KESHAV MEMORIAL INSTITUTE OF TECHNOLOGY
AN AUTONOMOUS INSTITUTION - ACCREDITED BY NAAC WITH 'A' GRADE
Narayanaguda, Hyderabad.

EMBEDDED LEARNING
DAY1-ML AND DL

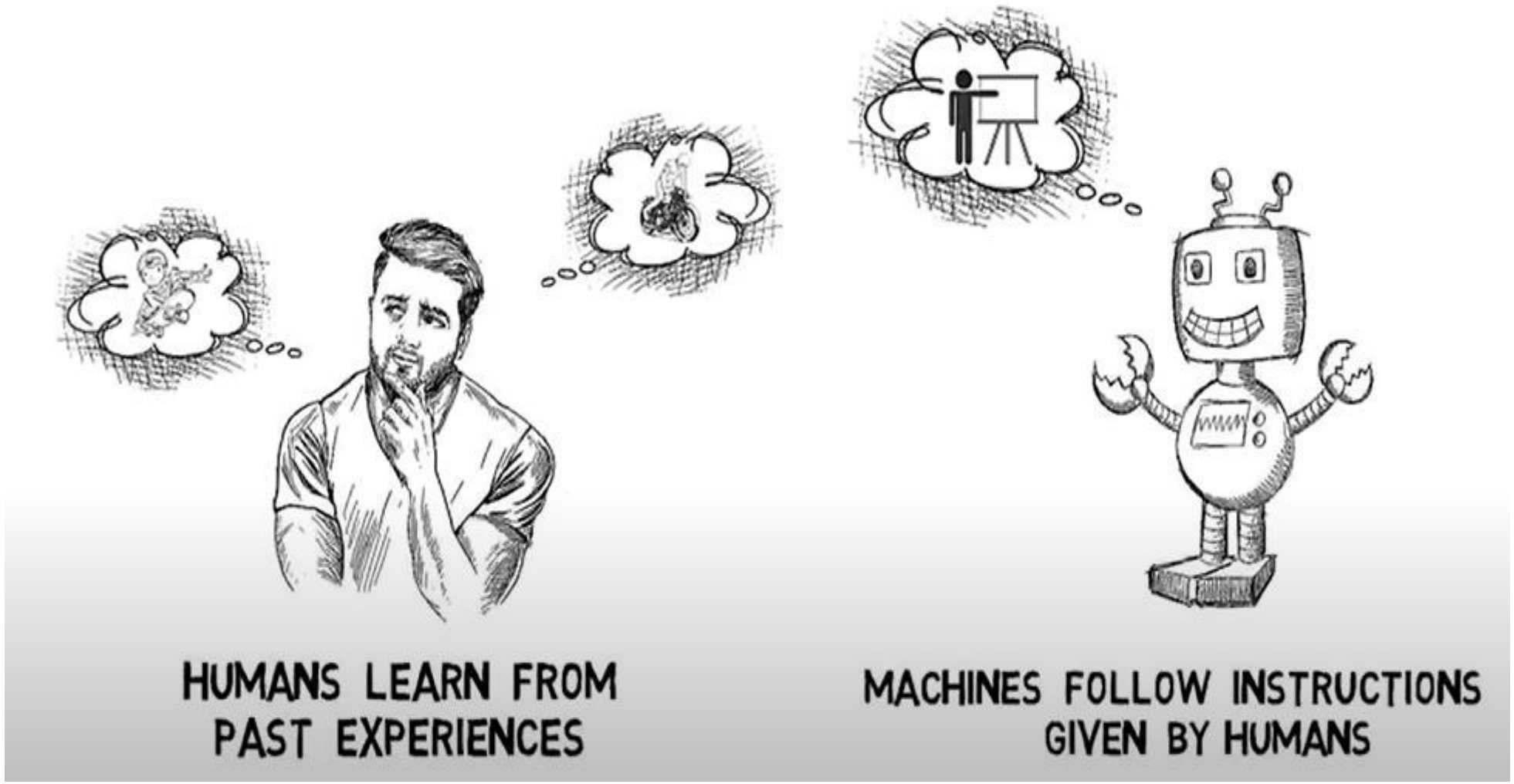
BY
ASHA M
ASSISTANT PROFESSOR
CSE(AI&ML)
KMIT

SESSION-1

20TH JANUARY 2025

CONTENTS

- What is Machine learning?
- Applications of Machine Learning
- Important Python Libraries
- Types of Machine Learning
- Model Evaluation and Tuning
- Data Preprocessing
- Introduction to Neural Networks


HUMANS LEARN FROM
PAST EXPERIENCES
MACHINES FOLLOW INSTRUCTIONS
GIVEN BY HUMANS

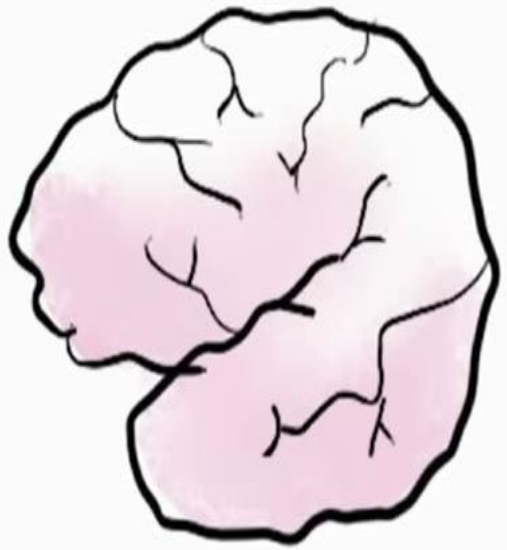
WHAT IF HUMANS CAN TRAIN THE MACHINES...

Kmit

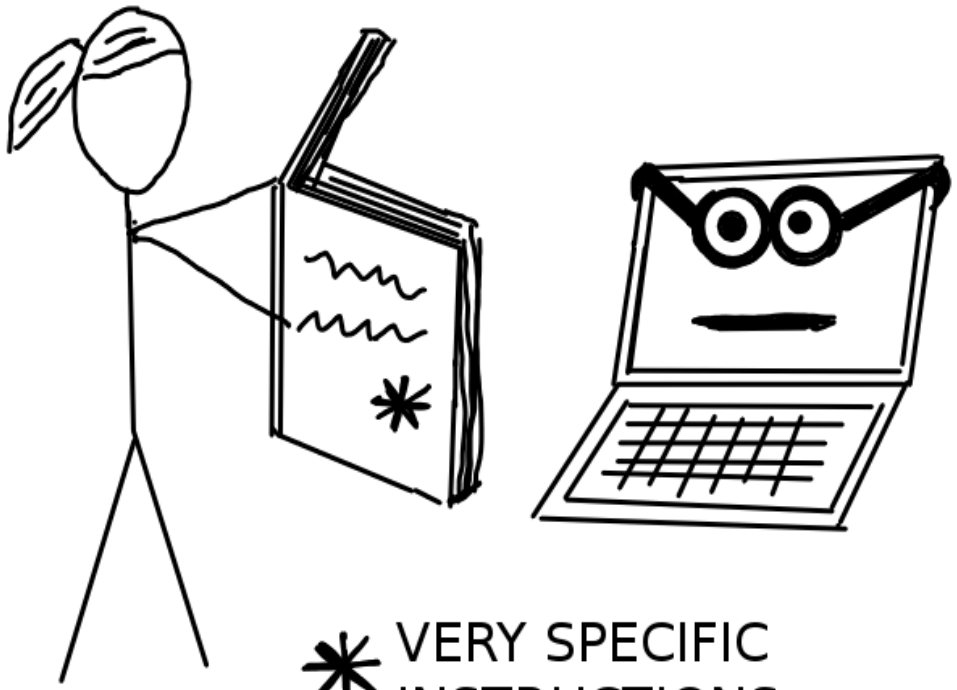
```
function withdraw(amount) {
    if (amount > balance) {
        fail("Hey you ain't got the cash!")
    } else {
        balance = balance - amount
    }
}
```

Execute Instructions

Learn + Think

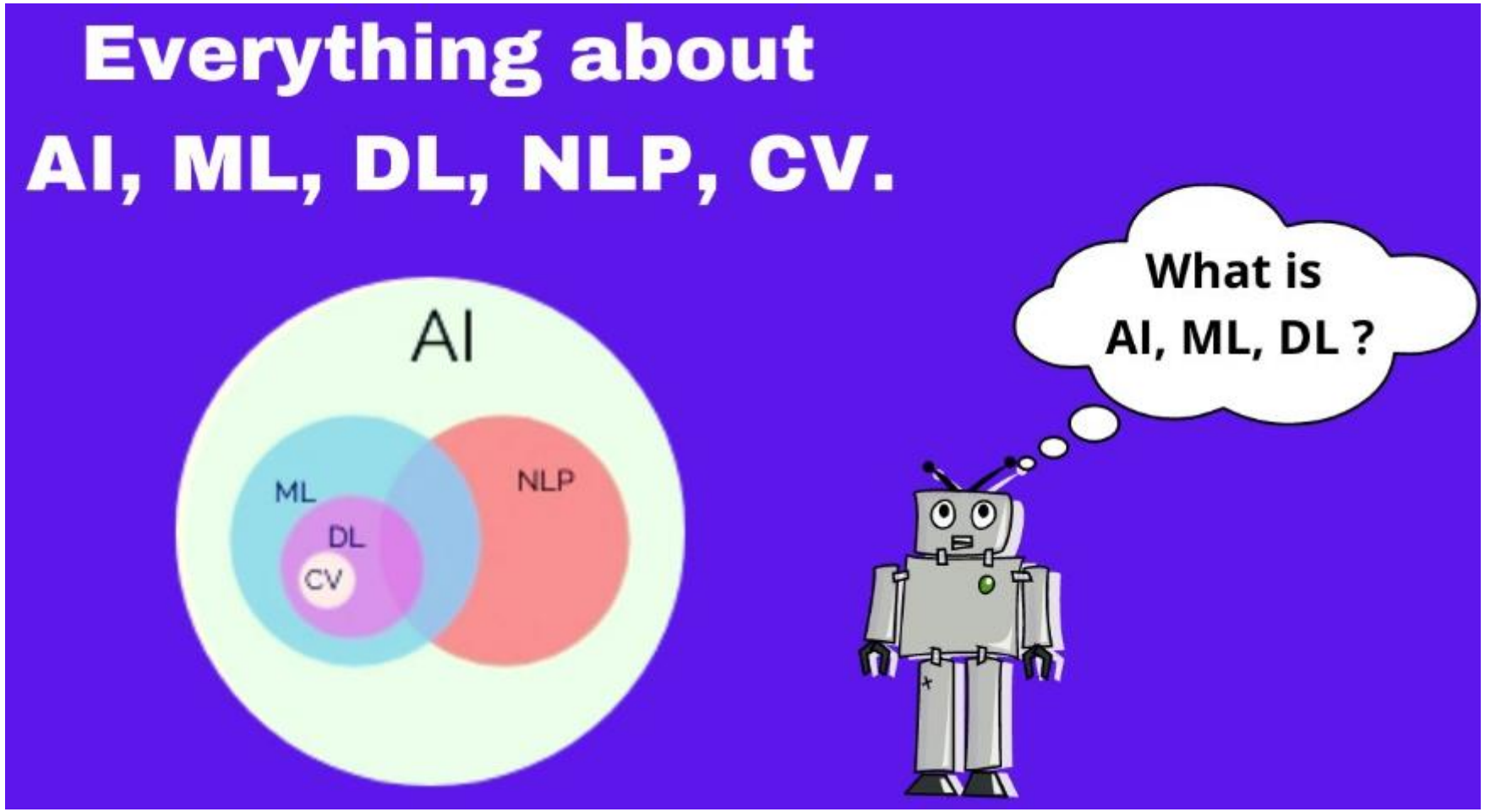


Without Machine Learning



* VERY SPECIFIC INSTRUCTIONS

With Machine Learning


Everything about
AI, ML, DL, NLP, CV.
AI
ML
DL
CV
NLP
What is
AI, ML, DL ?

ARTIFICIAL INTELLIGENCE

- ARTIFICIAL INTELLIGENCE REFERS TO THE DEVELOPMENT OF COMPUTER SYSTEMS THAT CAN PERFORM TASKS THAT TYPICALLY REQUIRE HUMAN INTELLIGENCE.
- THINK OF AI AS THE VIRTUAL ASSISTANT ON YOUR SMARTPHONE THAT CAN UNDERSTAND YOUR VOICE COMMANDS, PROVIDE RECOMMENDATIONS, AND LEARN FROM YOUR PREFERENCES OVER TIME.

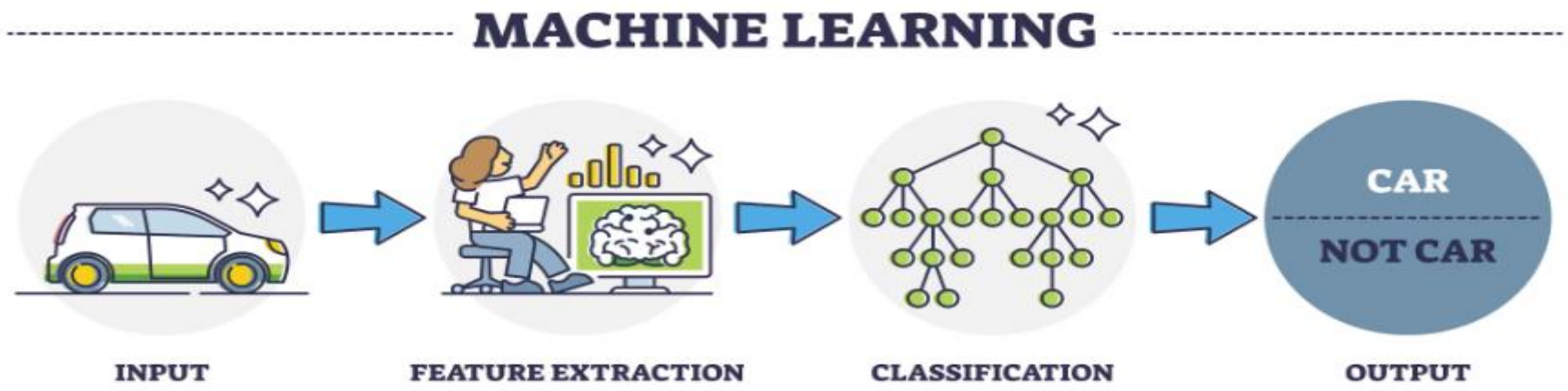
Tesla Autopilot Car

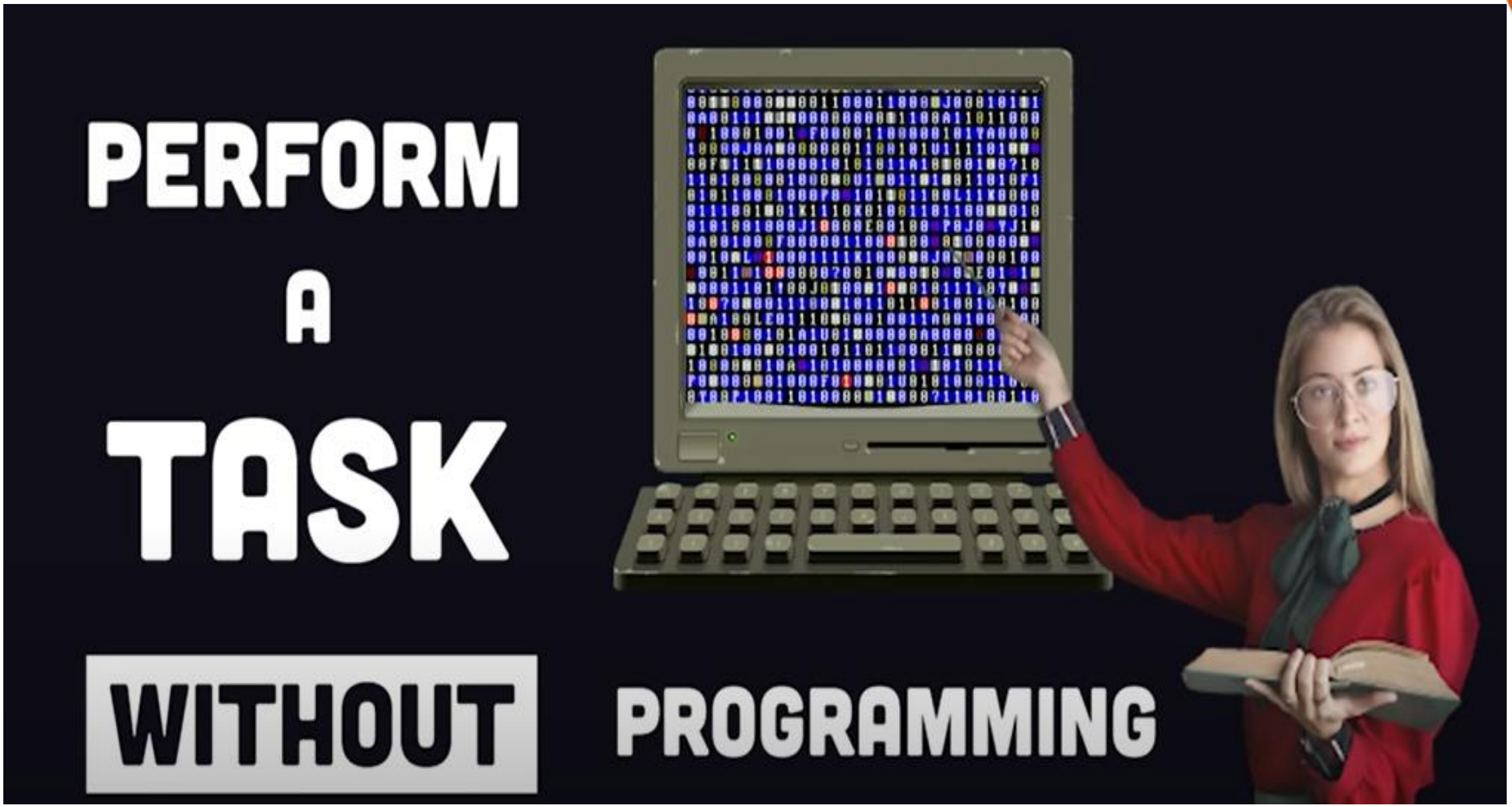
Sophia Humanoid Robot


Amazon Alexa

MACHINE LEARNING

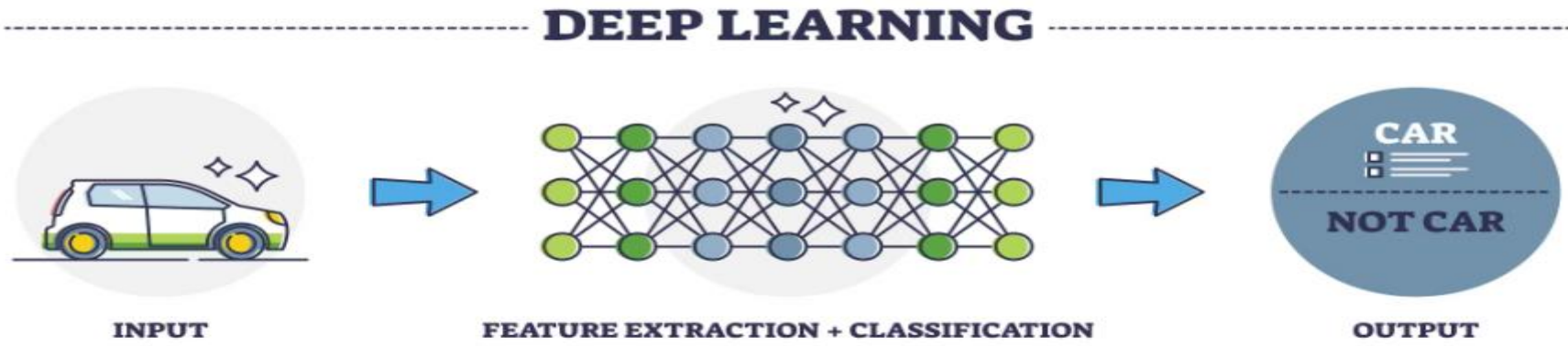
- MACHINE LEARNING INVOLVES ALGORITHMS AND STATISTICAL MODELS THAT ENABLE COMPUTERS TO IMPROVE THEIR PERFORMANCE ON A SPECIFIC TASK WITHOUT EXPLICIT PROGRAMMING.
- IT FOCUSES ON PATTERN RECOGNITION AND LEARNING FROM DATA.
- MACHINE LEARNING IS A SUBSET OF ARTIFICIAL INTELLIGENCE.




PERFORM
A
TASK
WITHOUT
PROGRAMMING

DEEP LEARNING

- DEEP LEARNING IS A SUBSET OF MACHINE LEARNING THAT INVOLVES NEURAL NETWORKS WITH MULTIPLE LAYERS (DEEP NEURAL NETWORKS).
- THESE NETWORKS CAN AUTOMATICALLY LEARN TO EXTRACT FEATURES FROM DATA AND MAKE COMPLEX DECISIONS BASED ON LARGE AMOUNTS OF DATA.




DATA
85%
ACCURACY

Machine learning vs. deep learning

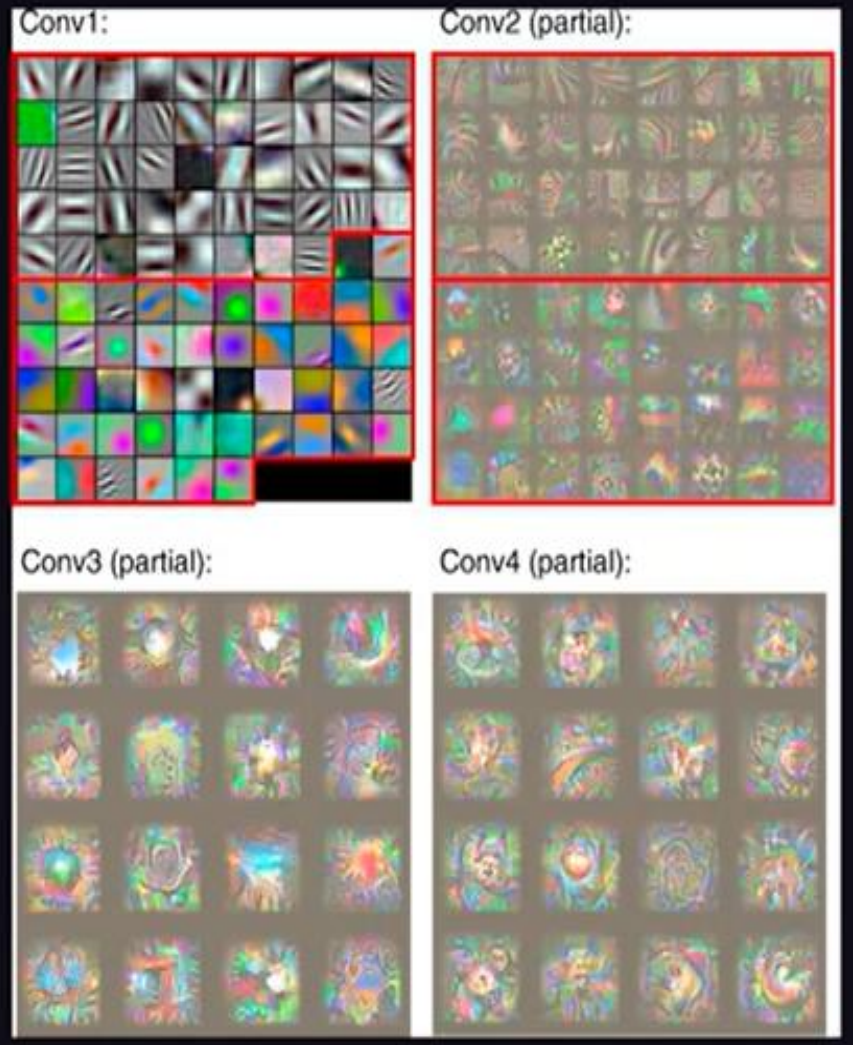
Machine learning

Uses algorithms and learns on its own but may need human intervention to correct errors

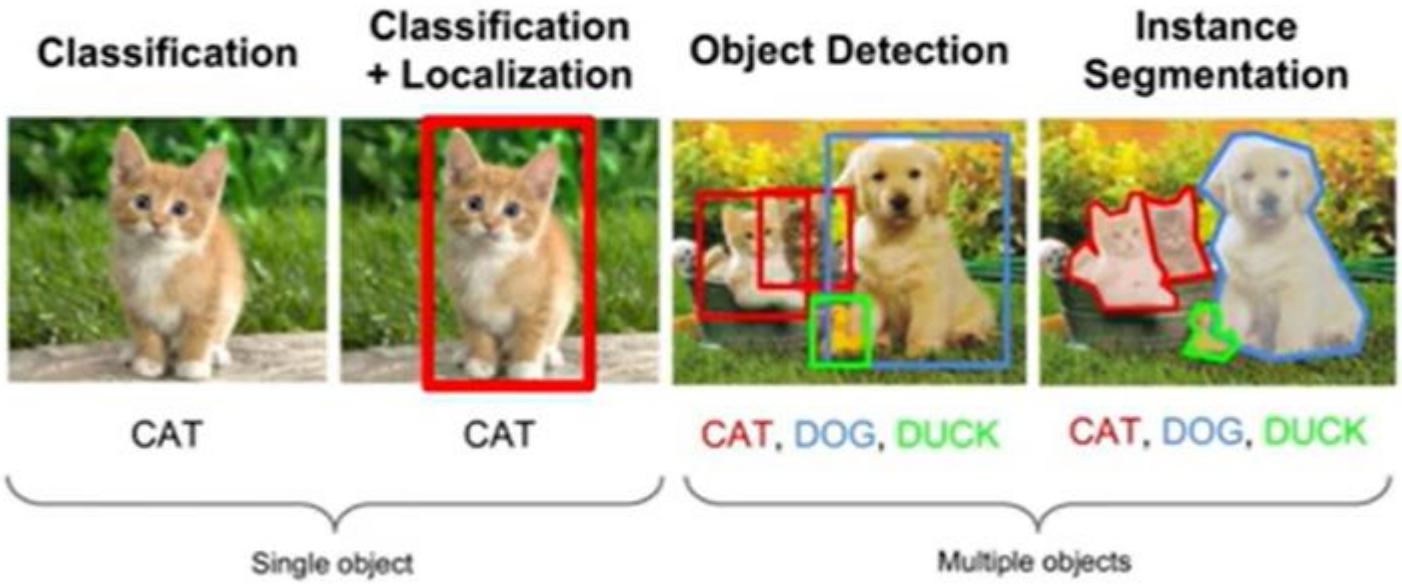
Deep learning

Uses advanced computing, its own neural network, to adapt with little to no human intervention

FEATURE ENGINEERING

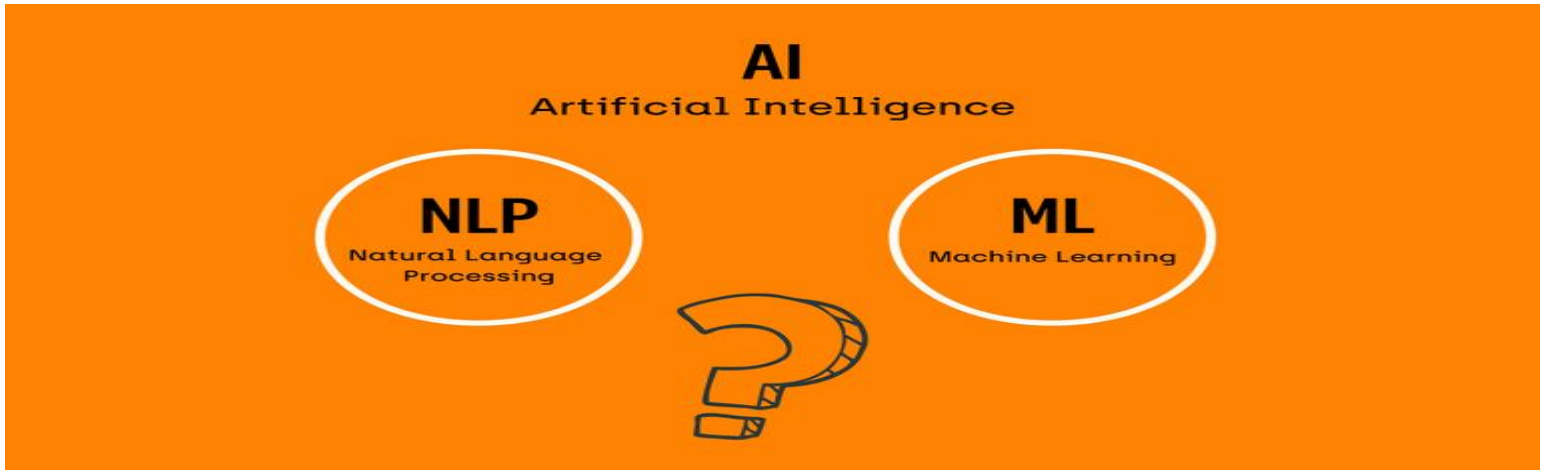


Computer vision (CV) is a field of artificial intelligence that enables computers to interpret and understand the visual world. Using digital images from cameras and videos and deep learning models, machines can accurately identify and classify objects, and then react to what they "see."

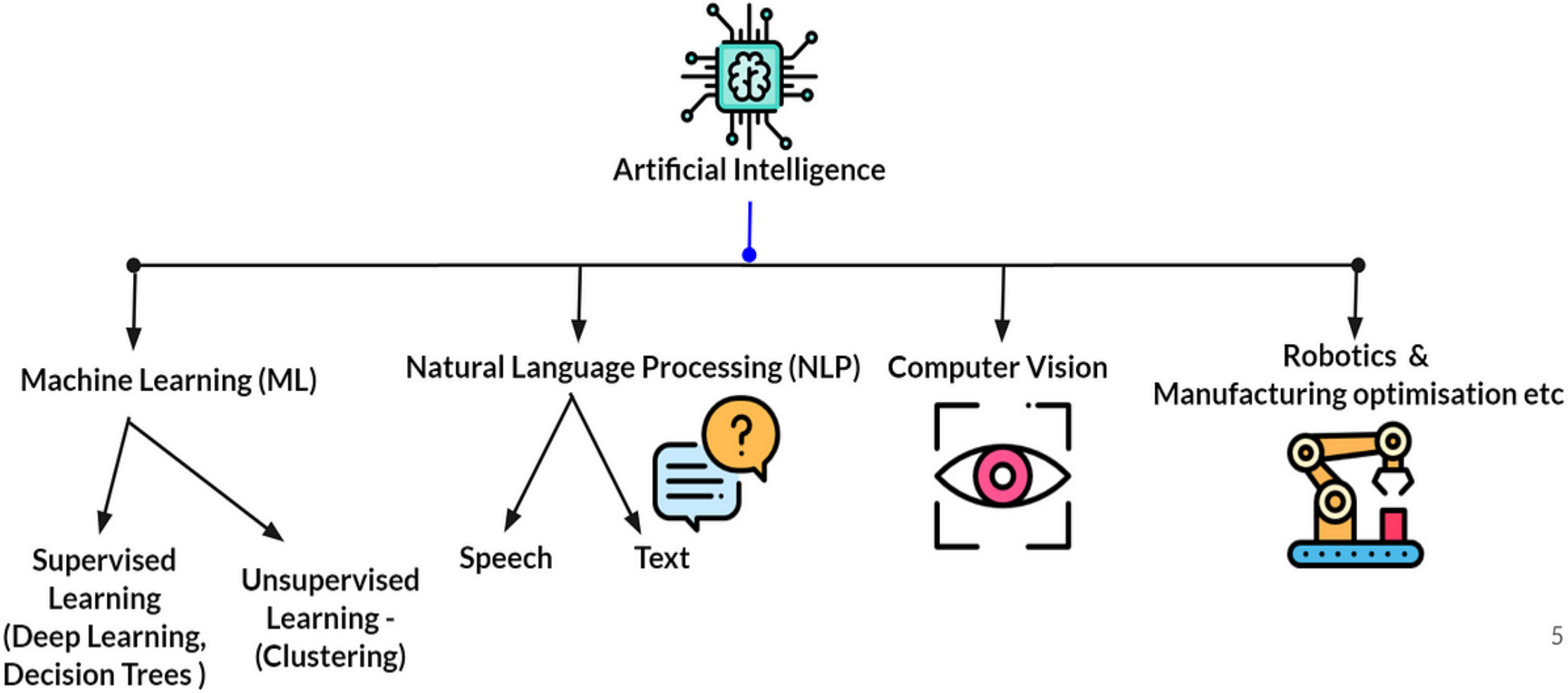



nest
68
IN 2+ HR
65
Kite
Person
Person
Person

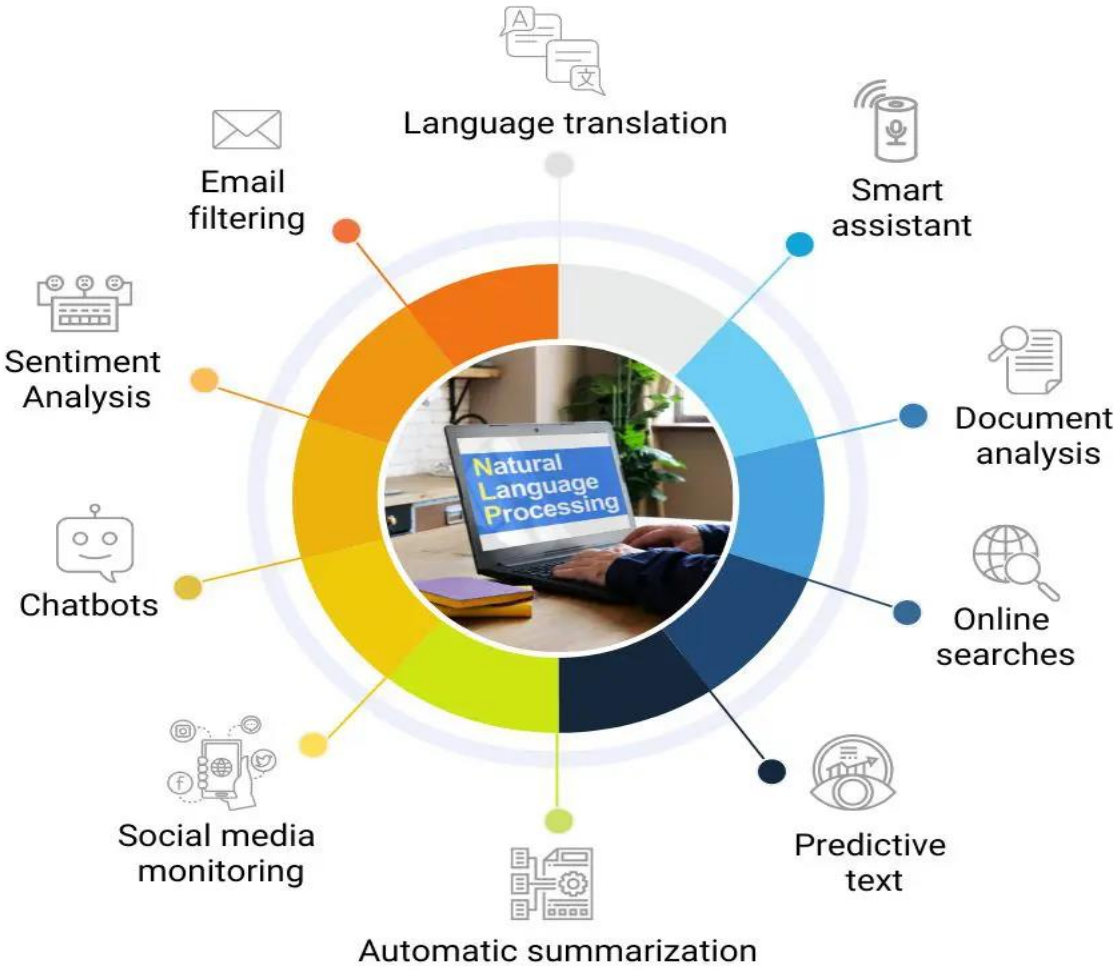
NATURAL LANGUAGE PROCESSING

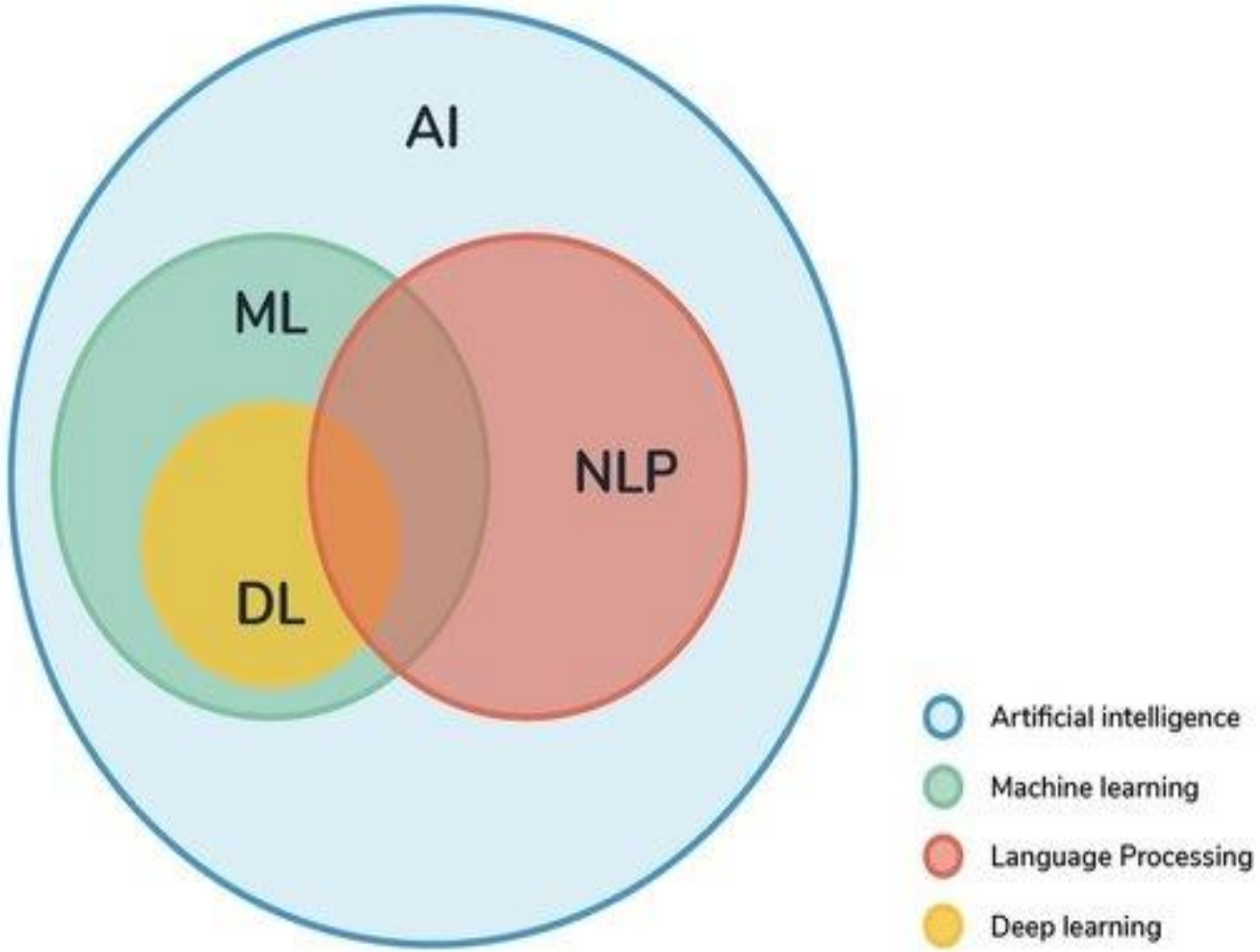


- NATURAL LANGUAGE PROCESSING (NLP) IS A SUBFIELD OF ARTIFICIAL INTELLIGENCE THAT FOCUSES ON THE INTERACTION BETWEEN COMPUTERS AND HUMANS THROUGH NATURAL LANGUAGE.
- THE ULTIMATE GOAL OF NLP IS TO ENABLE COMPUTERS TO UNDERSTAND, INTERPRET, AND GENERATE HUMAN LANGUAGES IN A VALUABLE WAY.


Artificial Intelligence
Machine Learning (ML)
Supervised Learning (Deep Learning, Decision Trees)
Unsupervised Learning - (Clustering)
Natural Language Processing (NLP)
Speech
Text
Computer Vision
Robotics & Manufacturing optimisation etc

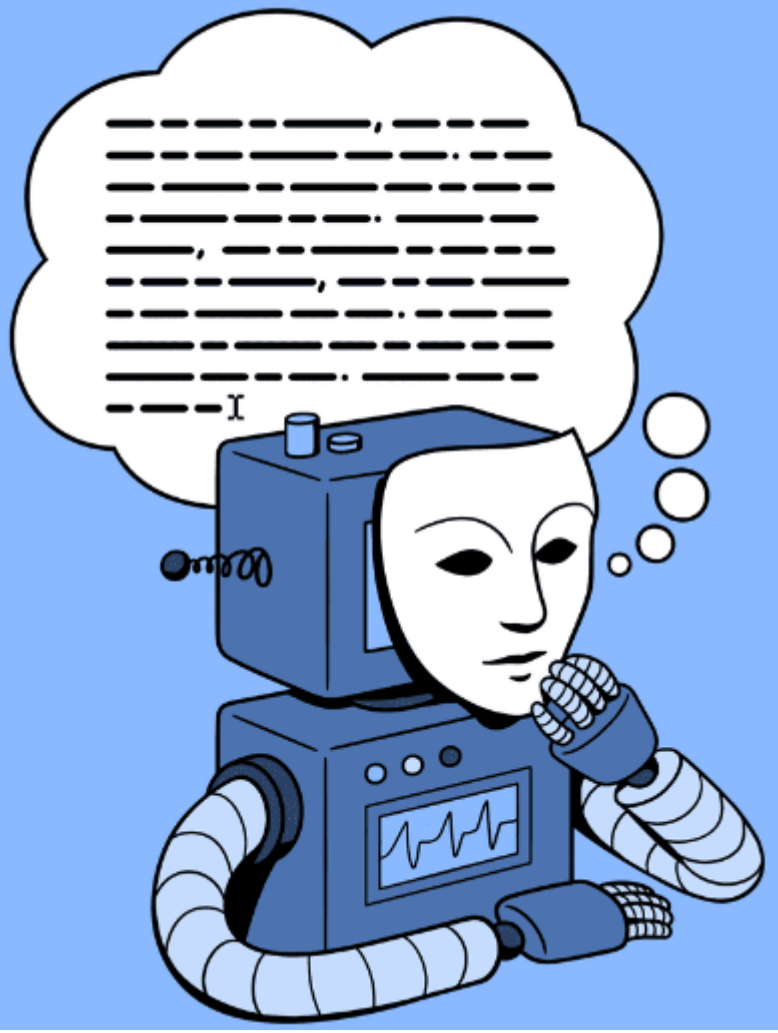
Applications of
Natural Language Processing




AI
ML
NLP
DL
Artificial intelligence
Machine learning
Language Processing
Deep learning

LARGE LANGUAGE MODEL

- LARGE LANGUAGE MODELS ARE ADVANCED AI MODELS TRAINED ON VAST AMOUNTS OF TEXT DATA, ENABLING THEM TO UNDERSTAND AND GENERATE HUMAN-LIKE LANGUAGE.
- VIRTUAL ASSISTANTS LIKE SIRI OR ALEXA UTILIZE LARGE LANGUAGE MODELS TO UNDERSTAND AND RESPOND TO **NATURAL LANGUAGE QUERIES**.
- LARGE LANGUAGE MODELS ARE A PRODUCT OF DEEP LEARNING AND ARE PART OF THE BROADER FIELD OF ARTIFICIAL INTELLIGENCE.



Large Language Model (LLM)

['lärj 'laŋ-gwij 'mä-dᵊl]

A deep learning algorithm that's equipped to summarize, translate, predict, and generate human-sounding text to convey ideas and concepts.

Large Language Models are becoming very large indeed

GENERATIVE AI

- GENERATIVE AI REFERS TO ARTIFICIAL INTELLIGENCE SYSTEMS THAT ARE CAPABLE OF CREATING NEW CONTENT SUCH AS TEXT, IMAGES, OR MUSIC.
- THESE SYSTEMS LEARN FROM EXISTING DATA PATTERNS AND GENERATE FRESH, ORIGINAL CONTENT.
- GENERATIVE AI IS BEHIND TOOLS THAT CAN CREATE REALISTIC-LOOKING IMAGES, OR WRITING ASSISTANT TOOLS THAT HELP TO CREATE CONTENT BASED ON A TOPIC, SUCH AS CHATGPT.


Model Types
Text → Text
"opposite of yellow" → "Purple"
Text → Image
"Einstein sitting in the basement" →
"Ugly cat" →
Image → Image
Image → Text
"Fusion of a human and a cat, seated in an armchair"
Speech → Text
Text → Audio
"People talking in a busy restaurant"
Charcoal
Crayon
Mosaic window
Text → Video
"Darth Vader surfing" →

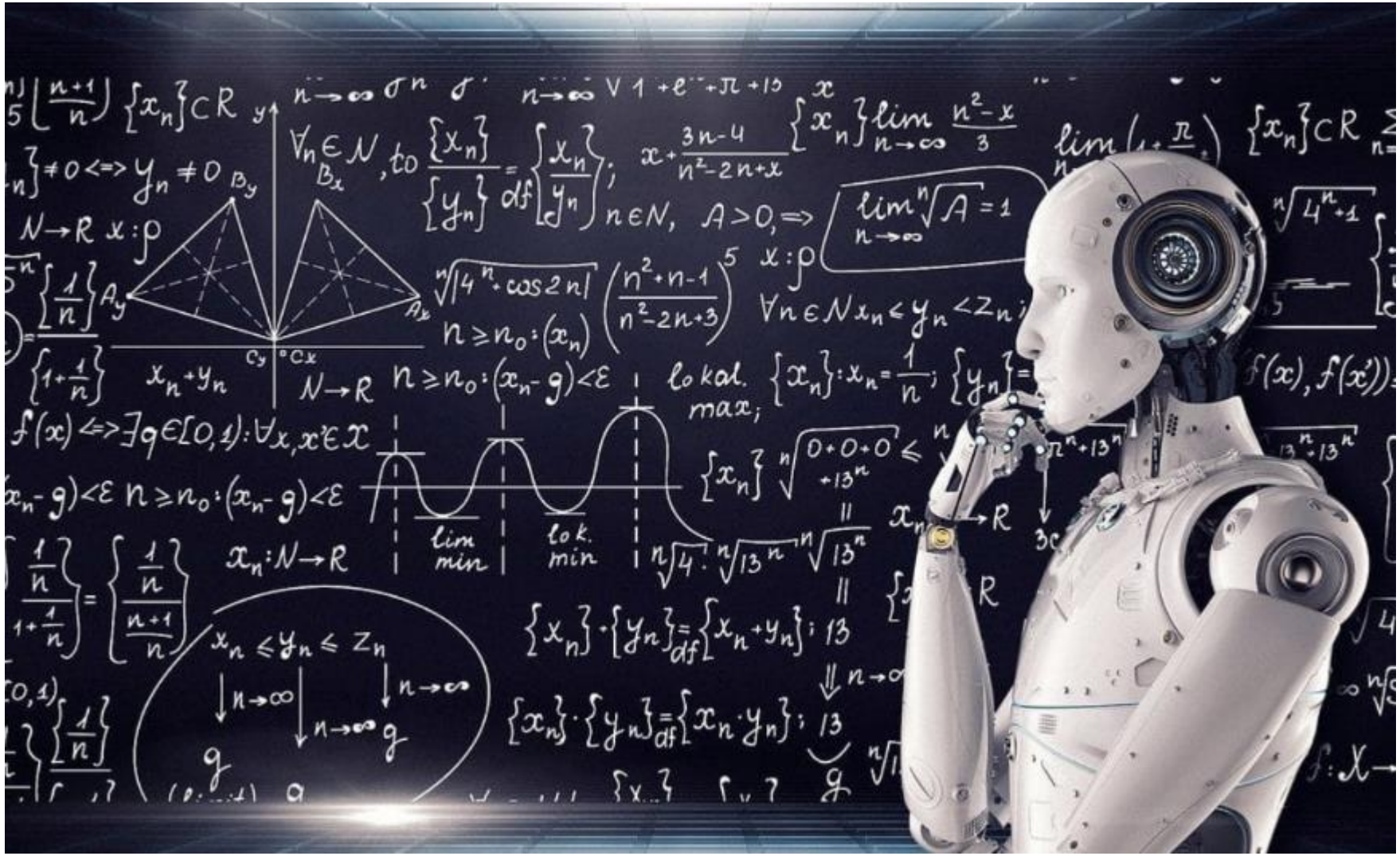
GENERATIVE AI

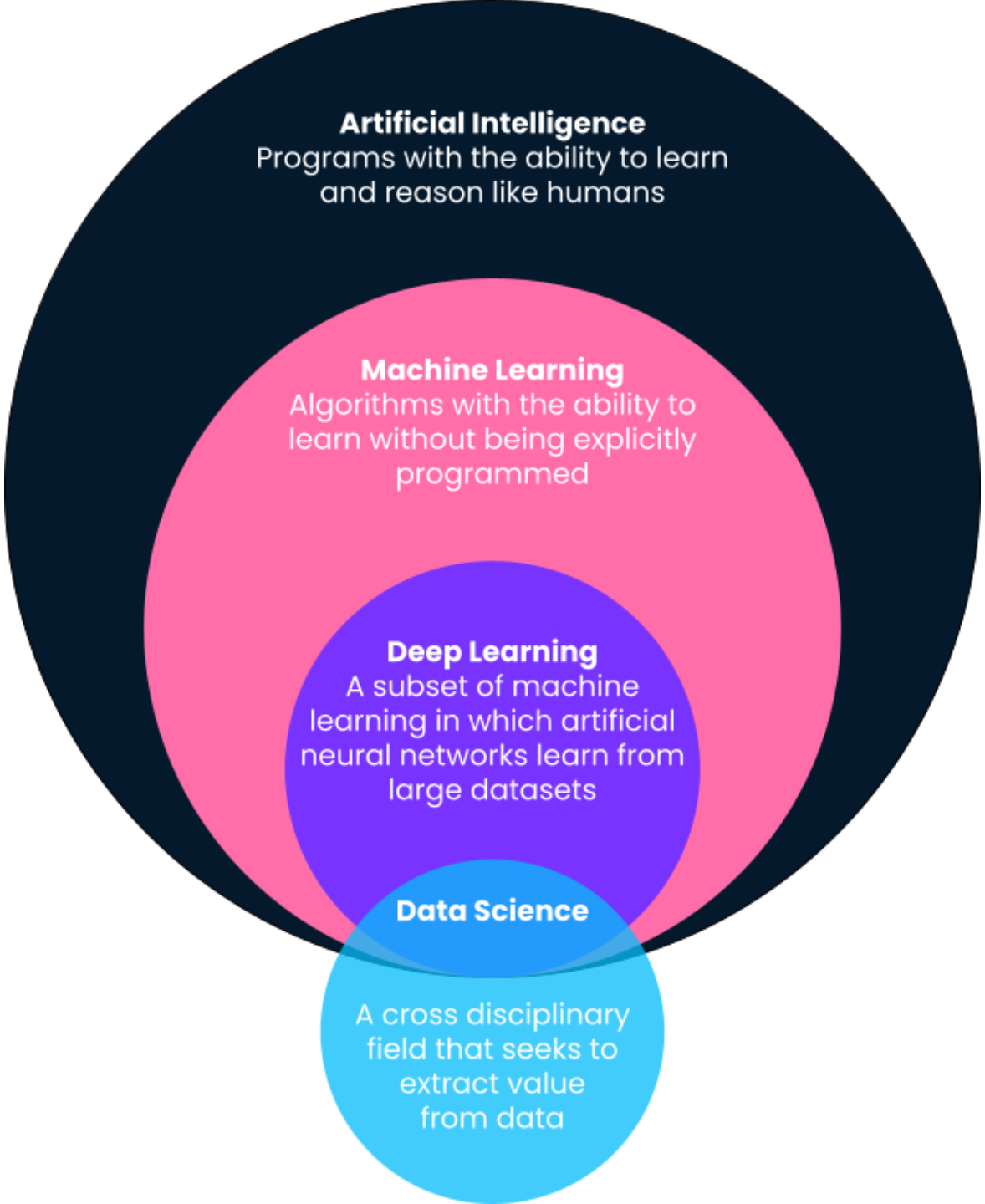
who are you? explain it to a 10 year old kid

Hi there! I'm ChatGPT, and I'm like a super-smart robot that can talk to you and help answer your questions. Imagine I'm a really big library that knows a lot about almost everything, but I can talk and understand what you say. Here's a simple way to understand me:


City Bank
.com

DATA SCIENCE




Artificial Intelligence
Programs with the ability to learn and reason like humans
Machine Learning
Algorithms with the ability to learn without being explicitly programmed
Deep Learning
A subset of machine learning in which artificial neural networks learn from large datasets
Data Science
A cross disciplinary field that seeks to extract value from data

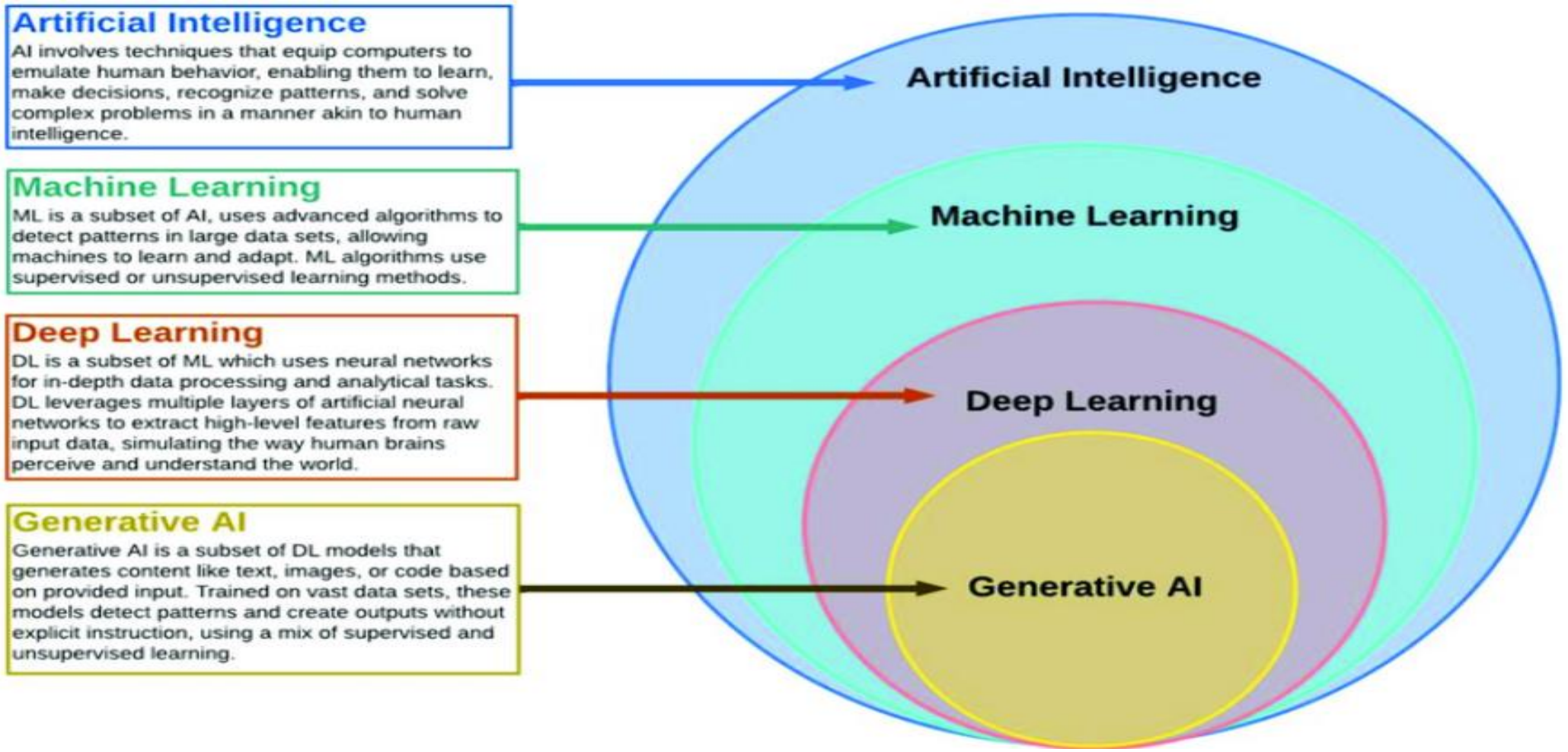
MACHINE LEARNING

In 1959, Arthur Samuel, a computer scientist who pioneered the study of artificial intelligence, described machine learning as "The study that gives computers the ability to learn."


ARTHUR SAMUEL
1959

Kmit

Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions.



MACHINE LEARNING IS A SUBSET OF ARTIFICIAL INTELLIGENCE THAT AIMS TO MIMIC HOW HUMAN BEINGS LEARN BY USING DATA.

A more technical definition given by **Tom M. Mitchell**'s (1997) : "**A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.**"

- **Task**: A task is defined as the main problem in which we are interested. This task/problem can be related to the predictions and recommendations and estimations, etc.

- **Experience**: It is defined as learning from historical or past data and used to estimate and resolve future tasks.

- **Performance**: It is defined as the capacity of any machine to resolve any machine learning task or problem and provide the best outcome for the same. However, performance is dependent on the type of machine learning problems.

What can machine do by learning?

01

It uses the data to *detect patterns* in a dataset and *adjust program actions accordingly*

It *focuses on* the *development of computer programs* that can teach themselves to *grow and change* when *exposed* to *new data*

02

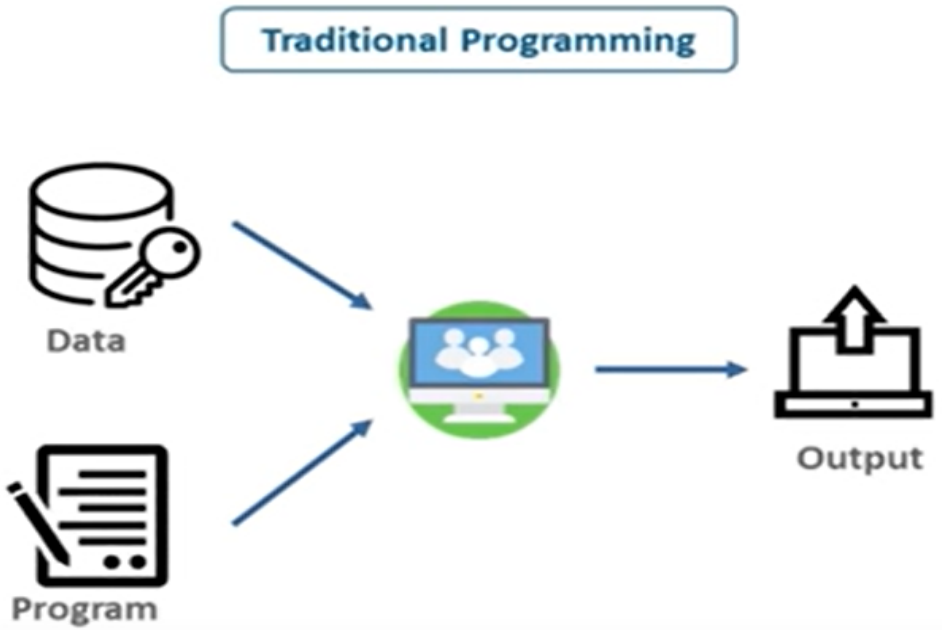
03

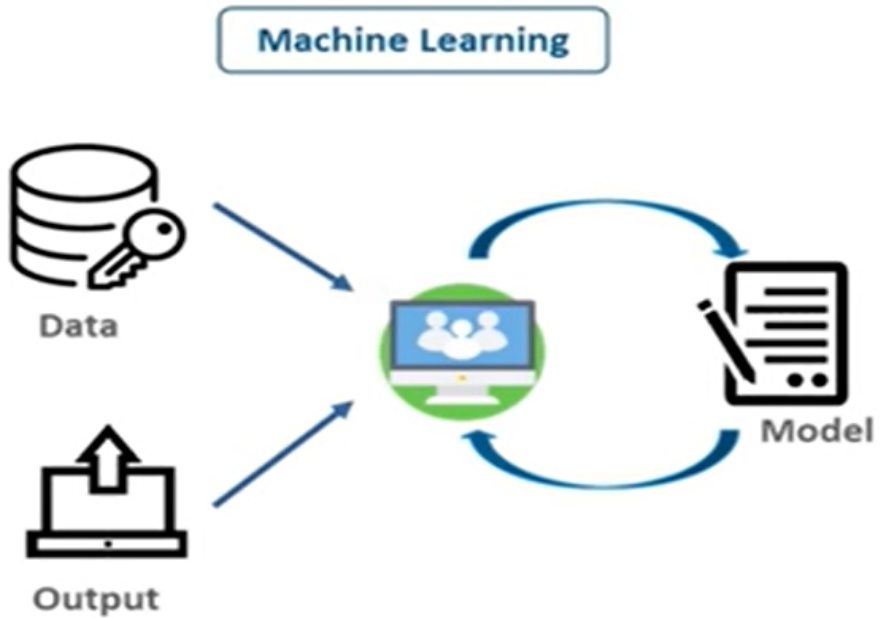
It enables computers to *find hidden insights using iterative algorithms without being explicitly programmed*

Machine learning is a *method* of *data analysis* that *automates analytical model building*

04

HOW MACHINE LEARNING WORKS?





Learn from Data

Find Hidden Insights

Train and Grow

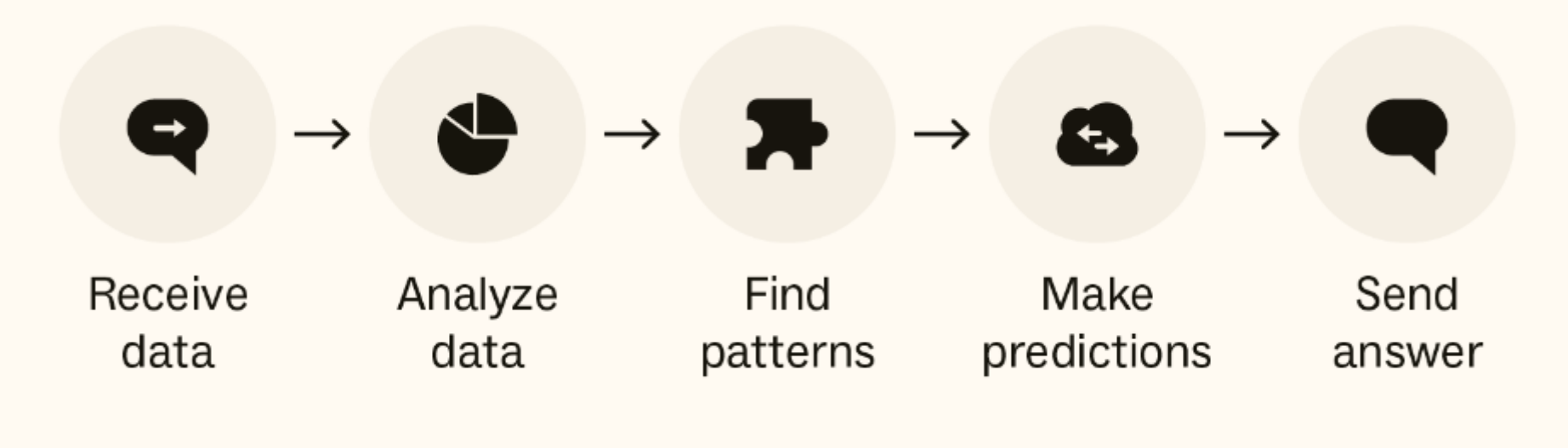
- Say I have 100 pictures, and I want to filter all photos of a guy named SAM. I can do this by writing a set of conditions that will look for things like glasses, skin color, or other facial characteristics. These will be the rules or logic or my computer program.
- For example:

1. If the current picture has a face, proceed to step 2. Else, move on to the next one.
2. Is the face white? If yes, proceed to step 3. If not, move on to the next picture.
3. Does it have glasses on? If yes, proceed to step 4. If not, move on to the next image.
4. Does it have black hair? If yes, proceed to step 5. If not, move on to the next picture.
5. Are the eyes brown? If yes, it is SAM. If not, move on to the next image.

- Let’s continue the previous example where we had to filter SAM’s photos from a set of 100 images. If we want to achieve the same task with machine learning, the first step would be to give it different photos of SAM so that the computer may learn what he looks like. These pictures are the training data (as our model gets trained with this dataset), and the rules used to identify SAM are called ‘feature sets.
- So in our oversimplified example, the computer will learn that SAM:

1. Wears rectangular glasses
2. Has short black hair
3. Has a long white face
4. Has big brown eyes
5. The first step is to feed SAM’s photos to the computer, making sure that they include and highlight the above features. After that, the machine would be able to recognize and filter SAM’s pictures on its own.

How Machine Learning Works?




DATASET
DATASET
DATASET

DATASET

- A dataset is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table.
- A tabular dataset can be understood as a database table or matrix, where each column corresponds to a particular variable, and each row corresponds to the fields of the dataset. The most supported file type for a tabular dataset is "Comma Separated File," or CSV.

NEED OF DATASET

To work with machine learning projects, we need a huge amount of data. Collecting and preparing the dataset is one of the most crucial parts while creating an ML/AI project.

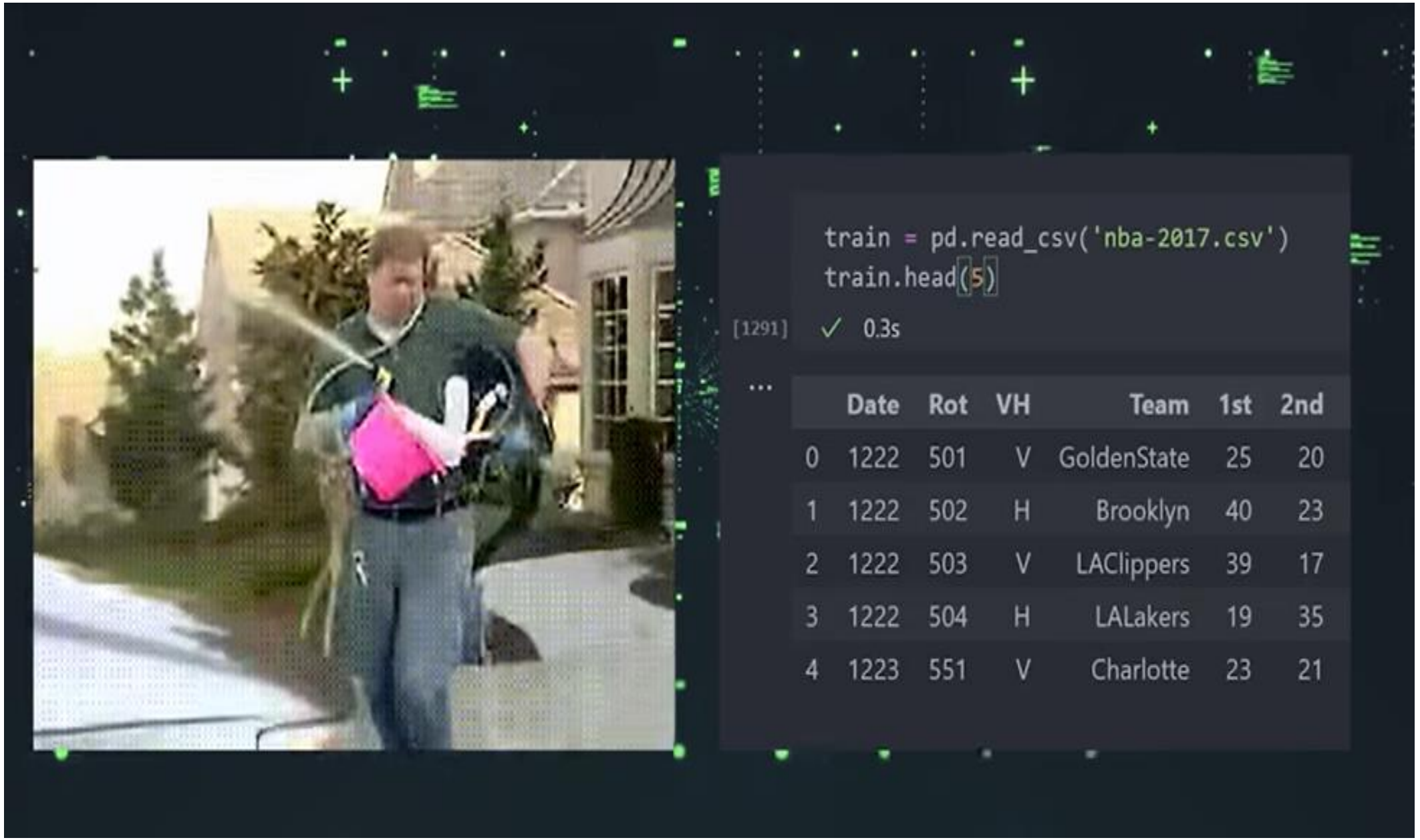
POPULAR SOURCES FOR MACHINE LEARNING DATASETS

- Kaggle Datasets
- UCI Machine Learning Repository
- Datasets via AWS
- Google's Dataset Search Engine
- Microsoft Datasets

DATA PREPROCESSING

DATA PRE-PROCESSING IS A PROCESS OF CLEANING THE RAW DATA I.E. THE DATA IS COLLECTED IN THE REAL WORLD AS MOST OF THE REAL-WORLD DATA IS MESSY, SOME OF THESE TYPES OF DATA ARE:

1. MISSING DATA
2. NOISY DATA
3. INCONSISTENT DATA



```
train = pd.read_csv('nba-2017.csv')
train.head(5)
```

[1291] ✓ 0.3s

...

	Date	Rot	VH	Team	1st	2nd
0	1222	501	V	GoldenState	25	20
1	1222	502	H	Brooklyn	40	23
2	1222	503	V	LAClippers	39	17
3	1222	504	H	LALakers	19	35
4	1223	551	V	Charlotte	23	21

Why is Data Preprocessing Important?

The majority of the real-world datasets for machine learning are highly susceptible to be missing, inconsistent, and noisy.

- Data processing is, therefore, important to improve the overall data quality.
- Duplicate or missing values may give an incorrect view of the overall statistics of data
- Outliers and inconsistent data points often tend to disturb the model's overall learning, leading to false predictions.

4 STEPS IN DATA PREPROCESSING

- DATA CLEANING
- DATA INTEGRATION
- DATA TRANSFORMATION
- DATA REDUCTION

DATA CLEANING

DATA CLEANING IS PARTICULARLY DONE AS PART OF DATA PREPROCESSING TO CLEAN THE DATA BY:

- MISSING VALUES
- NOISY DATA
- REMOVING OUTLIERS

DATA INTEGRATION

DATA INTEGRATION IS ONE OF THE DATA PREPROCESSING STEPS THAT ARE USED TO MERGE THE DATA PRESENT IN MULTIPLE SOURCES INTO A SINGLE LARGER DATA STORE LIKE A DATA WAREHOUSE.

DATA TRANSFORMATION

ONCE DATA CLEANING HAS BEEN DONE, WE NEED TO CONSOLIDATE THE QUALITY DATA INTO ALTERNATE FORMS BY CHANGING THE VALUE, STRUCTURE, OR FORMAT OF DATA USING THE DATA TRANSFORMATION STRATEGIES.

Data Reduction

- The size of the dataset in a data warehouse can be too large to be handled by data analysis and data mining algorithms.
- One possible solution is to obtain a reduced representation of the dataset that is much smaller in volume but produces the same quality of analytical results.

- **Handling Missing Values:** Techniques include removing instances with missing values, imputing missing values with the mean, median, or mode, or using advanced techniques like KNN imputation.
- **Removing Duplicates:** Identifying and removing duplicate instances to ensure the dataset is clean.
- **Feature Scaling:**
- **Normalization:** Rescaling the features to a range of [0, 1].
- **Standardization:** Rescaling the features to have a mean of 0 and a standard deviation of 1.

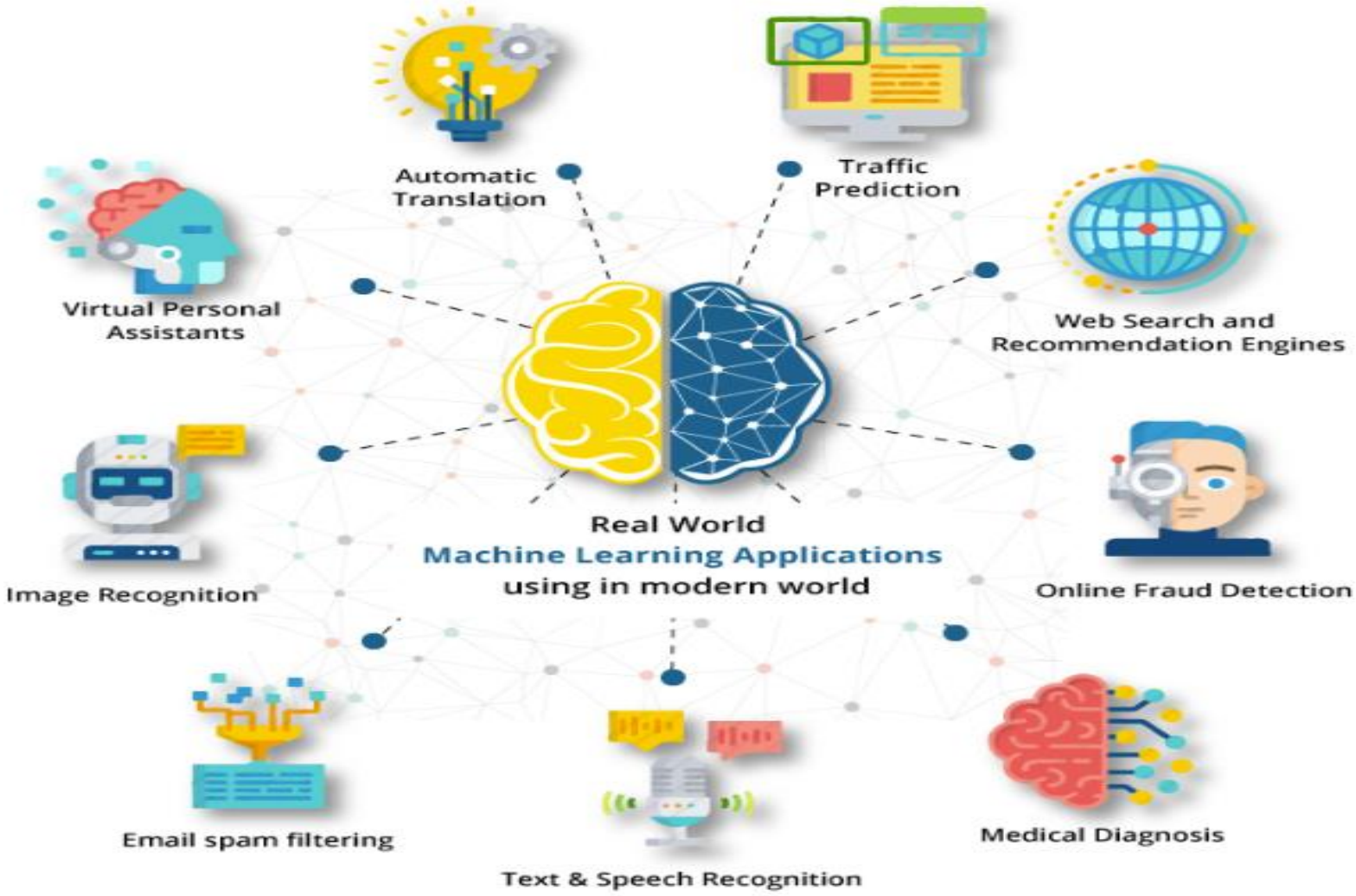
Encoding Categorical Data:

- **One-Hot Encoding:** Converting categorical variables into binary vectors.
- **Label Encoding:** Converting categorical variables into integer values.

Splitting Data:

- Dividing the dataset into training and testing sets to evaluate the model's performance.

WHERE IS MACHINE LEARNING USED?

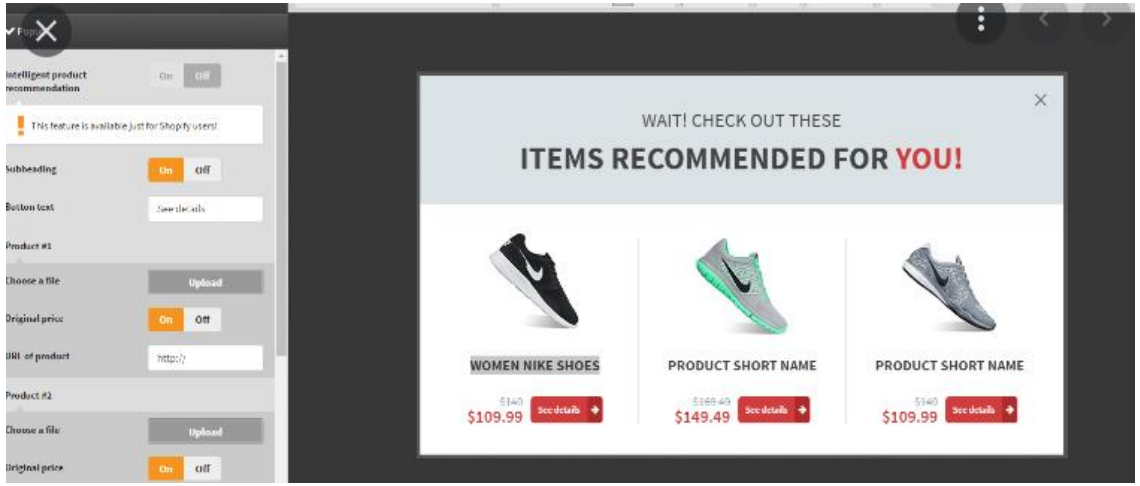


Traffic Alerts (Maps)

Now, **Google Maps** is probably **THE** app we use whenever we go out and require assistance in directions and traffic.

Products Recommendations

- Suppose you check an item on Amazon, but you do not buy it then and there.
- But the next day, you're watching videos on YouTube and suddenly you see an ad for the same item.
- You switch to Facebook, there also you see the same ad.



Virtual Personal Assistants

Virtual Personal Assistants assist in finding useful information, when asked via text or voice. Few of the major applications of Machine Learning here are:

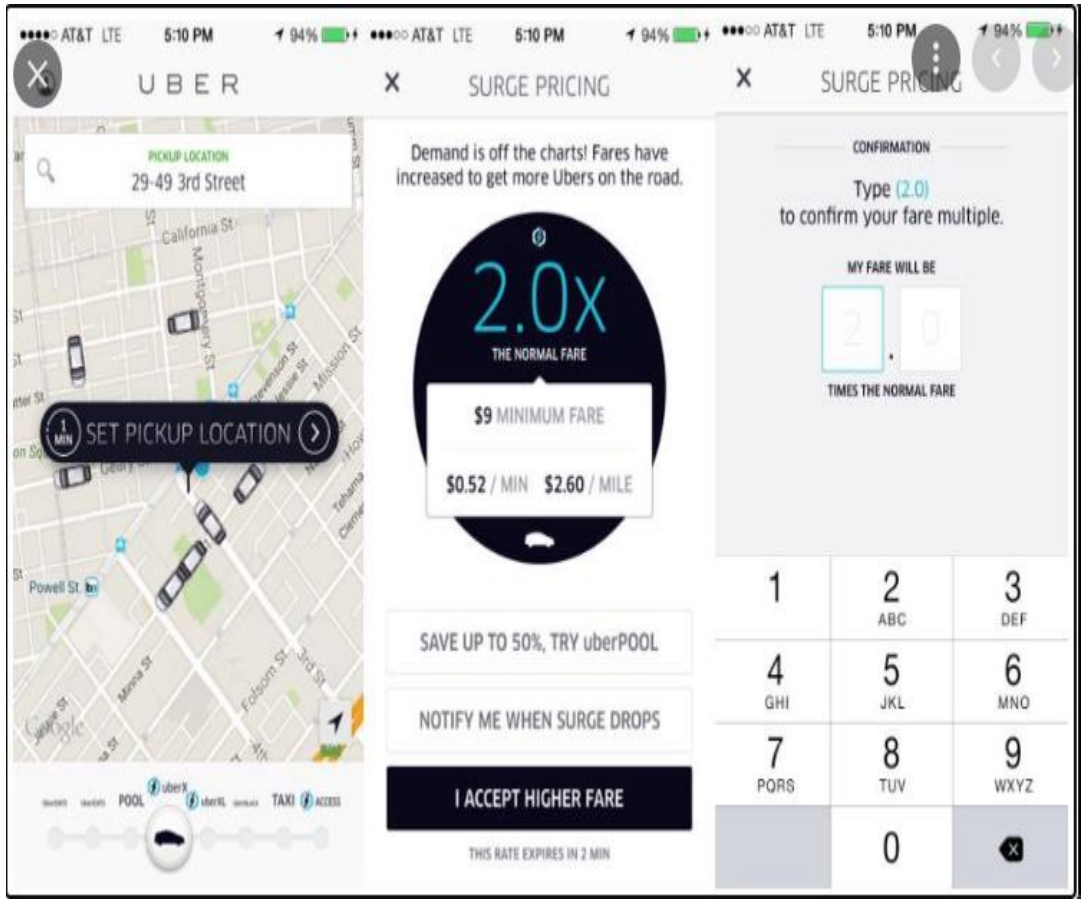
- Speech Recognition
- Speech to Text Conversion
- Text to Speech Conversion

Dynamic Pricing

- How does Uber determine the price of your ride?

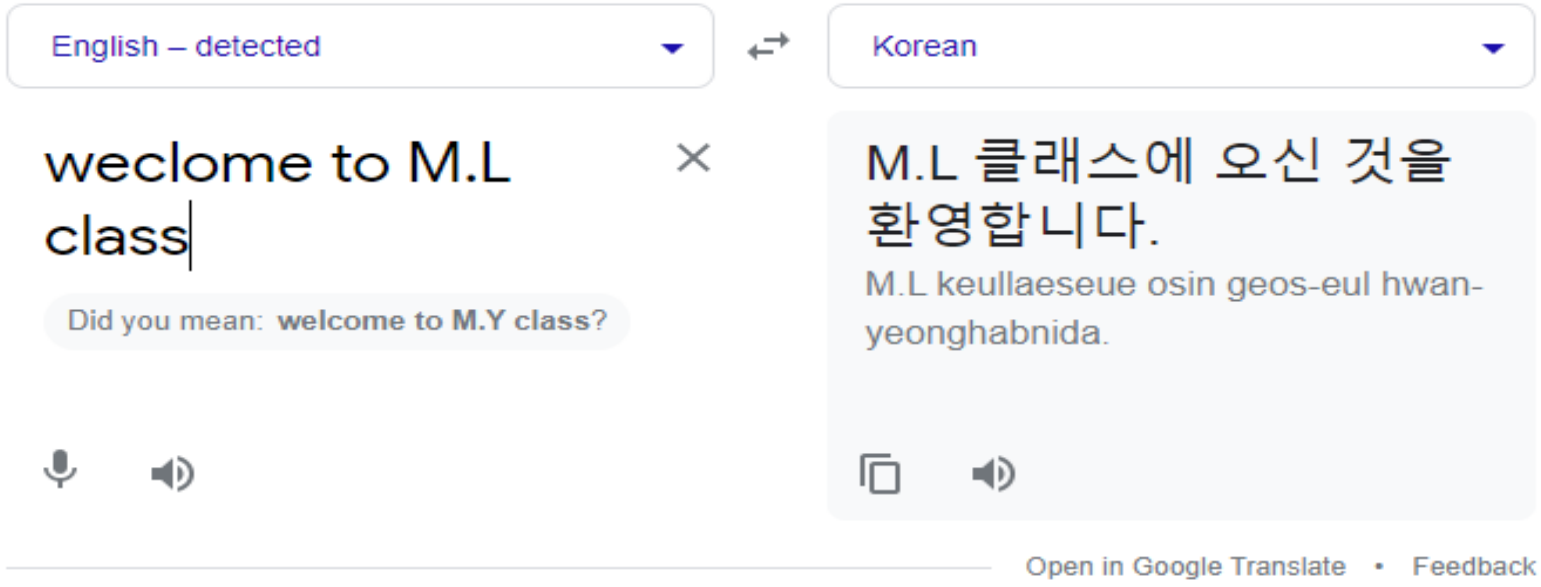
When to implement dynamic pricing

- Based on timing
- Based on demand
- Based on competition
- Based on quantity



Google Translate

- Google's **GNMT**(Google Neural Machine Translation) is a Neural Machine Learning that works on thousands of languages and dictionaries, uses ***Natural Language Processing*** to provide the most accurate translation.

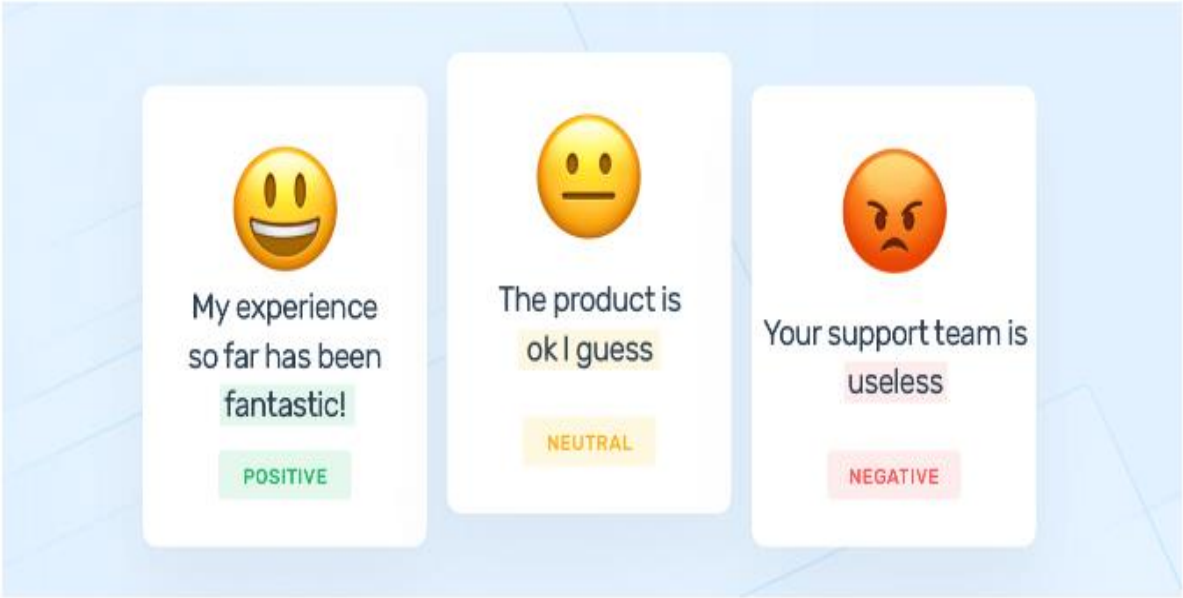


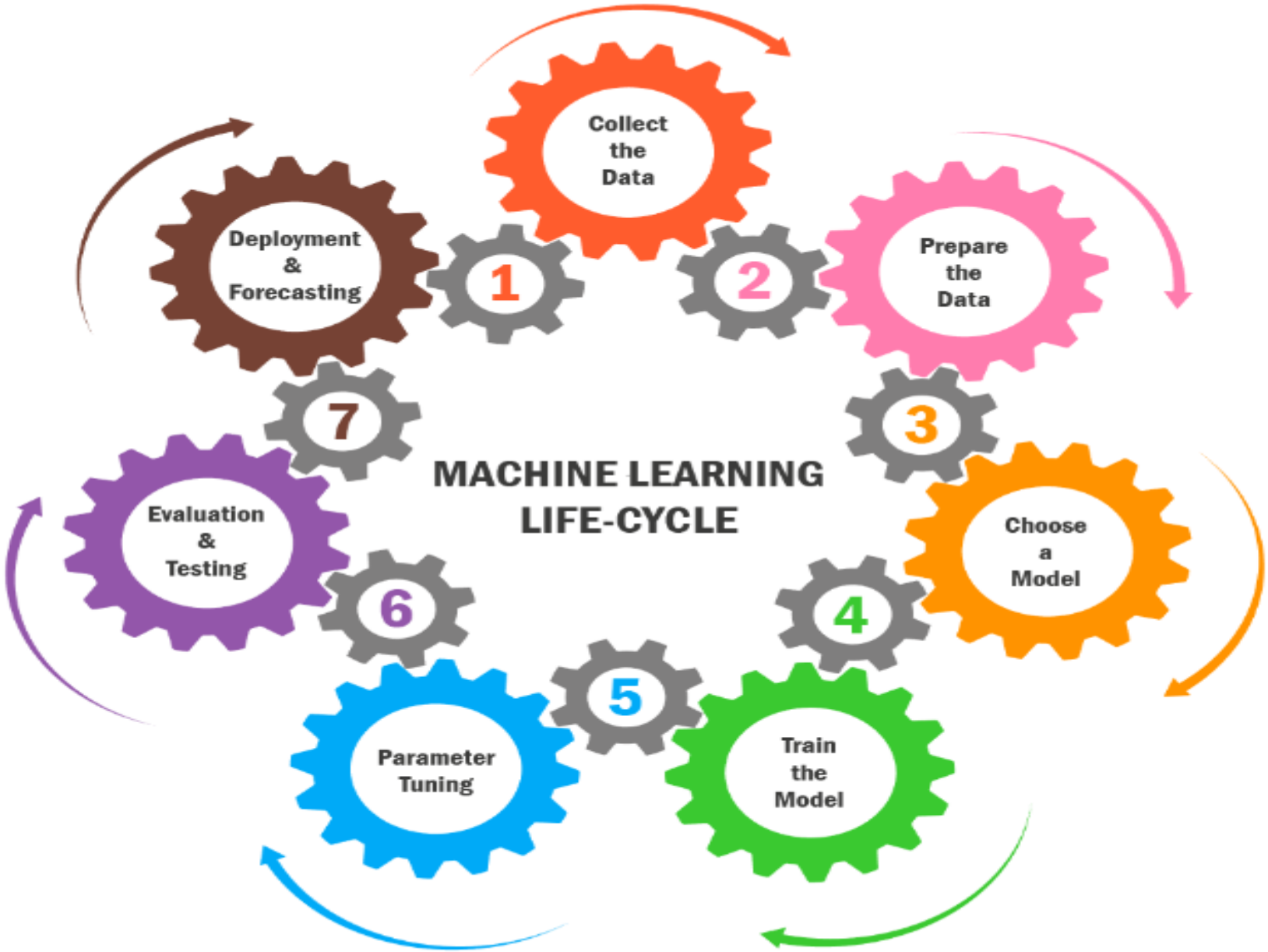
Social Media (Facebook)

- One of the most common applications of Machine Learning is **Automatic Friend Tagging Suggestions** in Facebook or any other social media platform.
- Facebook uses **face detection** and **Image recognition** to automatically find the face of the person which matches it's database and hence suggests us to tag that person.

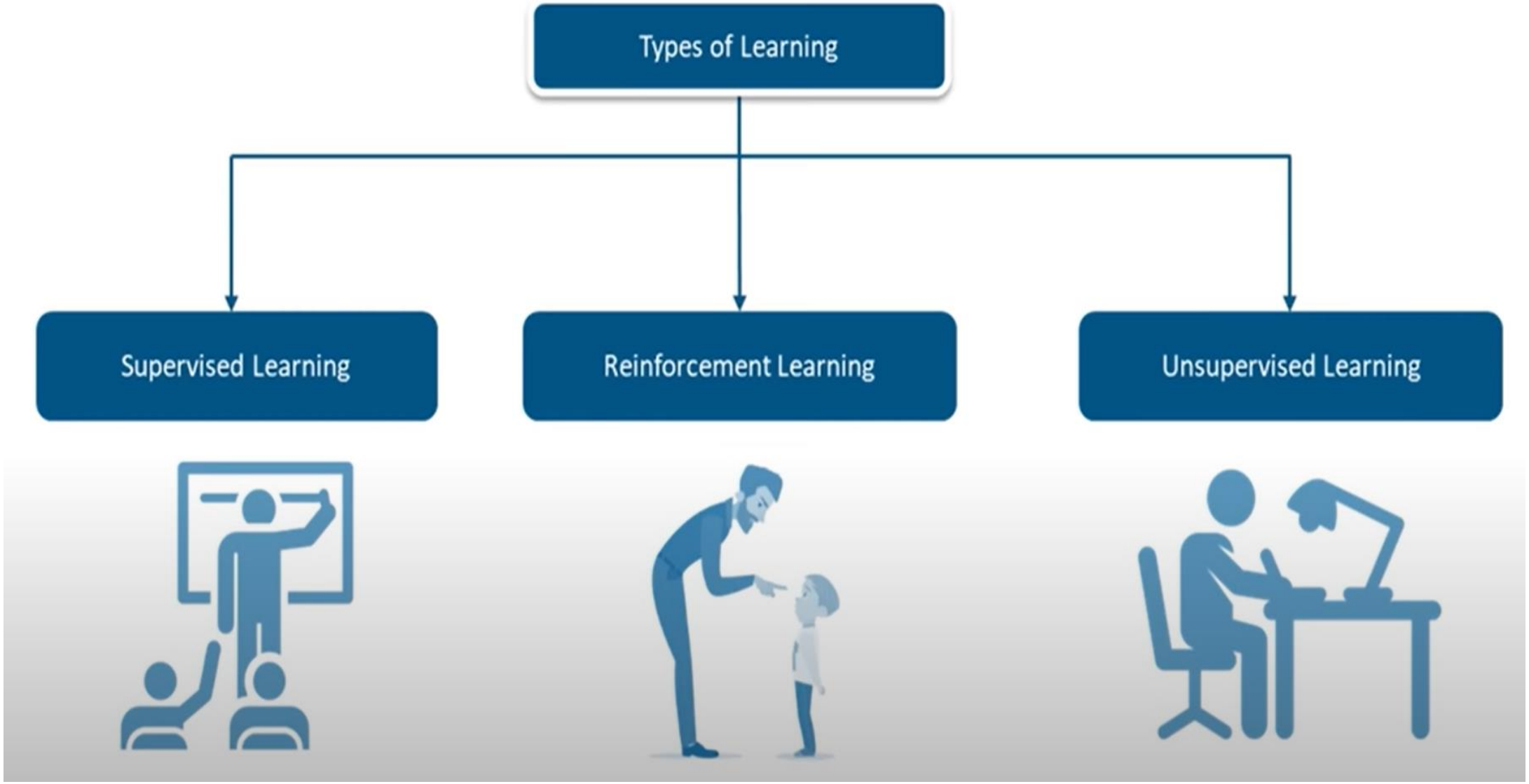
Sentiment Analysis

- Sentiment analysis is a real-time machine learning application that determines the emotion or opinion of the speaker or the writer.

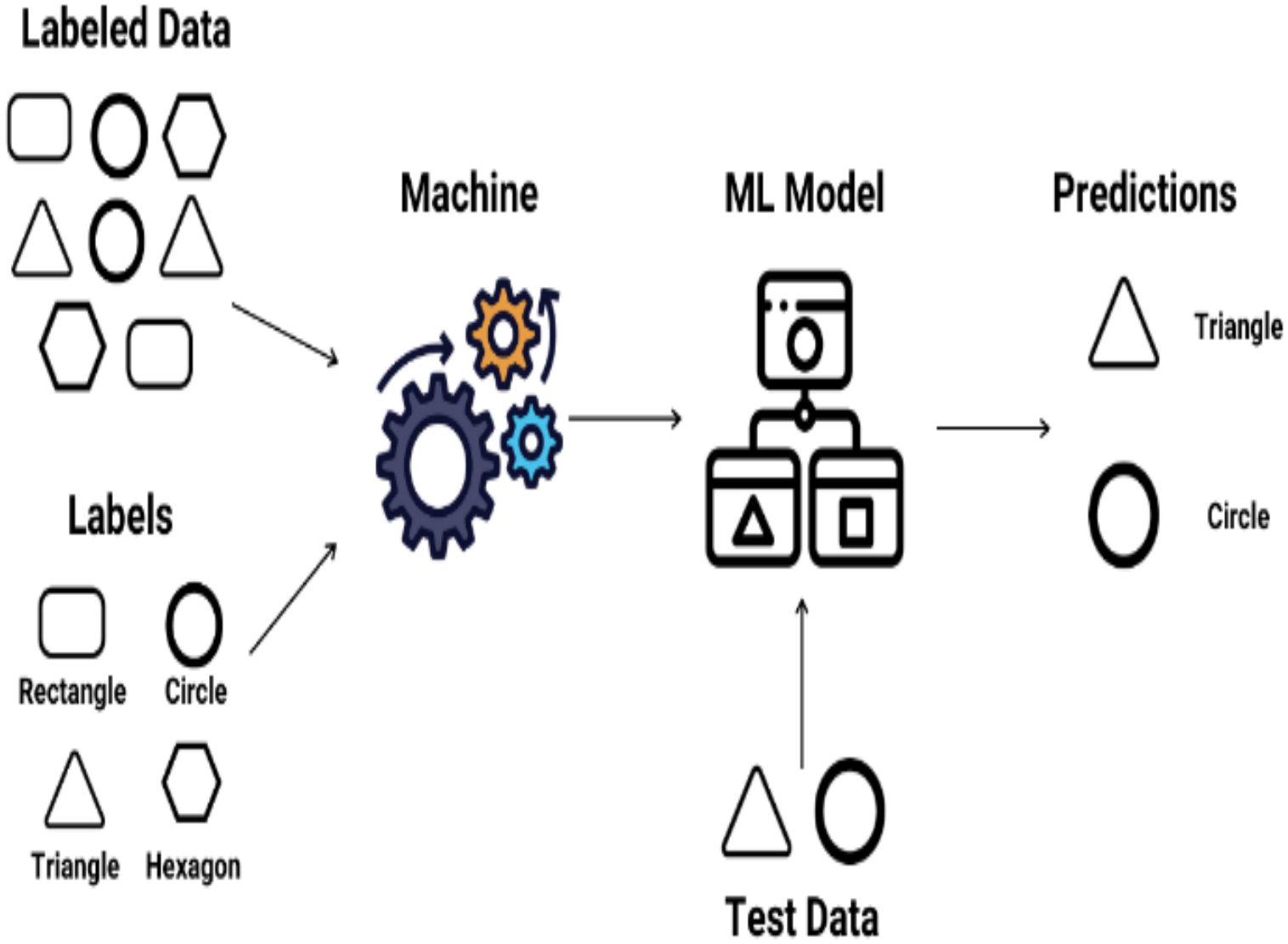


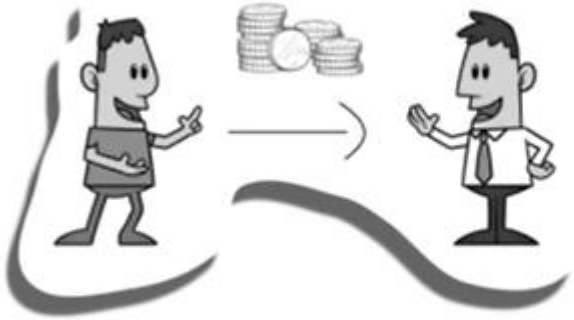

MACHINE LEARNING
LIFE-CYCLE
1
Collect the Data
2
Prepare the Data
3
Choose a Model
4
Train the Model
5
Parameter Tuning
6
Evaluation & Testing
7
Deployment & Forecasting

MACHINE LEARNING TYPES ?



1.SUPERVISED LEARNING





3 GRAMS

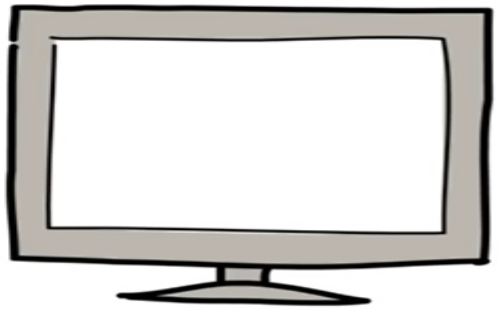
7 GRAMS

4 GRAMS

WEIGHT = FEATURE

CURRENCY = LABEL

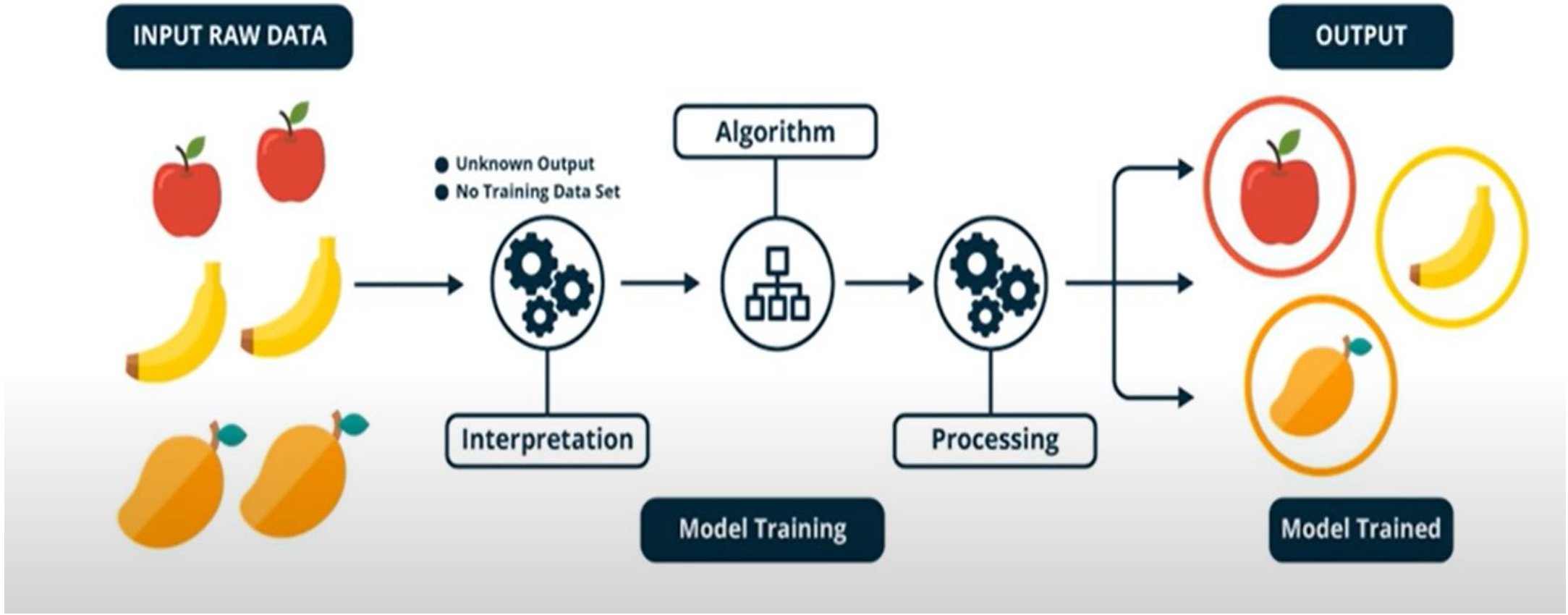
3 GRAMS = 1 RUPEE COIN

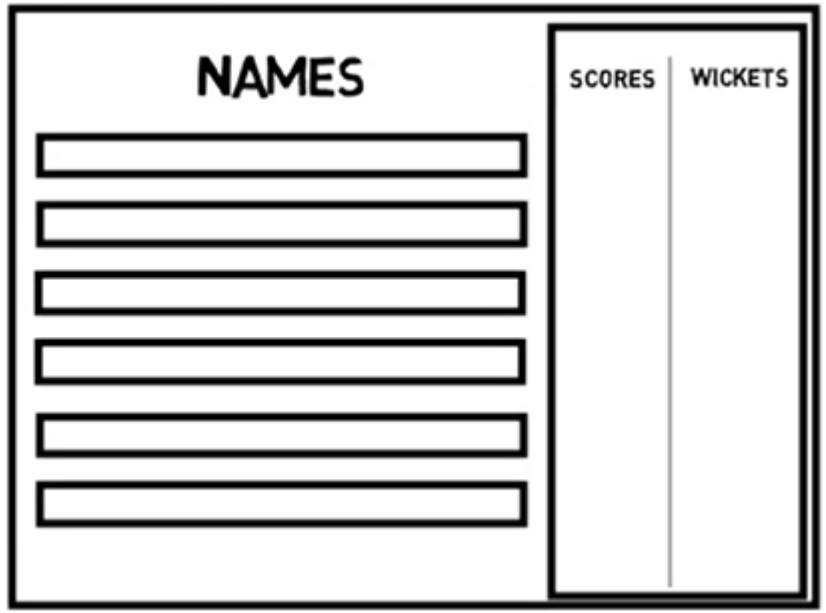


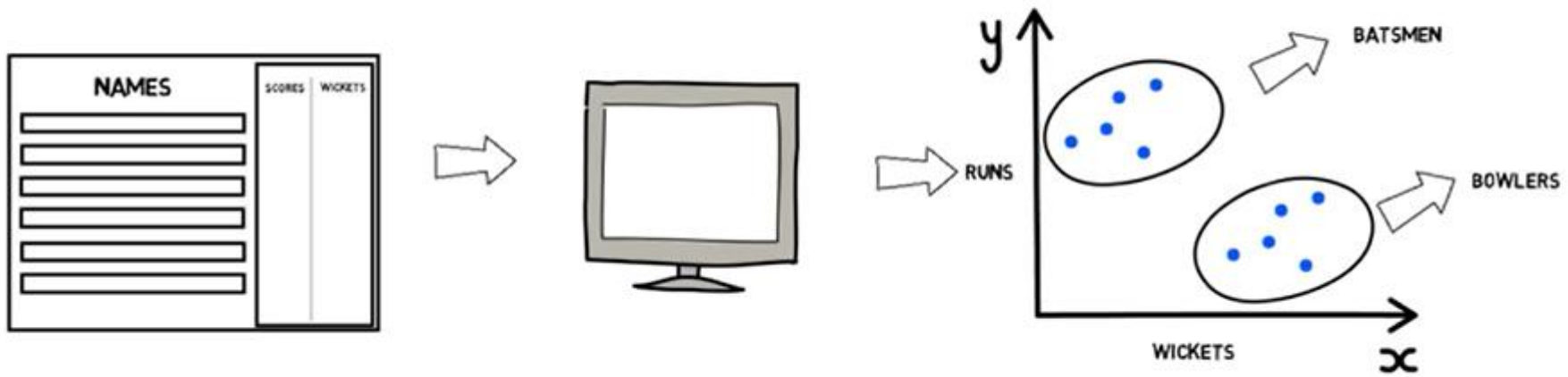
1 RUPEE

LABELED DATA

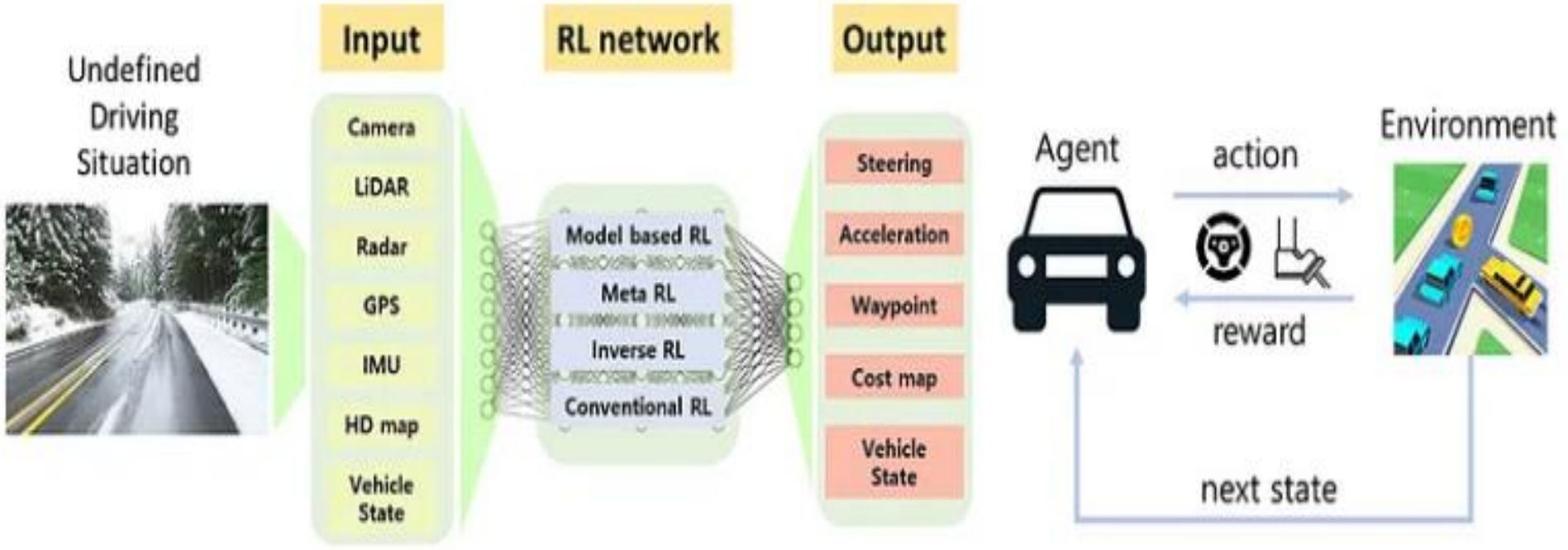
2.UNSUPERVISED LEARNING

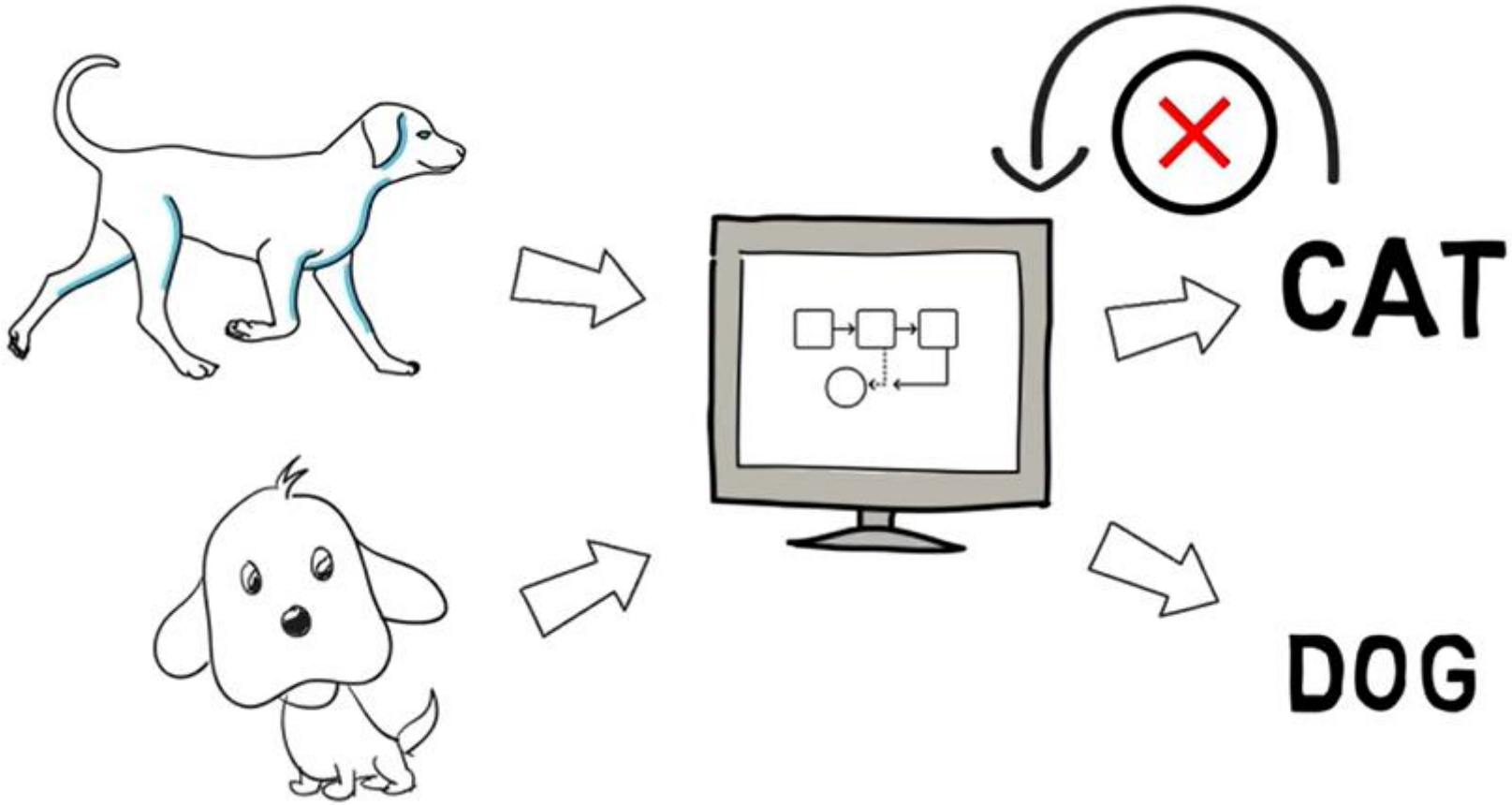



NAMES
SCORES
WICKETS


NAMES
SCORES
WICKETS
y
RUNS
BATSMEN
BOWLERS
WICKETS
x

3.REINFORCEMENT LEARNING




CAT
DOG

BASIC TERMINOLOGY

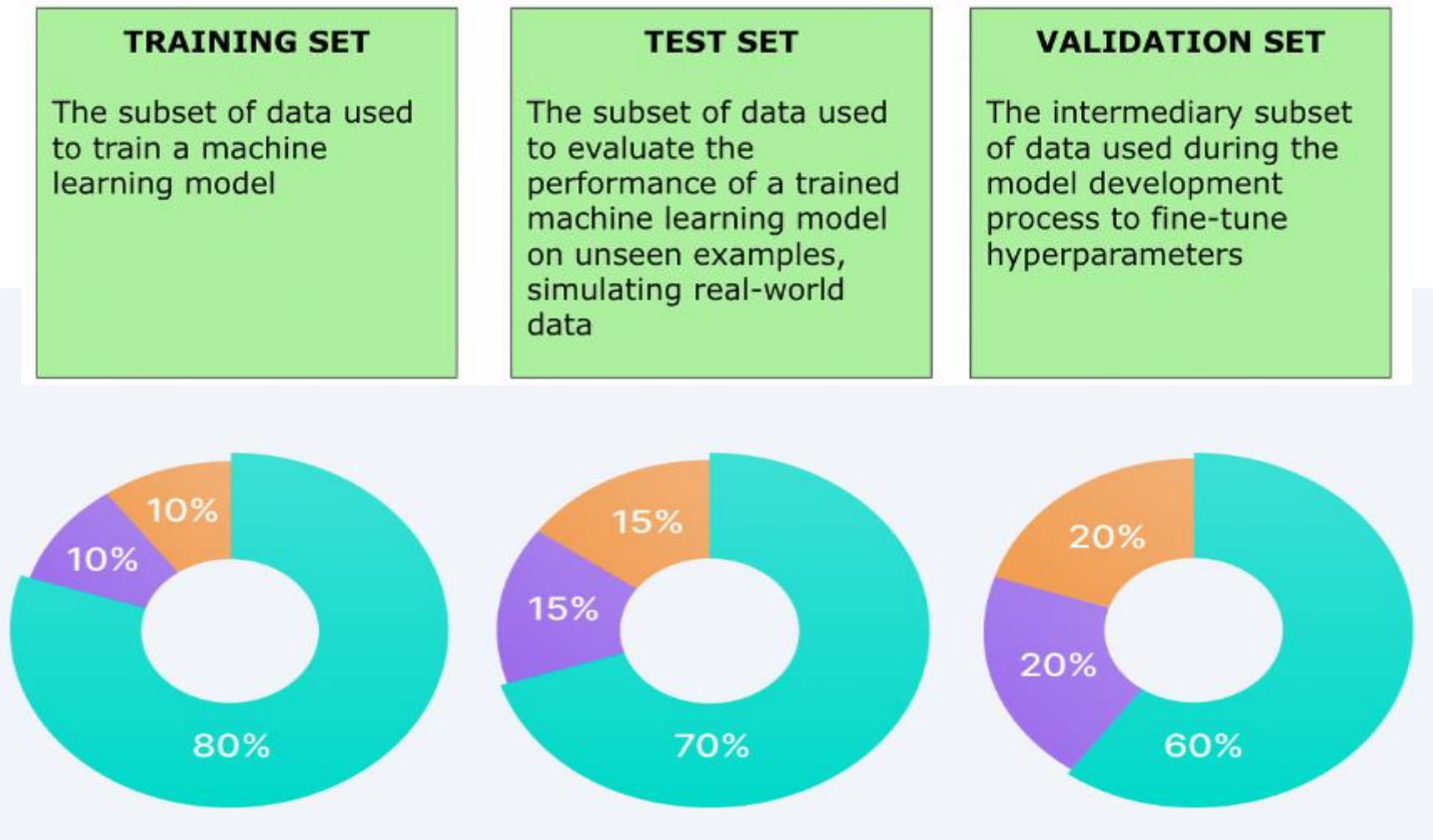
FEATURES AND LABELS:

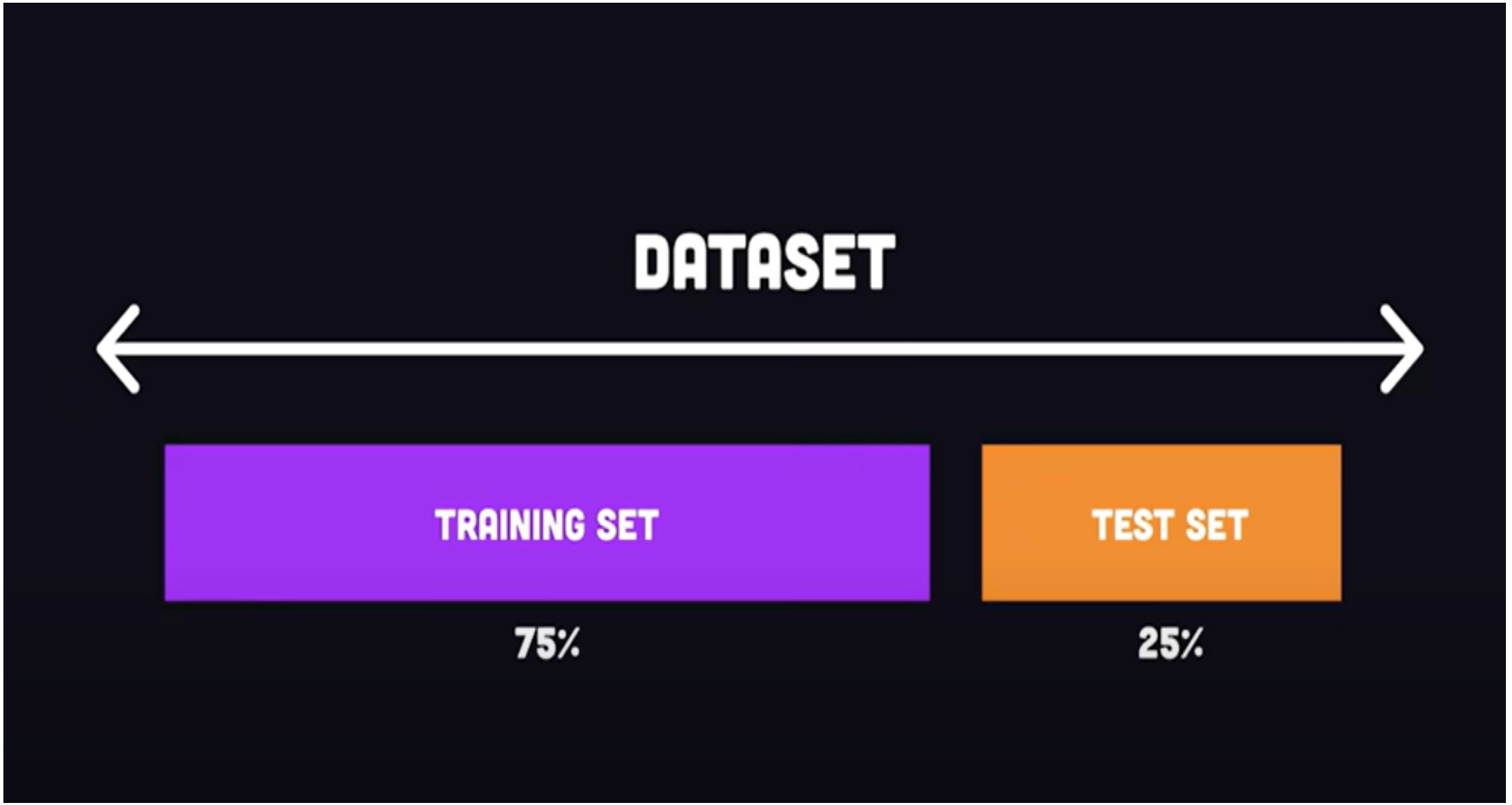
- **FEATURES:** THE INPUT VARIABLES (INDEPENDENT VARIABLES) USED BY THE MODEL TO MAKE PREDICTIONS.
- **LABELS:** THE OUTPUT VARIABLE (DEPENDENT VARIABLE) THAT THE MODEL IS TRYING TO PREDICT.

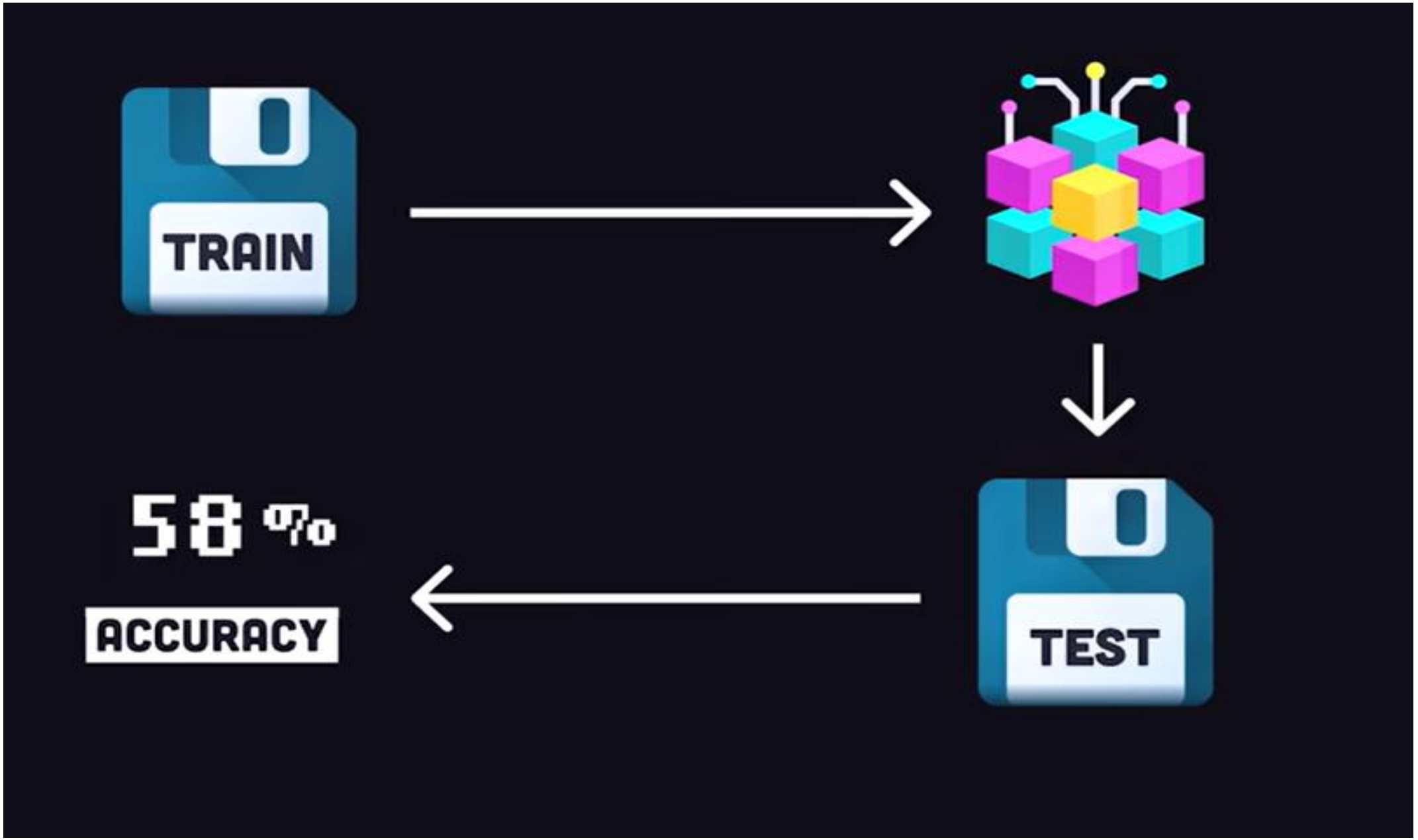
← Features →					Label
Position	Experience	Skill	Country	City	Salary ($)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Training and Testing:

- **Training set:** A subset of the dataset used to train the model.
- **Testing set:** A subset of the dataset used to evaluate the model's performance.

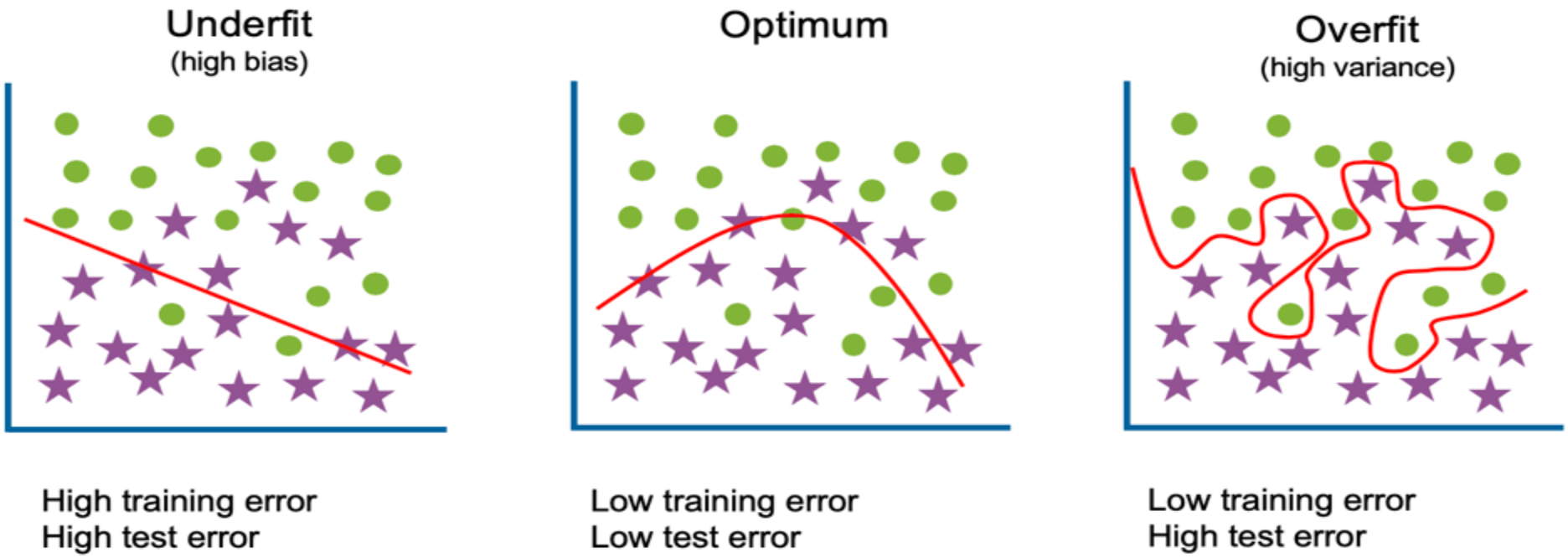



DATASET
TRAINING SET
TEST SET
75%
25%


TRAIN
TEST
58 %
ACCURACY

Overfitting and Underfitting:

- **Overfitting:** When the model performs well on the training data but poorly on the testing data because it has learned noise and details from the training data.
- **Underfitting:** When the model performs poorly on both the training and testing data because it is too simple to capture the underlying patterns in the data.



Machine learning algorithms are a set of methods used to enable computers to learn from data and make predictions or decisions without being explicitly programmed for specific tasks.


CLASSIFICATION

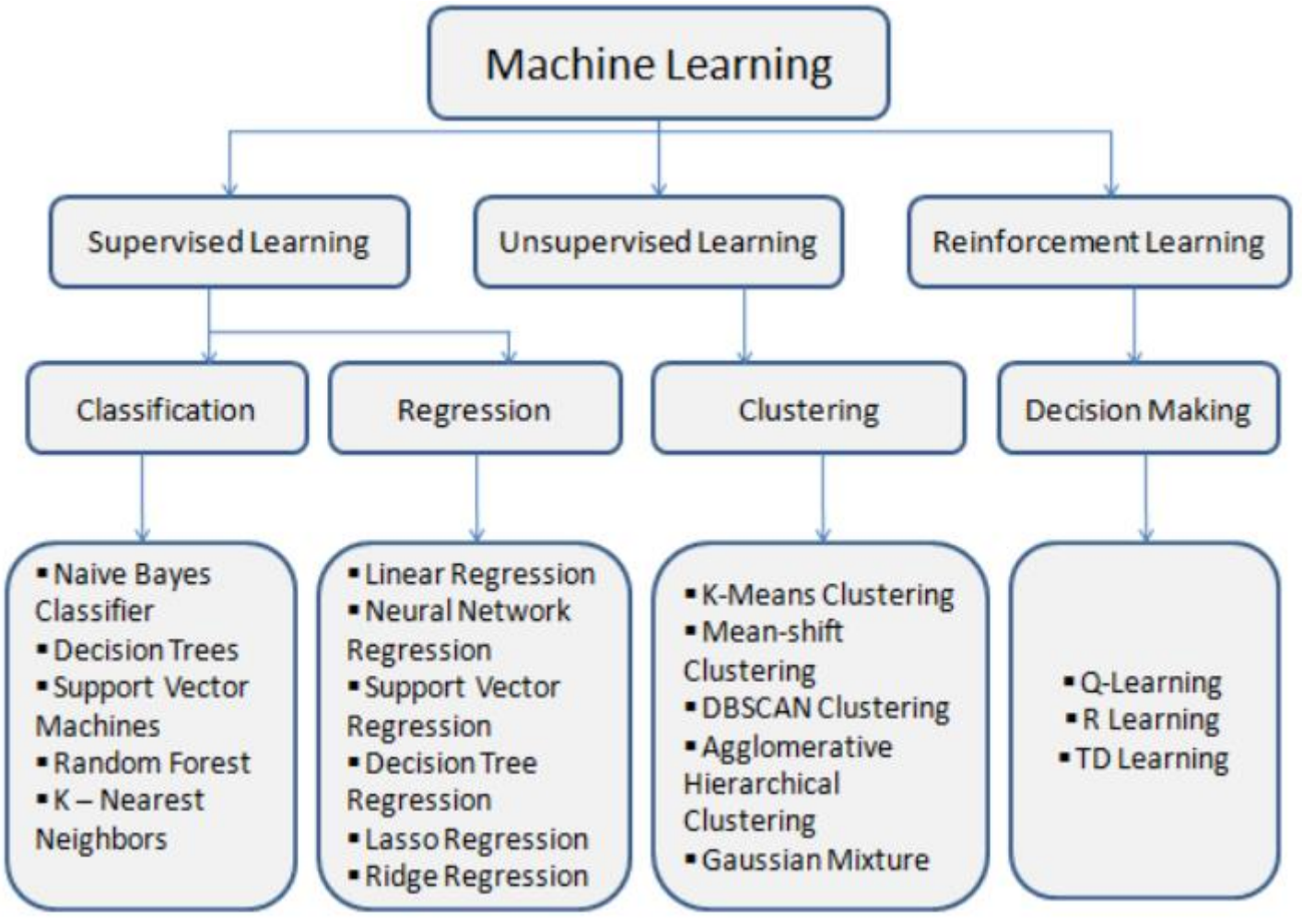
CLASSIFICATION

?


PREDICTION


REGRESSION
$ 43.01
Sara Lee
Iron Kids
White
Bread

MACHINE LEARNING ALGORITHMS



	ALGORITHM	DESCRIPTION	APPLICATIONS	ADVANTAGES	DISADVANTAGES
Linear Models	**Linear Regression**	A simple algorithm that models a linear relationship between inputs and a continuous numerical output variable	USE CASES 1. Stock price prediction 2. Predicting housing prices 3. Predicting customer lifetime value	1. Explainable method 2. Interpretable results by its output coefficients 3. Faster to train than other machine learning models	1. Assumes linearity between inputs and output 2. Sensitive to outliers 3. Can underfit with small, high-dimensional data
	Logistic Regression	A simple algorithm that models a linear relationship between inputs and a categorical output (1 or 0)	USE CASES 1. Credit risk score prediction 2. Customer churn prediction	1. Interpretable and explainable 2. Less prone to overfitting when using regularization 3. Applicable for multi-class predictions	1. Assumes linearity between inputs and outputs 2. Can overfit with small, high-dimensional data
	Ridge Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients closer to zero. Can be used for classification or regression	USE CASES 1. Predictive maintenance for automobiles 2. Sales revenue prediction	1. Less prone to overfitting 2. Best suited where data suffer from multicollinearity 3. Explainable & interpretable	1. All the predictors are kept in the final model 2. Doesn't perform feature selection
	Lasso Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients to zero. Can be used for classification or regression	USE CASES 1. Predicting housing prices 2. Predicting clinical outcomes based on health data	1. Less prone to overfitting 2. Can handle high-dimensional data 3. No need for feature selection	1. Can lead to poor interpretability as it can keep highly correlated variables

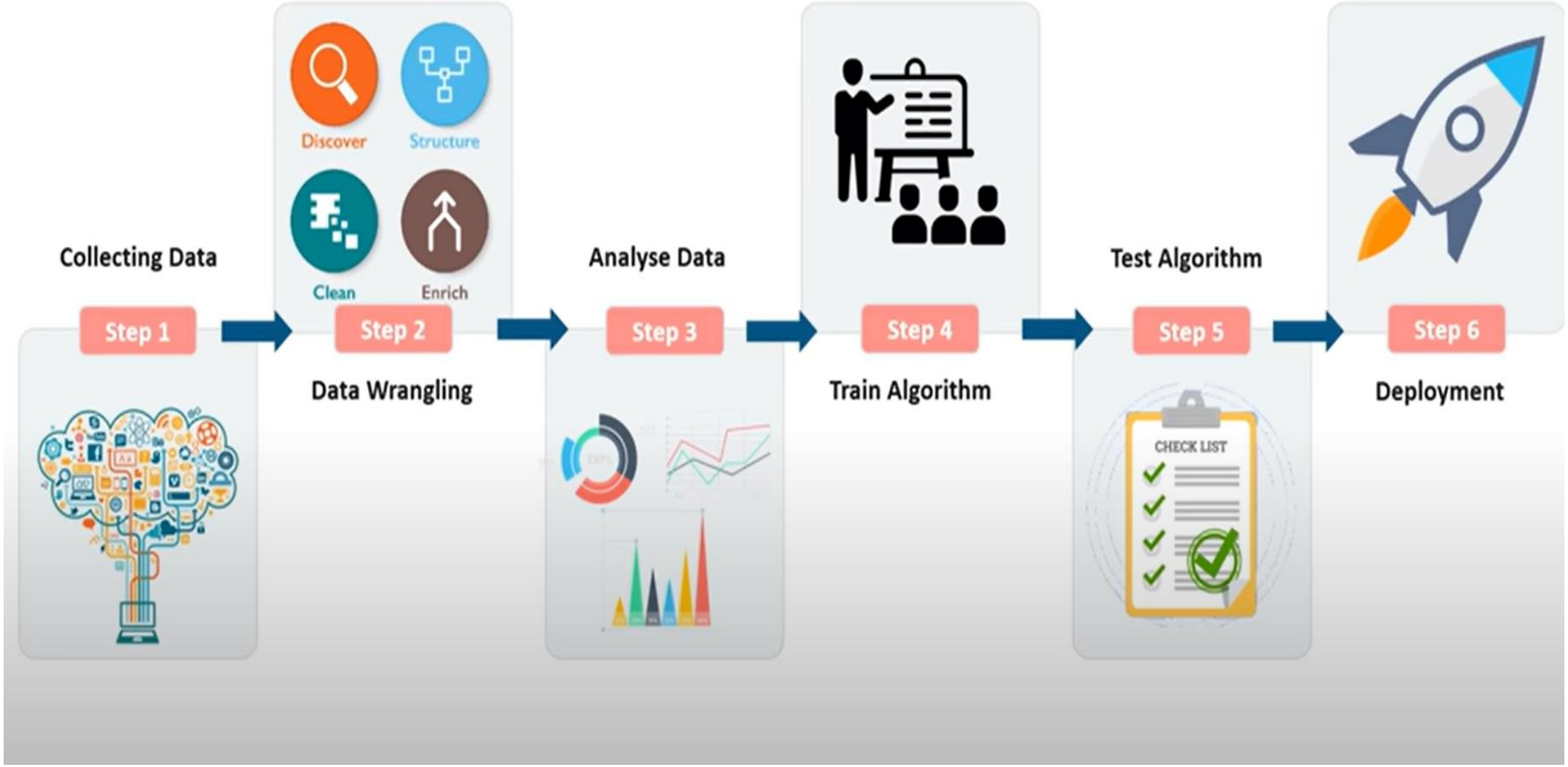
Kmit

Tree-Based Models

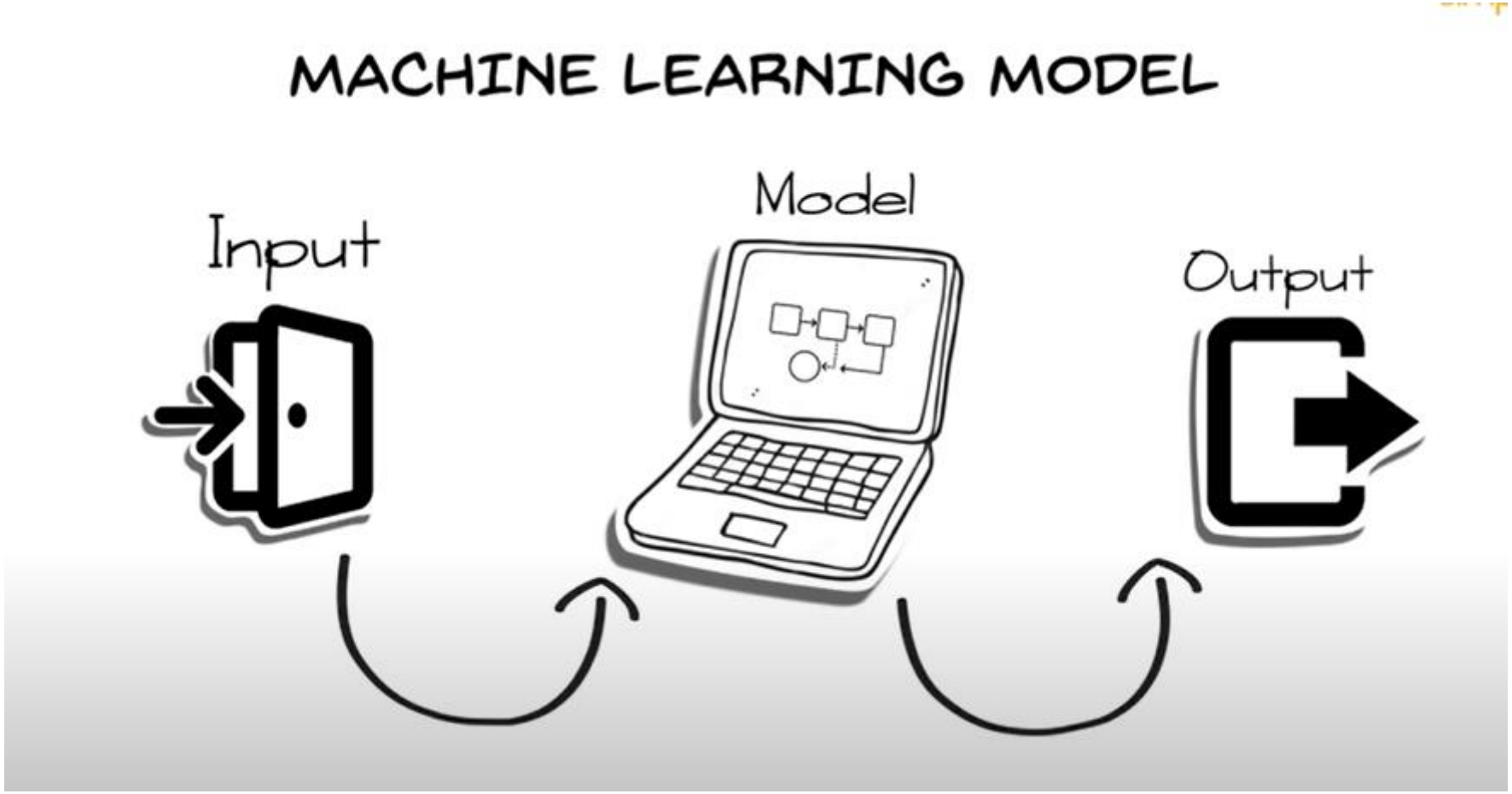
Decision Tree	Decision Tree models make decision rules on the features to produce predictions. It can be used for classification or regression	**USE CASES** 1. Customer churn prediction 2. Credit score modeling 3. Disease prediction	1. Explainable and interpretable 2. Can handle missing values	1. Prone to overfitting 2. Sensitive to outliers
Random Forests	An ensemble learning method that combines the output of multiple decision trees	**USE CASES** 1. Credit score modeling 2. Predicting housing prices	1. Reduces overfitting 2. Higher accuracy compared to other models	1. Training complexity can be high 2. Not very interpretable
Gradient Boosting Regression	Gradient Boosting Regression employs boosting to make predictive models from an ensemble of weak predictive learners	**USE CASES** 1. Predicting car emissions 2. Predicting ride hailing fare amount	1. Better accuracy compared to other regression models 2. It can handle multicollinearity 3. It can handle non-linear relationships	1. Sensitive to outliers and can therefore cause overfitting 2. Computationally expensive and has high complexity
XGBoost	Gradient Boosting algorithm that is efficient & flexible. Can be used for both classification and regression tasks	**USE CASES** 1. Churn prediction 2. Claims processing in insurance	1. Provides accurate results 2. Captures non linear relationships	1. Hyperparameter tuning can be complex 2. Does not perform well on sparse datasets
LightGBM Regressor	A gradient boosting framework that is designed to be more efficient than other implementations	**USE CASES** 1. Predicting flight time for airlines 2. Predicting cholesterol levels based on health data	1. Can handle large amounts of data 2. Computational efficient & fast training speed 3. Low memory usage	1. Can overfit due to leaf-wise splitting and high sensitivity 2. Hyperparameter tuning can be complex

Clustering	K-Means	K-Means is the most widely used clustering approach—It determines K clusters based on euclidean distances	USE CASES 1. Customer segmentation 2. Recommendation systems	1. Scales to large datasets 2. Simple to implement and interpret 3. Results in tight clusters	1. Requires the expected number of clusters from the beginning 2. Has troubles with varying cluster sizes and densities
	Hierarchical Clustering	A "bottom-up" approach where each data point is treated as its own cluster—and then the closest two clusters are merged together iteratively	USE CASES 1. Fraud detection 2. Document clustering based on similarity	1. There is no need to specify the number of clusters 2. The resulting dendrogram is informative	1. Doesn't always result in the best clustering 2. Not suitable for large datasets due to high complexity
	Gaussian Mixture Models	A probabilistic model for modeling normally distributed clusters within a dataset	USE CASES 1. Customer segmentation 2. Recommendation systems	1. Computes a probability for an observation belonging to a cluster 2. Can identify overlapping clusters 3. More accurate results compared to K-means	1. Requires complex tuning 2. Requires setting the number of expected mixture components or clusters
Association	Apriori algorithm	Rule based approach that identifies the most frequent itemset in a given dataset where prior knowledge of frequent itemset properties is used	USE CASES 1. Product placements 2. Recommendation engines 3. Promotion optimization	1. Results are intuitive and Interpretable 2. Exhaustive approach as it finds all rules based on the confidence and support	1. Generates many uninteresting itemsets 2. Computationally and memory intensive. 3. Results in many overlapping item sets

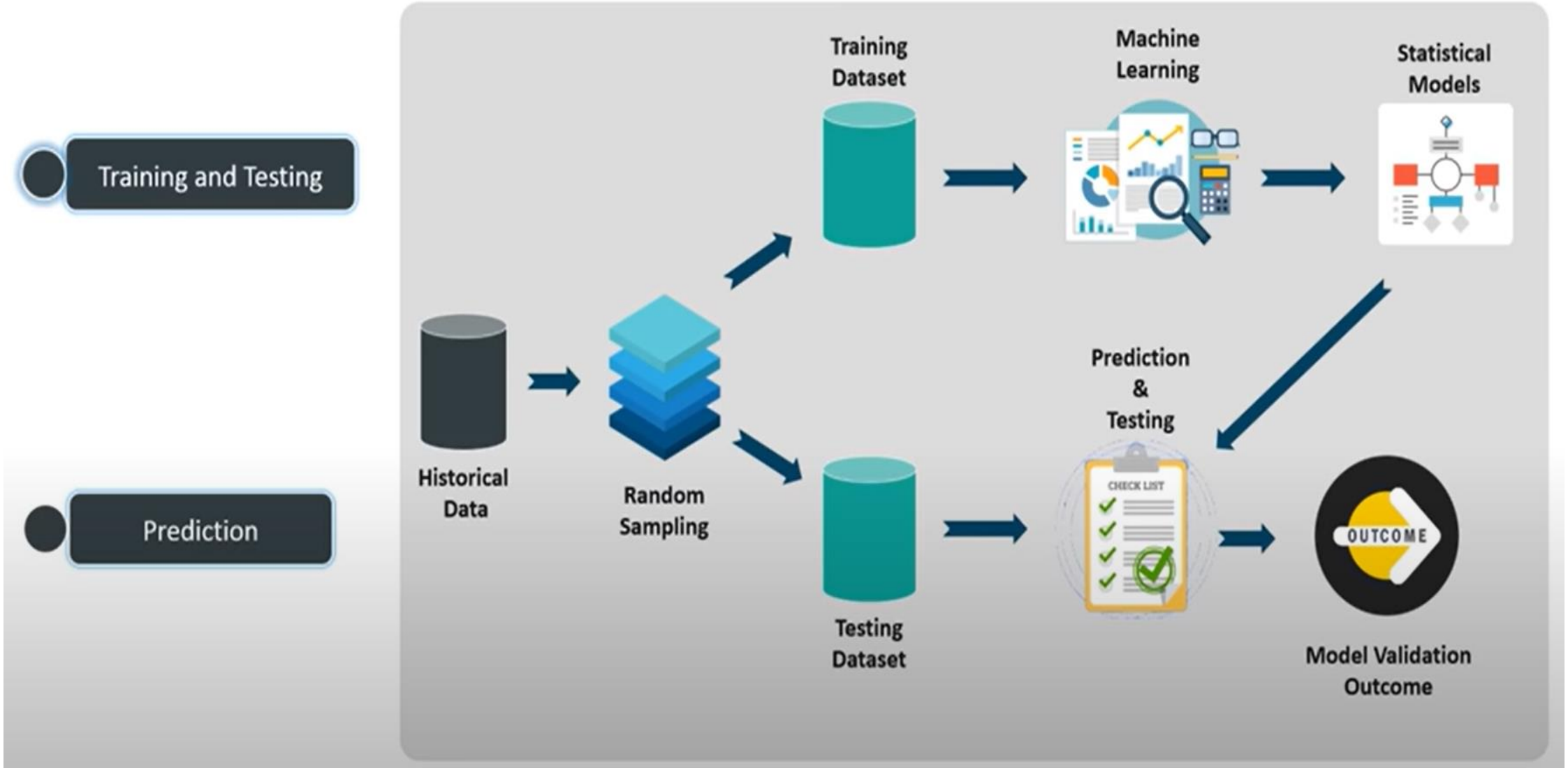
ALGORITHM DEVELOPMENT STEPS



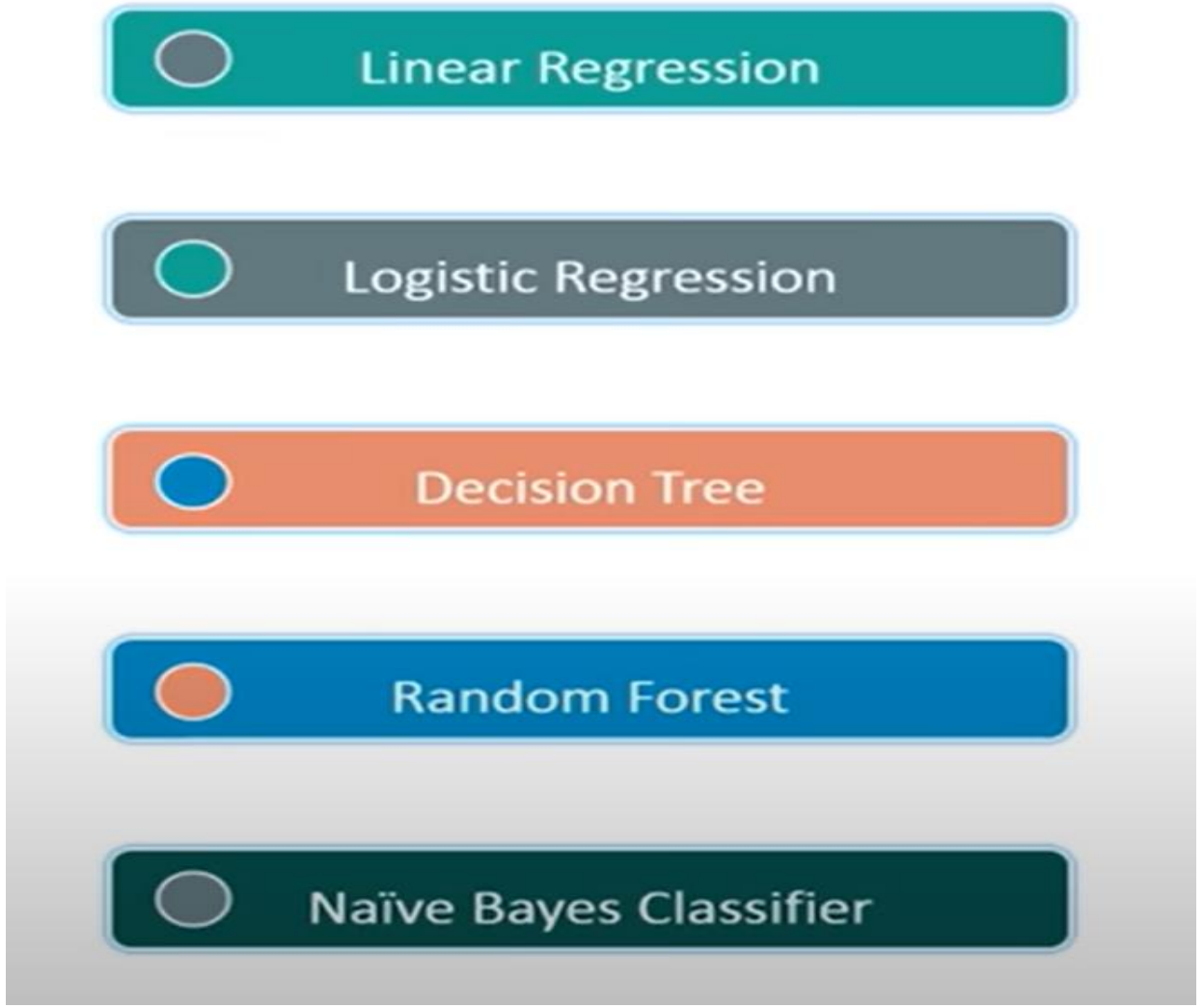

MODEL


MACHINE LEARNING MODEL
Input
Model
Output

1. SUPERVISED LEARNING

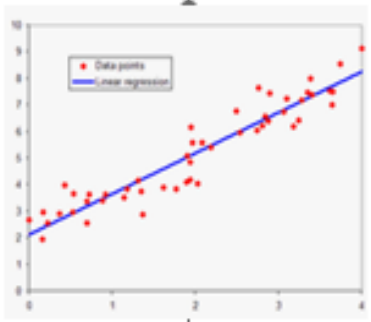


SUPERVISED LEARNING ALGORITHMS



Machine Learning-Popular Algorithms

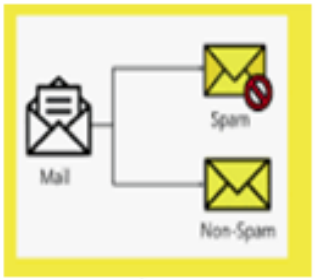
01



REGRESSION

predict the continuous values

02



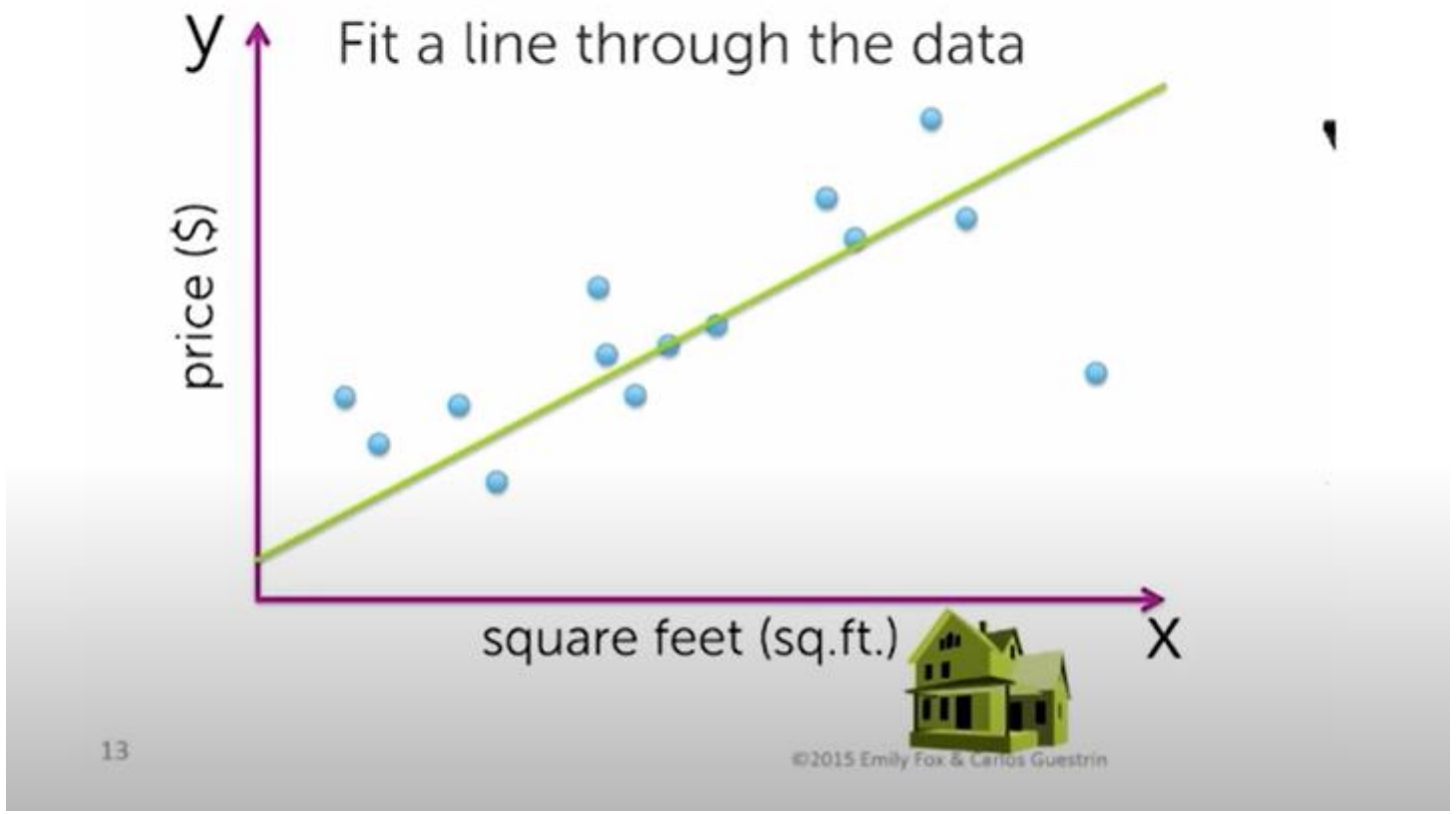
CLASSIFICATION

predict the categorical values

REGRESSION:

PREDICT A CONTINUOUS NUMERIC TARGET VARIABLE

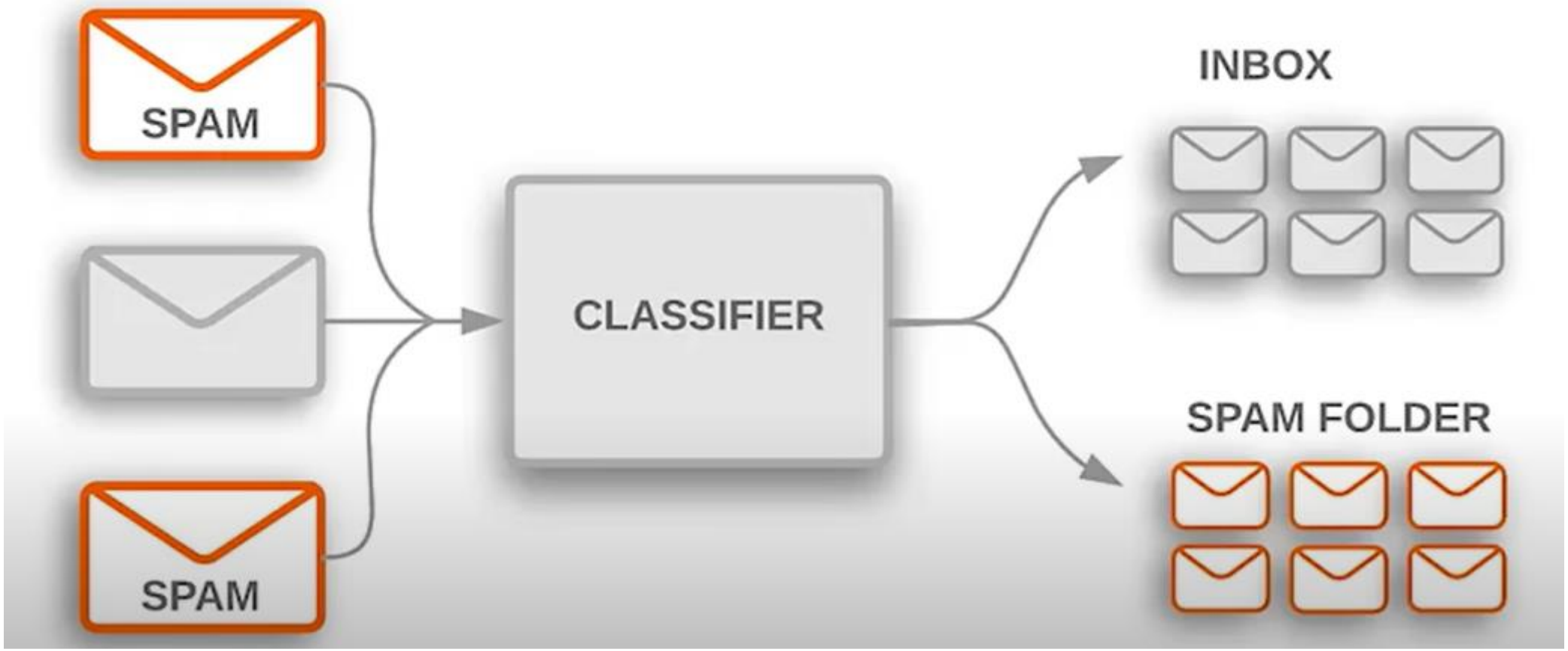
Use a **linear** regression model



Kmit

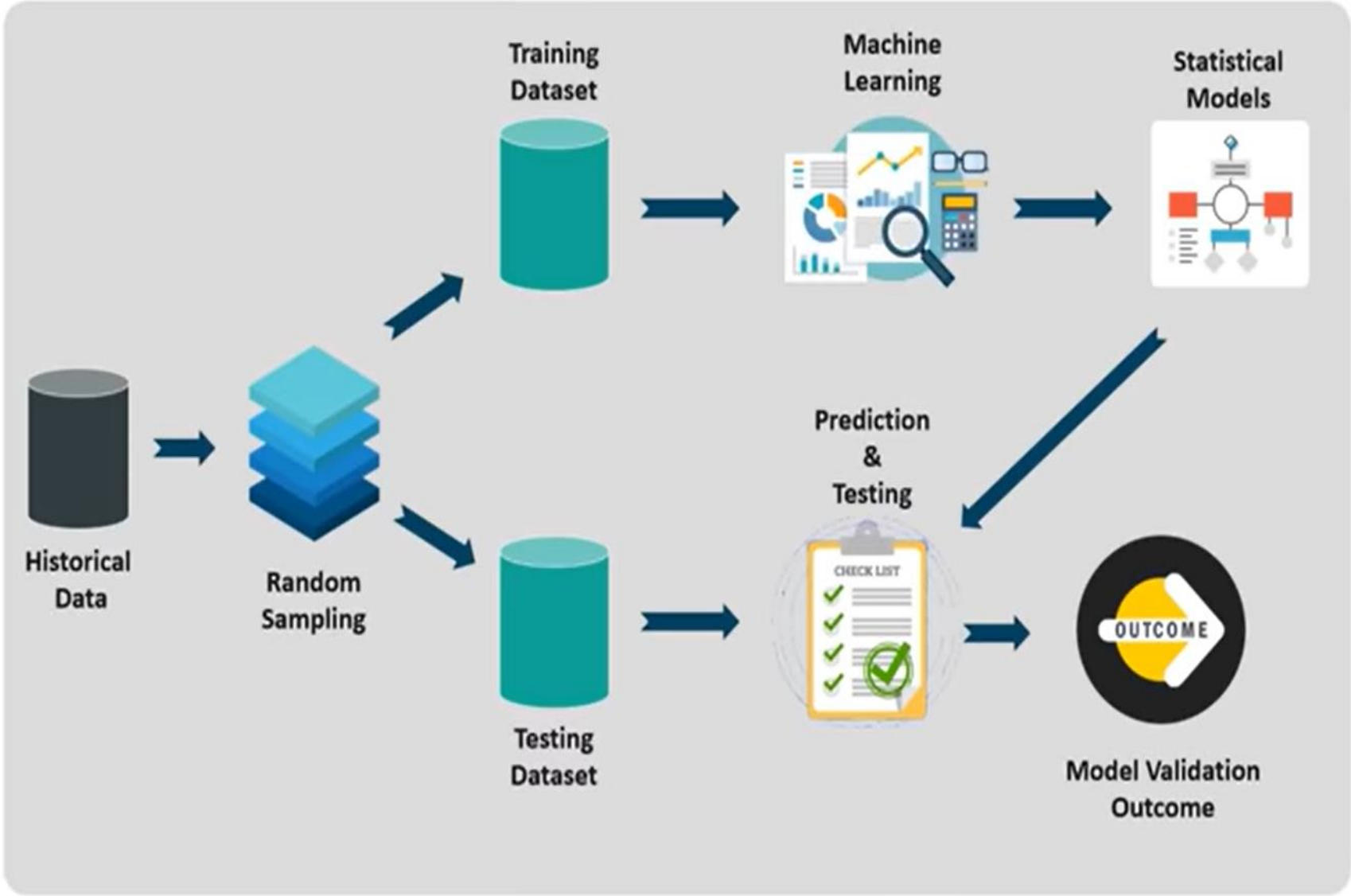
CLASSIFICATION:

PREDICT DISCRETE CATEGORICAL VARIABLE (“LABEL” / “CLASS”)

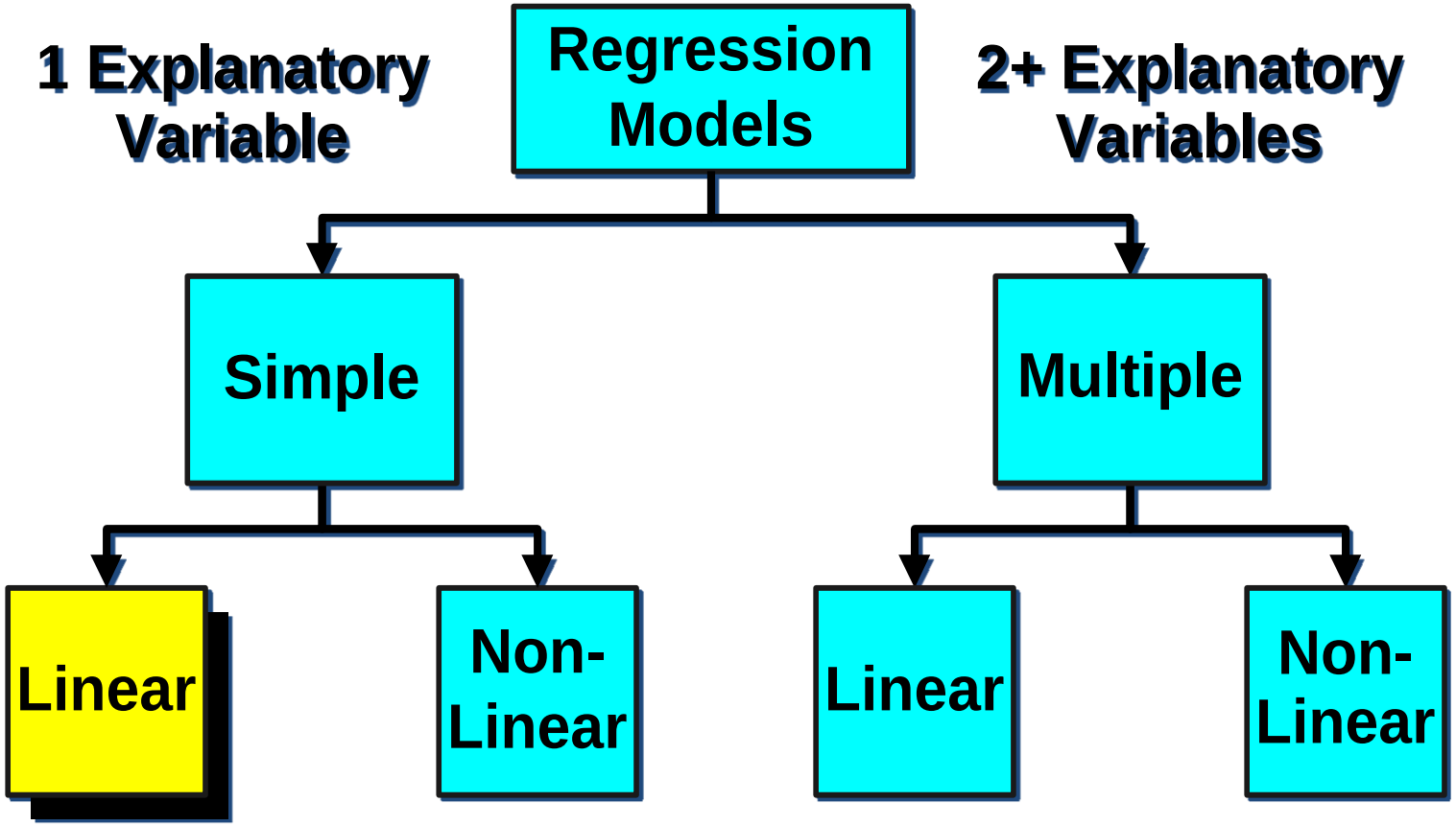

SPAM
SPAM
CLASSIFIER
INBOX
SPAM FOLDER

Kmit

- Training and Testing
- Prediction



TYPES OF REGRESSION MODELS



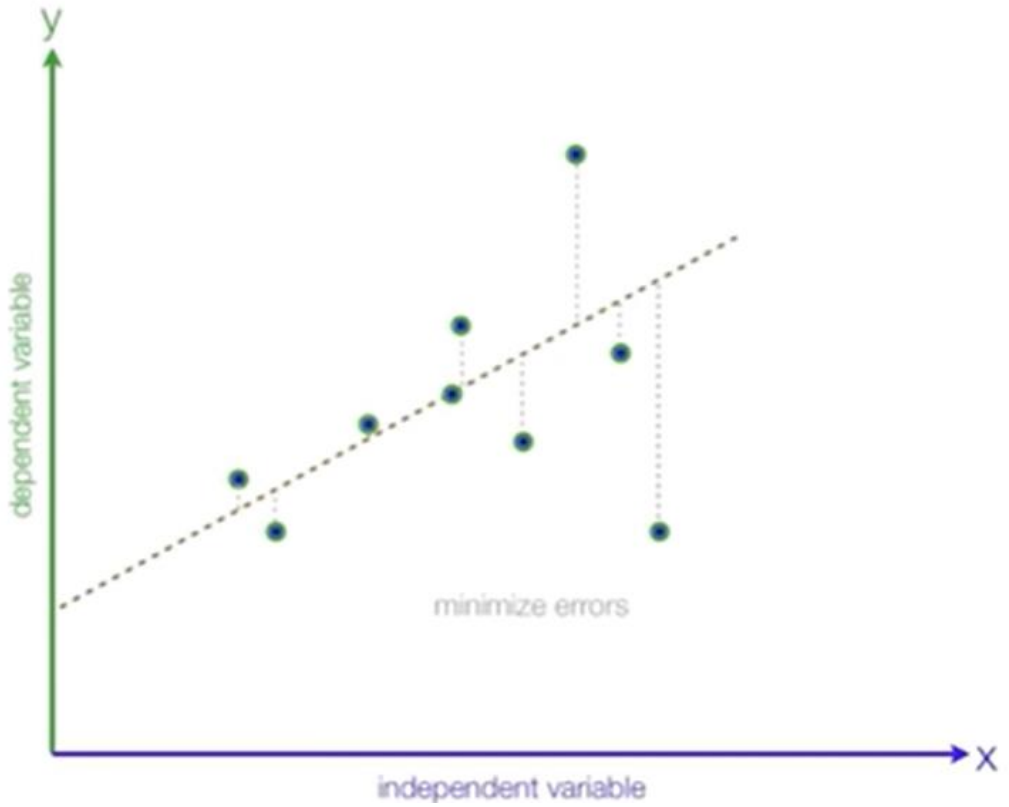
Kmit

LINEAR REGRESSION

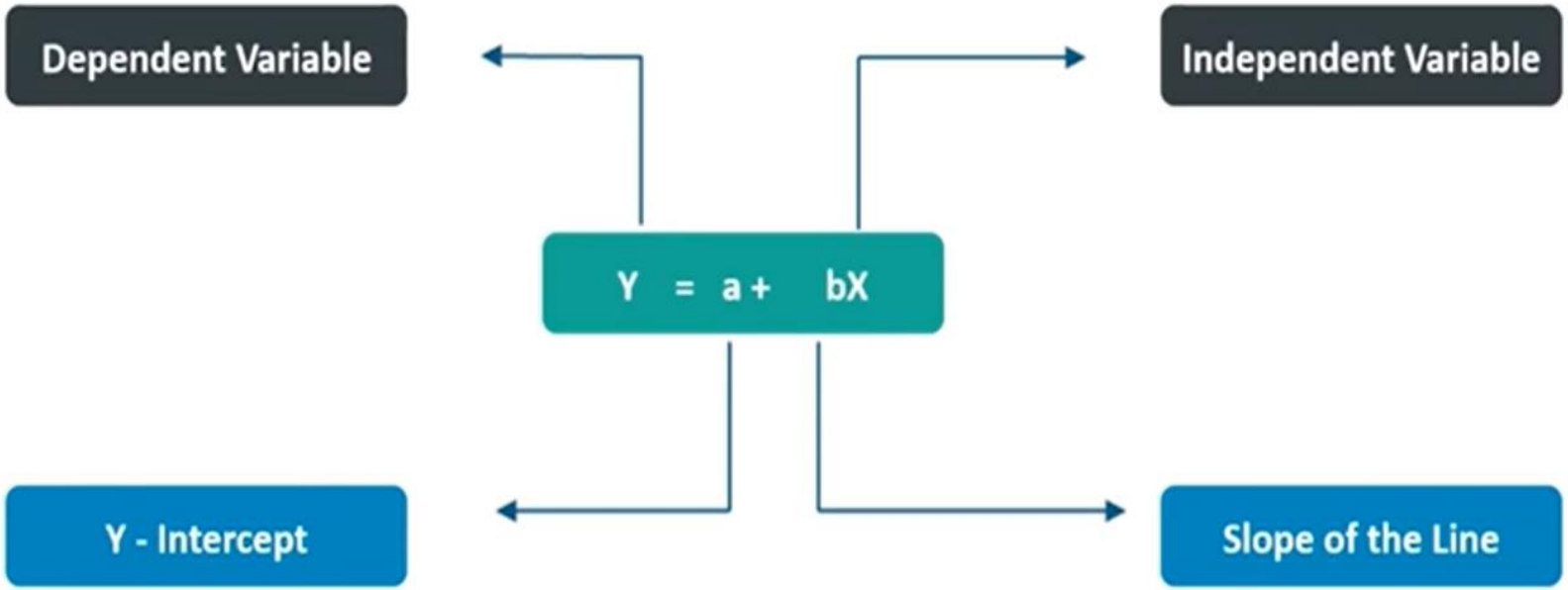
Linear Regression

Linear Regression Analysis is a powerful technique used for predicting the unknown value of a variable (**Dependent Variable**) from the known value of another variables (**Independent Variable**)

- A **Dependent Variable(DV)** is the variable to be predicted or explained in a regression model
- An **Independent Variable(IDV)** is the variable related to the dependent variable in a regression equation

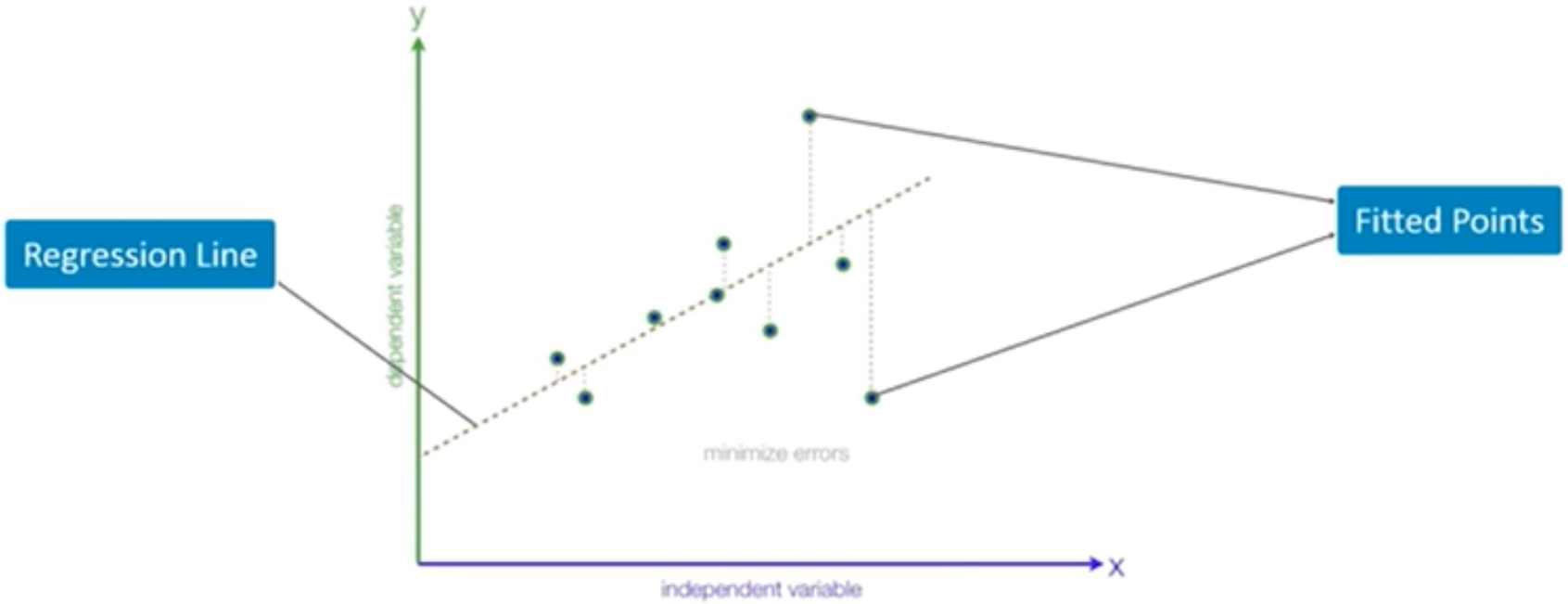


Simple Linear Regression



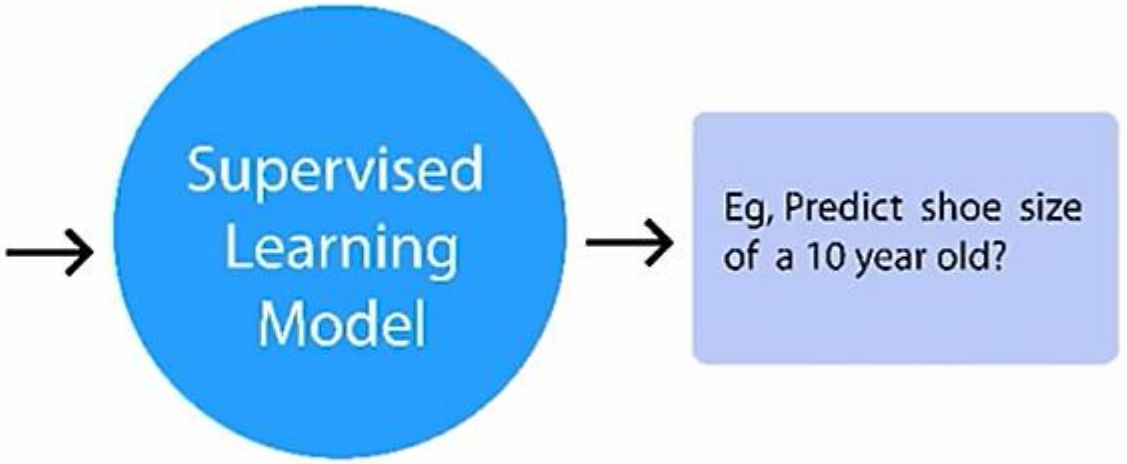
Regression Line

Linear Regression Analysis is a powerful technique used for predicting the unknown value of a variable (**Dependent Variable**) from The regression line is simply a single line that best fits the data (In terms of having the smallest overall distance from the line to the points)



EXAMPLE

Age (input)	Shoe size (inch) (output)
5	7
8	8.25
11	9
17	9.75

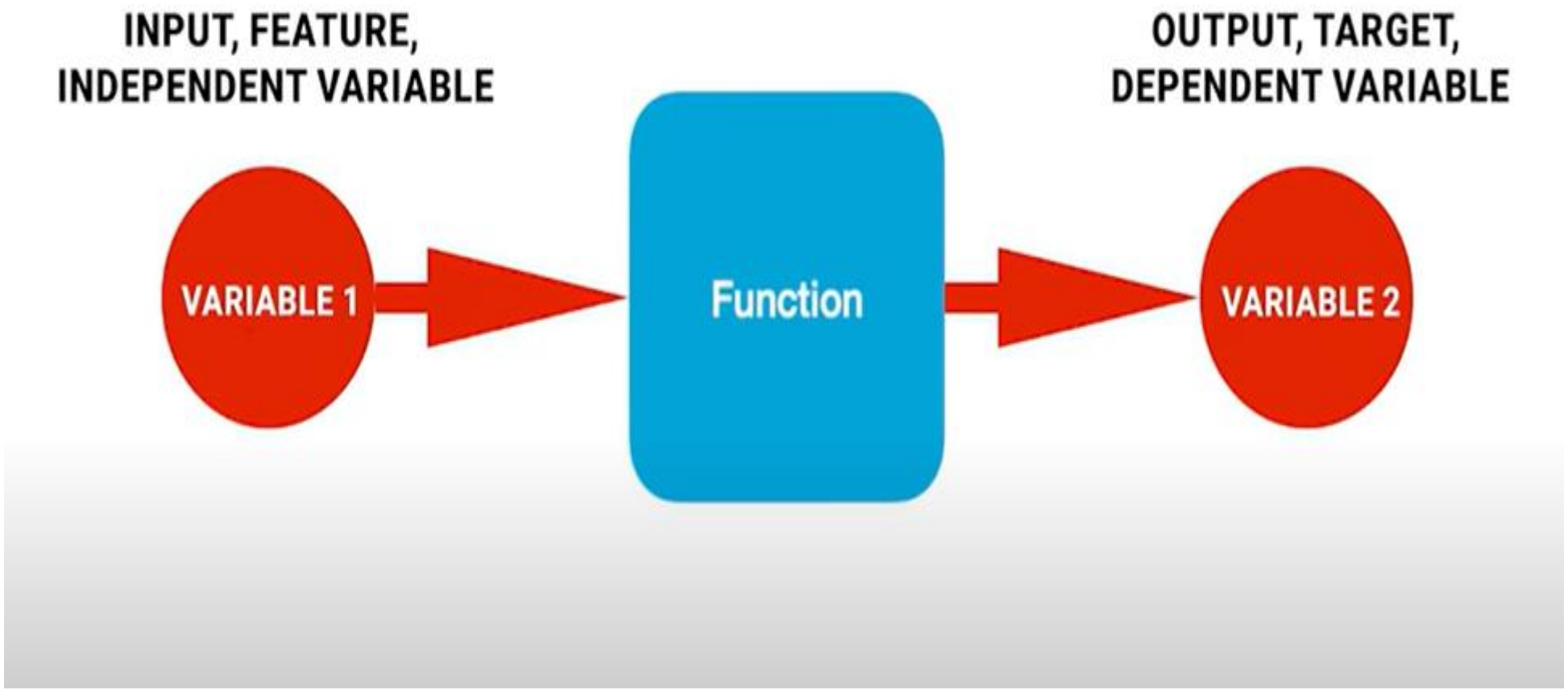


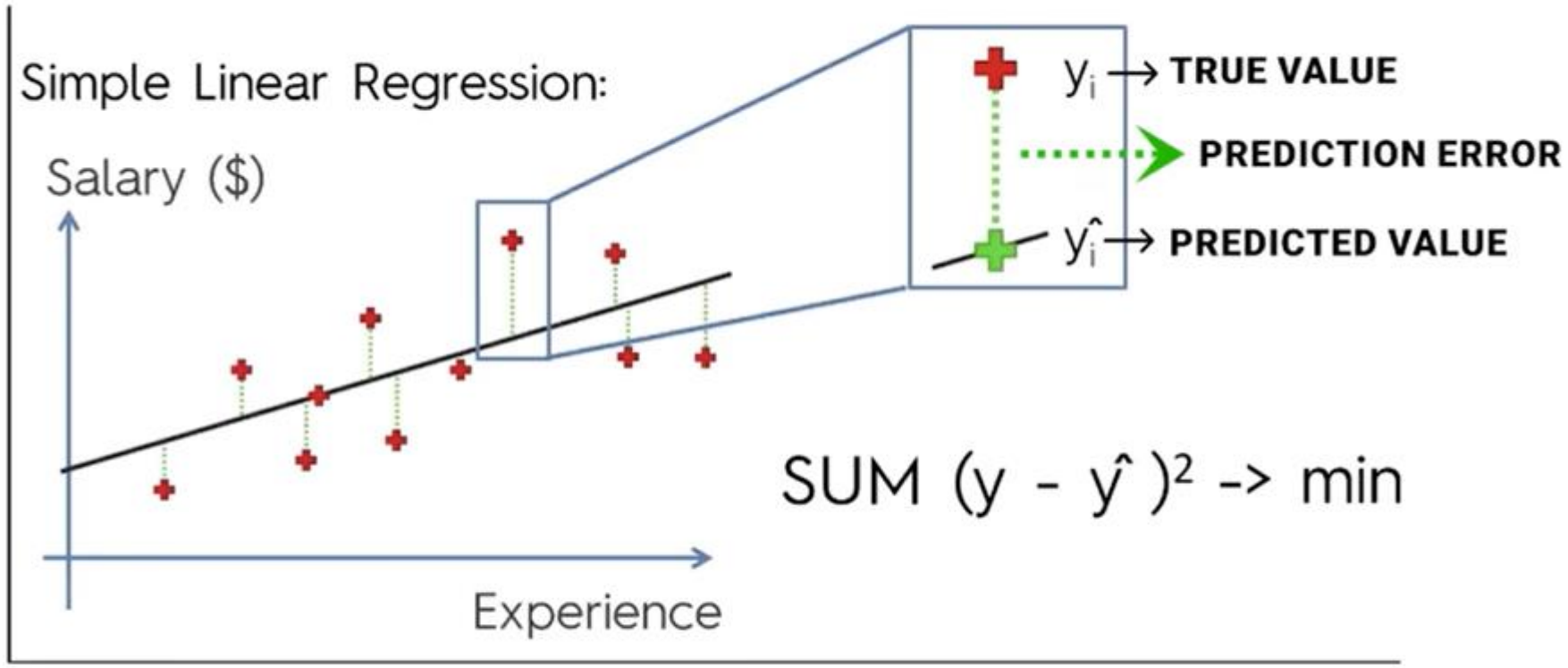
LINEAR REGRESSION & ITS TYPES

SIMPLE LINEAR REGRESSION

If there is only one input variable (x), then such linear regression is called **simple linear regression.**

AGE	SALARY


INPUT, FEATURE,
INDEPENDENT VARIABLE
OUTPUT, TARGET,
DEPENDENT VARIABLE
VARIABLE 1
Function
VARIABLE 2

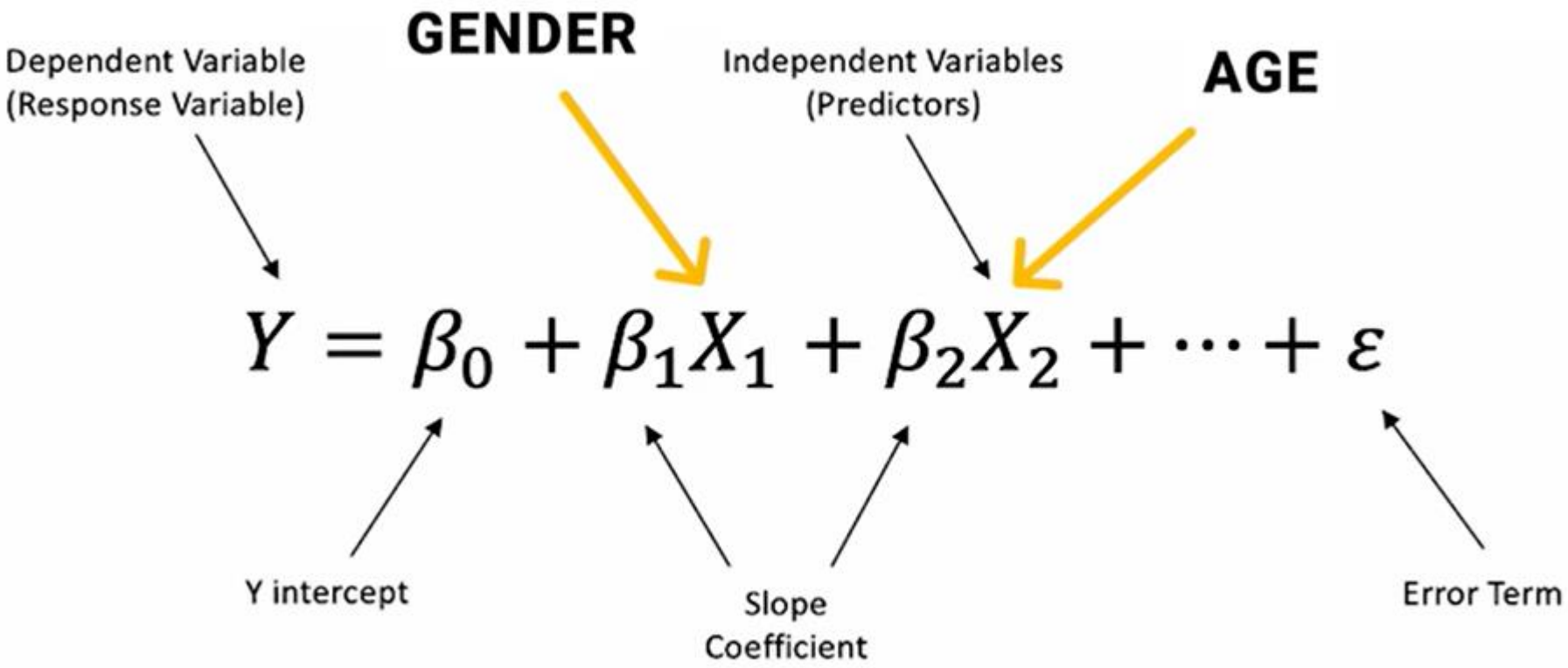

Simple Linear Regression:
Salary ($)
Experience
y_i → TRUE VALUE
PREDICTION ERROR
$\hat{y}_i$ → PREDICTED VALUE
SUM $(y - \hat{y})^2$ -> min

LINEAR REGRESSION & ITS TYPES

MULTIPLE LINEAR REGRESSION

if there is more than one input variable, then such linear regression is called **multiple linear regression**.

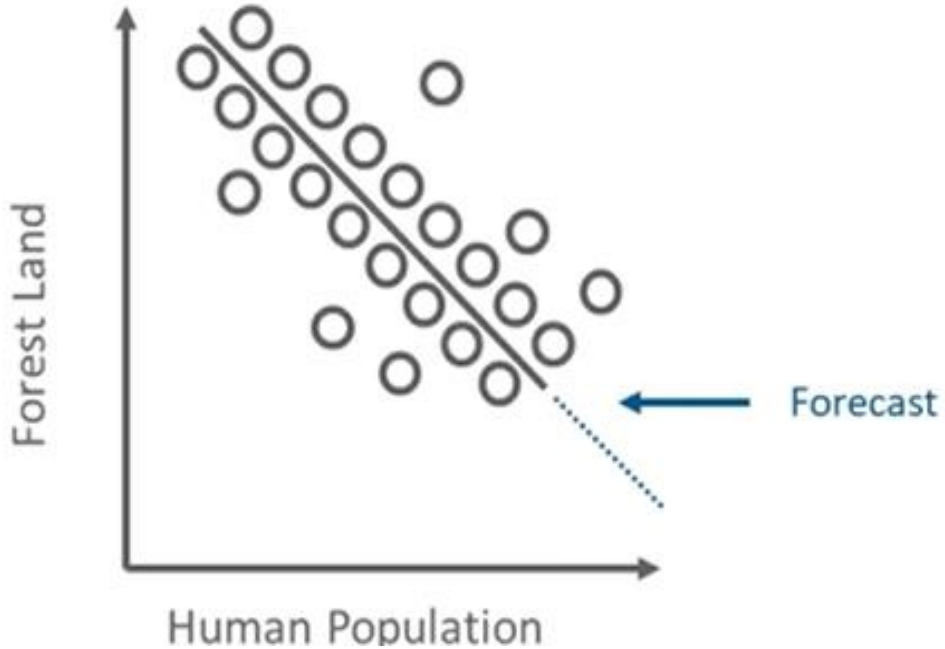
AGE	WEIGHT	HEART DISEASE


GENDER
Dependent Variable
(Response Variable)
Independent Variables
(Predictors)
AGE
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$
Y intercept
Slope
Coefficient
Error Term

Model Fitting

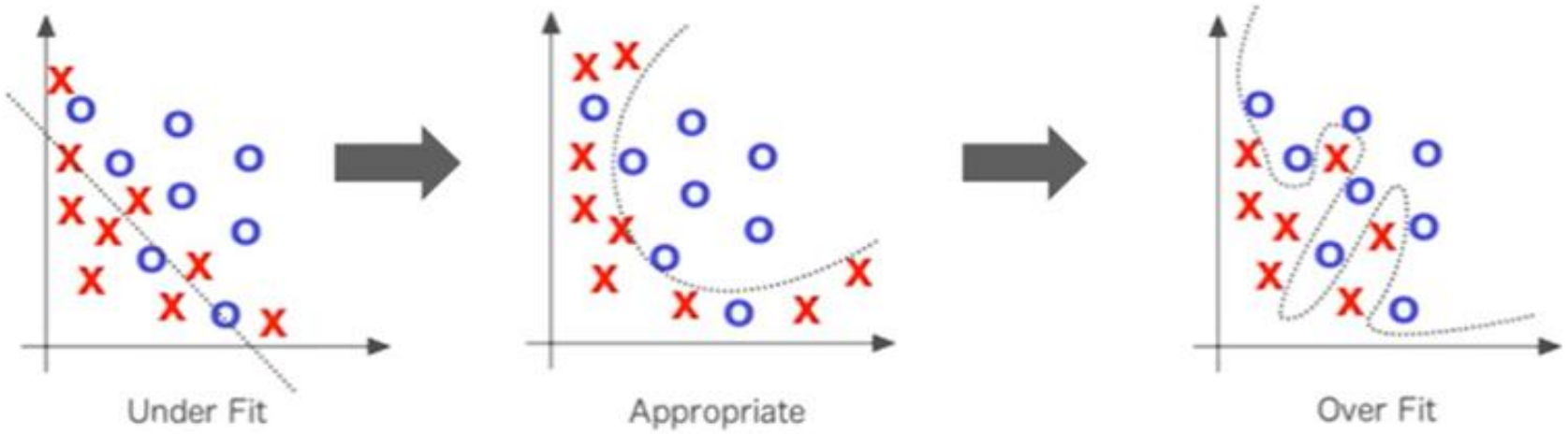
Fitting a model means that you're making your algorithm learn the relationship between predictors and outcome so that you can predict the future values of the outcome .

So the best fitted model has a specific set of parameters which best defines the problem at hand



Types of Fitting

Machine Learning algorithms first attempt to solve the problem of under-fitting; that is, of taking a line that does not approximate the data well, and making it to approximate the data better.

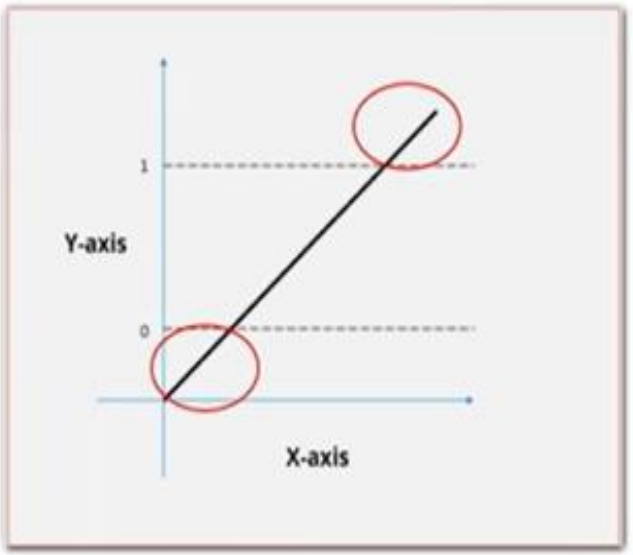


LOGISTIC REGRESSION

Need For Logistic Regression

WHO WILL WIN ?

Here, the best fit line in linear regression is going below 0 and above 1



What is Logistic Regression?

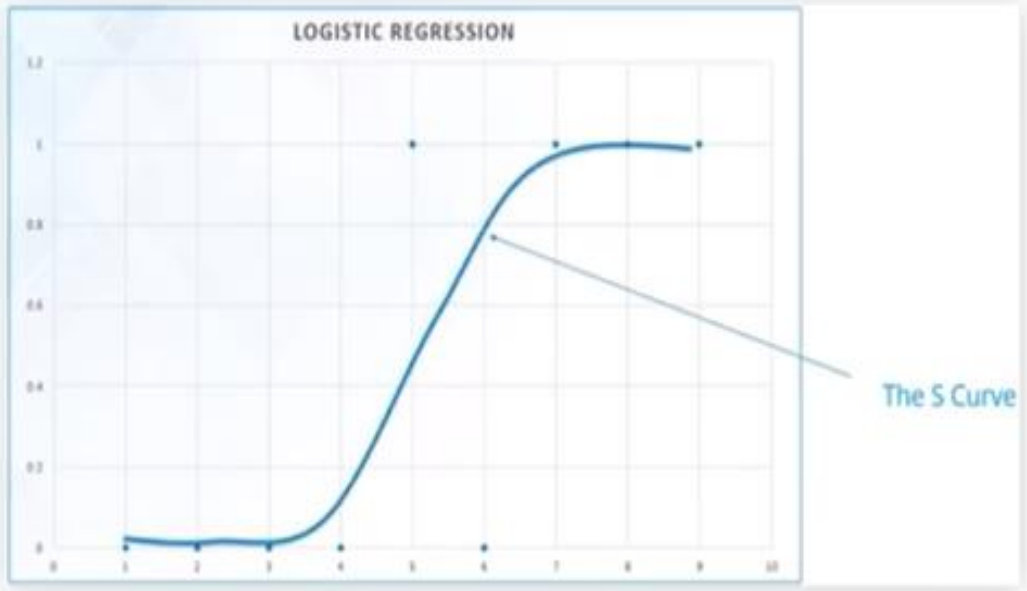
Logistic Regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome.
Outcome is a binary class type.

The outcome(result) will be binary(0/1)

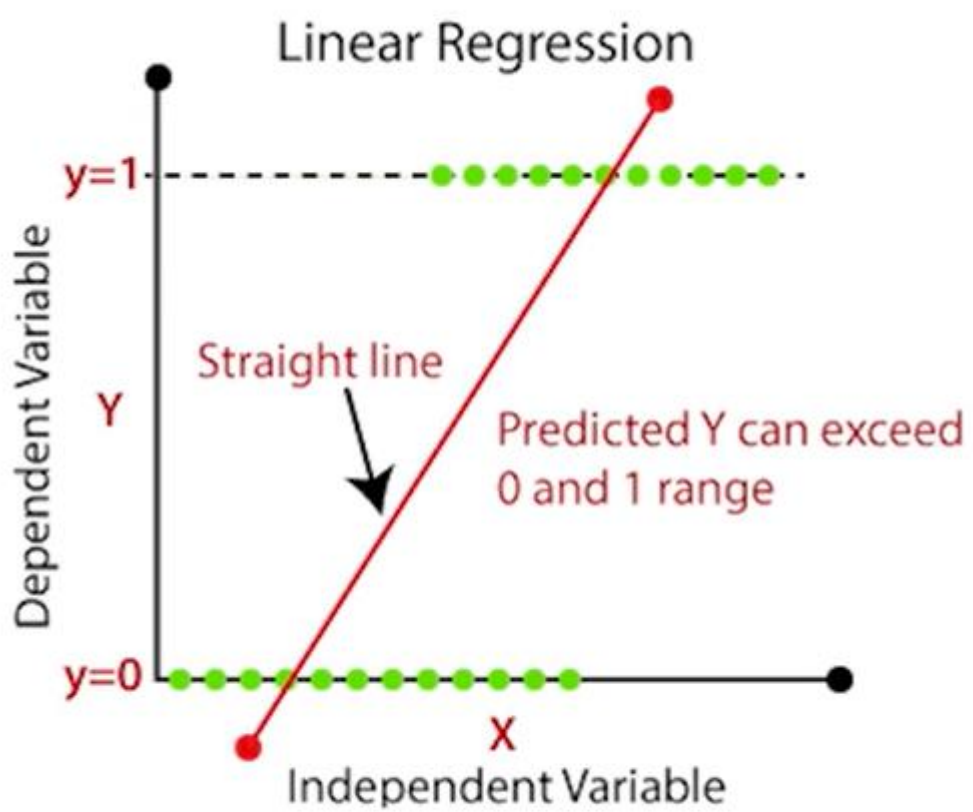
0- If malignant
1- If benign

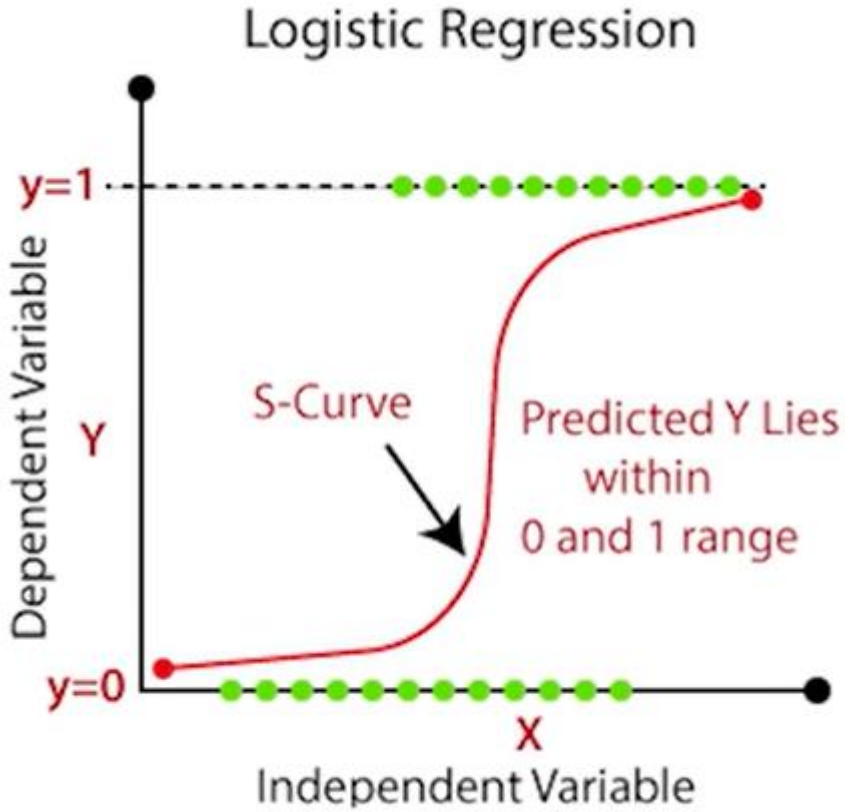
What is Logistic Regression?

The Logistic Regression Curve is called as "Sigmoid Curve", also known as S-Curve



Based on the threshold value set, we decide the output from the function

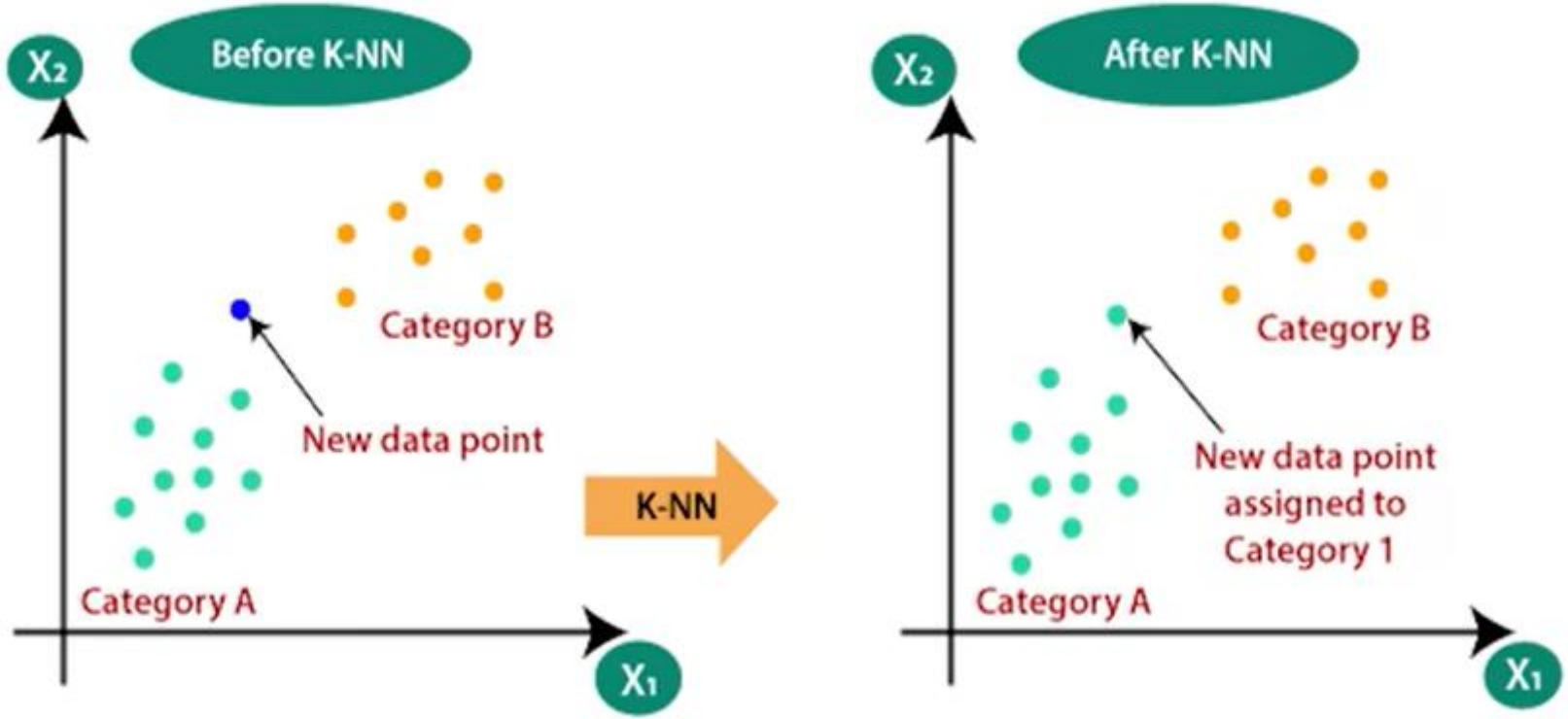




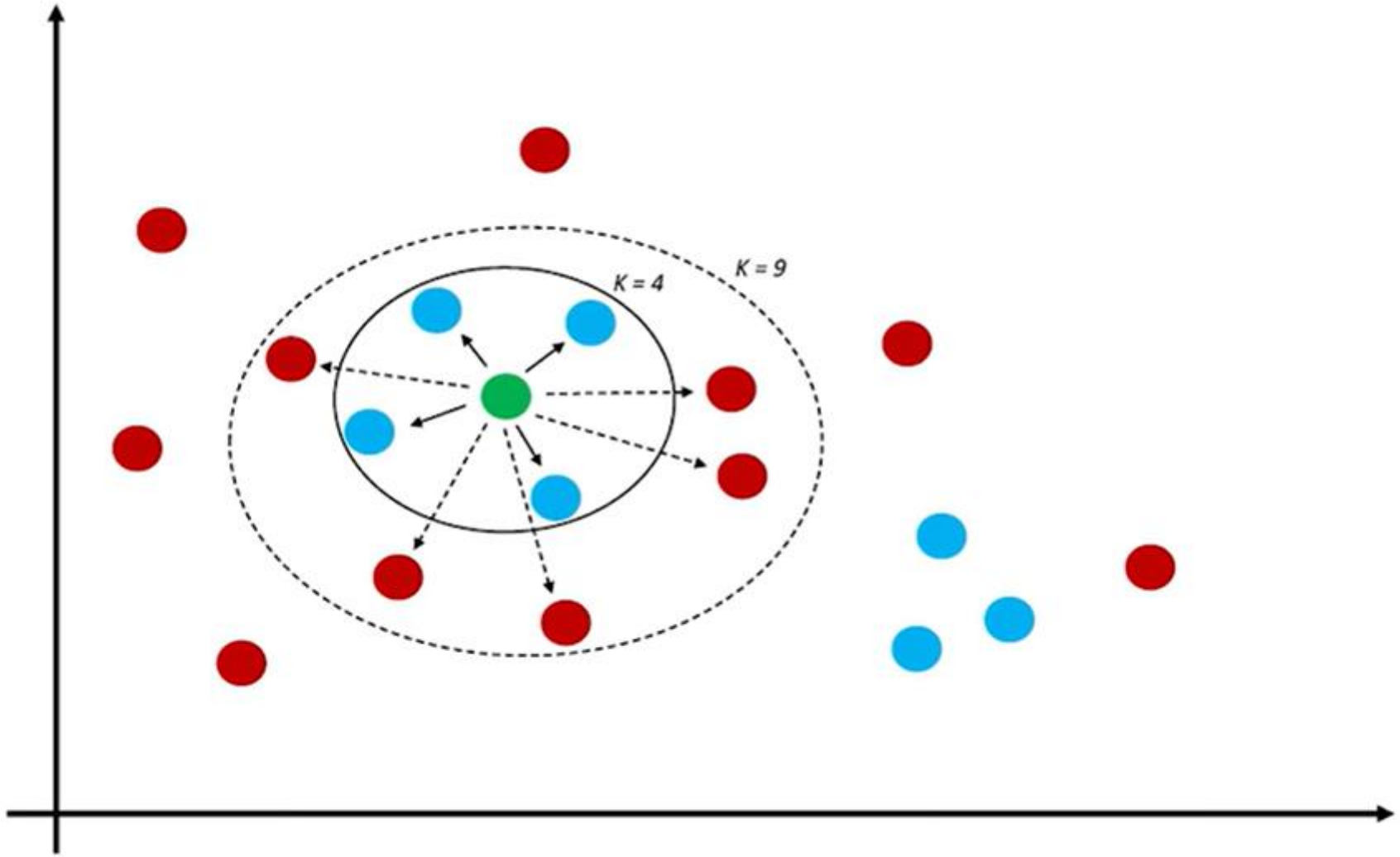
SIGMOID FUNCTION

K NEAREST NEIGHBORS ALGORITHM (KNN)

K NEAREST NEIGHBORS (KNN)



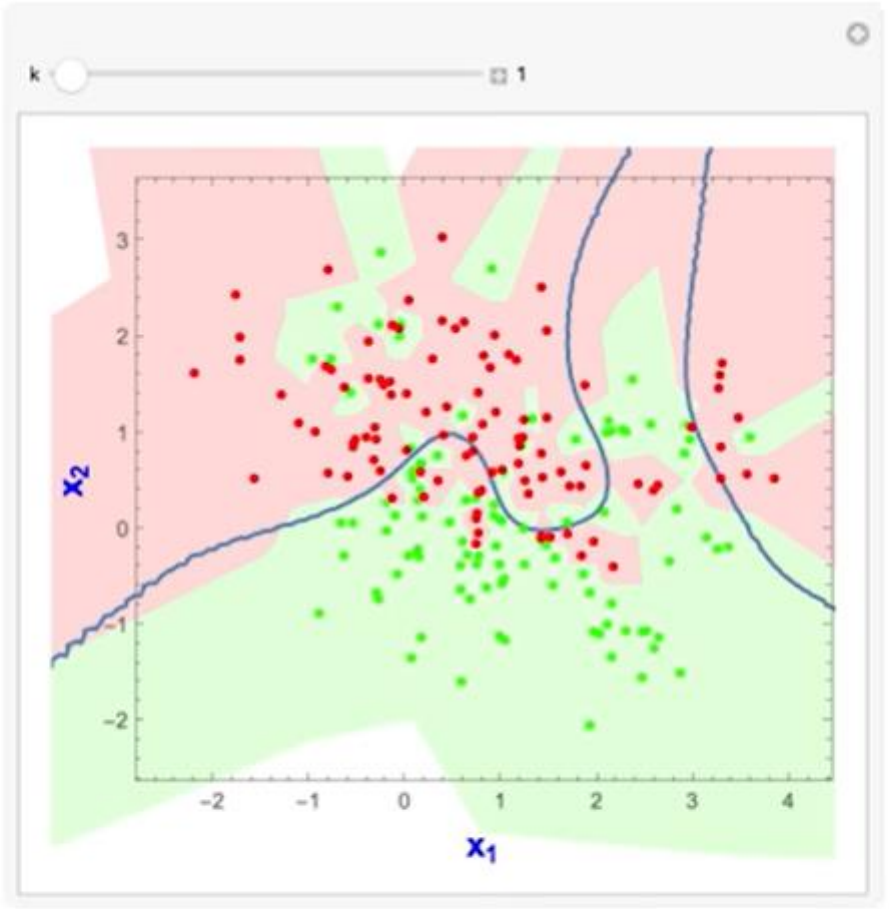
NON-PARAMETRIC ALGORITHM


K = 4
K = 9

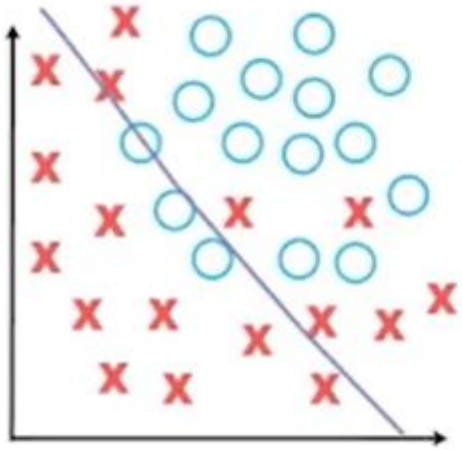
K NEAREST NEIGHBORS (KNN)

REGRESSION

CLASSIFICATION



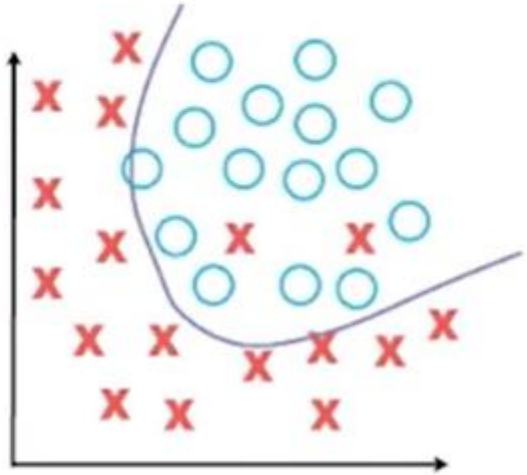
COMPLICATED NON-LINEAR DECISION BOUNDARY



Under-Fitting

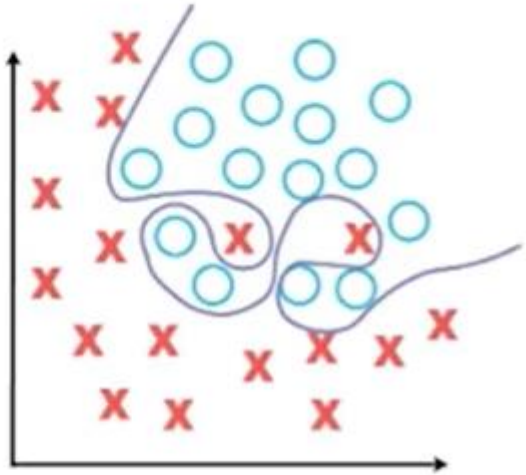
(too simple to explain the variance)

K=1000



Appropriate-Fitting

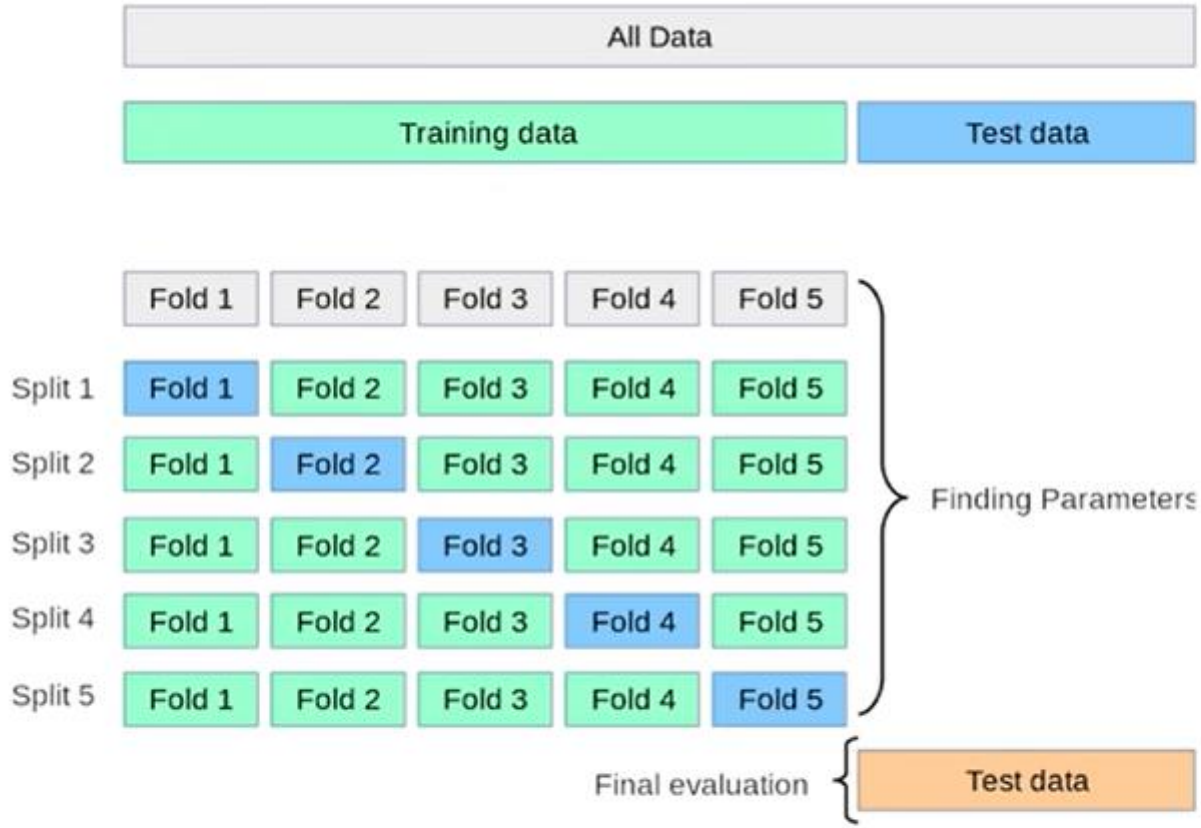
K=5



Over-Fitting

(force-fitting – too good to be true)

K=1



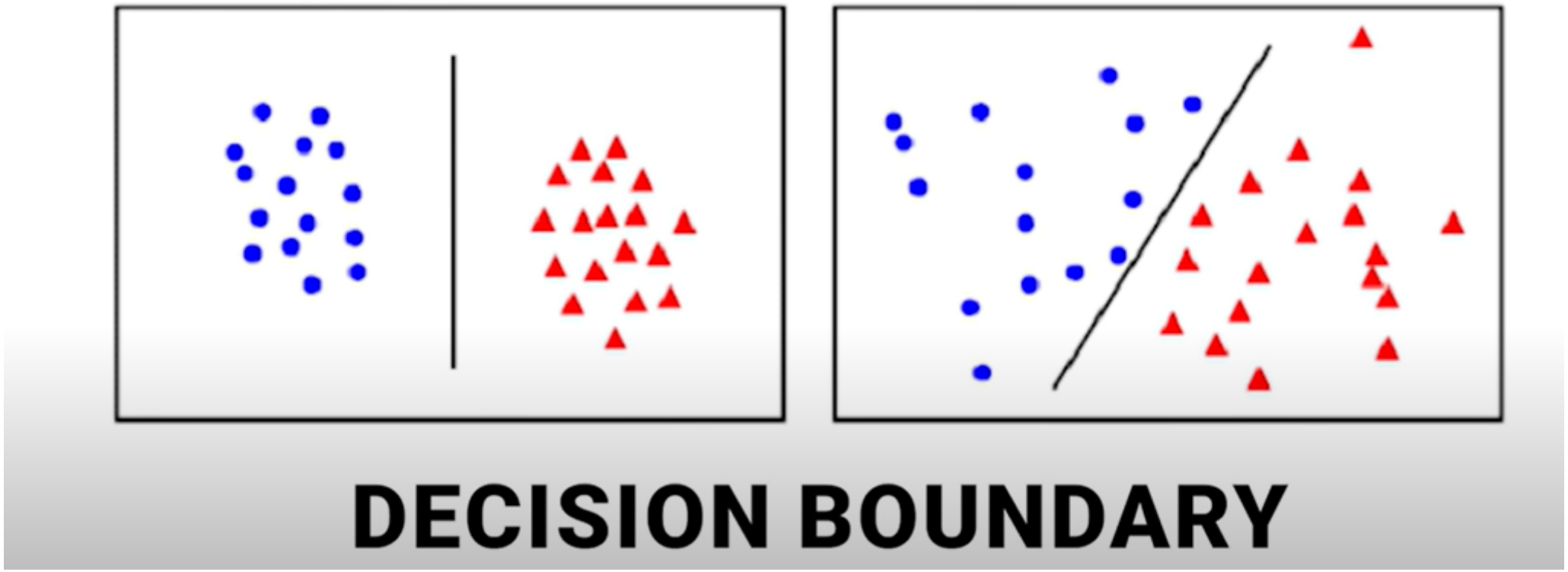
CROSS VALIDATION

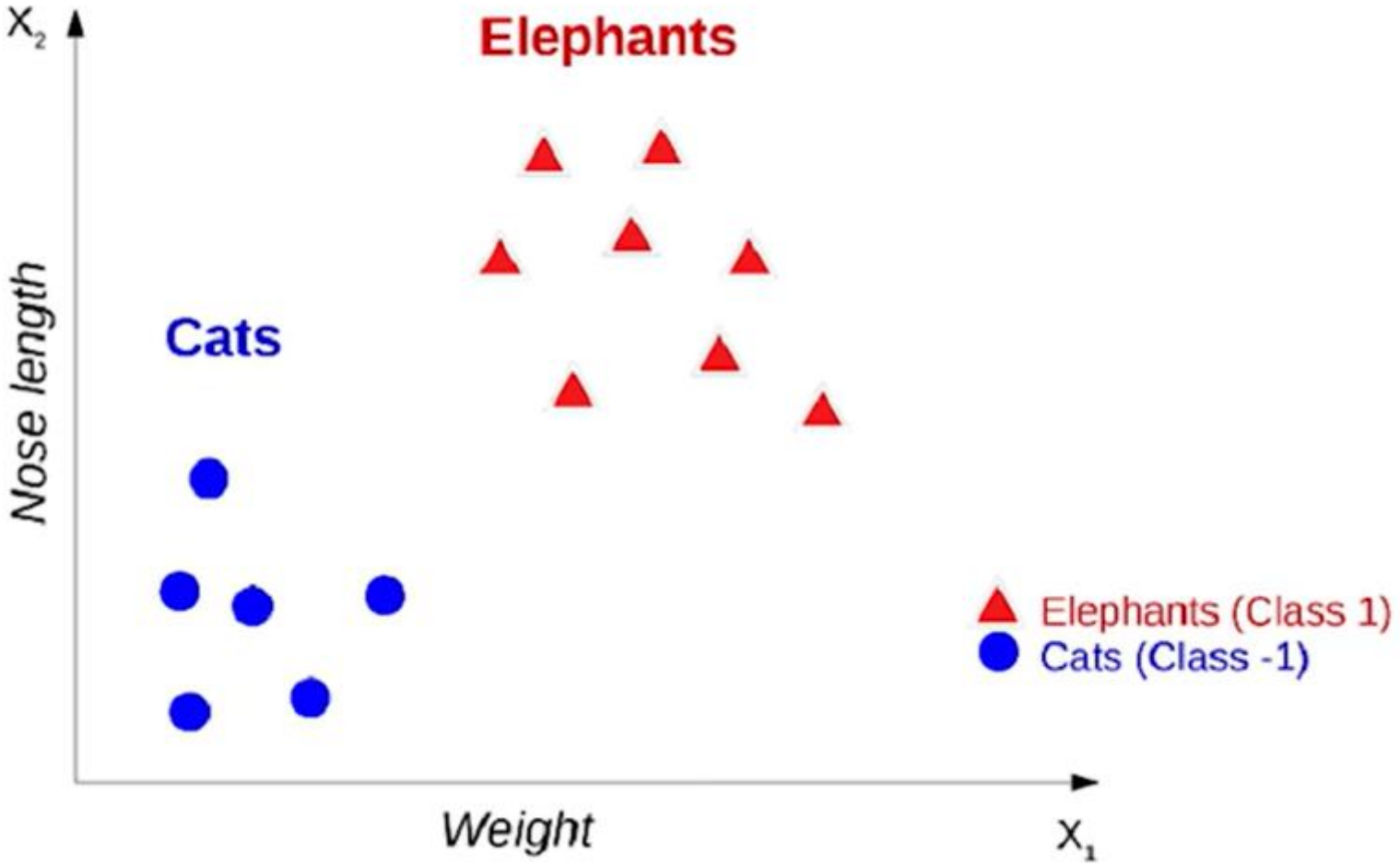
SUPPORT VECTOR MACHINE (SVM)

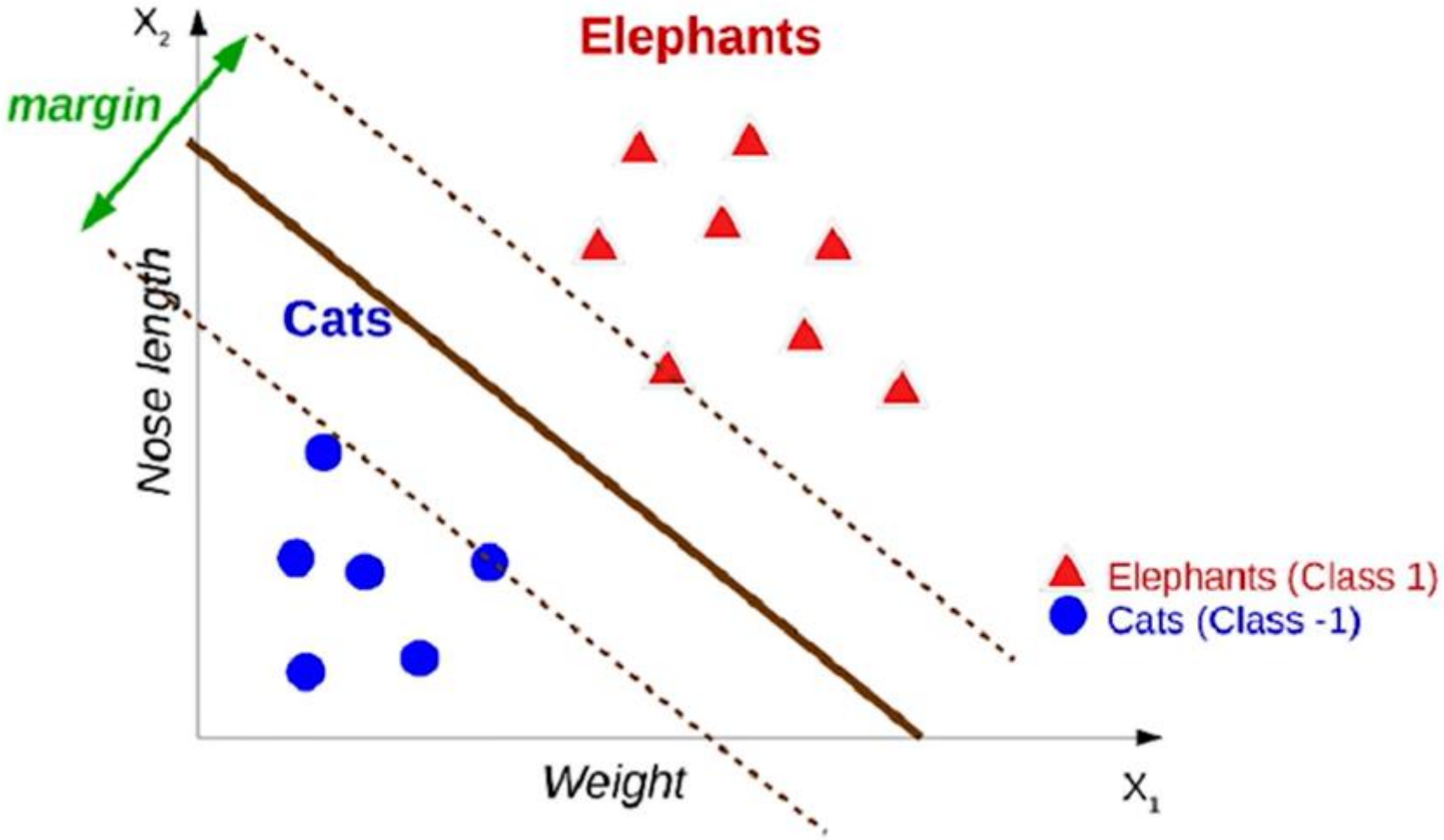
Kmit

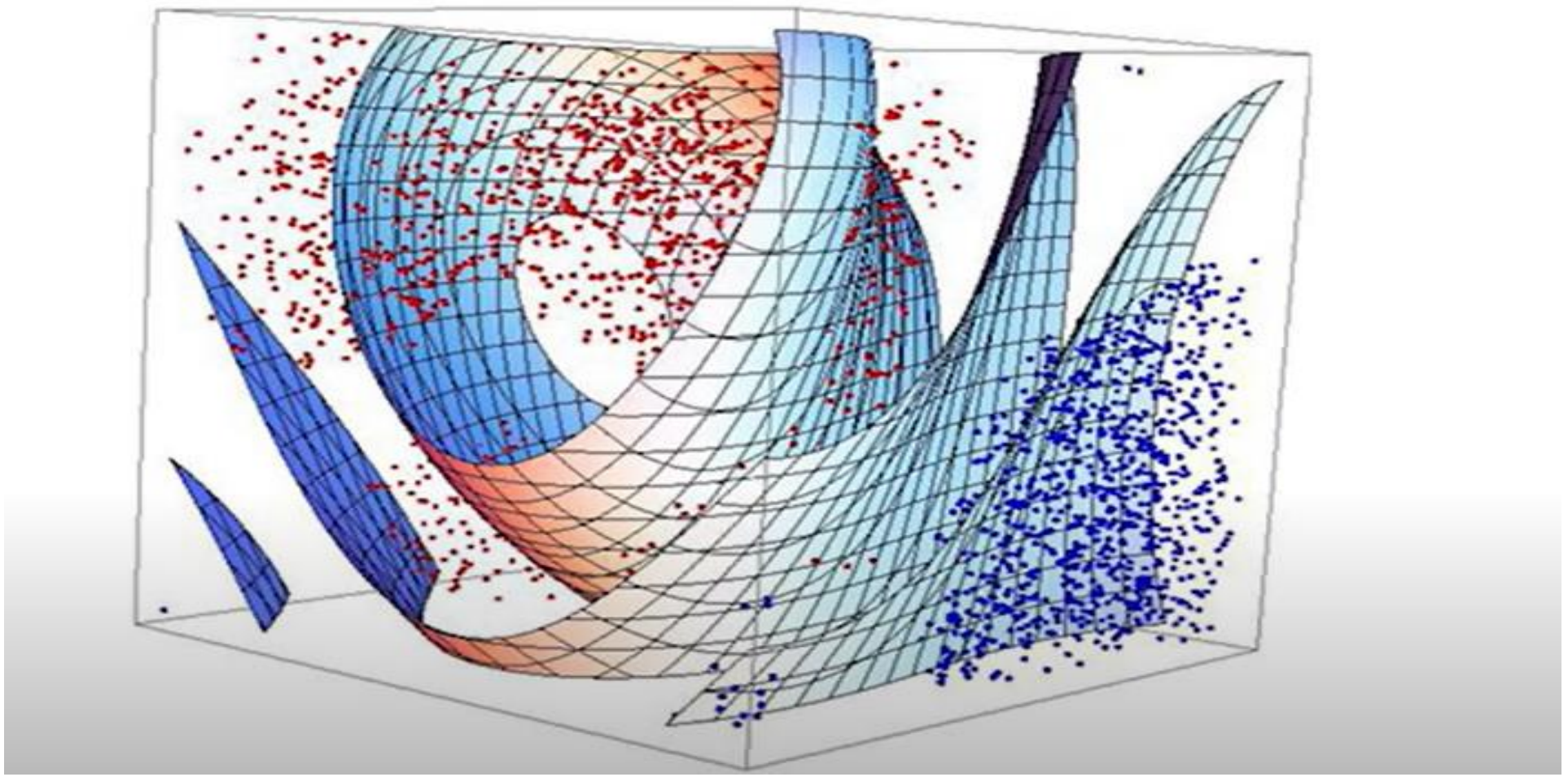
CLASSIFICATION

REGRESSION

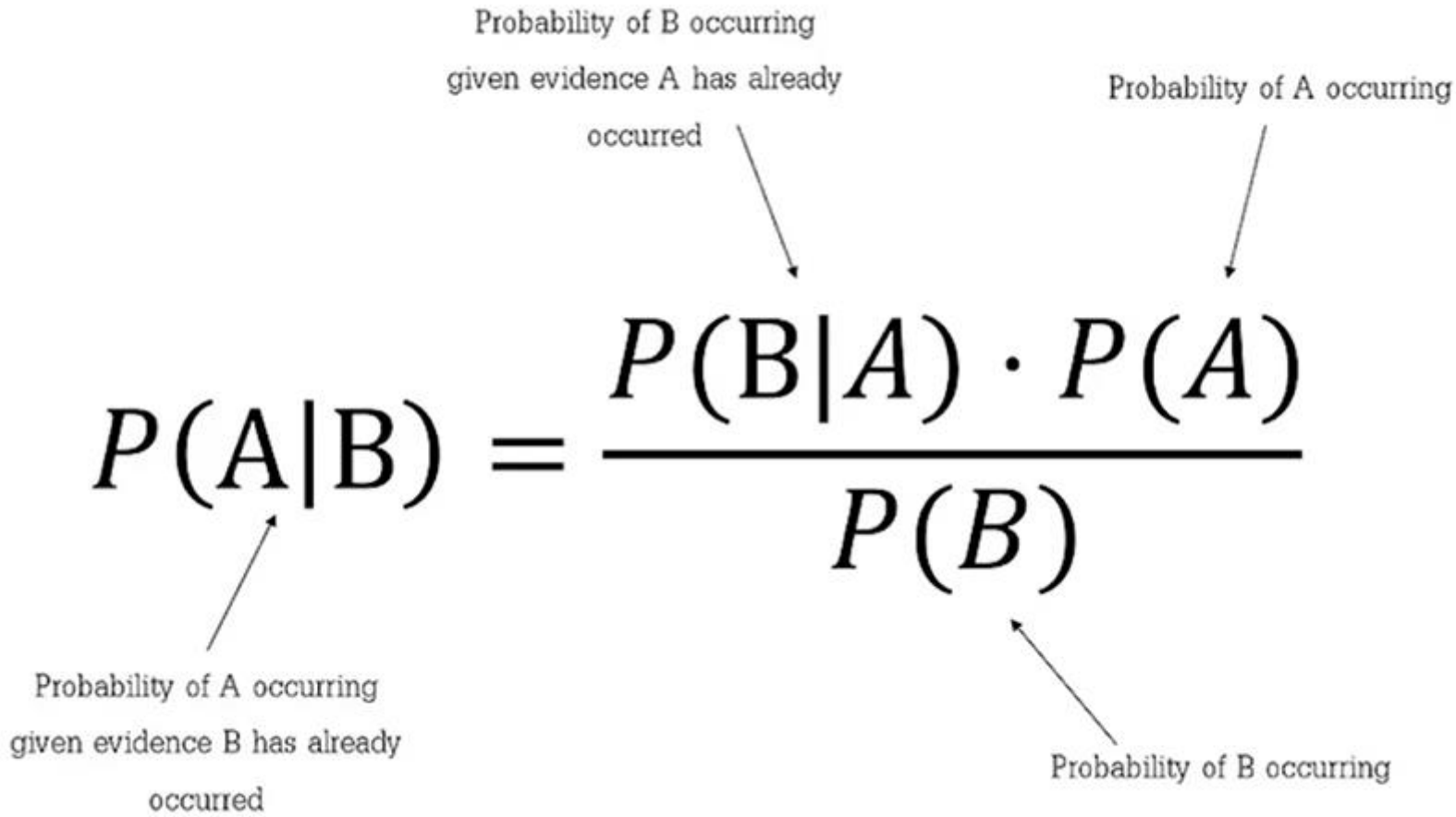

DECISION BOUNDARY

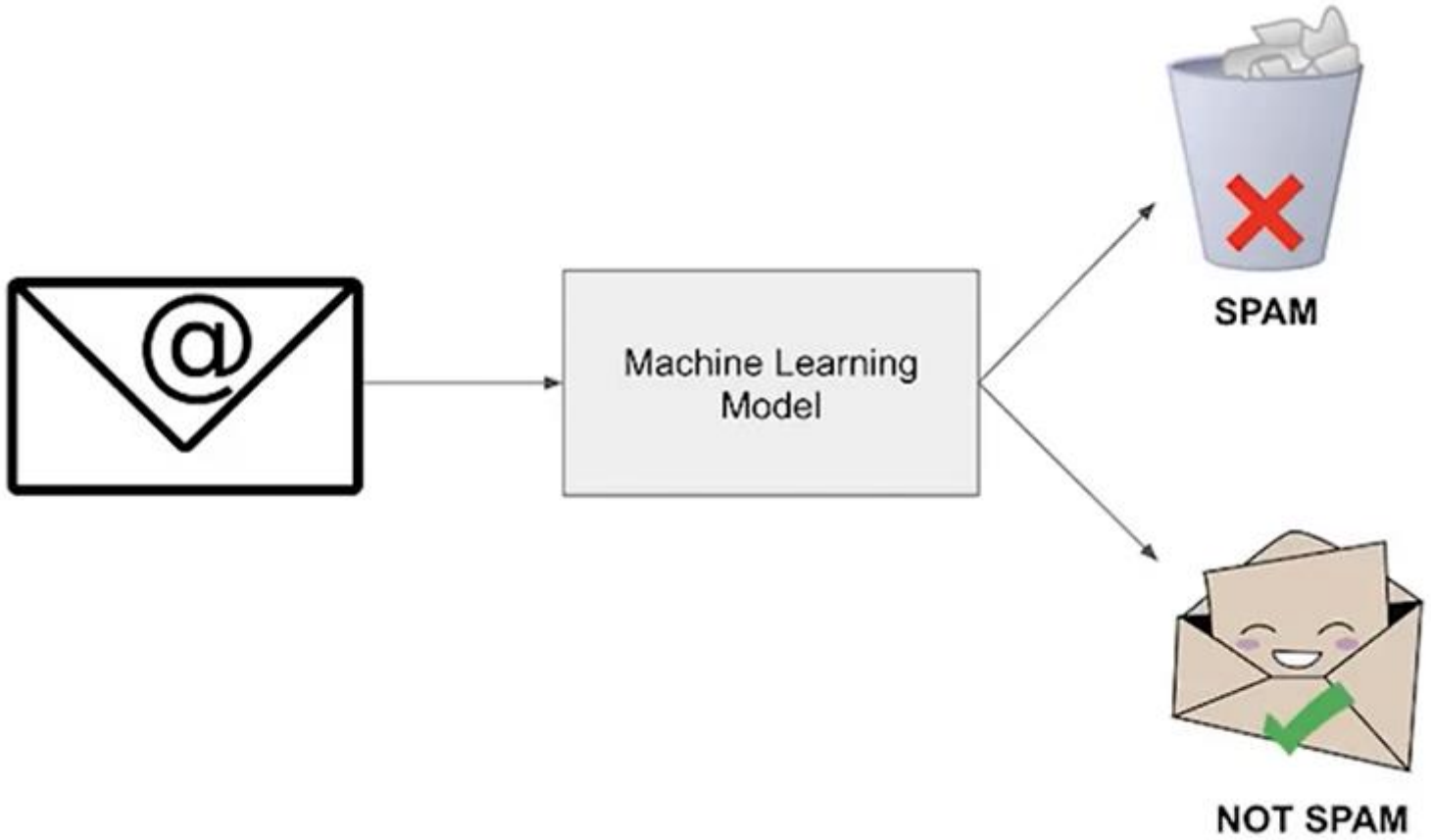

Elephants
Cats
X_2
Nose length
Weight
X_1
Elephants (Class 1)
Cats (Class -1)


X2
margin
Elephants
Cats
Nose length
Weight
X1
Elephants (Class 1)
Cats (Class -1)



NAIVE BAYES CLASSIFIER


Probability of B occurring given evidence A has already occurred
Probability of A occurring
$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$
Probability of A occurring given evidence B has already occurred
Probability of B occurring


Machine Learning
Model
SPAM
NOT SPAM

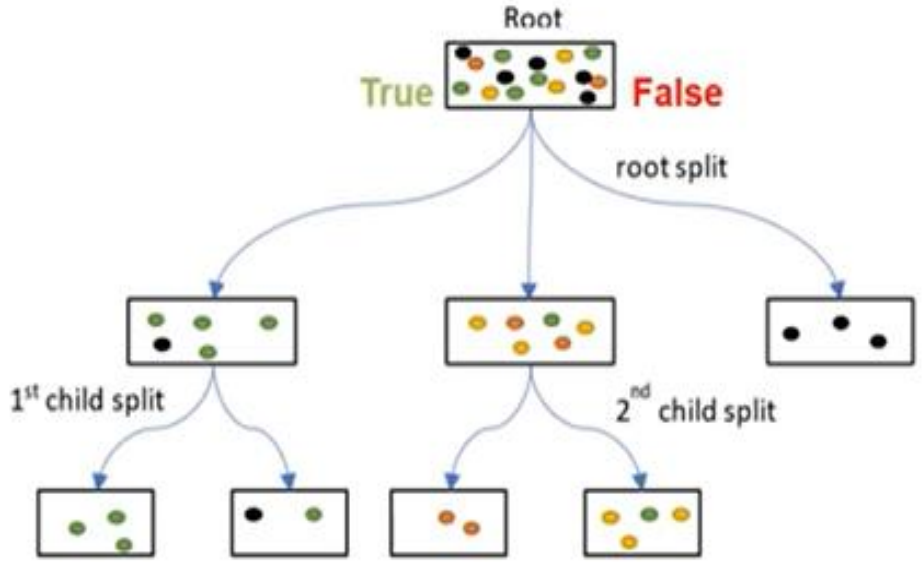
	Label	SMS
0	spam	SECRET PRIZE! CLAIM SECRET PRIZE NOW!!
1	ham	Coming to my secret party?
2	spam	Winner! Claim secret prize now!

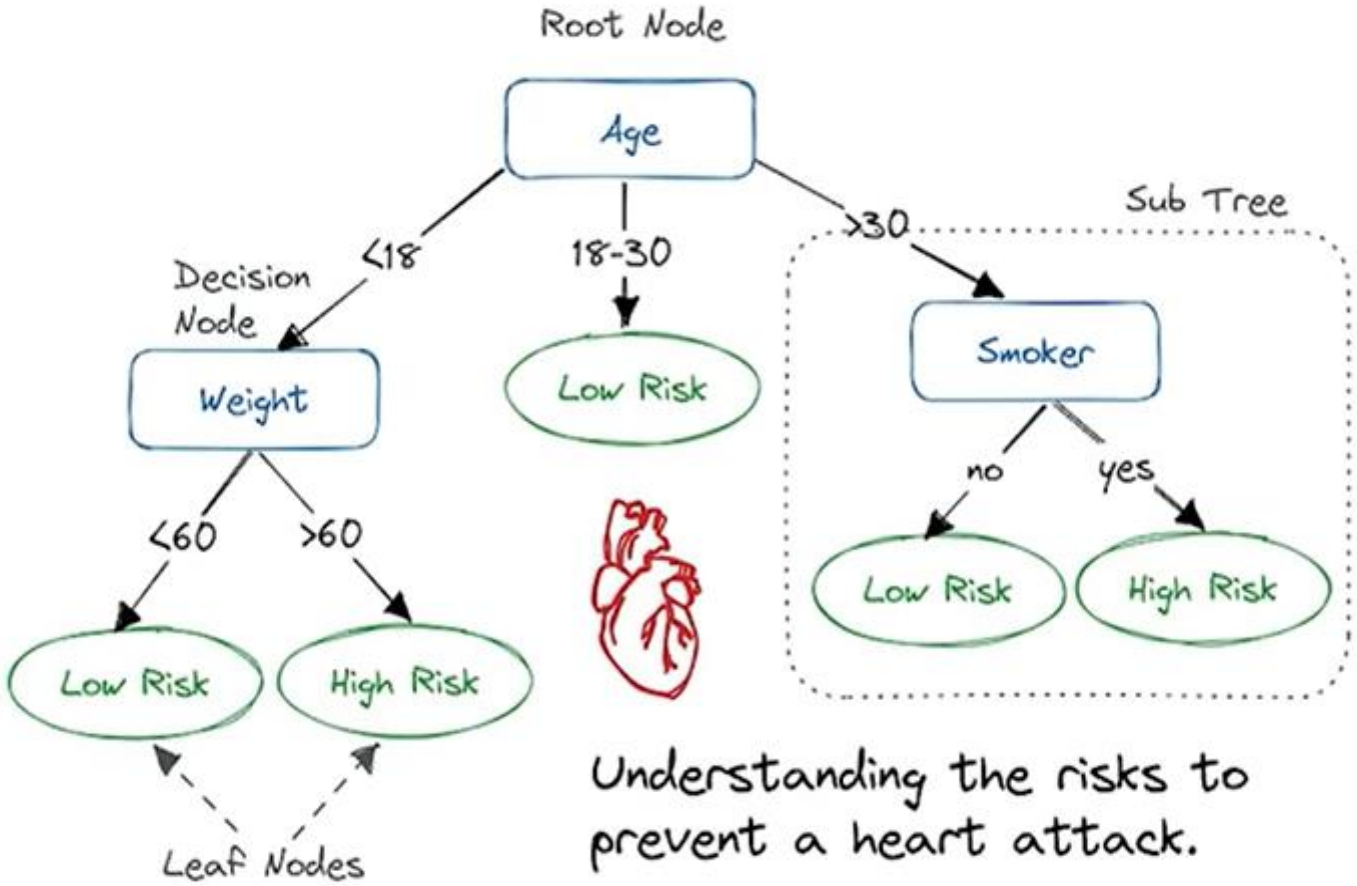
	Label	secret	prize	claim	now	coming	to	my	party	winner
0	spam	2	2	1	1	0	0	0	0	0
1	ham	1	0	0	0	1	1	1	1	0
2	spam	1	1	1	1	0	0	0	0	1

What is a Decision Tree?

A decision tree is a tree-like structure in which internal node represents test on an attribute

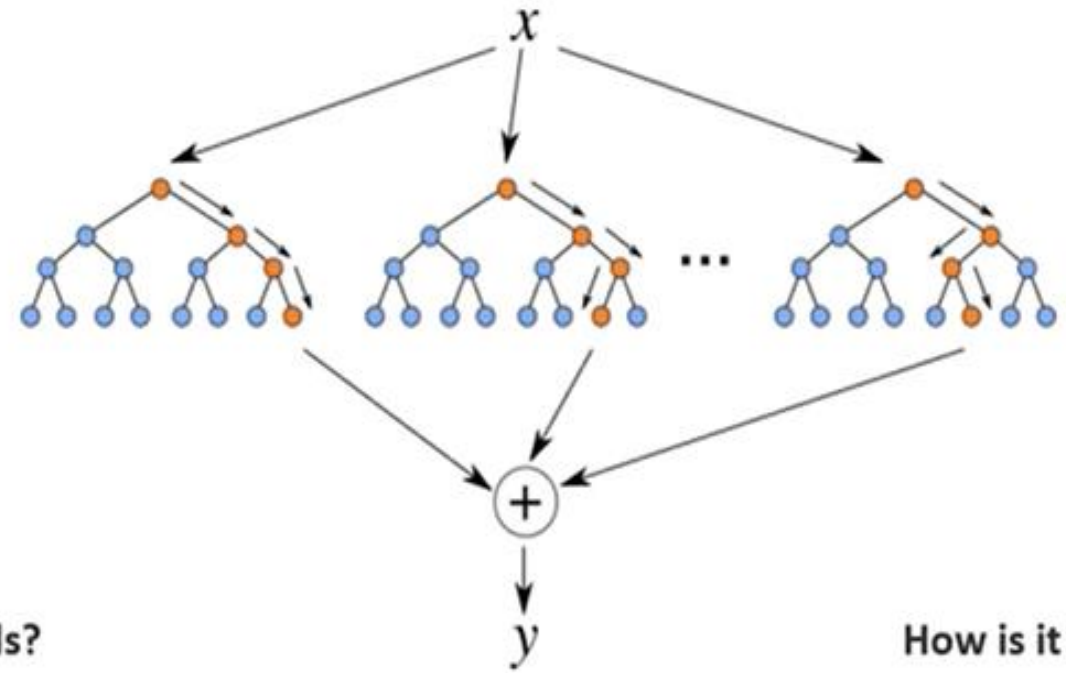
- Each branch represents outcome of test and each leaf node represents class label (decision taken after computing all attributes)
- A path from root to leaf represents classification rules.




Root Node
Age
<18
18-30
>30
Sub Tree
Decision Node
Weight
Low Risk
Smoker
<60
>60
no
yes
Low Risk
High Risk
Low Risk
High Risk
Leaf Nodes
Understanding the risks to prevent a heart attack.

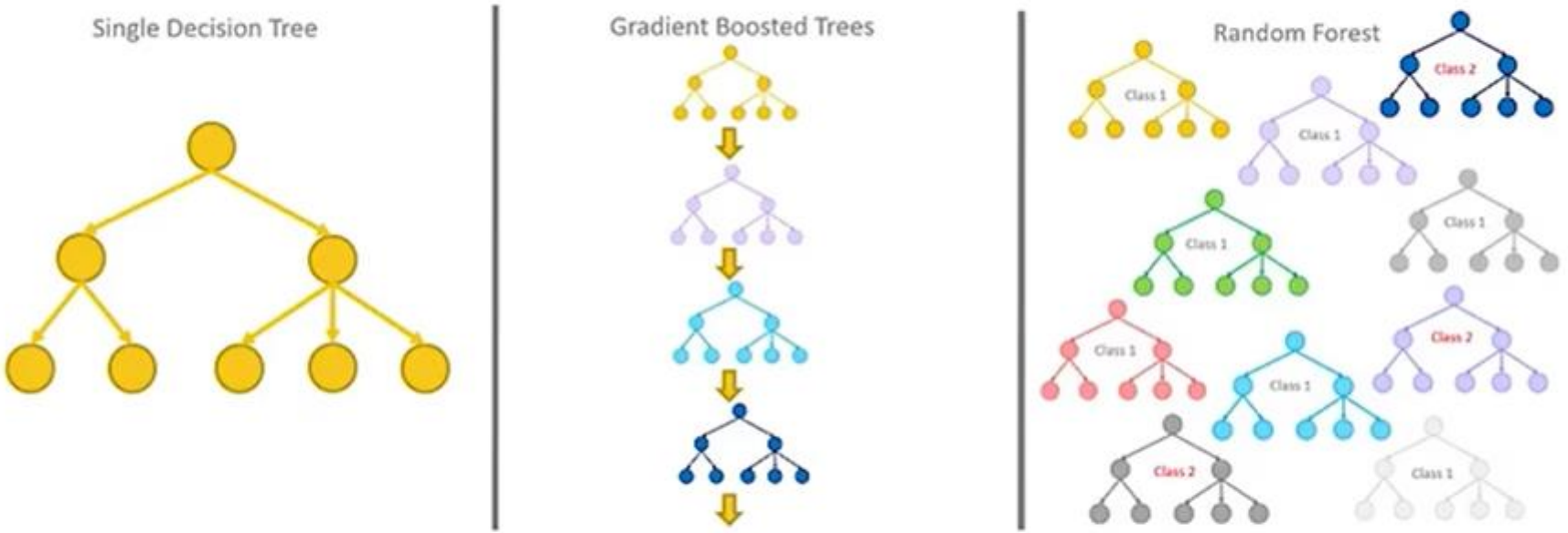
What is Random Forest?

Random Forest is an ensemble classifier made using many Decision tree models



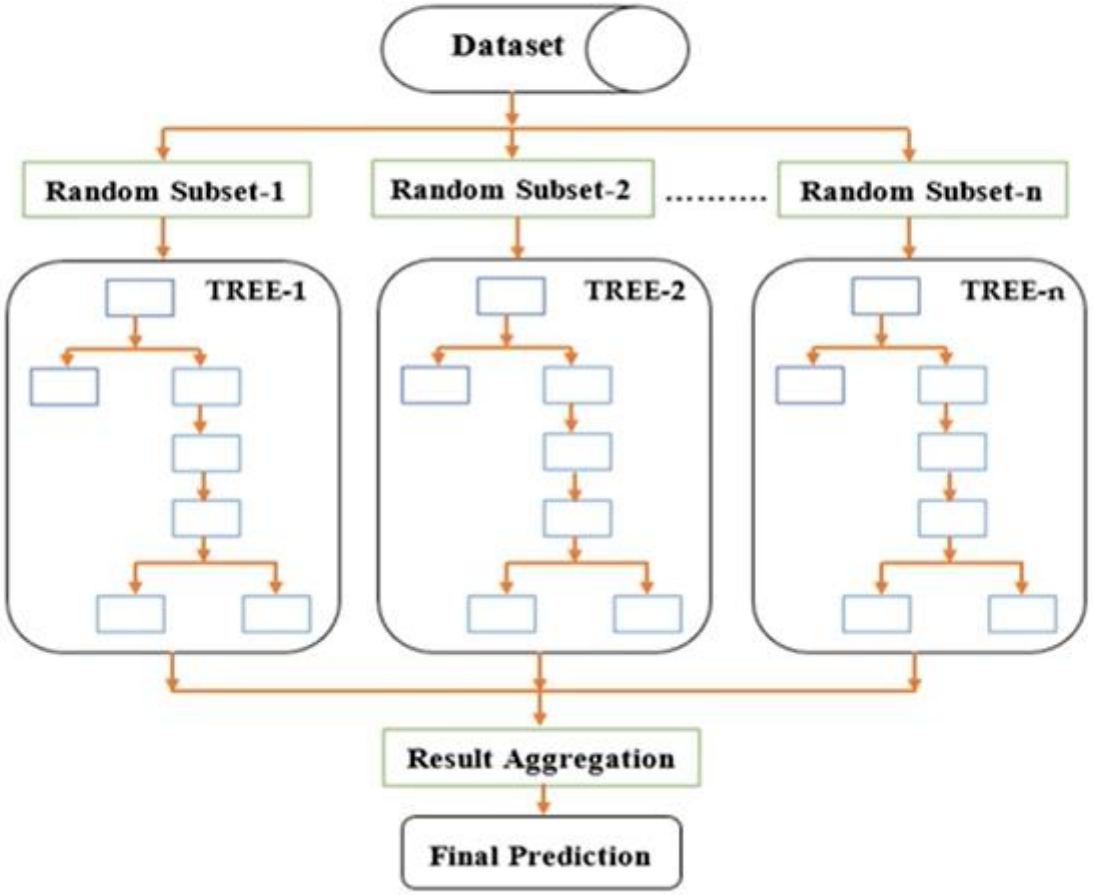
What are Ensemble models?

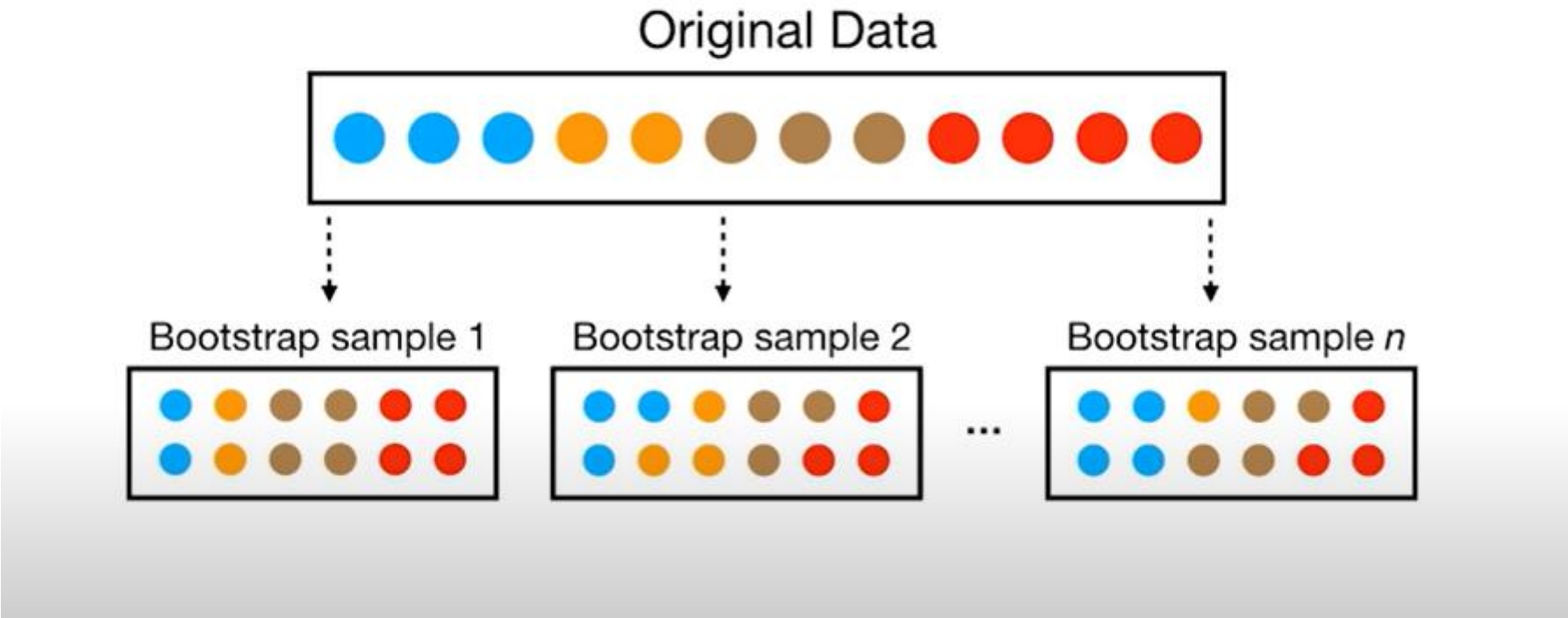
How is it better from Decision Trees ?

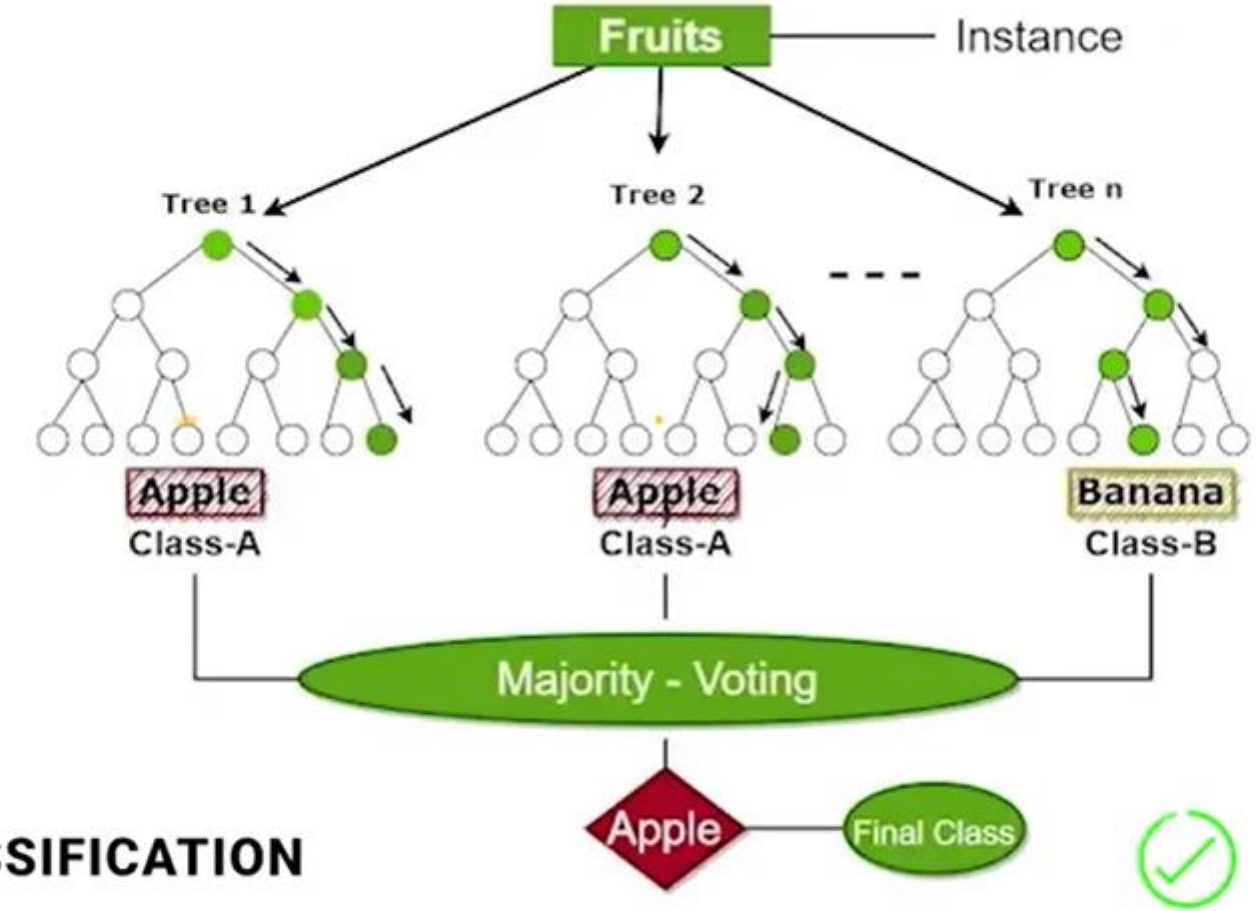


ENSEMBLE ALGORITHM

BAGGING


Dataset
Random Subset-1
Random Subset-2
.....
Random Subset-n
TREE-1
TREE-2
TREE-n
Result Aggregation
Final Prediction

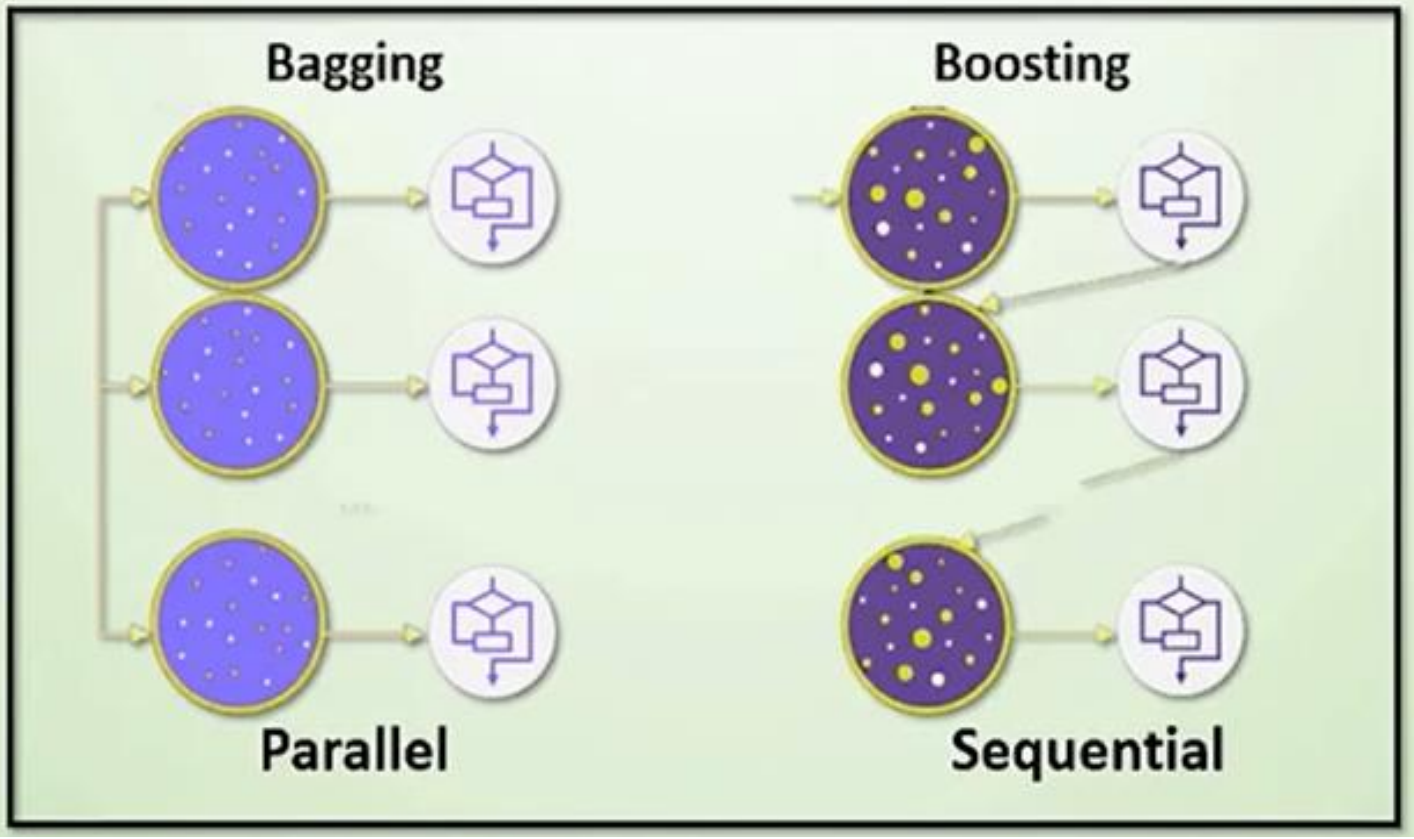

Original Data
Bootstrap sample 1
Bootstrap sample 2
...
Bootstrap sample n



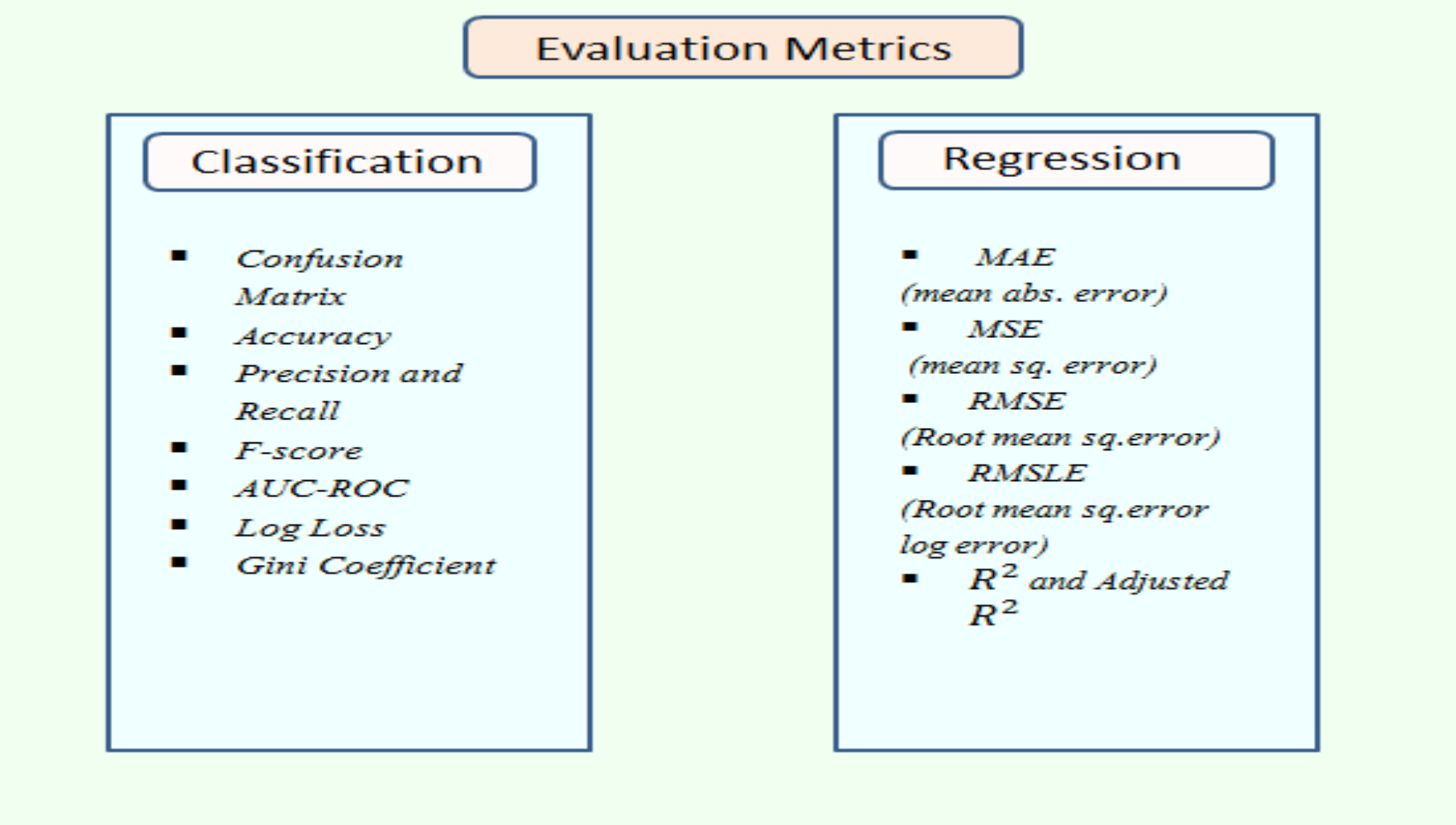
CLASSIFICATION

REGRESSION

Bagging and Boosting

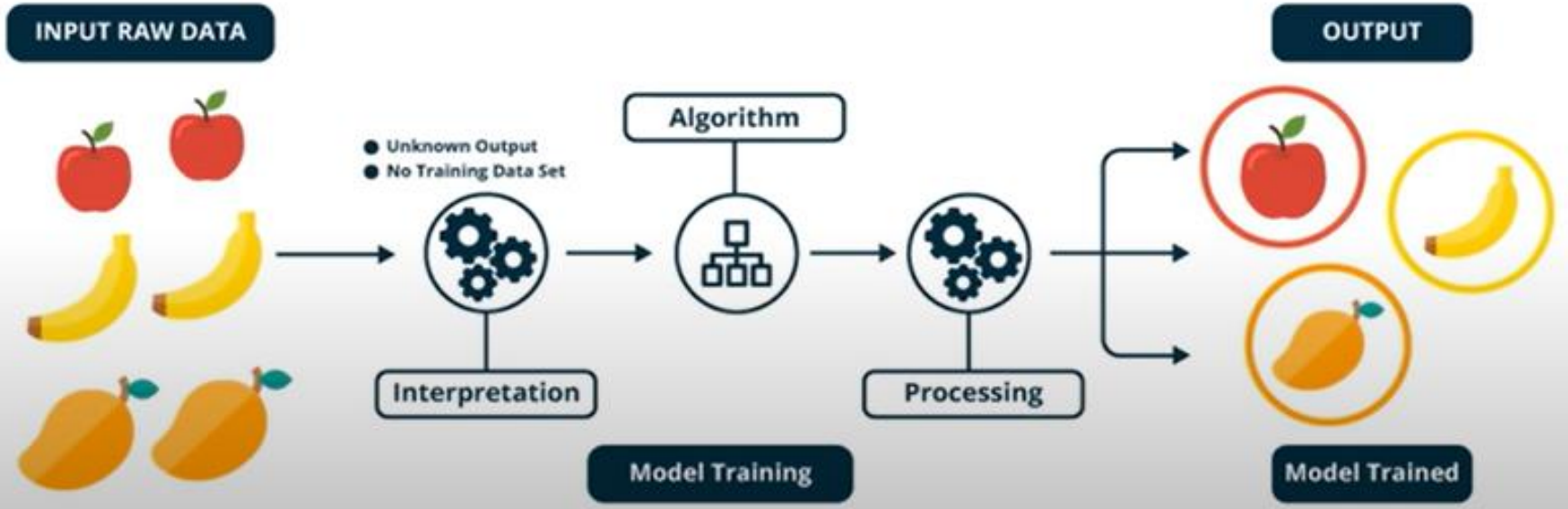


EVALUATION METRICS REGRESSION AND CLASSIFICATION



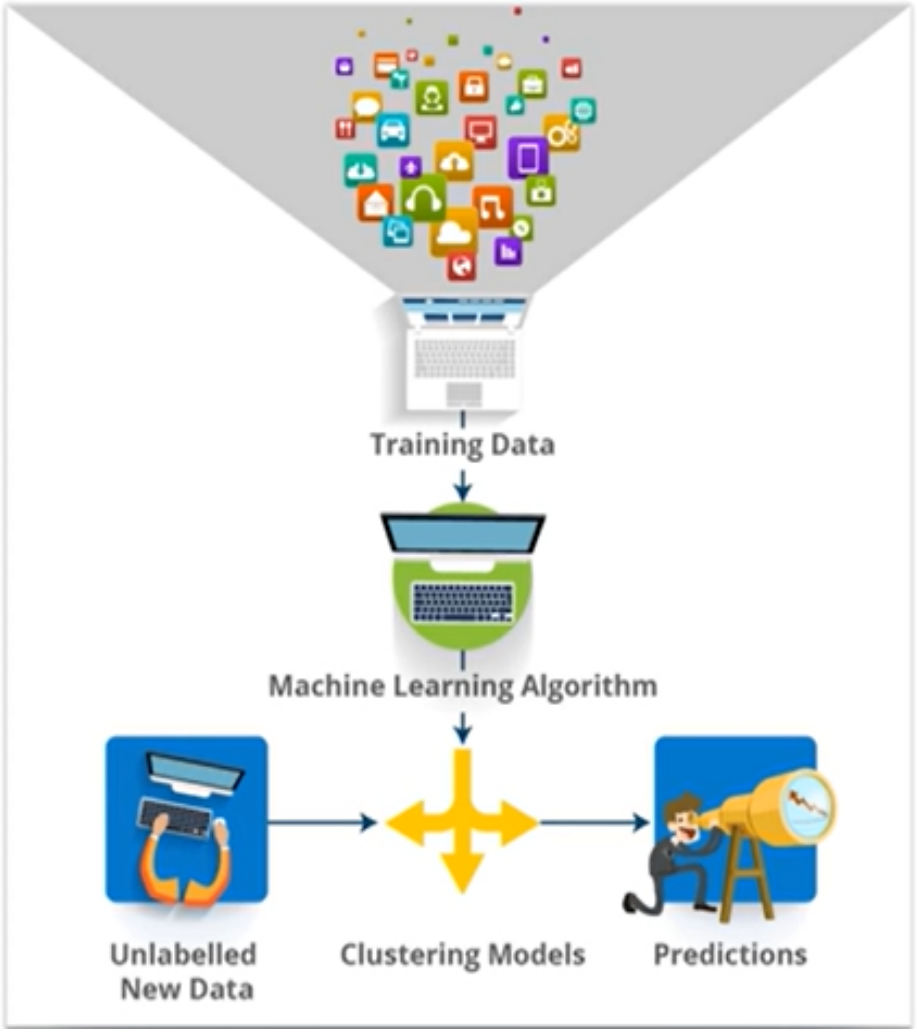
Unsupervised Learning

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data **Without labelled responses**



Unsupervised Learning: Process Flow

Training data is collection of information without any label



What is Clustering?

"Clustering is the process of dividing the datasets into groups, consisting of similar data-points"

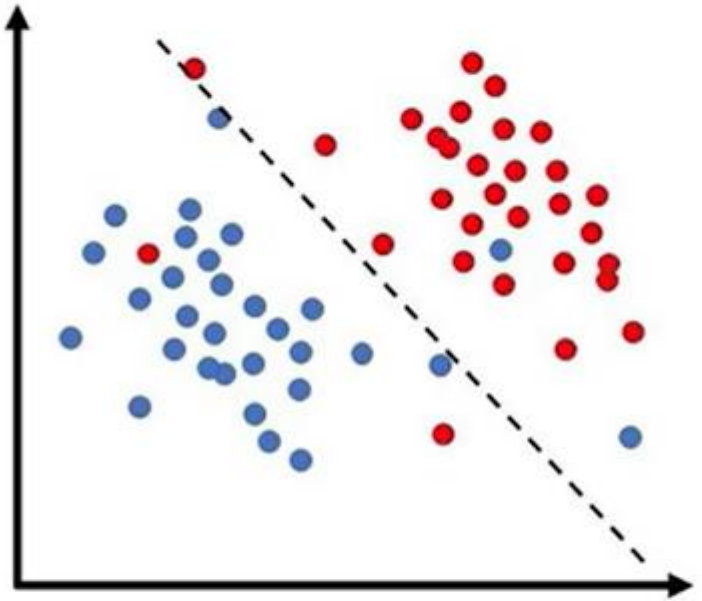
It means grouping of objects based on the information found in the data, describing the objects or their relationship

Why is Clustering Used?

The goal of clustering is to determine the intrinsic grouping in a set of **Unlabelled Data**

Sometimes, Partitioning is the goal

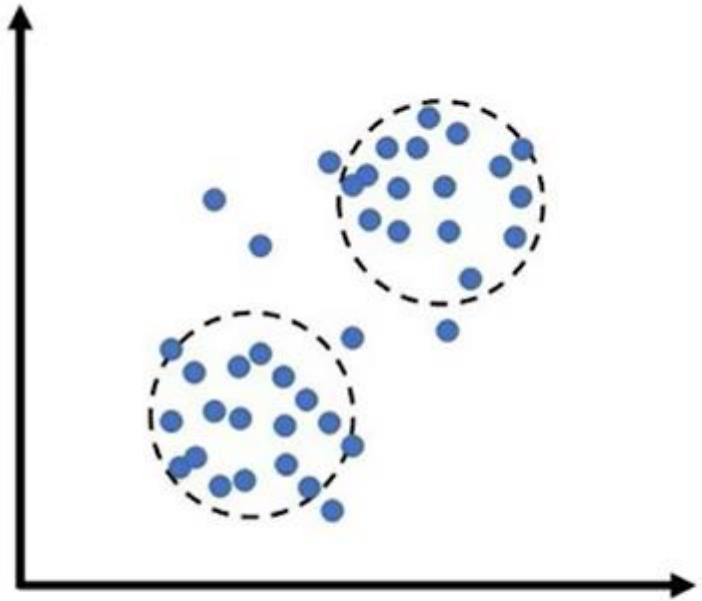
CLASSIFICATION



Supervised Learning
(a)

CLASSES KNOWN

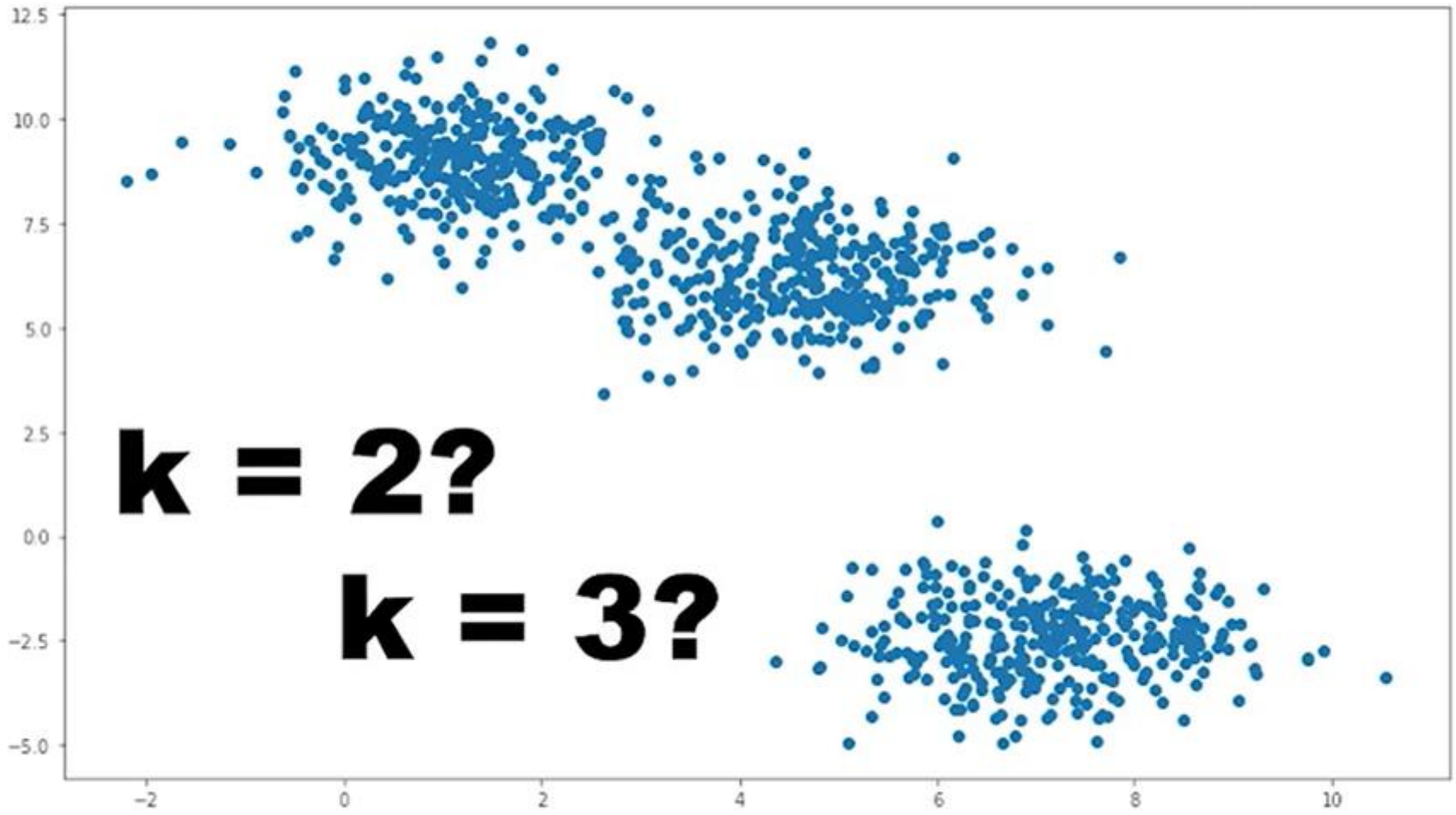
CLUSTERING



Unsupervised Learning
(b)

CLASSES UNKNOWN

K-MEANS CLUSTERING


12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−2
0
2
4
6
8
10
k = 2?
k = 3?

Where is it used?

Retail Store

Banking

Insurance Companies

amazon

NETFLIX

Recommended Movies

flickr

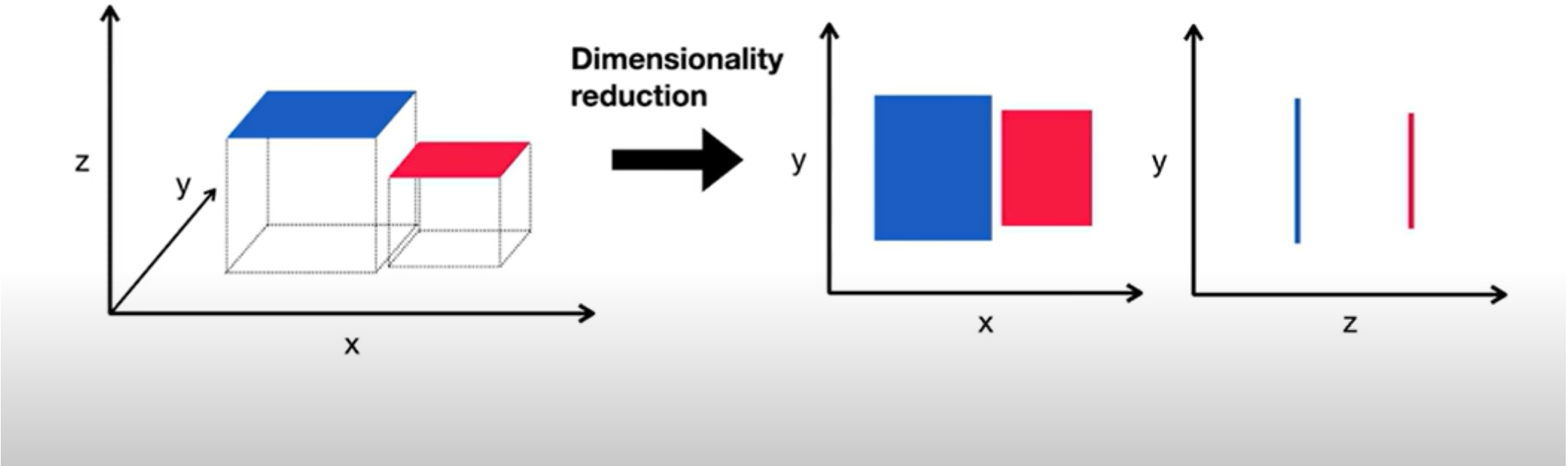
Flickr's Photos

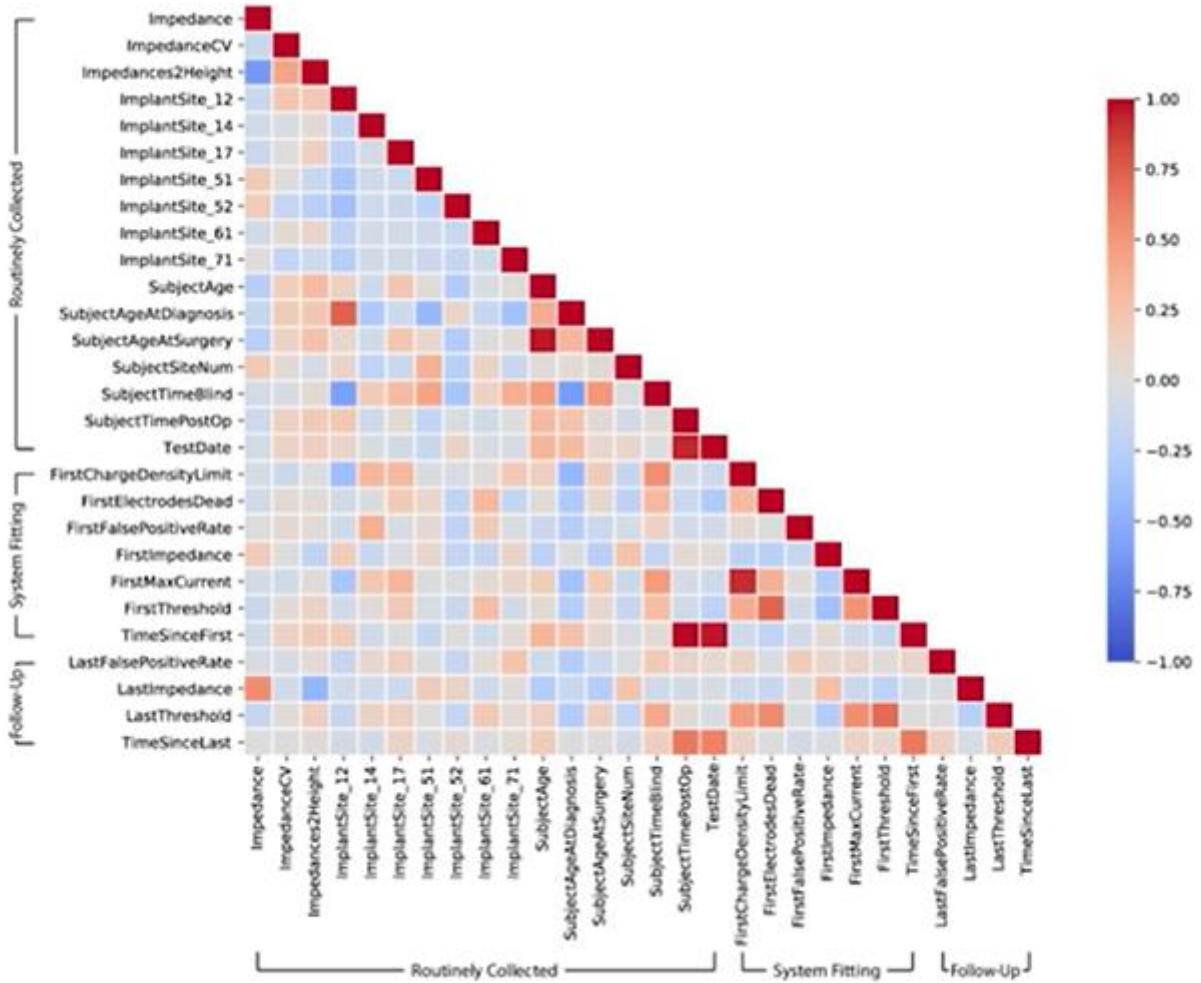
Types of Clustering

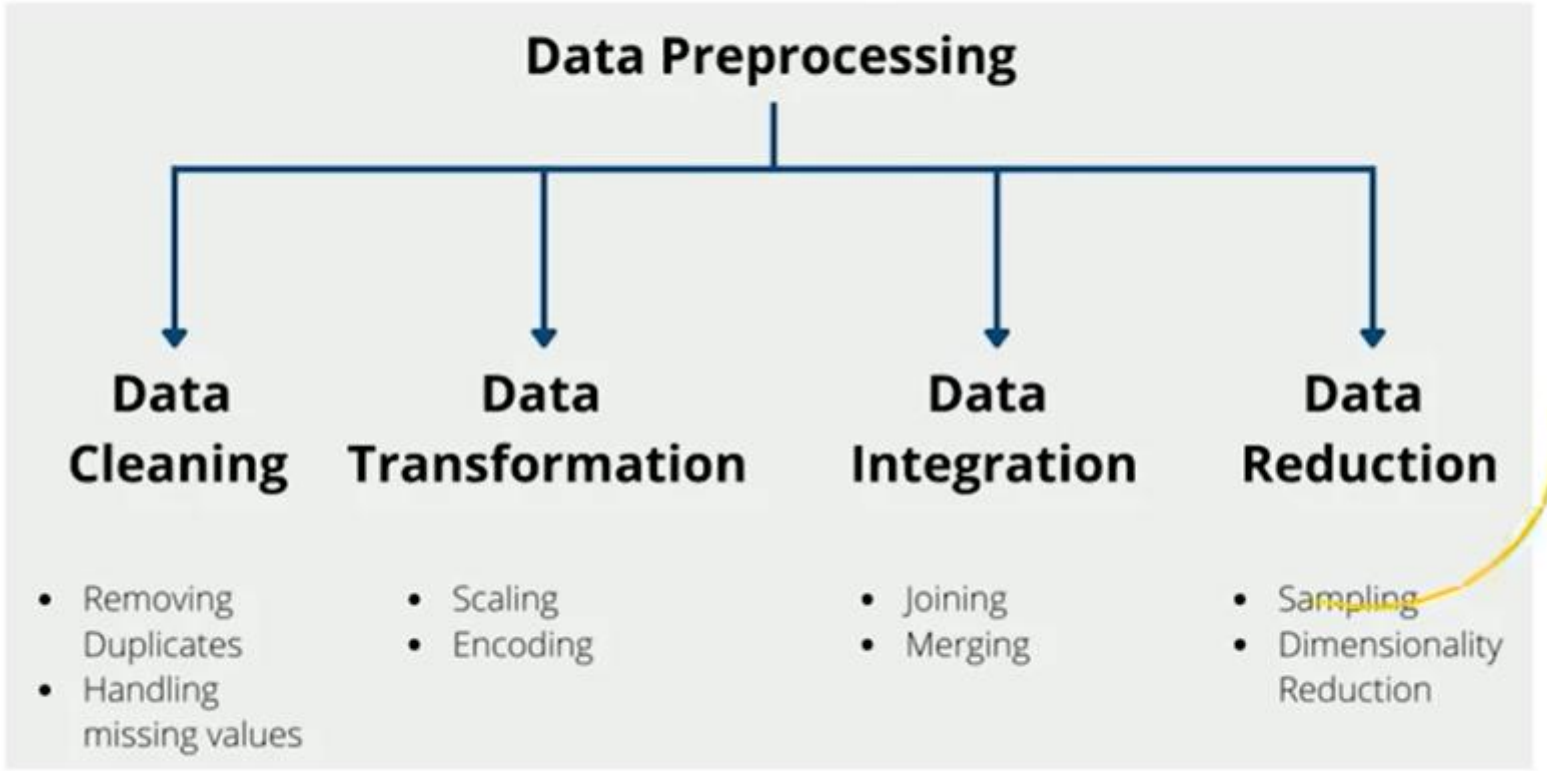
- **Exclusive Clustering**
- Overlapping Clustering
- Hierarchical Clustering

K-Means Clustering

DIMENSIONALITY REDUCTION


z
y
x
Dimensionality reduction
y
x
y
z

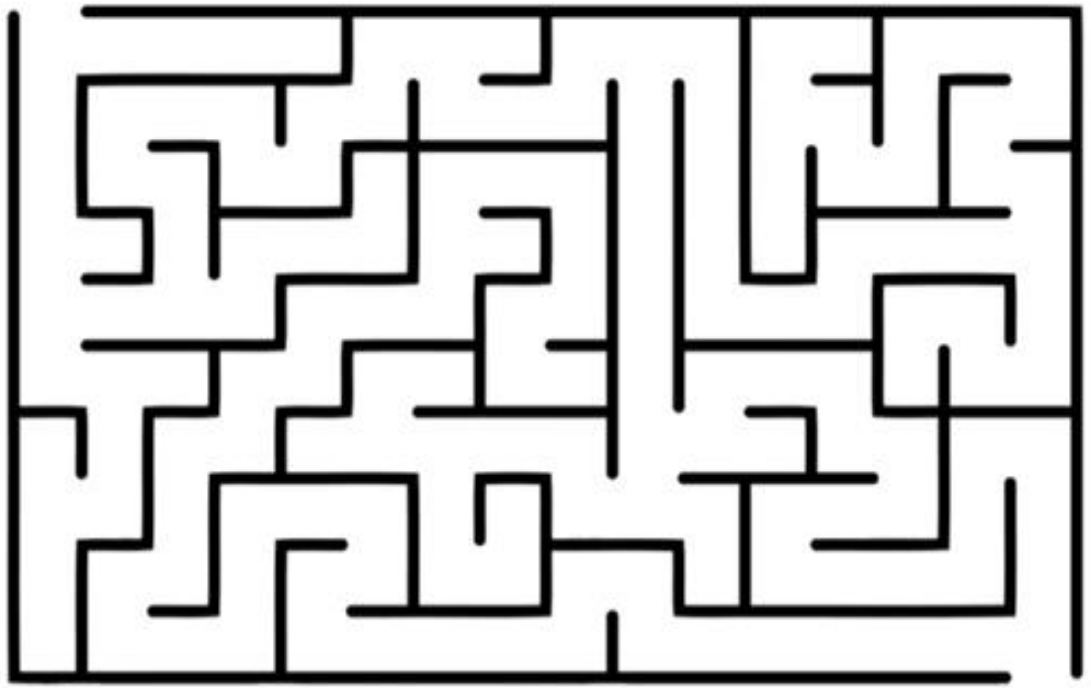

Impedance
ImpedanceCV
Impedances2Height
ImplantSite_12
ImplantSite_14
ImplantSite_17
ImplantSite_51
ImplantSite_52
ImplantSite_61
ImplantSite_71
SubjectAge
SubjectAgeAtDiagnosis
SubjectAgeAtSurgery
SubjectSiteNum
SubjectTimeBlind
SubjectTimePostOp
TestDate
FirstChargeDensityLimit
FirstElectrodesDead
FirstFalsePositiveRate
FirstImpedance
FirstMaxCurrent
FirstThreshold
TimeSinceFirst
LastFalsePositiveRate
LastImpedance
LastThreshold
TimeSinceLast
Routinely Collected
System Fitting
Follow-Up
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Data Preprocessing
Data Cleaning
Data Transformation
Data Integration
Data Reduction
• Removing Duplicates
• Handling missing values
• Scaling
• Encoding
• Joining
• Merging
• Sampling
• Dimensionality Reduction

PRINCIPAL COMPONENT ANALYSIS (PCA)

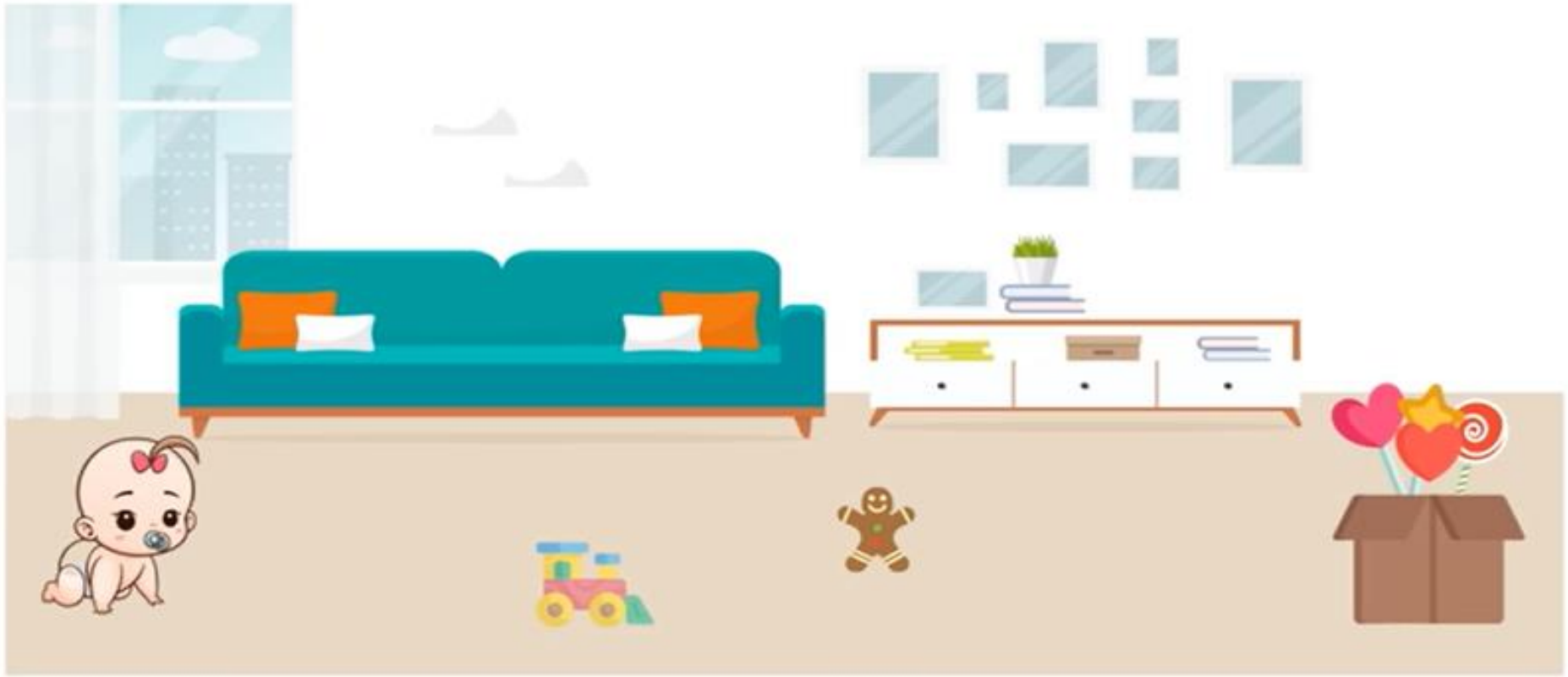
What is Reinforcement Learning?

Reinforcement learning is a type of Machine Learning where an agent learns to behave in an environment by performing actions and seeing the results



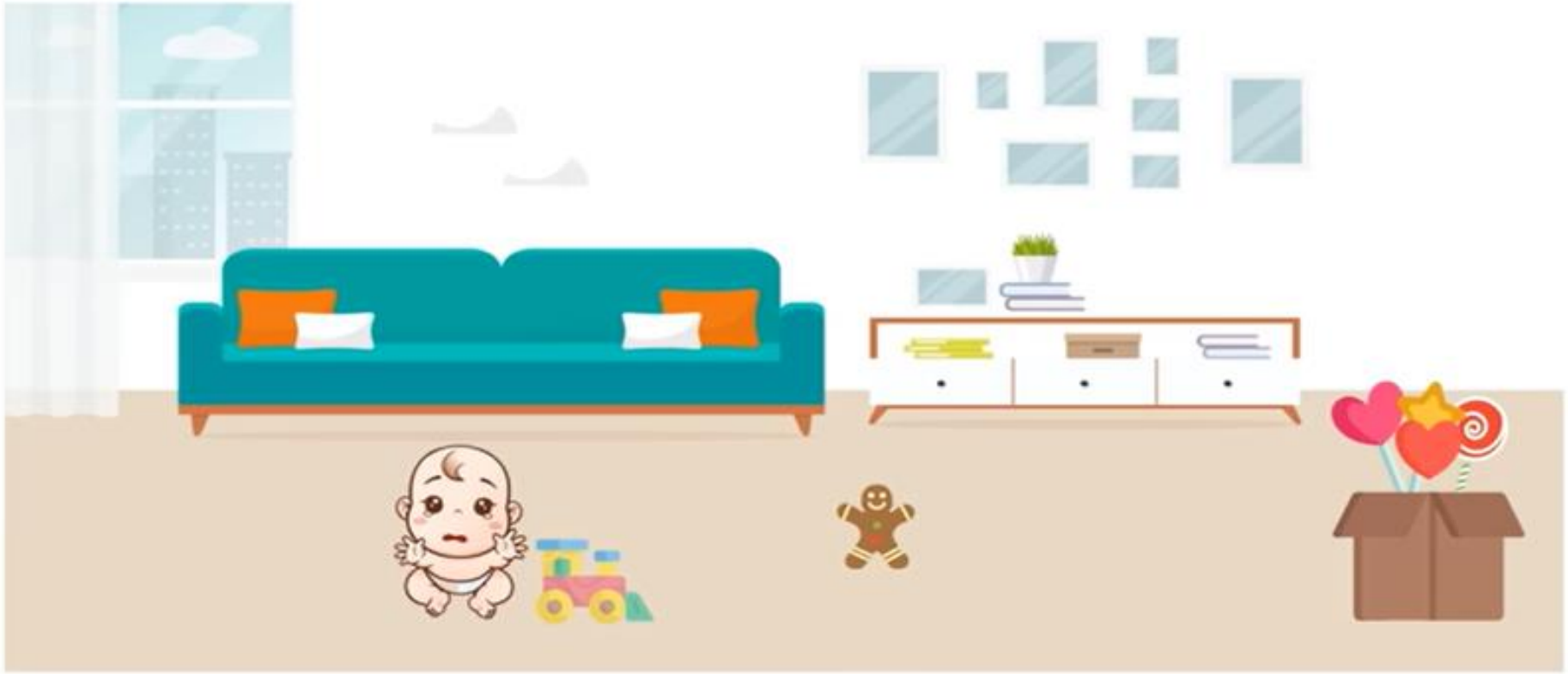
Analogy

Scenario 1: Baby starts crawling and makes it to the candy



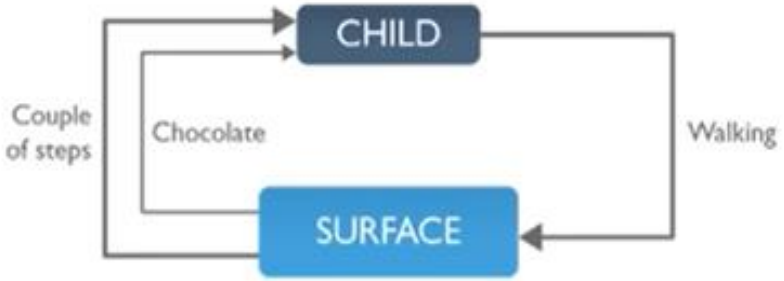
Analogy

Scenario 2: Baby starts crawling but falls due to some hurdle in between



Analogy

- **The child is an agent** trying to manipulate the **environment (which is the surface on which it walks)**
- **Taking actions (viz walking)** and he/she tries to go from one **state (viz each step he/she takes)** to another
- The child gets a **reward (let's say chocolate)** when he/she accomplishes a **submodule of the task (viz taking couple of steps)**



- Will not receive any chocolate **(negative reward)** when he/she is not able to walk

Reinforcement Learning Definitions

Agent

Intelligent programs

Model of the Environment

Used for planning & if know the current state and action then predict the resultant next state and next reward

Environment

An external condition

Value Function

Value of a state is the total amount of reward an agent can expect to accumulate over the future

Policy

A mapping from state to actions defining agent's behavior at a particular time

Reward Function

Could be +1 or any other value, indicating, what's good in an immediate sense

Reinforcement Learning Definitions

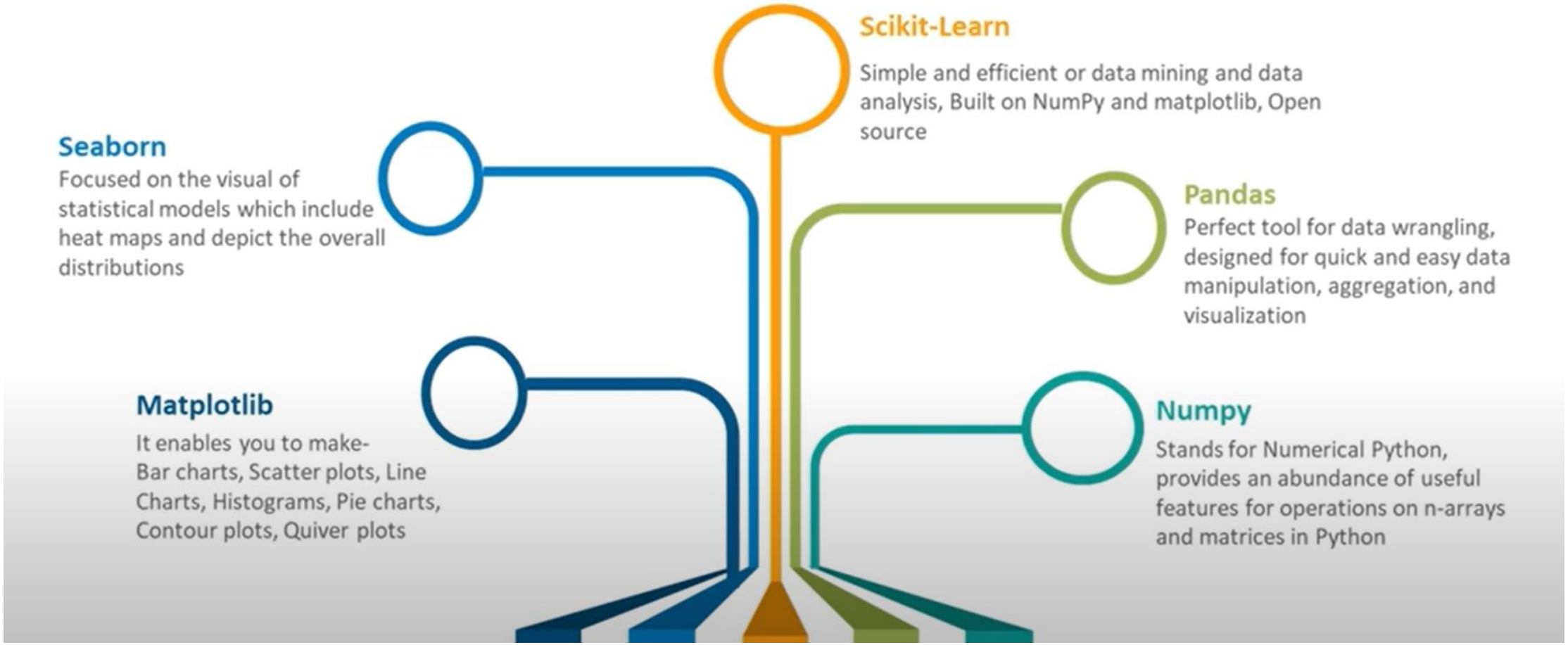
Reward (R): An instant return from the environment to appraise the last action

Policy (π): The approach that the agent uses to determine the next action based on the current state

Value (V): The expected long-term return with discount, as opposed to the short-term reward R

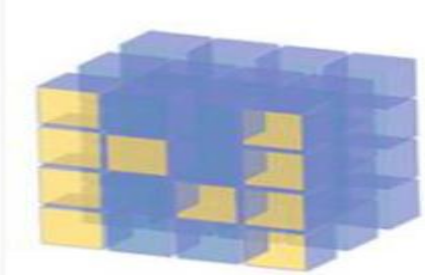
Action-value (Q): This similar to Value, except, it takes an extra parameter, the current action (A)

What libraries do we use for machine learning?




NumPy
scikit
learn
PyTorch
pandas

NumPy



- NumPy – Numerical python is a very popular python library for array and matrix processing, with the help of a large collection of high-level mathematical functions.
- It is very useful for fundamental scientific computations in Machine Learning.

Pandas

- Pandas-Panel data is a popular Python library for data analysis.
- It is not directly related to Machine Learning but the dataset must be prepared before training for which Pandas are useful as it is developed specifically for data extraction and preparation.
- It provides data structures and wide variety tools for data analysis. It provides many inbuilt methods for filtering, combining and grouping data.

Matplotlib

- Matplotlib is a Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It is needed when a programmer wants to visualize the patterns in the data.
- A module named pyplot makes it easy for programmers for plotting .

Scikit-learn

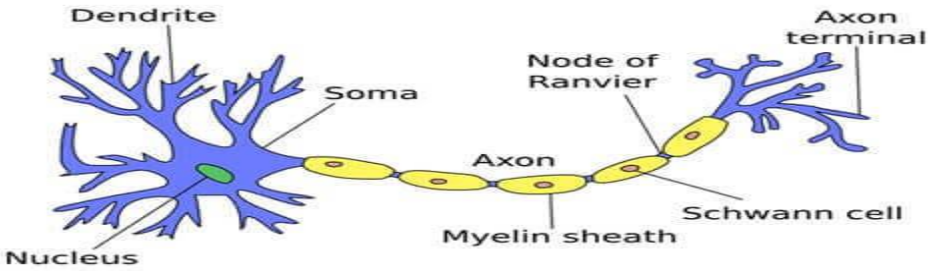
- Scikit-learn is one of the most popular ML libraries for classical ML algorithms.
- Scikit-learn supports most of the supervised and unsupervised learning algorithms

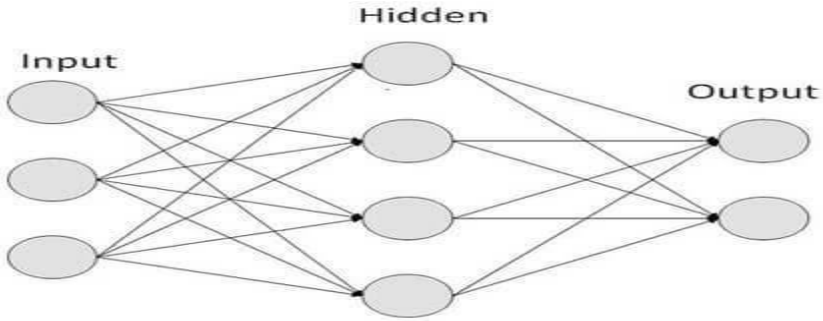
TensorFlow

- TensorFlow is a popular open-source library for high performance numerical computation.
- It can train and run deep neural networks that can be used to develop several AI applications. TensorFlow is widely used in the field of deep learning research and application.

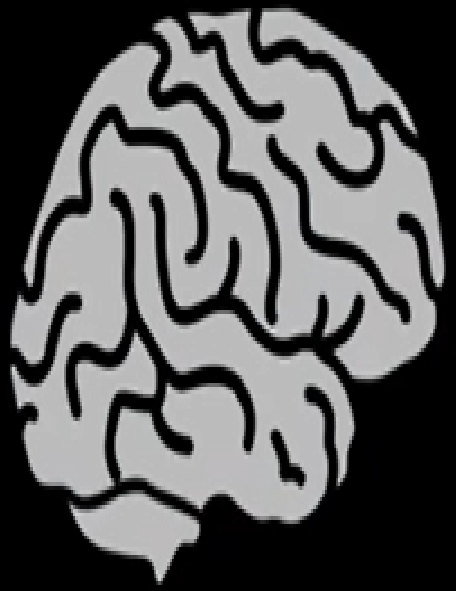
Introduction to Neural Networks

- Neural networks are computational models inspired by the human brain. They consist of interconnected units or nodes called neurons, organized into layers.
- Neural networks are capable of learning from data and making predictions or decisions without being explicitly programmed.



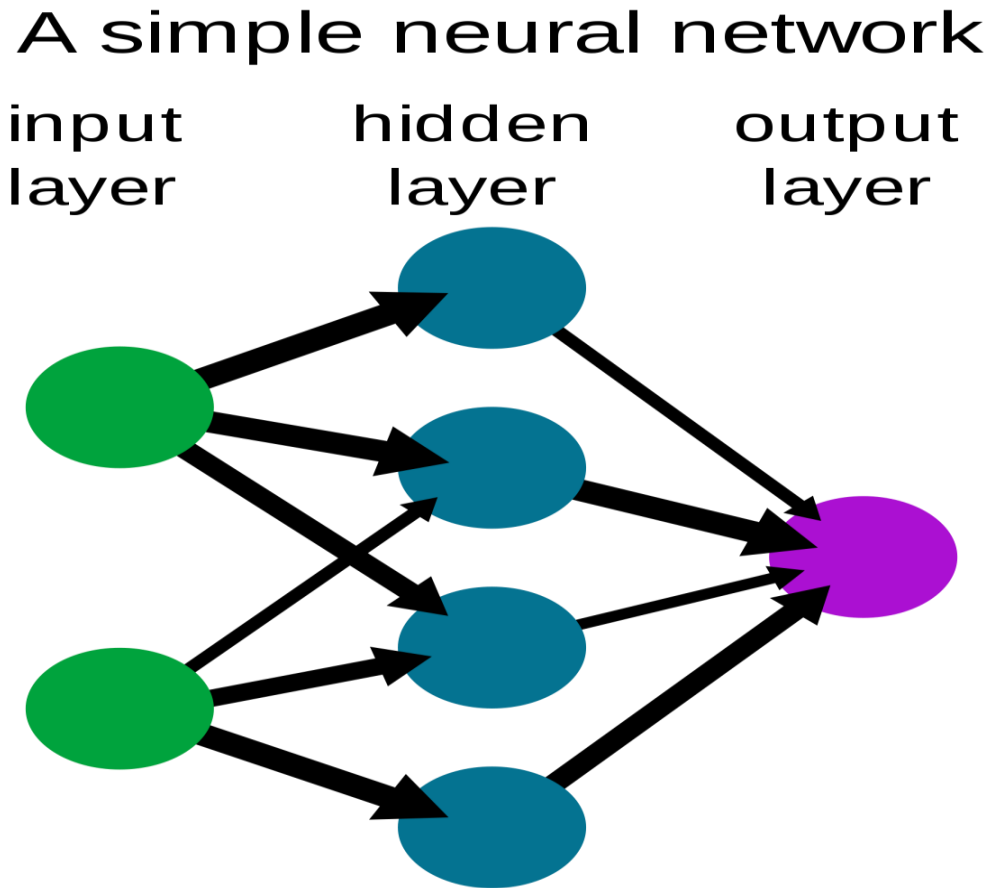


Neural network



What are the neurons?

How are they connected?



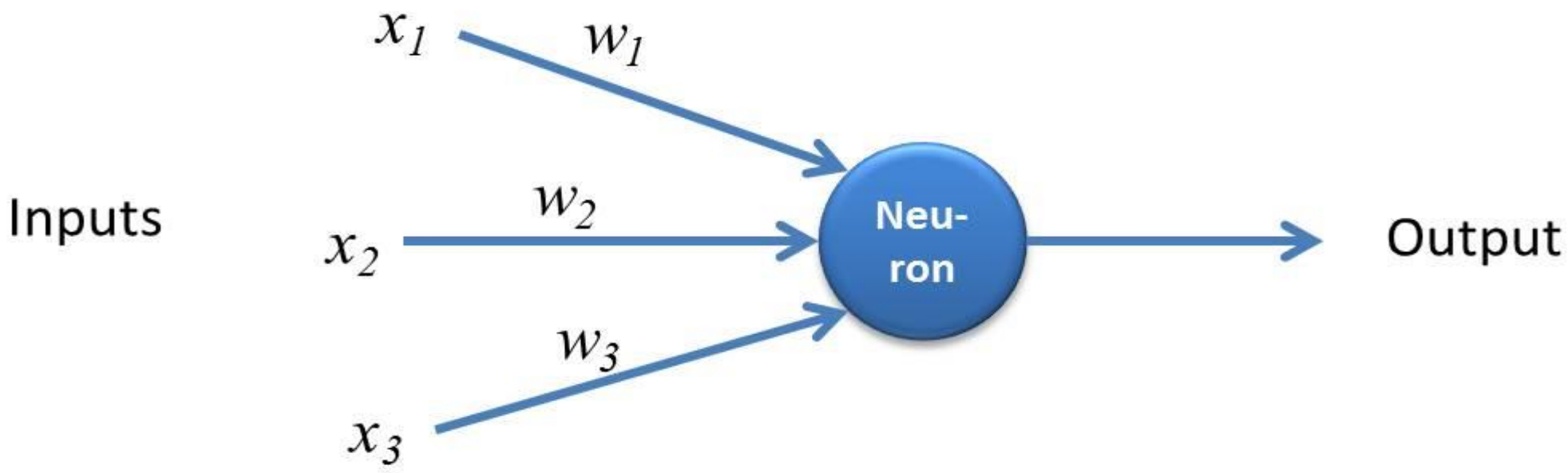
A simple neural network consists of three components :

- Input layer
- Hidden layer
- Output layer

- **Input Layer:** Also known as Input nodes are the inputs/information from the outside world is provided to the model. Input nodes pass the information to the next layer i.e Hidden layer.
- **Hidden Layer:** Hidden layer is the set of neurons where all the computations are performed on the input data. There can be any number of hidden layers in a neural network. The simplest network consists of a single hidden layer.
- **Output layer:** The output layer is the output/conclusions of the model derived from all the computations performed. There can be single or multiple nodes in the output layer. If we have a binary classification problem the output node is 1 but in the case of multi-class classification, the output nodes can be more than 1.

Perceptron

- **Perceptron** is a simple form of Neural Network and consists of a single layer where all the mathematical computations are performed.

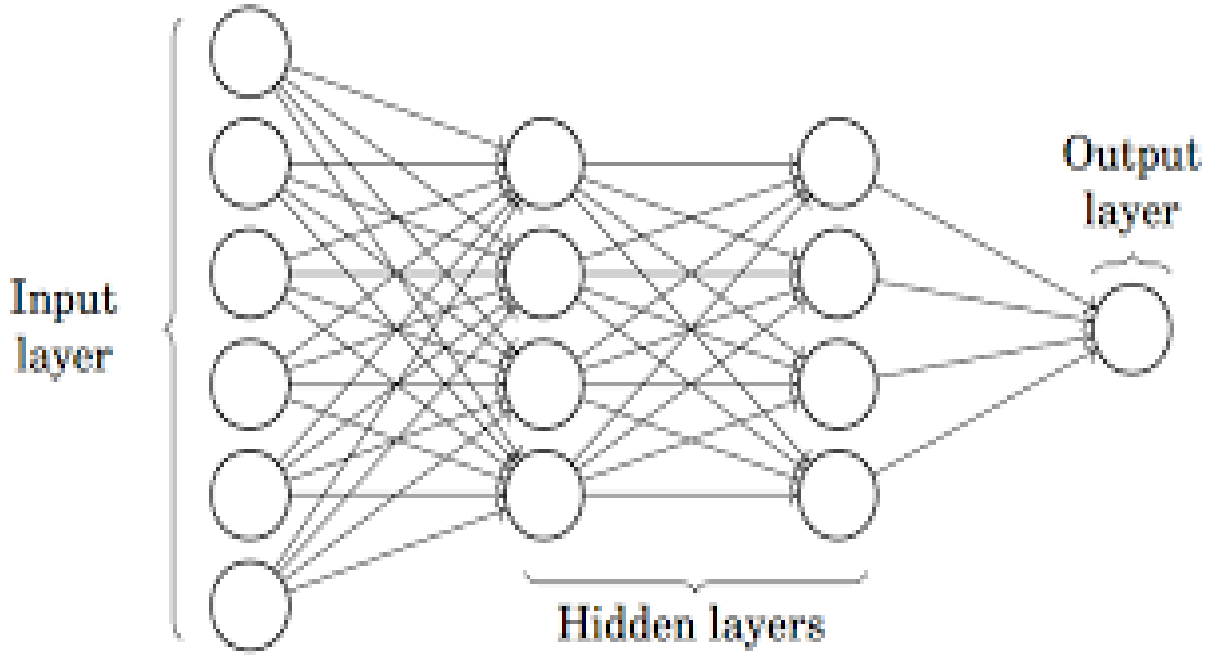


Types of neural networks

The three important types of neural networks in deep learning are:

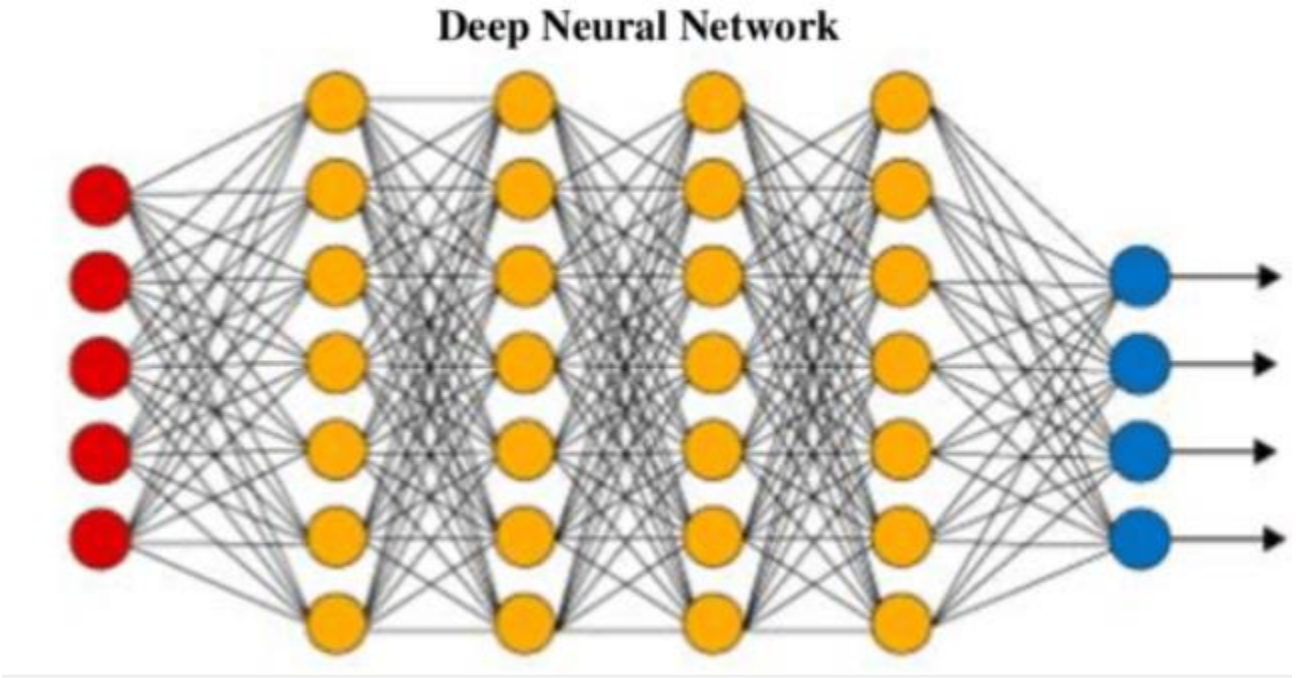
- Artificial Neural Networks (ANN)
- Convolution Neural Networks (CNN)
- Recurrent Neural Networks (RNN)

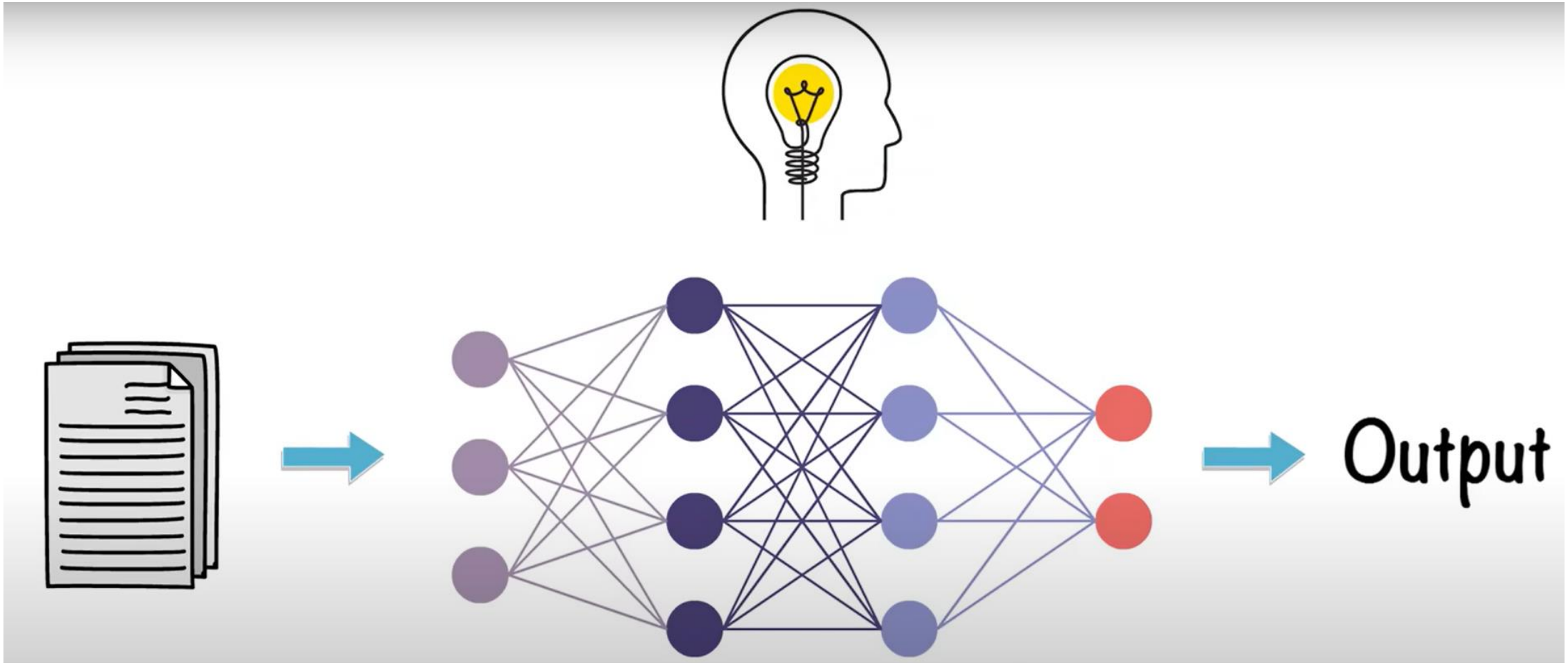
- Whereas, **Multilayer Perceptron** also known as **Artificial Neural Networks** consists of more than one perception which is grouped together to form a multiple layer neural network.



DEEP NEURAL NETWORK

A Deep Neural Network is a simple neural network that consist of multiple hidden layers.




Output


ANY
QUESTIONS

Thank You