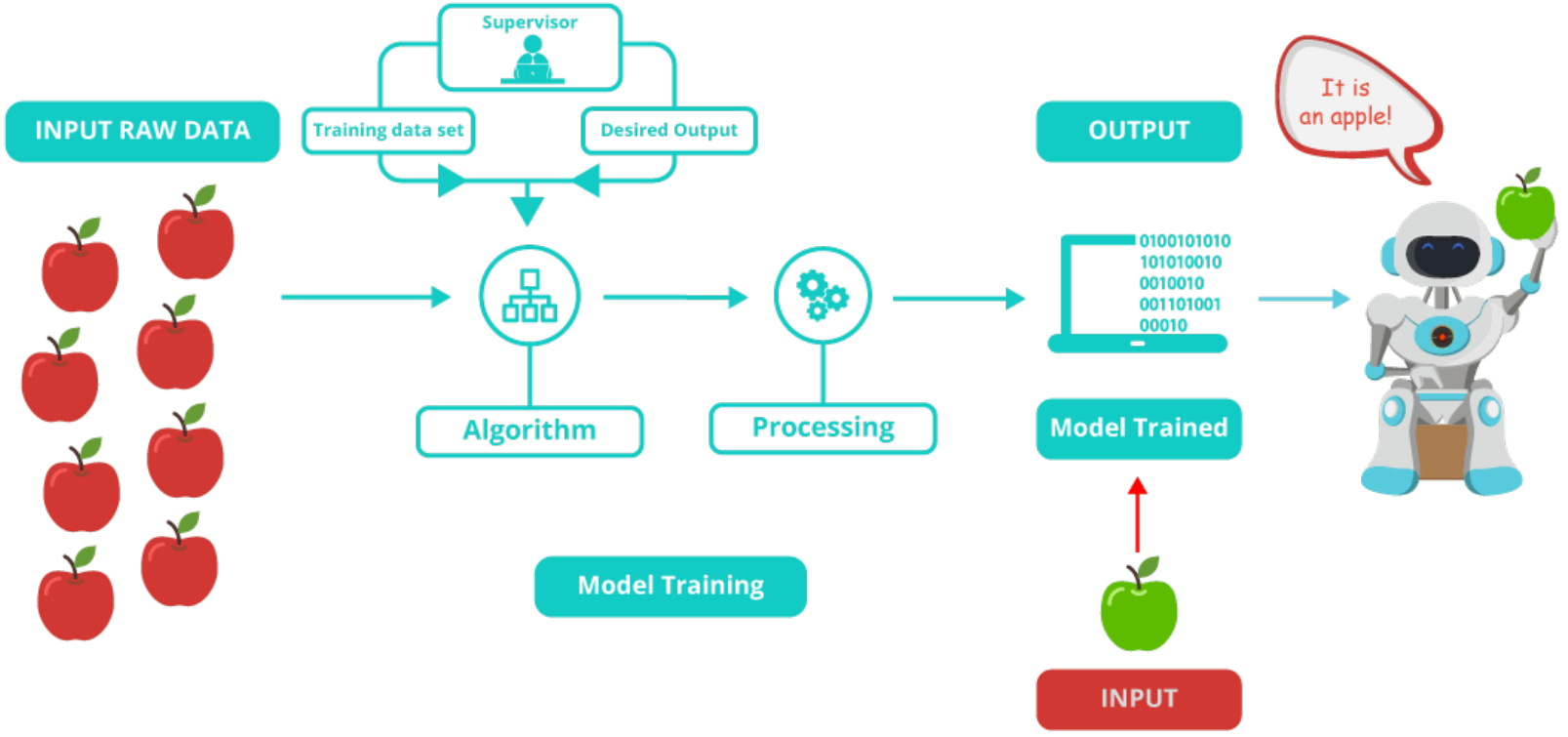
DL, ML, AI?


Artificial Intelligence
Any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)
Machine Learning
A subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning
Deep Learning
The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.

Type of Machine Learning

- Supervised Learning.
- Semi-Suervised Learning.
- Unsupervised Learning.
- Reinforcement Learning.
- Continuous Learning.

Supervised Learning

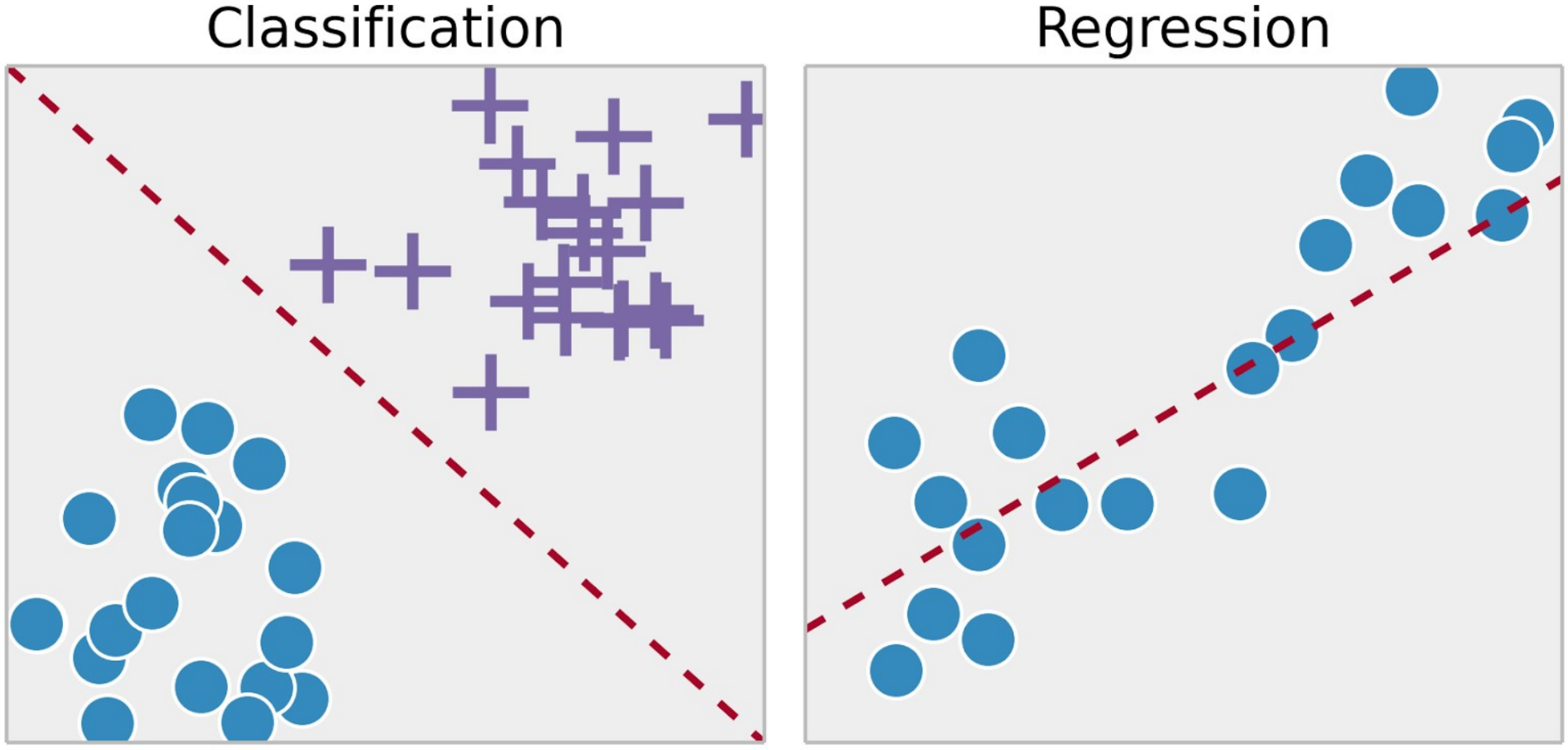


Supervised Learning

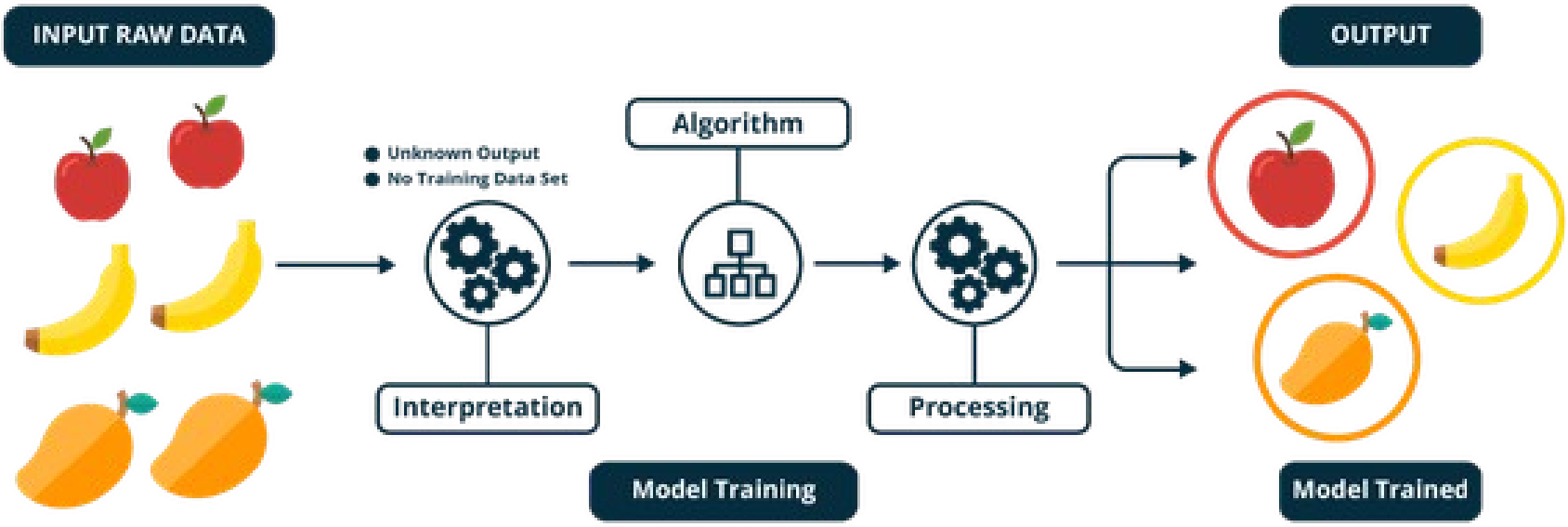
Supervised Learning approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples. Some classification problem are:

- Classification: A classification problem is when the output variable is a category or a group, such as “black” or “white” or “spam” and “no spam”.

- Regression: A regression problem is when the output variable is a real value, such as “weight” or “height.”


Classification
Regression

Unsupervised Learning

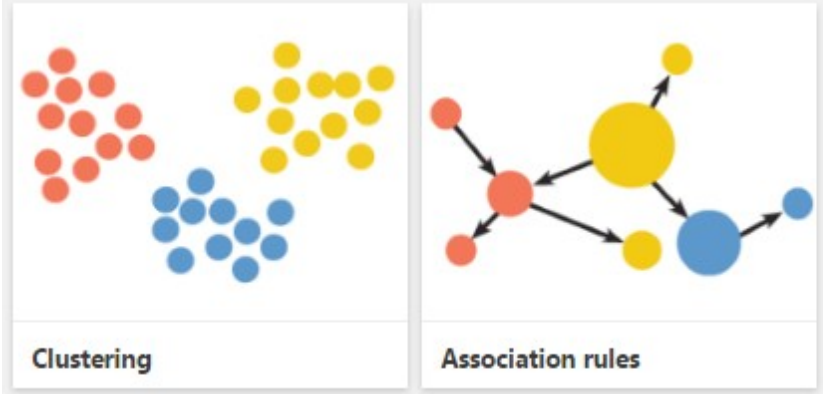


Unsupervised Learning

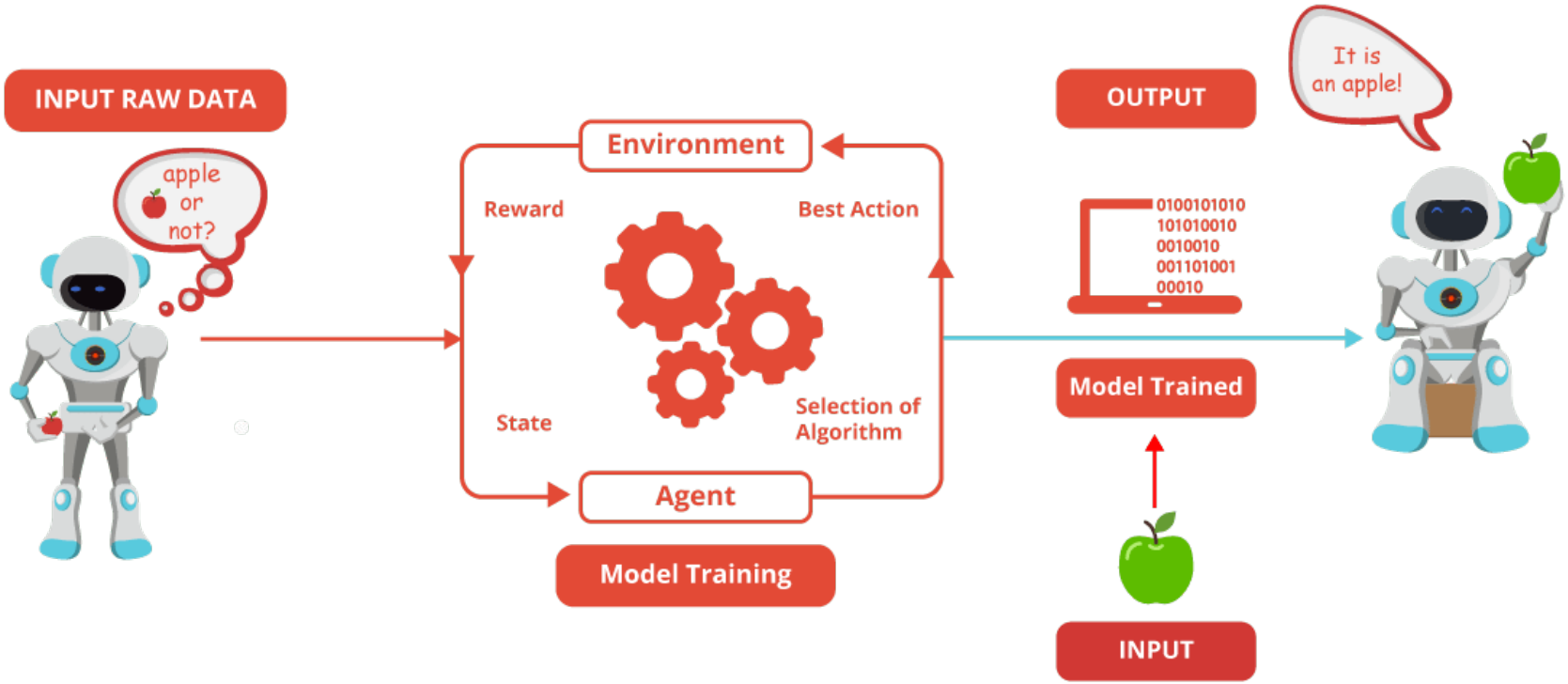
Mathematically, Unsupervised learning is where you only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

- Clustering: Grouping a set of objects in such a manner that objects in the same group are more similar than to those object belonging to other groups.

- Association: Finding associations amongst items within large commercial databases.

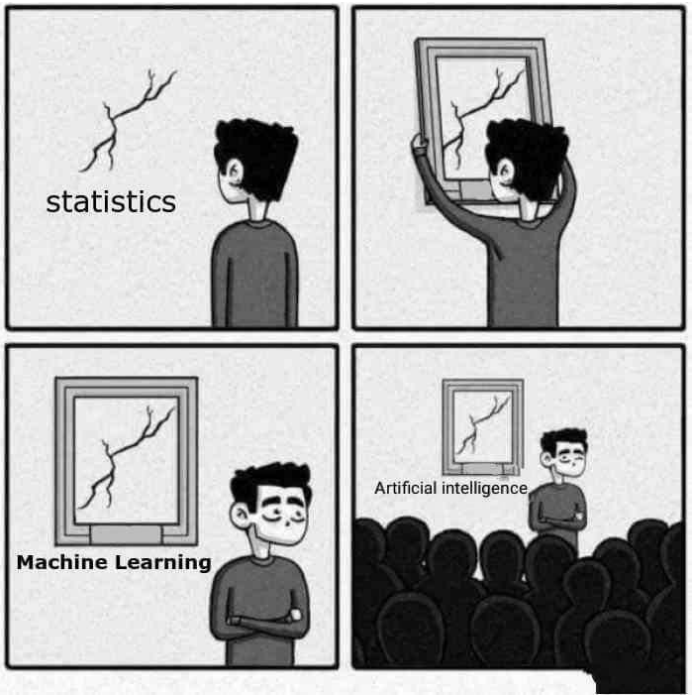


Reinforcement Learning

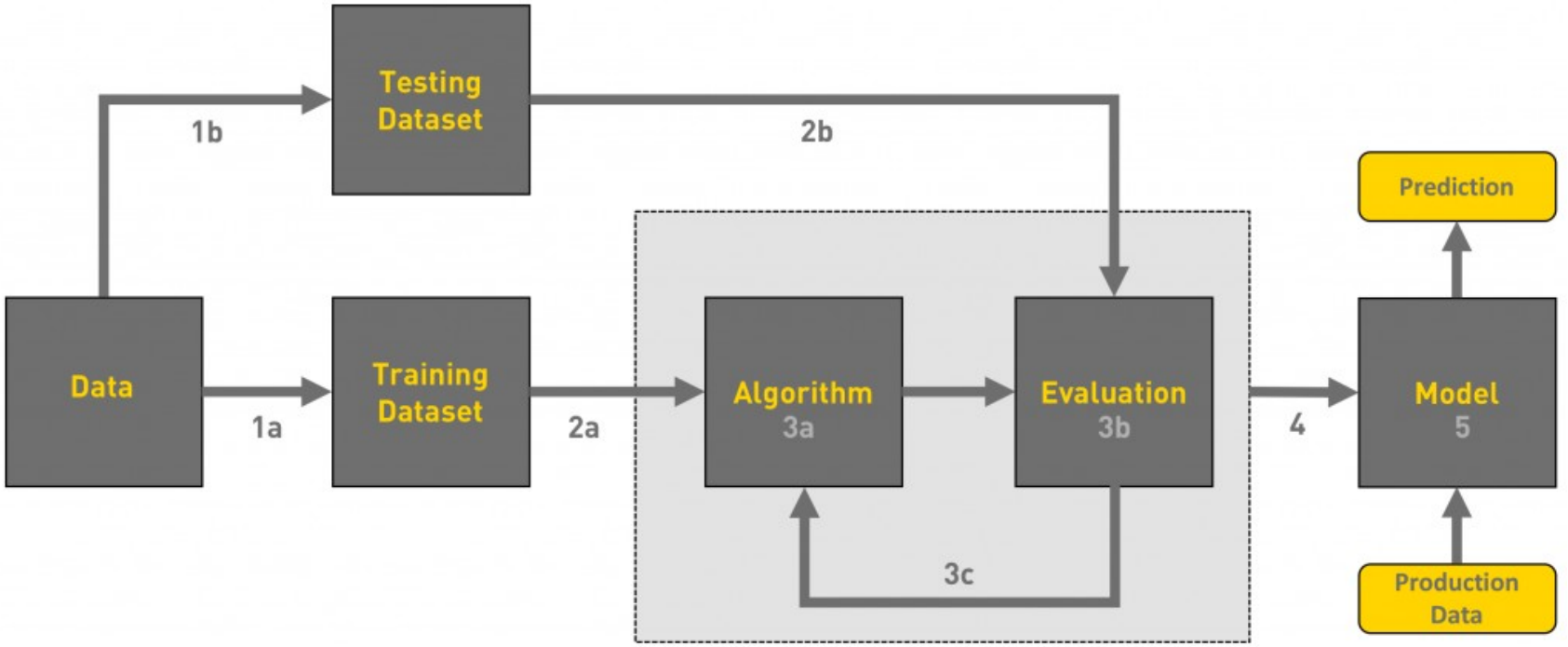


Some Useful Information

- In machine learning, a target is called a label.
- In statistics, a target is called a dependent variable.
- A variable in statistics is called a feature in machine learning.
- A transformation in statistics is called feature creation in machine learning.



Machine Learning Workflow



Machine Learning Workflow

- We can define the machine learning workflow in 5 stages:
  - Gathering data
  - Data pre-processing
  - Researching the model that will be best for the type of data
  - Training and testing the model
  - Evaluation

1. Gathering Data

- Depends on the type of project.
- Source of the data can be:
  - Database
  - Sensor
  - Social Media
  - Kaggle
  - Etc.

2. Data pre-processing

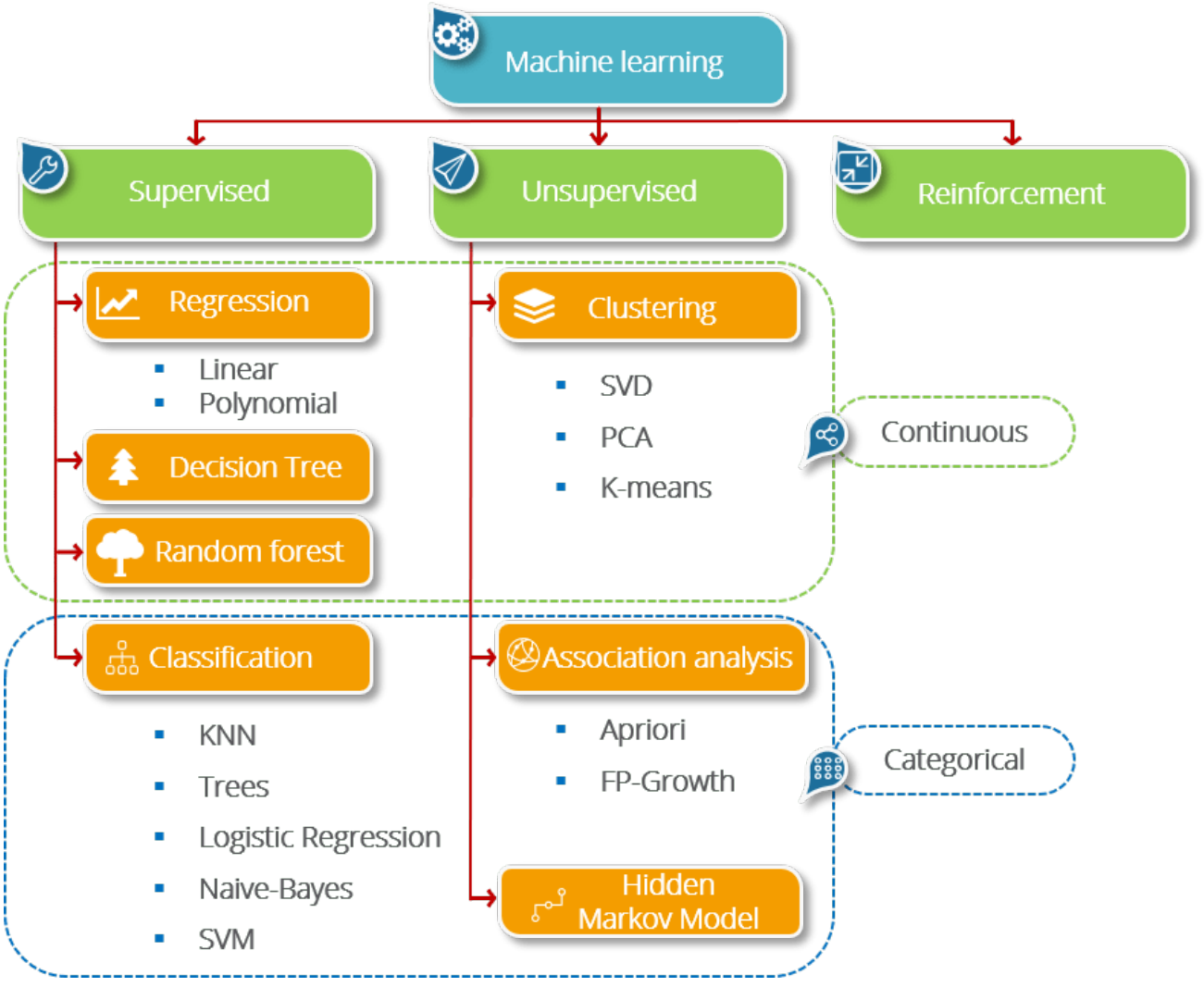
- Process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean/conditioned data set.
- The most important steps in machine learning.

Why do we need Data Pre-Processing?

- Most of the real-world data is messy.
- Type of messy data:
  - Missing data
  - Noisy data
  - Inconsistent data
- Garbage in Garbage Out

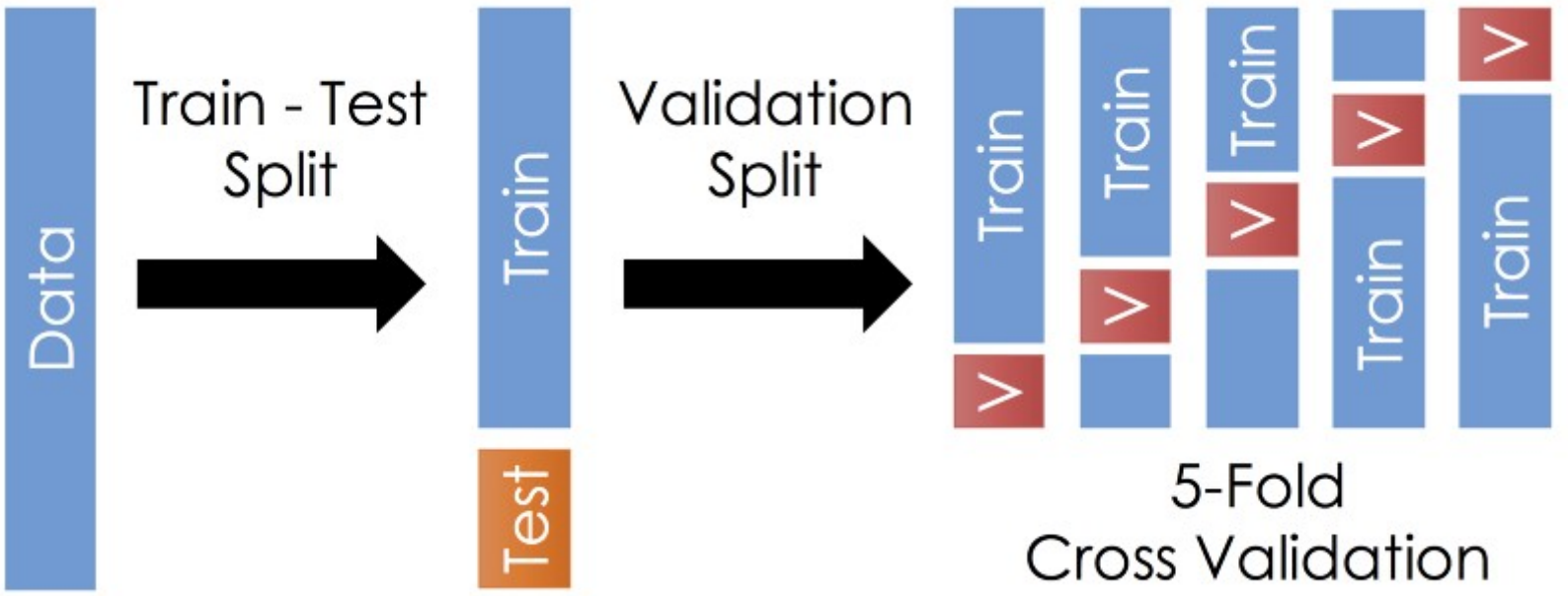
3. Searching the best model

- Depends on the purpose, and the condition of the data.
- No free lunch theorem
  - there is no one model that works best for every problem.
  - No algorithm is the best, No universal best algorithm.


Machine learning
Supervised
Unsupervised
Reinforcement
Regression
Linear
Polynomial
Decision Tree
Random forest
Clustering
SVD
PCA
K-means
Continuous
Classification
KNN
Trees
Logistic Regression
Naive-Bayes
SVM
Association analysis
Apriori
FP-Growth
Categorical
Hidden Markov Model

4. Training and testing the model on data

- To measure the performance of Machine Leaning Model.



5. Evaluation

- Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

Why use R?

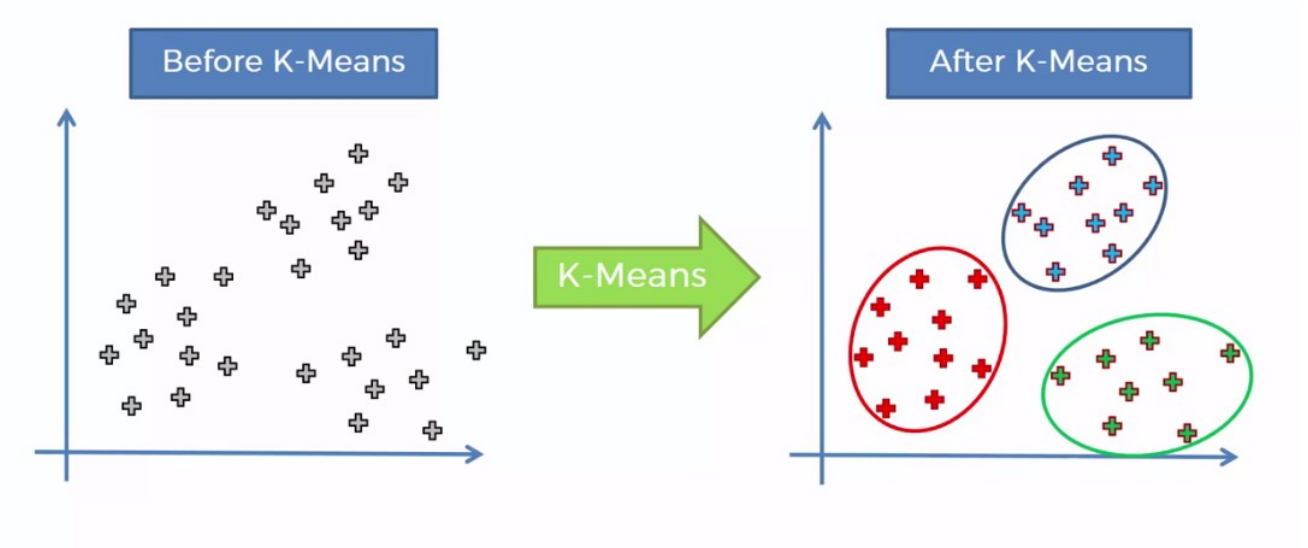
- R is **FREE!**
- R provides many machine learning packages and visualization function.
- All we need to know is how each algorithm work and use written package to generate prediction model on data with a few command lines.

Clustering

R-Academy: Ramadhan Edition

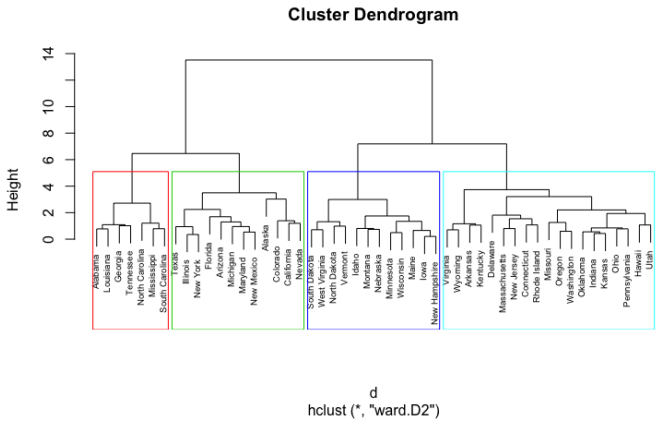
Clustering

- **Clustering** is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

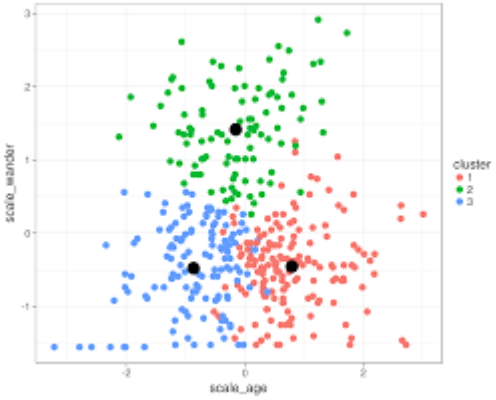


Types of Clustering

- Hierarchical algorithms: these find successive clusters using previously established clusters.
  - Agglomerative ("bottom-up"): Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters.
  - Divisive ("top-down"): Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

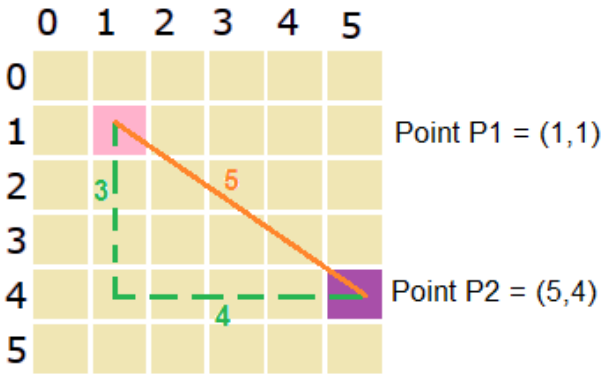


- Partitional clustering: Partitional algorithms determine all clusters at once They include:
  - *K*-means and derivatives
  - Fuzzy *c*-means clustering
  - QT clustering algorithm



Distance Measures

- Distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters. They include:
  - The Euclidean distance
  - The Manhattan distance

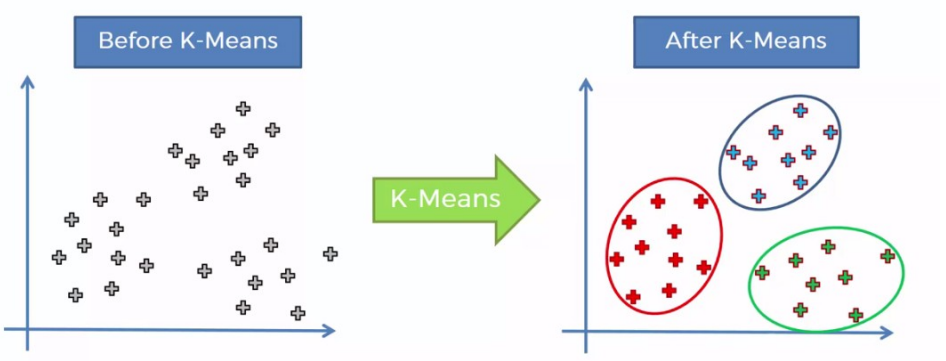


Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

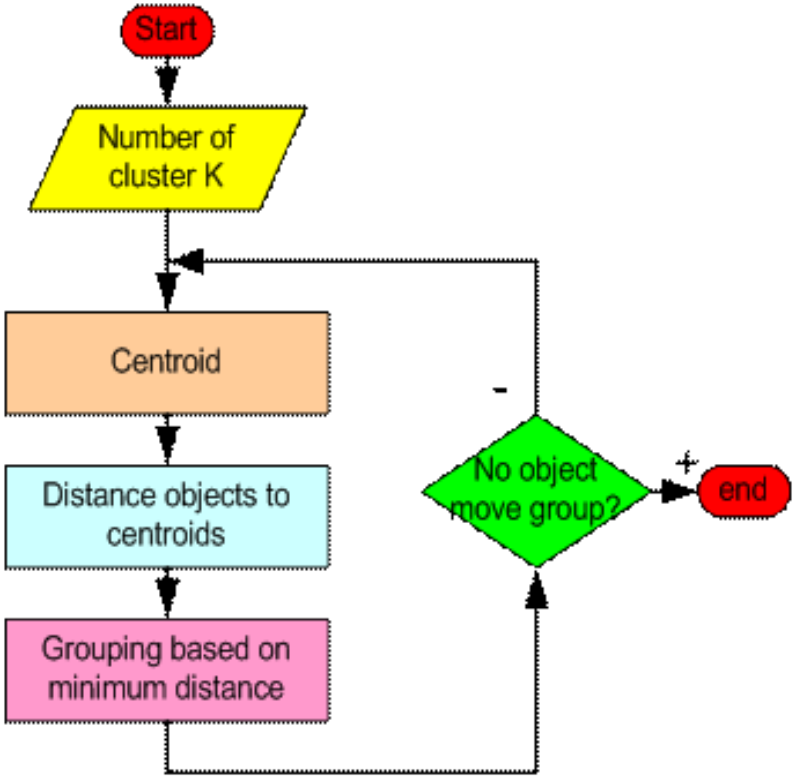
Manhattan distance = $|5-1| + |4-1| = 7$

K-Means Clustering

- The **k-means algorithm** is an algorithm to cluster *n* objects based on attributes into *k* partitions, where $k < n$.
- Simply speaking k-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group
- K is positive integer number.
- The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.



How K-Means Work



How K-Means Work

- Step 1: Begin with a decision on the value of k = Number of clusters.
- Step 2: Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following:
  - Take the first k training sample as single-element clusters
  - Assign each of the remaining (N-k) training sample to the cluster with the nearest centroid. After each assignment, recompute the centroid of the gaining cluster.

How K-Means Work (Cont.)

- Step 3: Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.
- Step 4: Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

Example

- A Simple example showing the implementation of k-means algorithm (using K=2)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Example

- Step 1: Initialization: Randomly we choose following two centroids (k=2) for two clusters.
- *In this case the 2 centroid are: m1=(1.0,1.0) and m2 =(5.0,7.0).*

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Example

Step 2:

Thus, we obtain two clusters containing: {1,2,3} and {4,5,6,7}. Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

$$d(m_1, 2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} = 1.12$$

$$d(m_2, 2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

Example

Step 3:

Now using these centroids we compute the Euclidean distance of each object, as shown in table.Therefore, the new clusters are:

{1,2} and {3,4,5,6,7}

Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

Example

Step 4 :

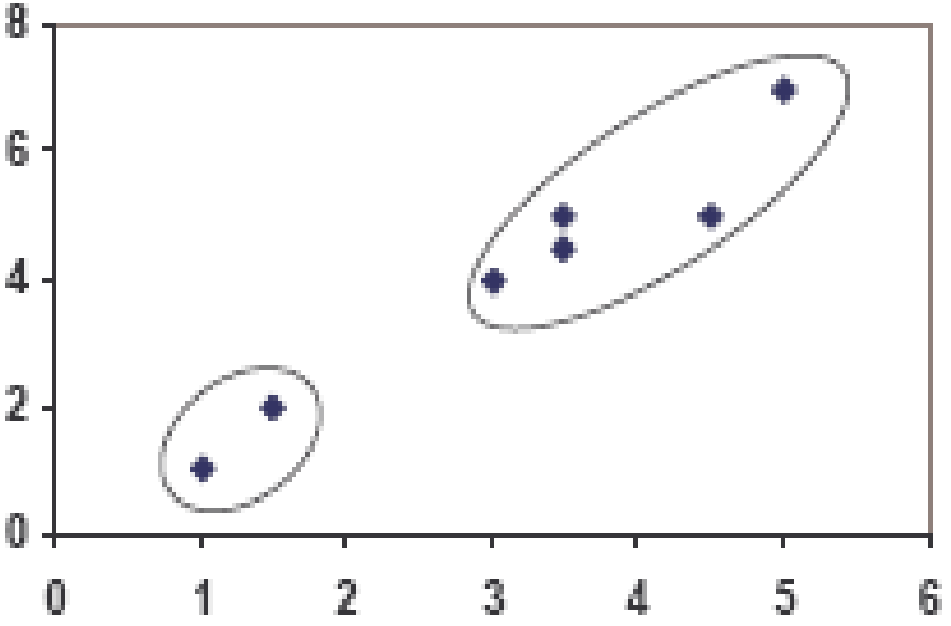
The clusters obtained are: {1,2} and {3,4,5,6,7}

Therefore, there is no change in the cluster.

Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.

Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

Example



Weaknesses of K-Mean Clustering

- When the numbers of data are not so many, initial grouping will determine the cluster significantly.
- The number of cluster, K, must be determined before hand. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.
- We never know the real cluster, using the same data, because if it is inputted in a different order it may produce different cluster if the number of data is few.
- It is sensitive to initial condition. Different initial condition may produce different result of cluster. The algorithm may be trapped in the local optimum.

Practice Time

Let's Build Clustering Model using R

Do it in group!

1. Make K-Means Clustering using agriculture dataset data(agriculture).
2. Visualize the model.
3. Determine the number of optimal cluster.
4. Make an Analysis.

Classification

R-Academy: Ramadhan Edition

Classification

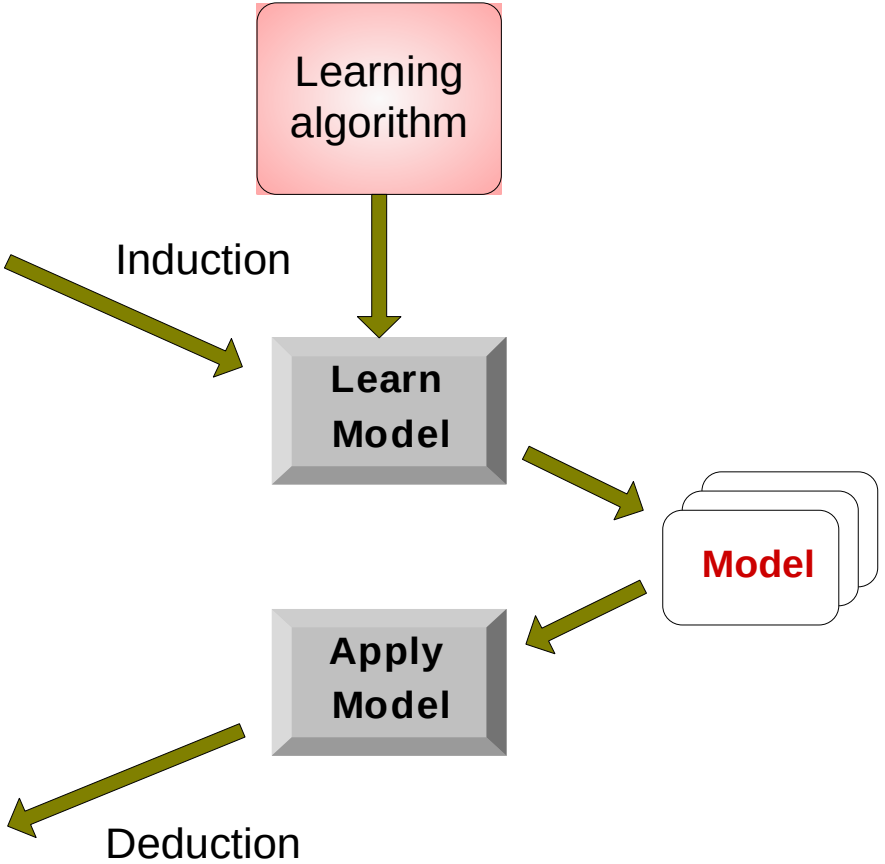
- In machine learning and statistic, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.
- Examples:
  - Assigning a given email to the “spam” or “non-spam” class,
  - Assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.)

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

Training 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	?

Test Set



Example of classification application

- **Handwriting recognition:** used to interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices
- **Speech recognition:** used for recognition and translation of spoken language into text by computers.
- **Biological classification:** used for classifying biological organism on the basis of shared characteristics (taxonomy)
- **Credit scores:** used to determine who qualifies for a loan, at what interest rate, and what credit limits.

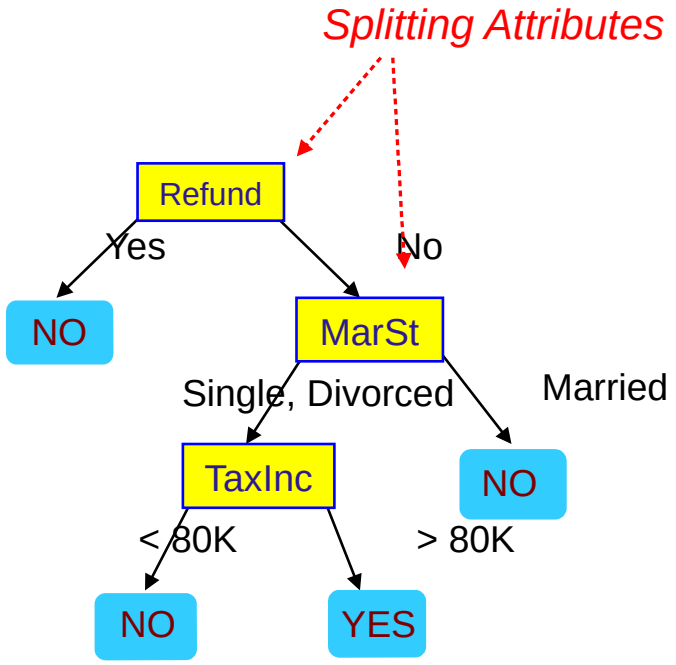
Decision Tree

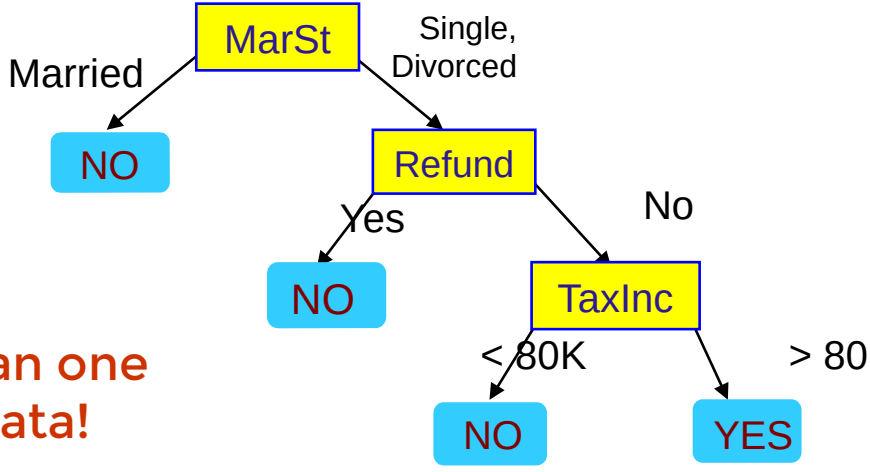
- Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions.
- It is one of the most widely used and practical methods for supervised learning.
- Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.

categorical categorical continuous 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

Training Data

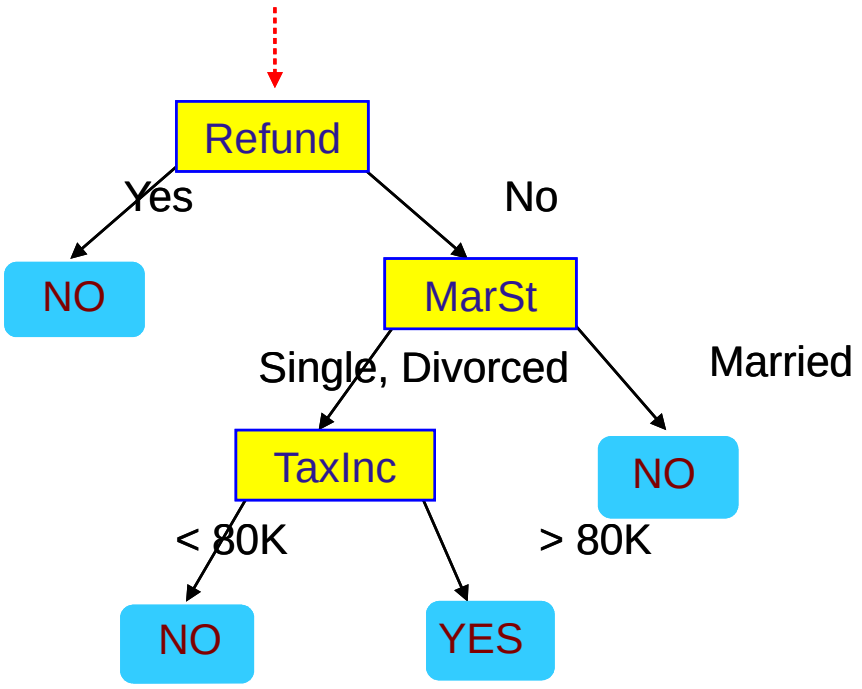




There could be more than one tree that fits the same data!

Apply to Test Data

Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Decision Tree Summary

- **Advantages:**
  - 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
- **Disadvantages:**
  - Overfitting when algorithm capture noise in the data
  - The model can get unstable due to small variation of data
  - Low biased tree: difficult for the model to work with new data

Naive Bayes

- Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling.
- Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- P(A|B) : the conditional probability that event A occurs , given that B has occurred. This is also known as the posterior probability.
- P(A) and P(B) : probability of A and B without regard of each other.
- P(B|A) : the conditional probability that event B occurs , given that A has occurred.

Example of Bayes Theorem

Given:

- A doctor knows that meningitis causes stiff neck 50% of the time
- Prior probability of any patient having meningitis is 1/50,000
- Prior probability of any patient having stiff neck is 1/20

If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

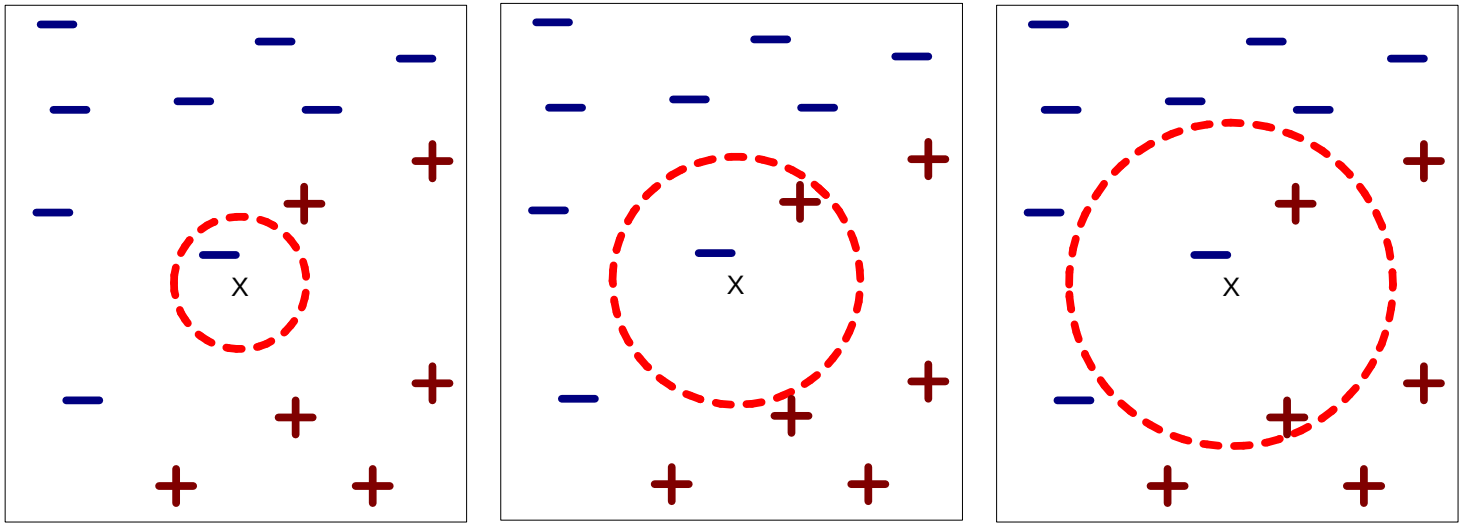
Types of Naive Bayes Classifier

- Multinomial Naive Bayes: This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc.
- Bernoulli Naive Bayes: This is similar to the multinomial naive bayes but the predictors are boolean variables.
- Gaussian Naive Bayes: When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

Naive Bayes

- Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, recommendation systems etc.
- Naive are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real life cases, the predictors are dependent, this hinders the performance of the classifier.

K-Nearest Neighborhood



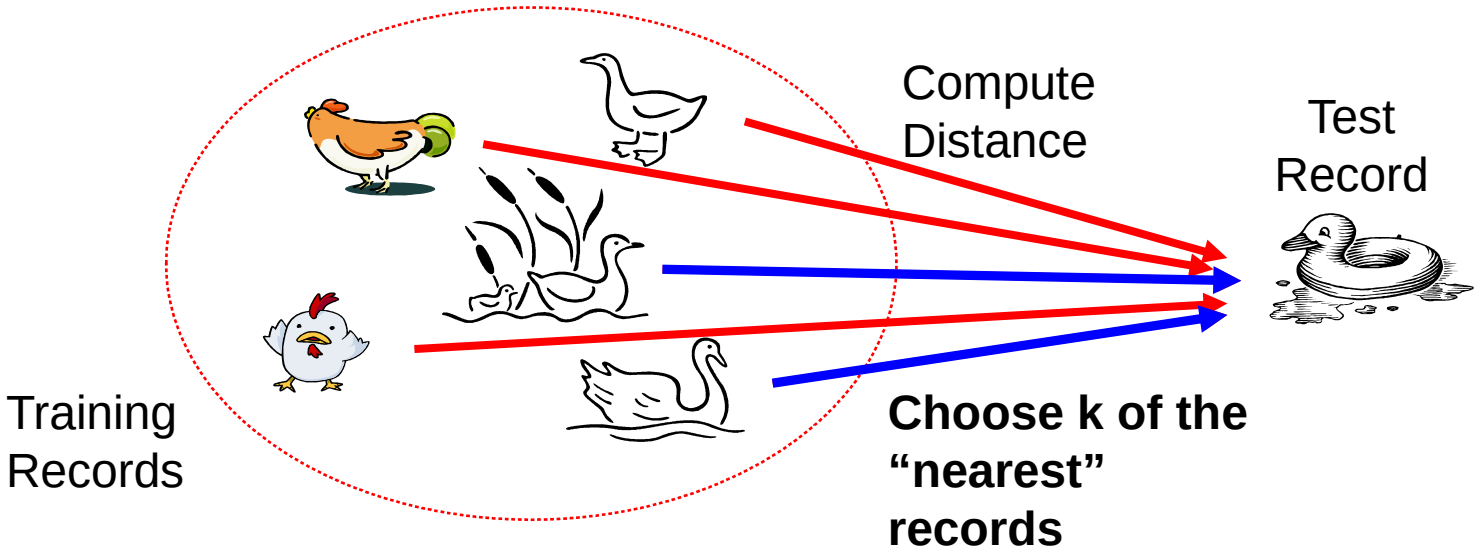
(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x at k data points that have smallest distance to x

Illustration Nearest Neighbor Classifiers

Basic idea:

- If it walks like a duck, quacks like a duck, then it's probably a duck.



K-NN

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes

- k-NN classifiers are lazy learners
  - It does not build models explicitly
  - Unlike eager learners such as decision tree induction and rule-based systems
  - Classifying unknown records are relatively expensive

Metrics for Performance Evaluation

- Focus on the predictive capability of a model Rather than how fast it takes to classify or build models, scalability, etc.

Confusion Matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	TP	FN
	Class=No	FP	TN

- TP (true positive)
- FN (false negative)
- FP (false positive)
- TN (true negative)

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Computing the Metrics

		PREDICTED CLASS	
ACTUAL CLASS	5,000 observations	Class=Yes	Class=No
	Class=Yes	TP = 4000 4000/5000 = 80%	FN = 50 50/5000=1%
	Class=No	FP = 200 200/5000 = 4%	TN = 750 750/5000 = 15%

- Precision = TP / (TP + FP) = 4000 / (4000+200) = 95%
- Recall or Sensitivity = TP / (TP + FN) = 4000 / (4000+50) = 99%
- Specificity = TN / (FP + TN) = 750 / (200+750) = 79%
- Accuracy = (TP + TN) / (TP + FP + TN + FN) = (4000+750) / (4000+200+750+50) = 95%

Practice Time

Let's Build Classification Model using R

Do it in group!

1. Make a Prediction of employee attrition using HR-Employee-Attrition.CSV
2. Free to use any algorithm.
3. Share Your Highest Accuracy with your Friend.

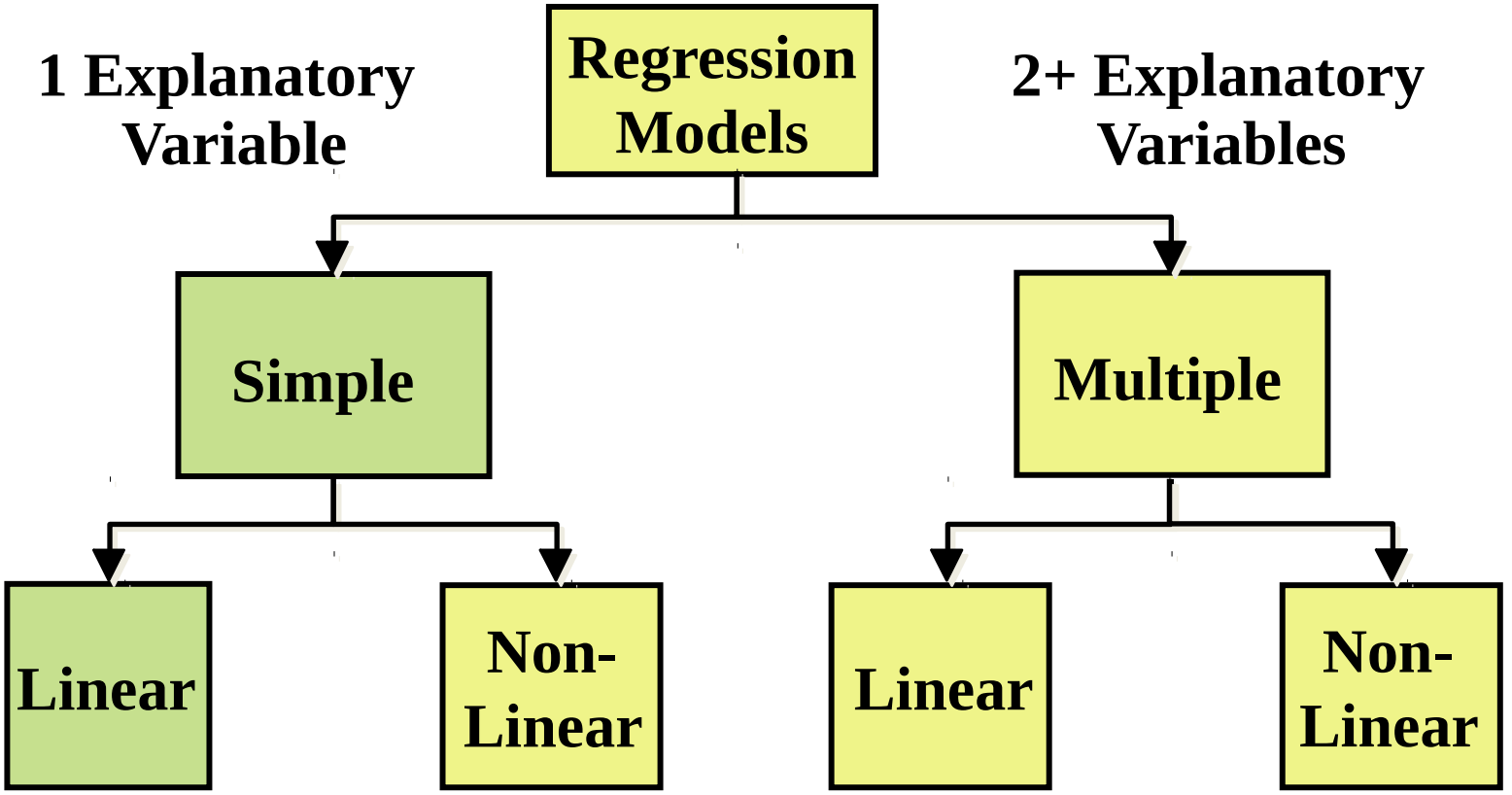
Regression

R-Academy: Ramadhan Edition

Regression Analysis

- Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). Regression analysis is an important tool for modelling and analyzing data.

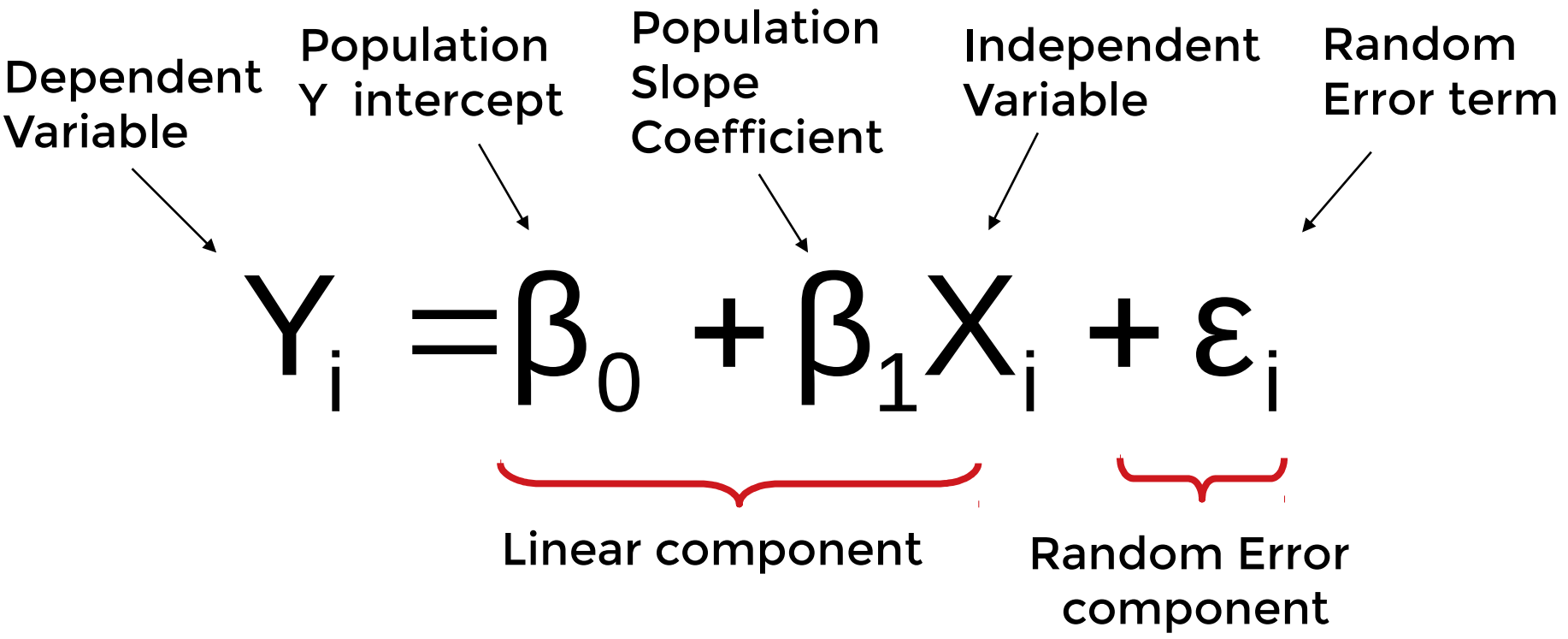
- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable


Regression Models
1 Explanatory Variable
2+ Explanatory Variables
Simple
Multiple
Linear
Non-Linear
Linear
Non-Linear

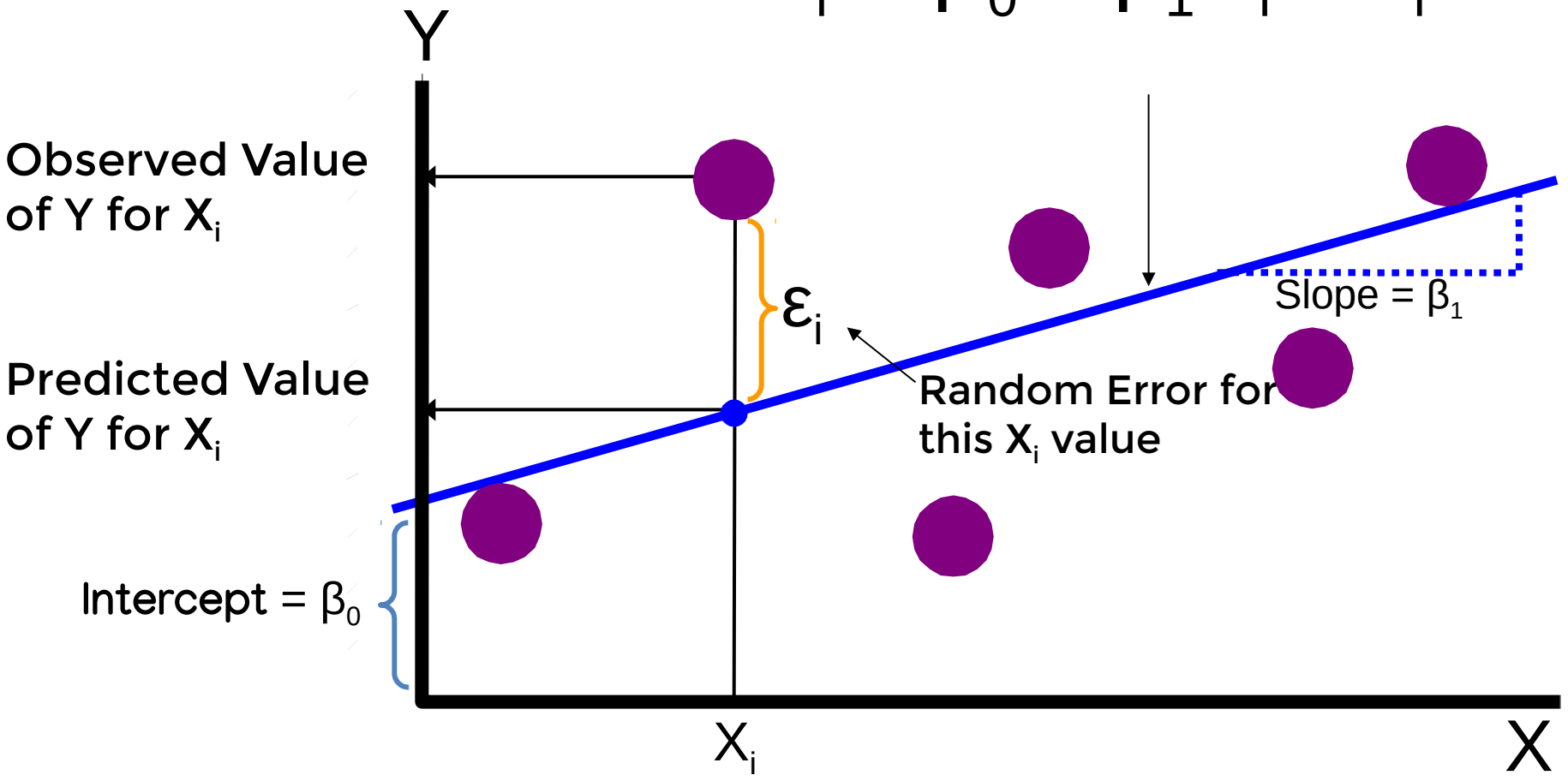
Simple Linear Regression Model

- Only one independent variable, X Relationship between X and Y is described by a linear function.
- Changes in Y are assumed to be caused by changes in X? Relationship between variables is a linear function

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$


Dependent Variable
Population Y intercept
Population Slope Coefficient
Independent Variable
Random Error term
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Linear component
Random Error component

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



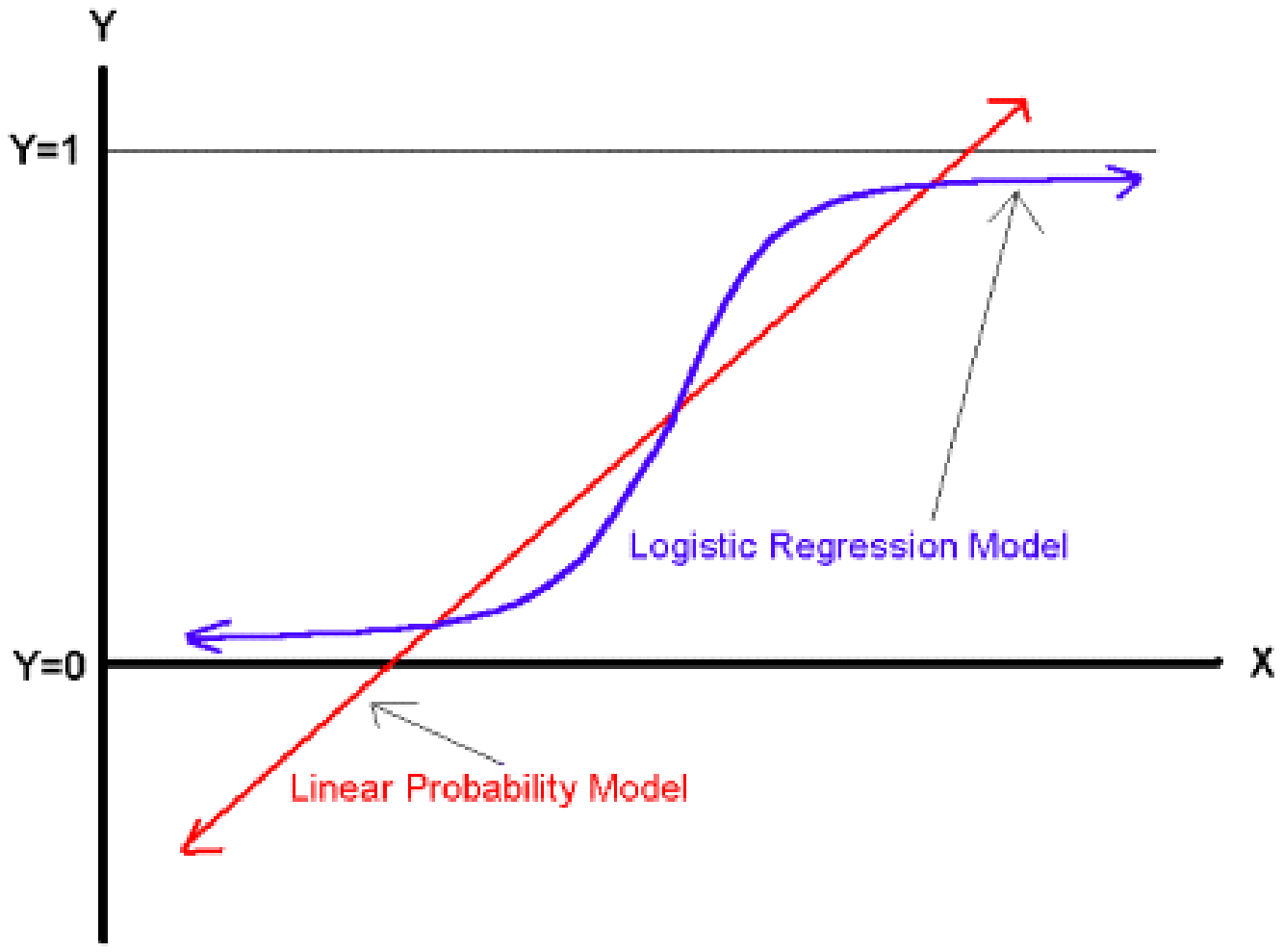
Logistic Regression

- A statistical method used to model dichotomous or binary outcomes (but not limited to) using predictor variables. Used when the research method is focused on whether or not an event occurred, rather than when it occurred (time course information is not used).

What is the “Logistic” component?

- Instead of modeling the outcome, Y, directly, the method models the log-odds(Y) using the logistic function.

- Why use logsitic regression?
  - There are many important research topics for which the dependent variable is "limited."
  - For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.
  - Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1(did vote)


Y
Y=1
Y=0
X
Logistic Regression Model
Linear Probability Model

Practice Time

Let's Build Simple Linear Regression Model using R

Capstone Project

- Group 1: Nadya Shafirah, Asniar, Nabilla Kalvina
- Group 2: Putu Giri Artha, Ibrahim Hasan, Andika Heru
- Group 3: Magdalena Karismayanti, Femi Yulianti, Muh. Rosidi
- Group 4: Widi Astuti, Silvia Latifah, Oman Fikri, Muhammad Naqi

Thank You

Every Ending Is Really Just A New Beginning