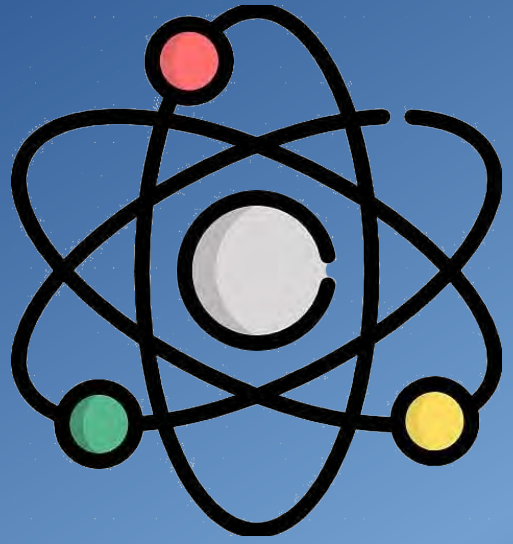
inixindo
"Continuous learning, keep up to date"



Data Science Fundamental

Agenda


Evolution of Data
Introduction To Data Science
Data Science Careers and Salary
Statistics for Data Science
What is Machine Learning?
Types of Machine Learning
What is Deep Learning?


Which Provisioner do you use the most ?

Evolution of Data

2.5×10^{18} Bytes Everyday

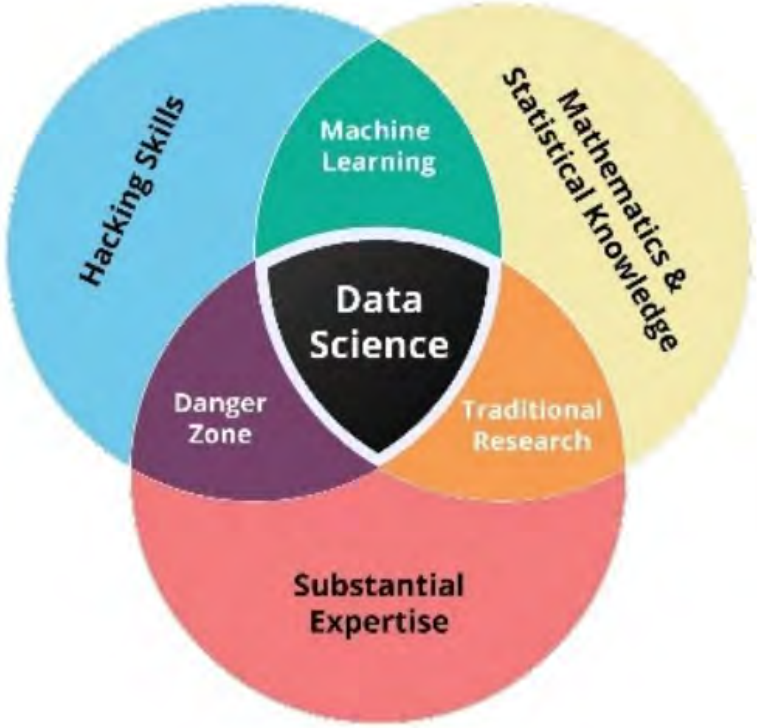
Evolution of Data

3 MILLION

4.3 MILLION

What is Data Science?

Data science, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured.



Career : Data Science

Data Engineer → Data Analyst → AI/ML Engineer → Data Scientist

Data Analyst

A Data Analyst takes data and uses it to help companies make better business decisions.

Data Scientist

Data scientists are those who crack complex data problems with their strong expertise in certain scientific disciplines. They work with several elements related to mathematics, statistics, computer science, etc

Machine Learning Engineer

Machine Learning engineers are sophisticated programmers who develop machines and systems that can learn and apply knowledge without specific direction.
Artificial intelligence is the goal of a machine learning engineer

Data Analyst Skills

Analytical skills

Communication skills

Critical thinking

Attention to detail

Statistical skills

Technical skills/tools

Data Scientist Skills

Analytics & Statistics

Machine Learning Algorithms

Problem Solving Skills

Deep Learning

Business Communication

Technical skills/tools

ML Engineer Skills

Programming Languages

Calculus & Statistics

Signal Processing

Applied Maths

Neural Networks

Language Processing

Data Science Peripherals

Statistics

Data Science Peripherals

Statistics

Prog Languages

Data Science Peripherals

Statistics

Prog Languages

Software

Data Science Peripherals

Statistics

Prog Languages

Software

Machine Learning

Data Science Peripherals

Statistics

Prog Languages

Software

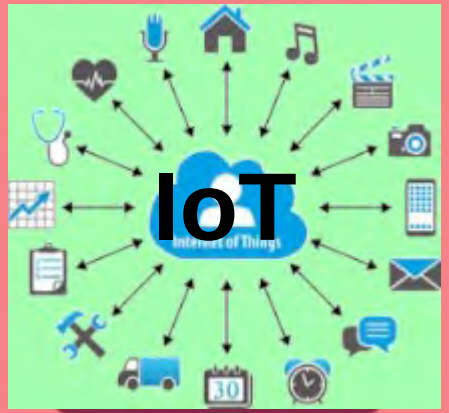
Machine Learning

Big Data

Data Science Peripherals

Statistics

Prog Languages



Machine Learning

Big Data

What is Data ?

Qualitative data

Nominal data

Type 1
Male
Female
Female
Male
Male

Ordinal data

Type 2	
Customer's id	Rating scale
101	Strongly agree
302	Disagree
013	Neutral
001	Agree
917	Strongly disagree

Binary data

Customer Id	Own a car?
192	Yes
271	No
122	Yes
111	Yes
321	No

What is Data ?

Quantitative data

Discreet data

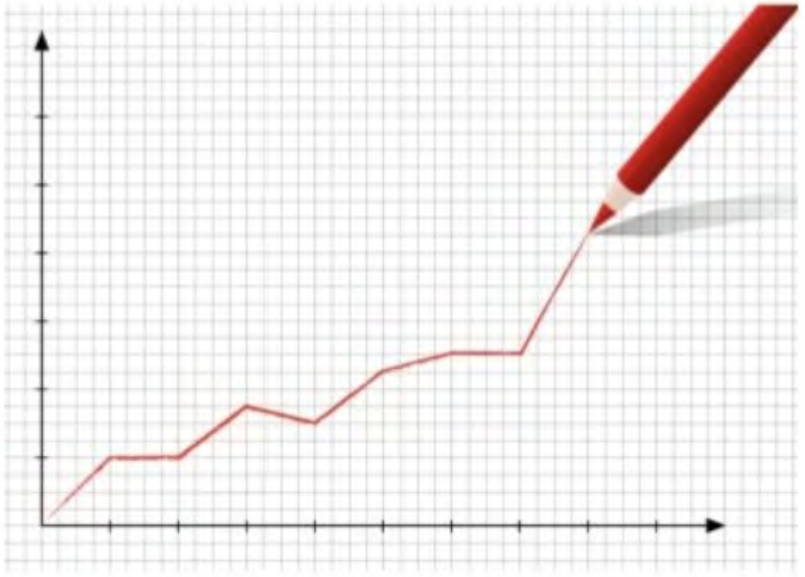
Organization	Number of Products
Samsung	500
Apple	30
Nokia	10
LG	450
Sony	200

Continuous data

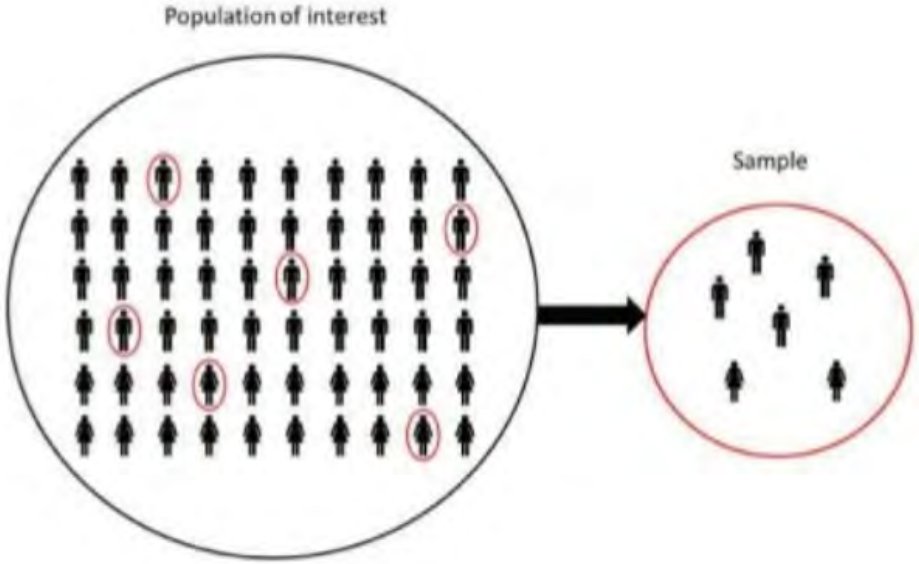
Patient id	Weight of Patient
12	86.5 kg.
15	91.3 kg.
11	56.1 kg.
7	70.9 kg.
5	60.34 kg.

Variables & Research

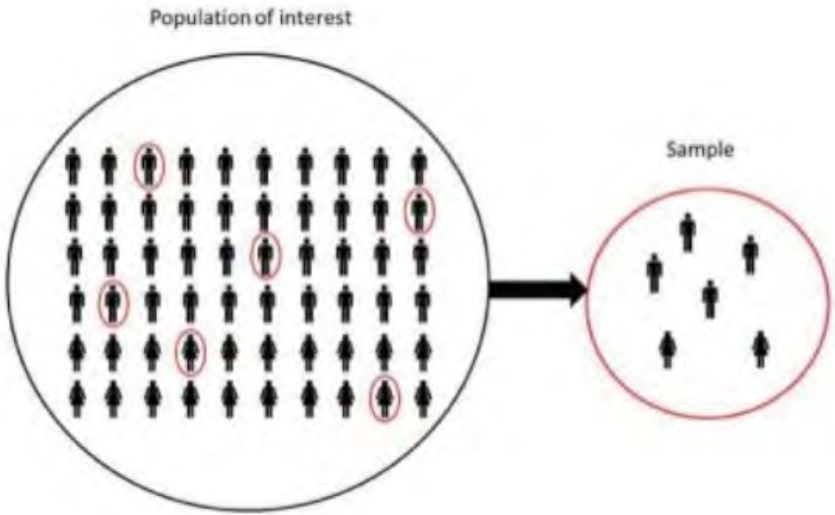
$\propto$



Population & Sampling



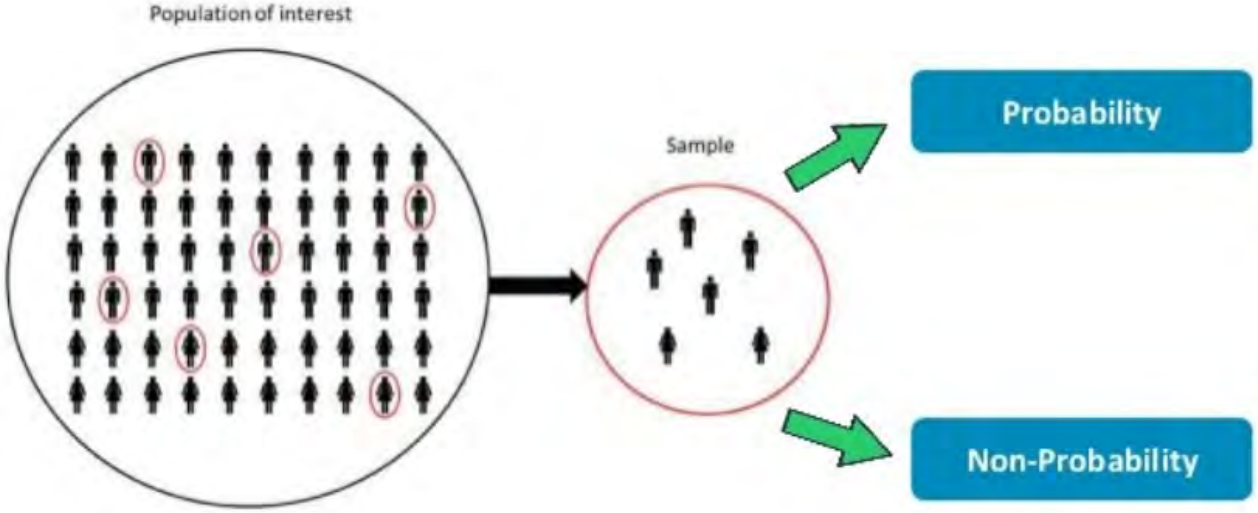
Population & Sampling



Probability

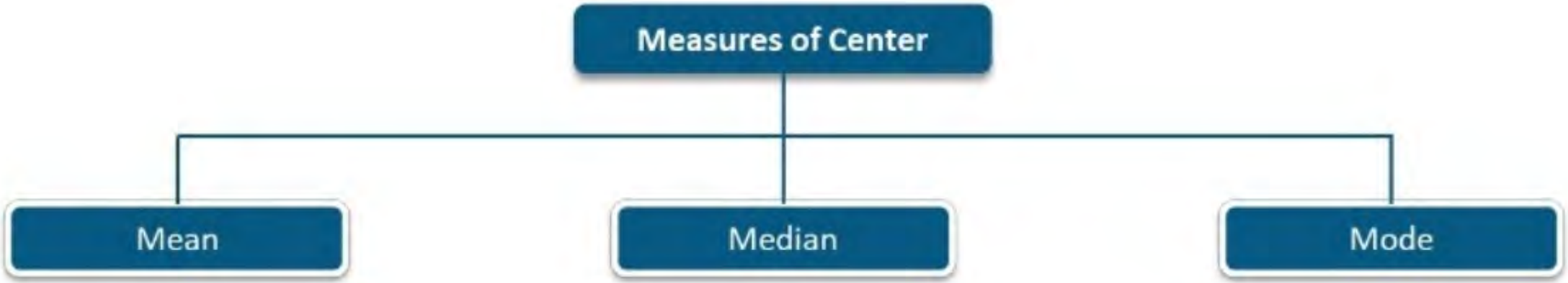
Non-Probability

Population & Sampling

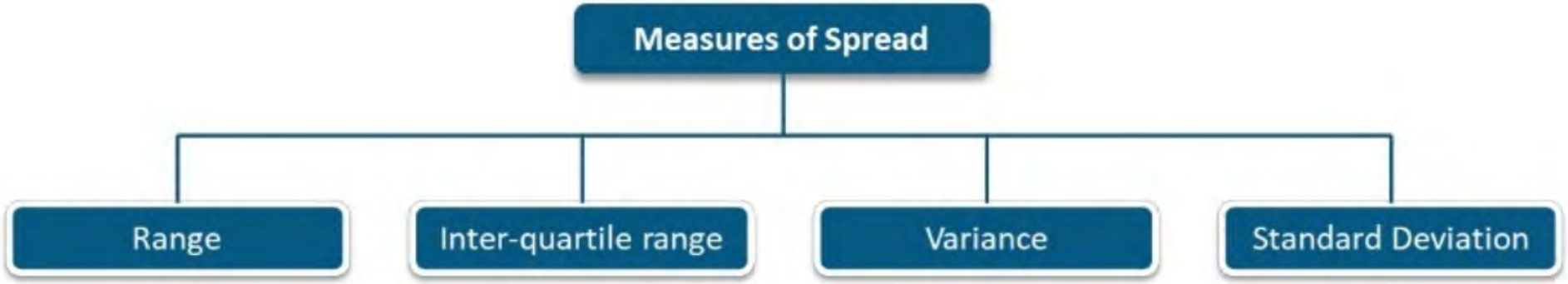


- Random sampling
- Systematic sampling
- Stratified sampling

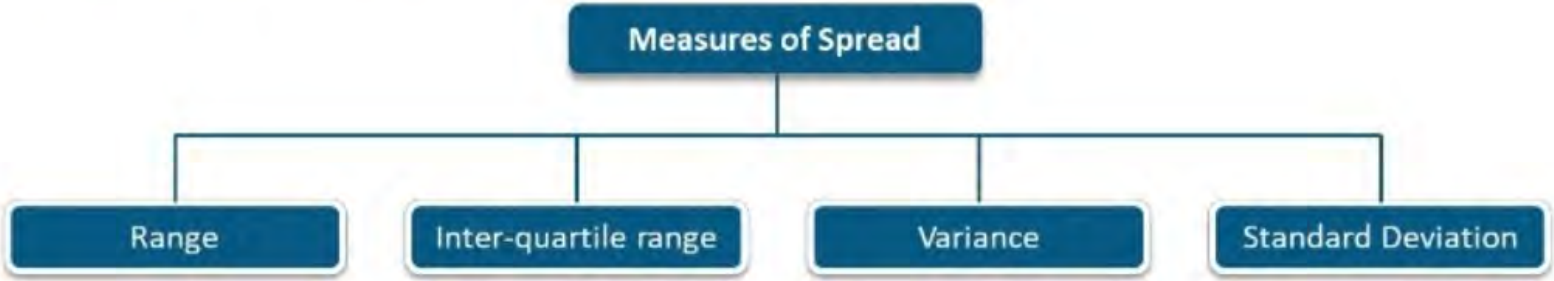
Measures of Center



Measures of Spread



Measures of Spread

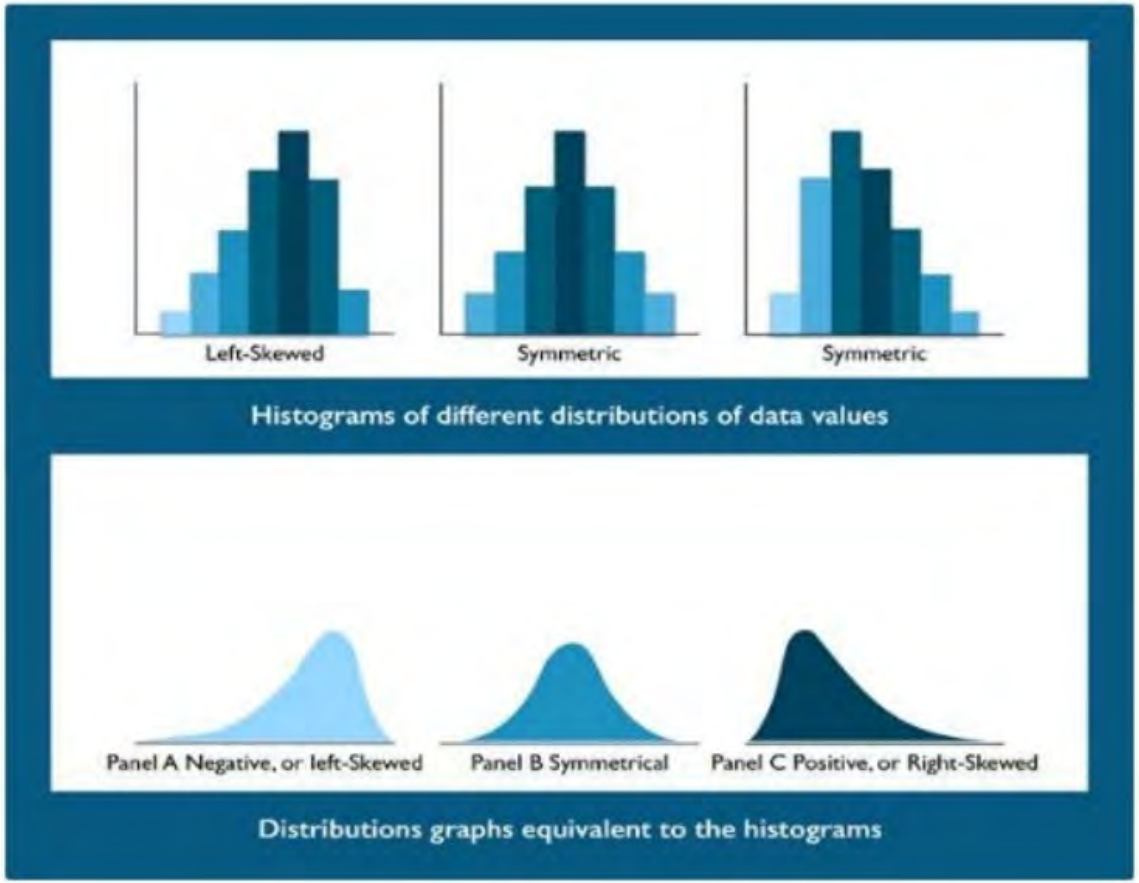


$$\sigma^2 = \sum (X_i - \bar{X})^2 / N$$

σ^2 = variance
X_i = the value of the ith element
$\bar{X}$ = the mean of X
N = the number of elements

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Skewness


Left-Skewed
Symmetric
Symmetric
Histograms of different distributions of data values
Panel A Negative, or left-Skewed
Panel B Symmetrical
Panel C Positive, or Right-Skewed
Distributions graphs equivalent to the histograms

Confusion Matrix

You can calculate the **accuracy** of your model with:

$$\frac{\text{True Positives + True Negatives}}{\text{True Positives + True Negatives + False Positives + False Negatives}}$$

Confusion Matrix

You can calculate the **accuracy** of your model with:

$$\frac{\text{True Positives + True Negatives}}{\text{True Positives + True Negatives + False Positives + False Negatives}}$$

Probability

Probability is the measure of how likely something will occur

Probability Density Function 01

Normal Distribution 02

Central Limit Theorem 03

Probability

***The equation** describing a continuous probability distribution is called a probability density function.*

Probability Density Function 01

Normal Distribution 02

Central Limit Theorem 03

Probability

The normal distribution *is a probability distribution that associates the normal random variable X with a cumulative probability .*

The normal distribution is defined by the following equation:

$$Y = [\ 1/\sigma * sqrt(2\pi)\ ] * e^{-(x-\mu)2/2\sigma 2}$$

Where,

X *is a normal random variable.*

μ *is the mean and*

σ *is the standard deviation.*

Probability Density Function 01

Normal Distribution 02

Central Limit Theorem 03

Probability

Probability Density Function 01

Normal Distribution 02

Central Limit Theorem 03

The central limit *theorem states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough.*

What is Machine Learning?

Machine Learning is a class of algorithms which is data-driven, i.e. unlike "normal" algorithms it is the data that "tells" what the "good answer" is

Getting computers to program themselves and also teaching them to make decisions using data

"Where writing software is the bottleneck, let the data do the work instead."

Features of Machine 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

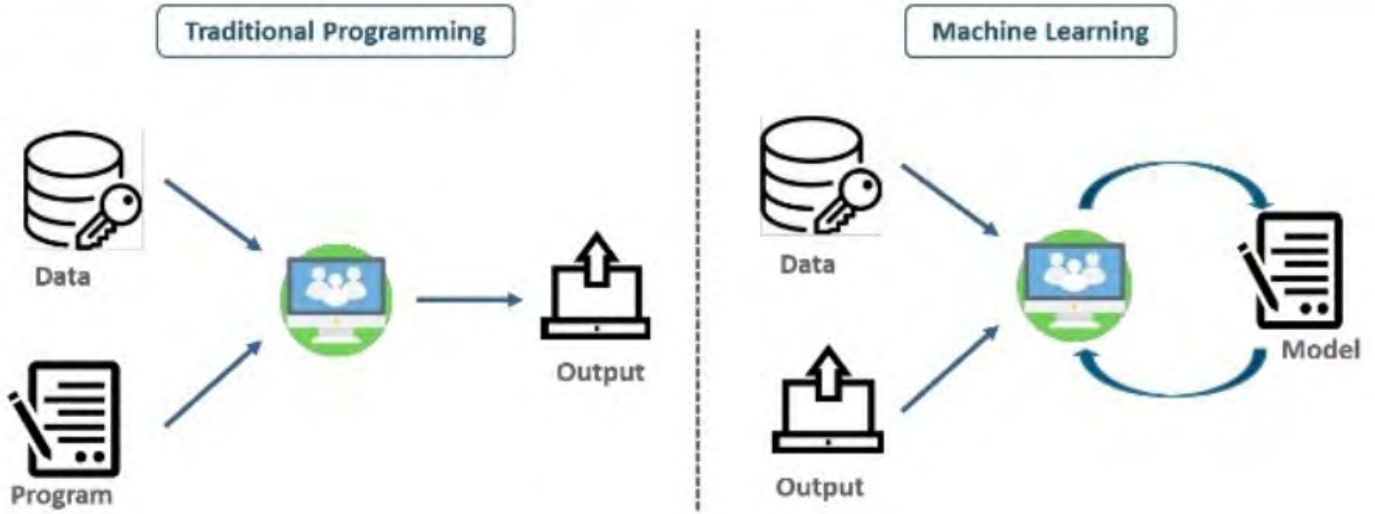
03

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 It Works?



Learn from Data

Find Hidden Insights

Train and Grow

Applications of Machine Learning

UBER

```
<img class="spotlight" alt="Image may contain: sky, grass, outdoor and nature"
```

Applications of Machine Learning

UBER

TESLA

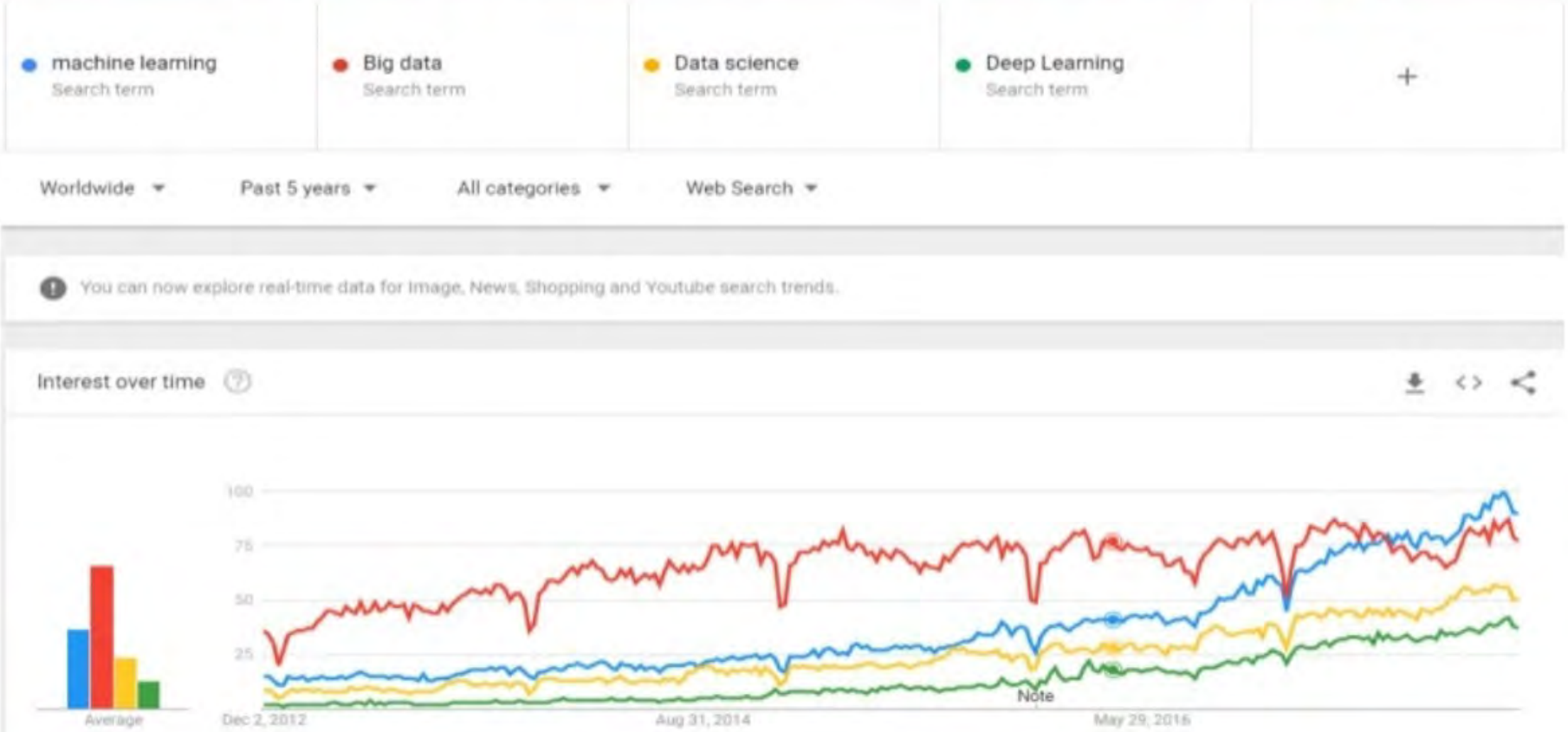
Applications of Machine Learning

NETFLIX

amazon video

hulu

Market Trend: Machine Learning



Machine Learning Life Cycle

Machine Learning Life Cycle

Data Wrangling

1

2

3

4

5

6

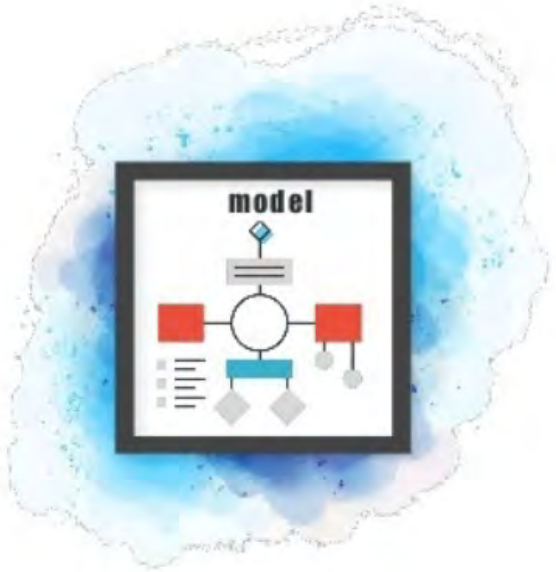
Data acquired from sources

Data filtering

Clean Data

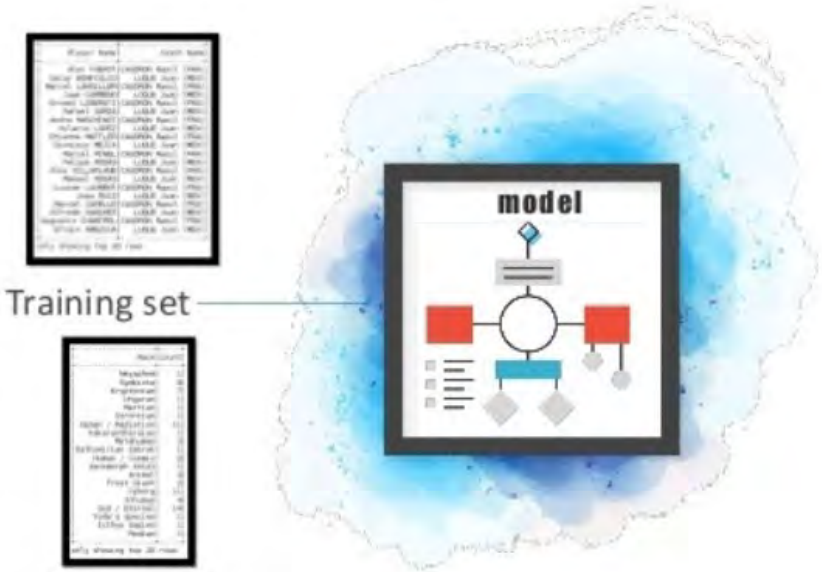
Analyse Data

1
2
3
4
5
6



Train Algorithm

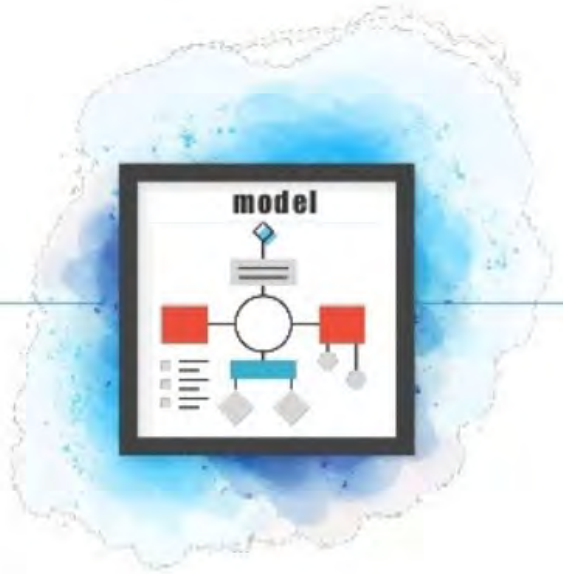
1
2
3
4
5
6



Test Algorithm

1
2
3
4
5
6

Test set

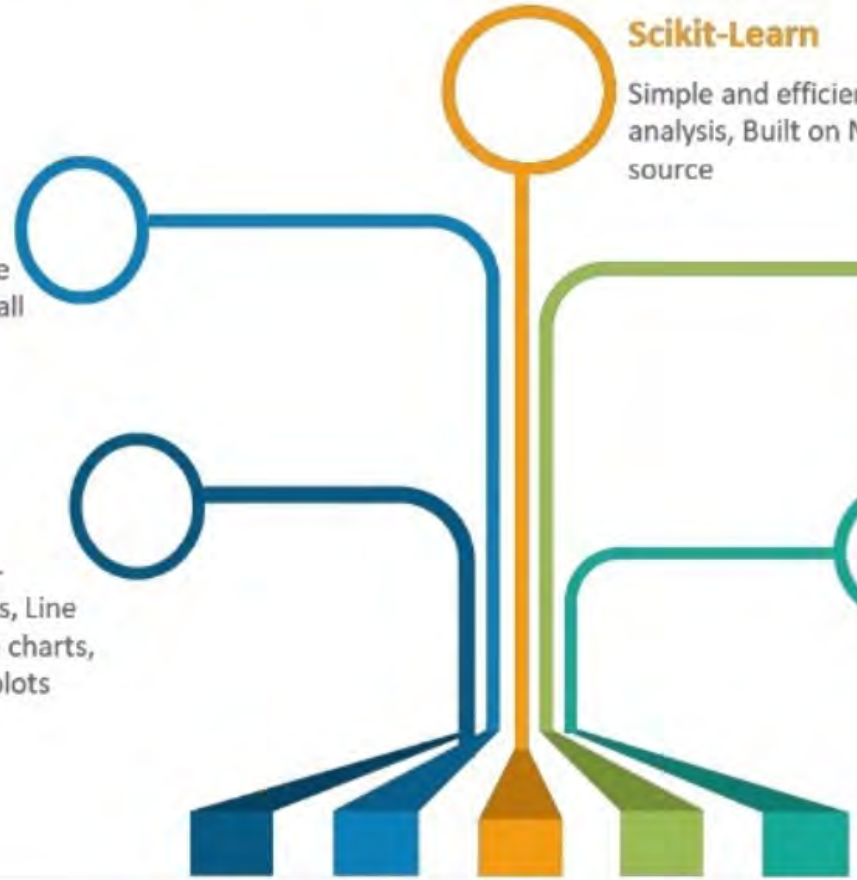


Accurate?

Operation and Optimization

1
2
3
4
5
6

Important Python Libraries



Scikit-Learn

Simple and efficient or data mining and data analysis, Built on NumPy and matplotlib, Open source

Seaborn

Focused on the visual of statistical models which include heat maps and depict the overall distributions

Pandas

Perfect tool for data wrangling, designed for quick and easy data manipulation, aggregation, and visualization

Matplotlib

It enables you to make-
Bar charts, Scatter plots, Line Charts, Histograms, Pie charts, Contour plots, Quiver plots

Numpy

Stands for Numerical Python, provides an abundance of useful features for operations on n-arrays and matrices in Python

Types of Machine Learning

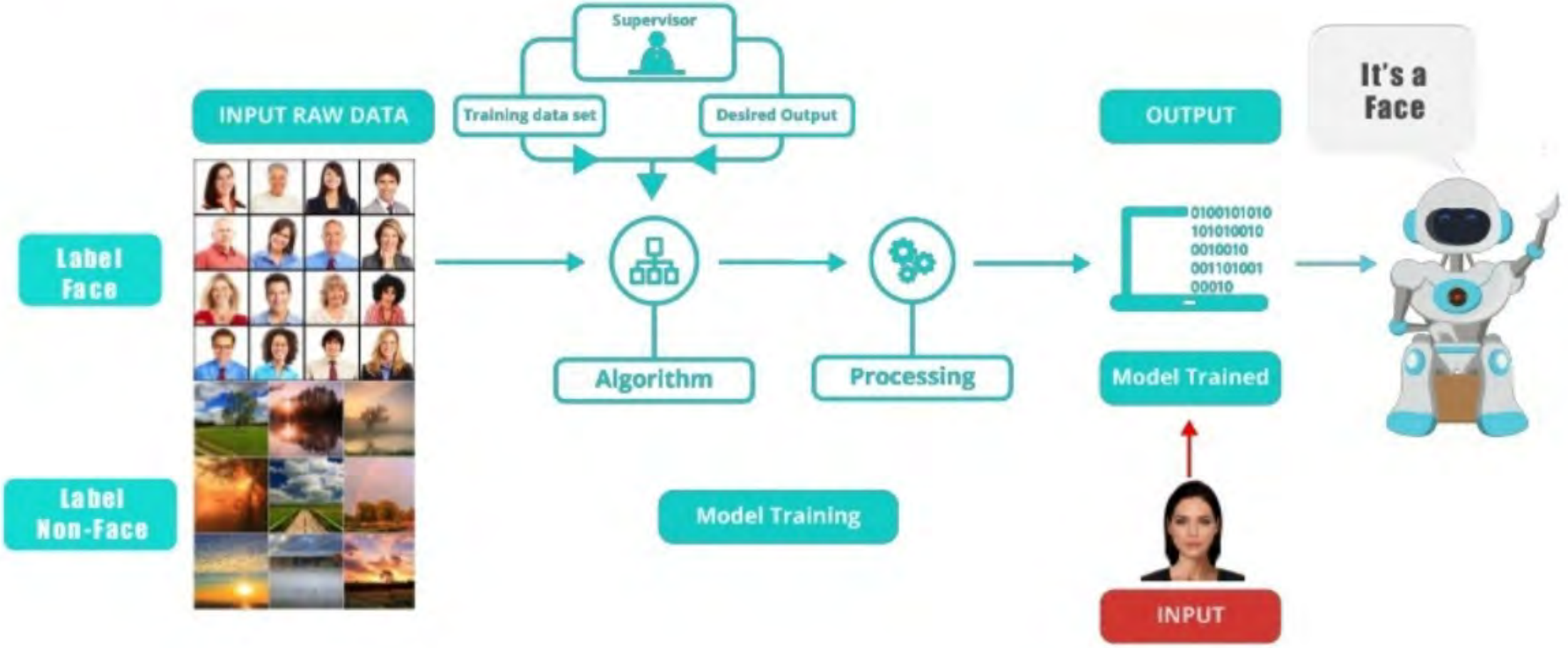
Types of Learning

Supervised Learning

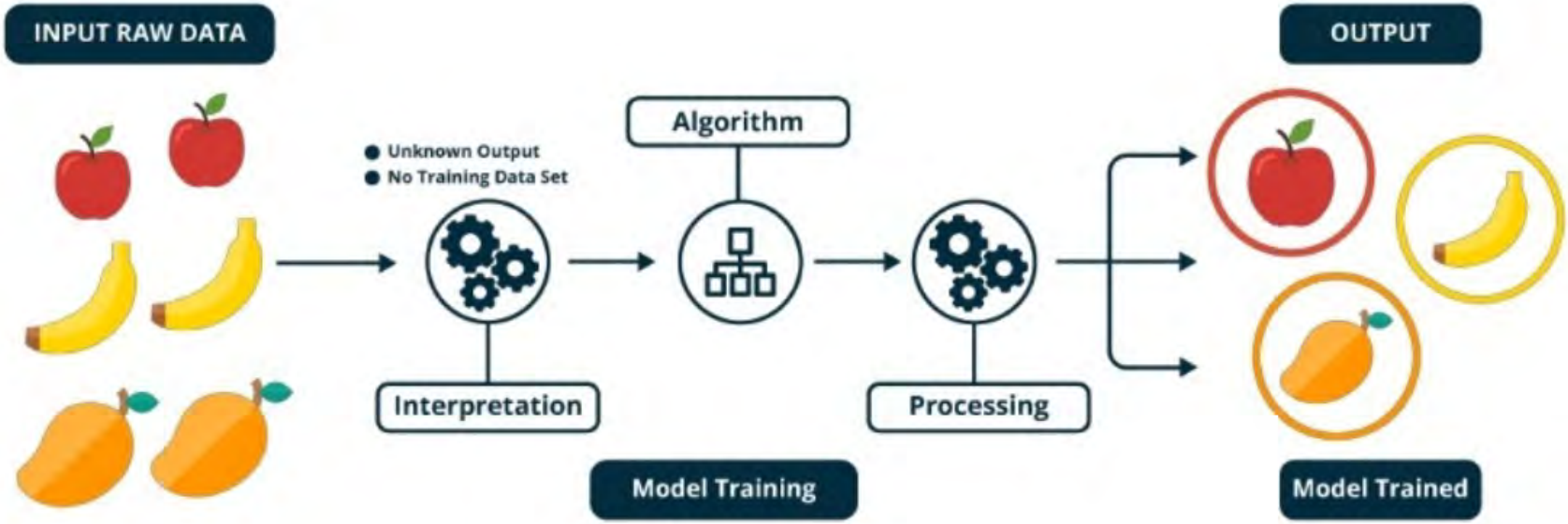
Reinforcement Learning

Unsupervised Learning

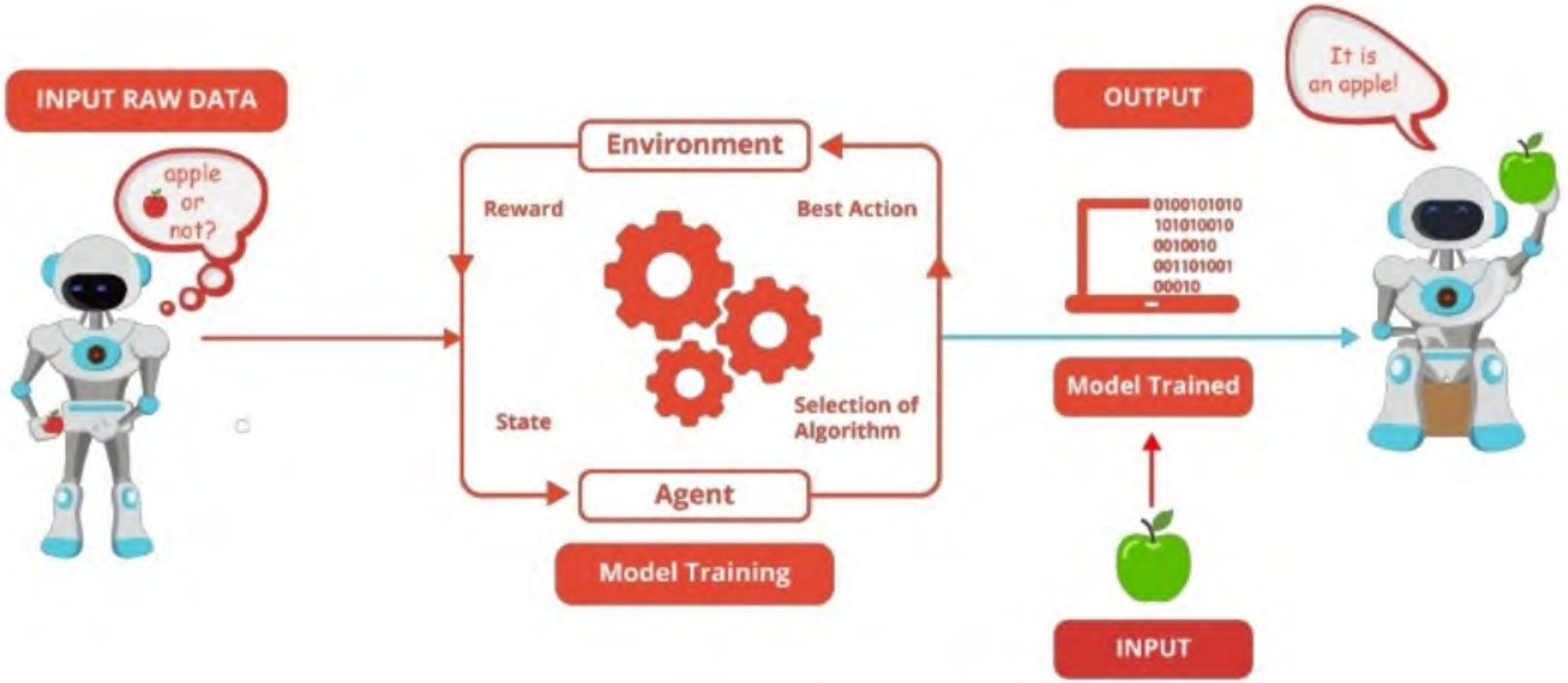
Supervised Learning



Unsupervised Learning



Reinforcement Learning



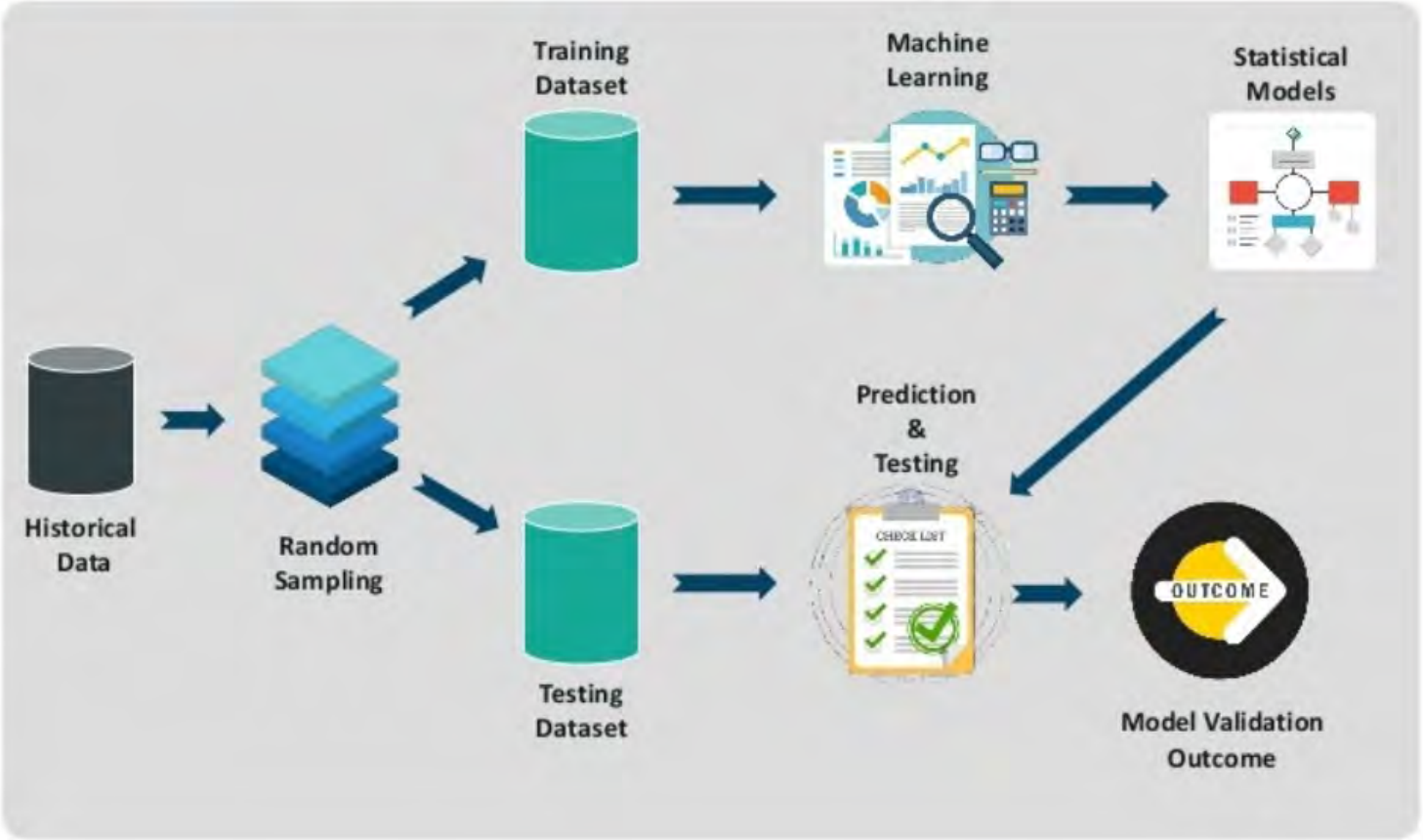
Supervised Learning

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output

It is called Supervised Learning because the process of an algorithm learning from the training dataset can be thought as a teacher supervising the learning process

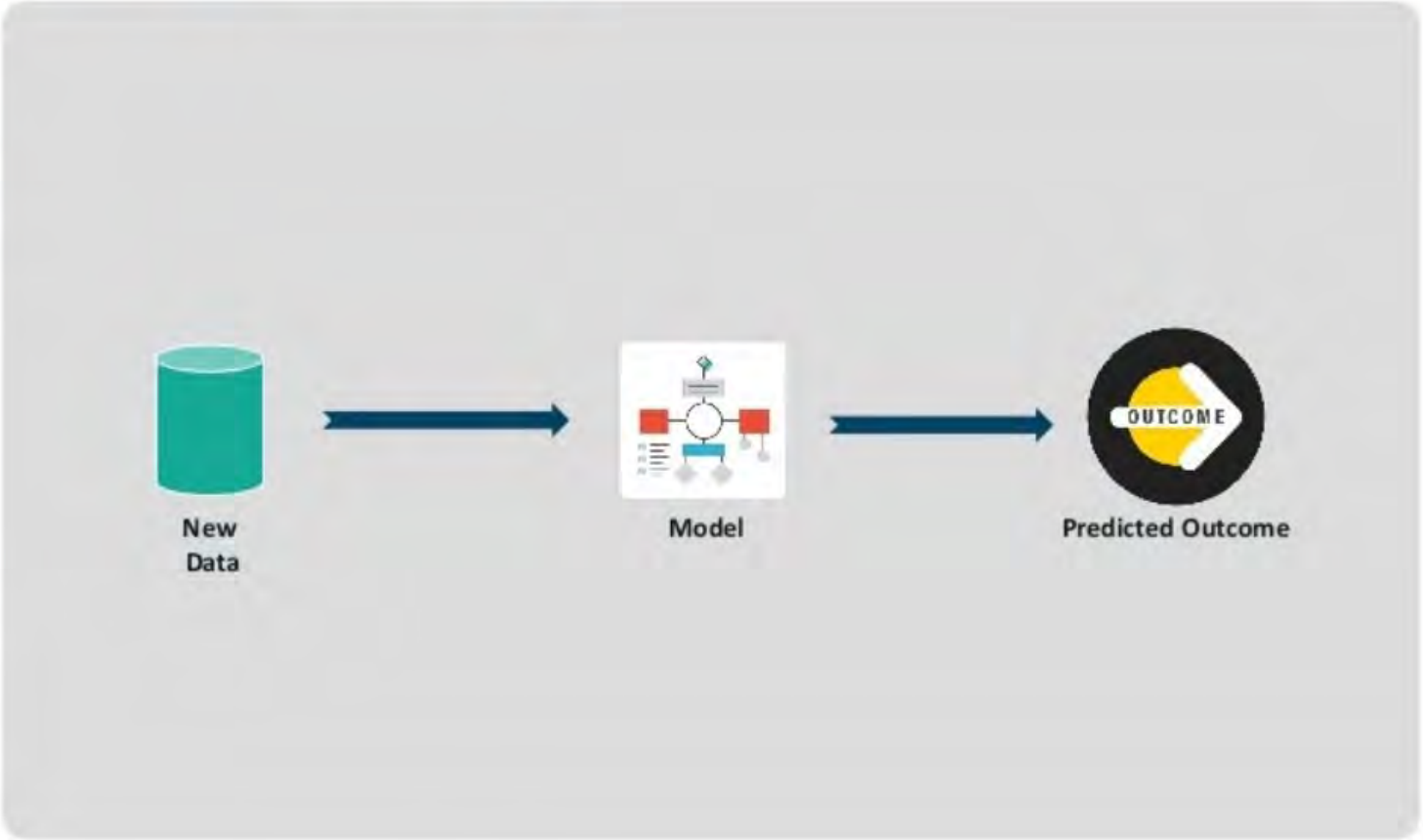
Supervised Learning

- Training and Testing
- Prediction



Supervised Learning

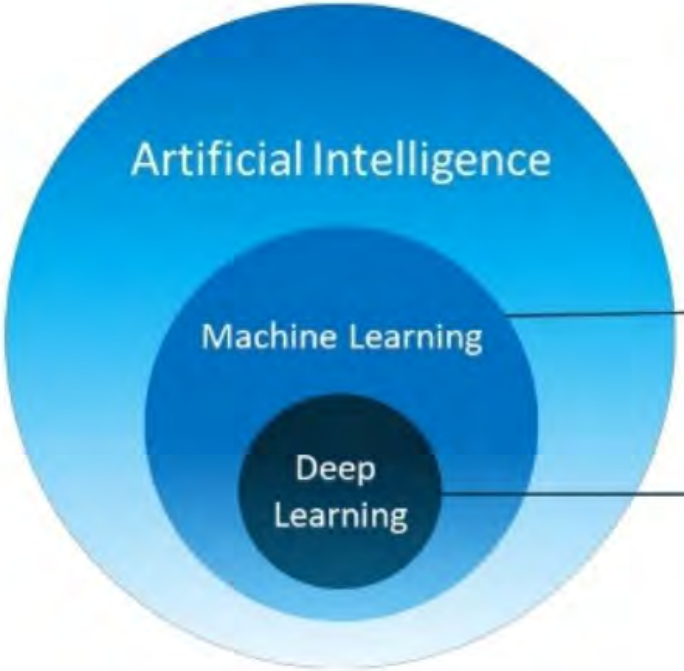
- Training and Testing
- Prediction



Supervised Learning Algorithms

- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- Naïve Bayes Classifier

AI and ML and DL



We'll learn more about Deep Learning when we discuss Deep networks and Neural networks in module 3 and 4

Machine Learning is a subset of AI

Deep Learning is a subset of Machine Learning

Deep Learning uses neural networks to simulate human like decision making

Limitations of Machine Learning

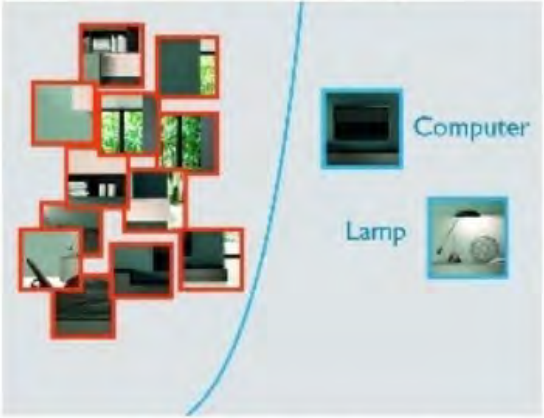
Traditional Machine Learning algorithms have failed to solve crucial problems of AI, such as Natural language processing, image recognition and so on.

Deep Learning to Rescue

- Deep Learning is one of the only methods by which we can circumvent the challenges of feature extraction.
- This is because Deep Learning models are capable of learning to focus on the right features by themselves, requiring little guidance from the programmer.

Where are the electric appliances?

Bag of image patches

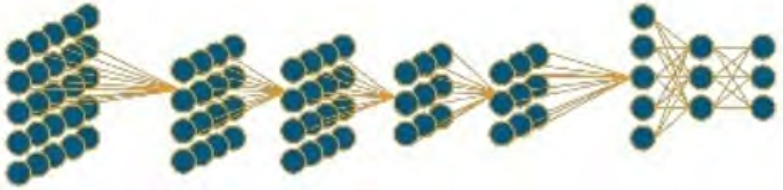


In some feature space

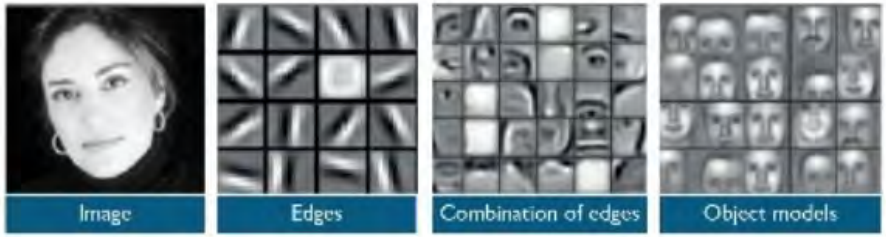
What is Deep Learning ?

Deep Learning is a subset of Machine Learning where similar Machine Learning Algorithms are used to train Deep Neural Networks so as to achieve better accuracy in those cases where the former was not performing up to the mark.

Basically, Deep learning mimics the way our brain functions i.e. it learns from experience.



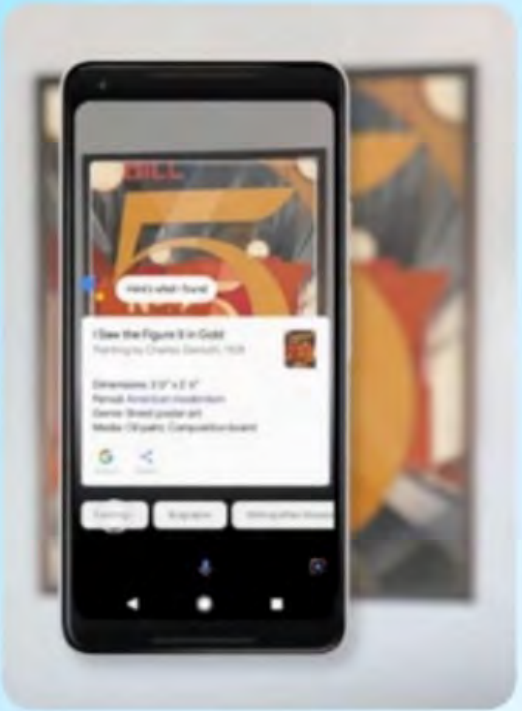
"Saha"



Applications of Deep Learning

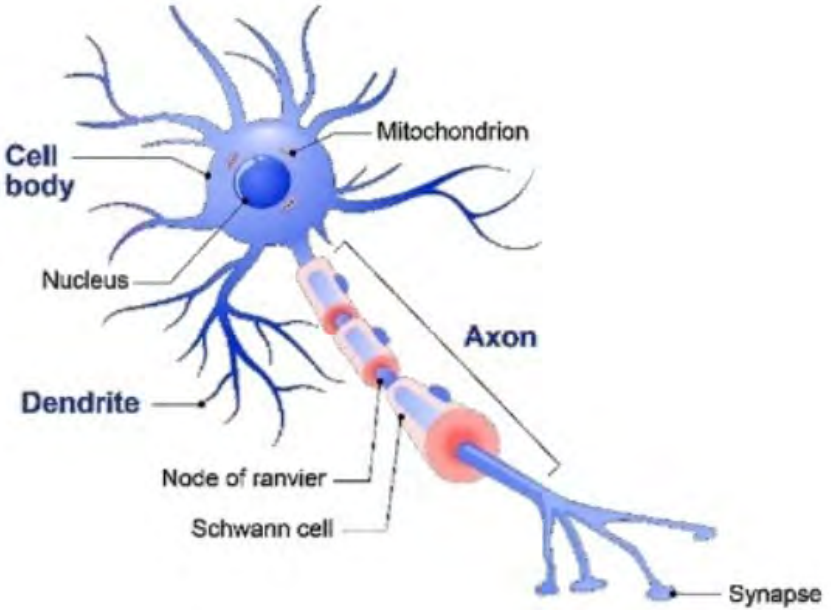
- Automatic Machine Translation
- Object Classification in Photographs
- Automatic Handwriting Generation
- Character Text Generation
- Image Caption Generation
- Colorization of Black and White Images
- Automatic Game Playing

Applications of Deep Learning



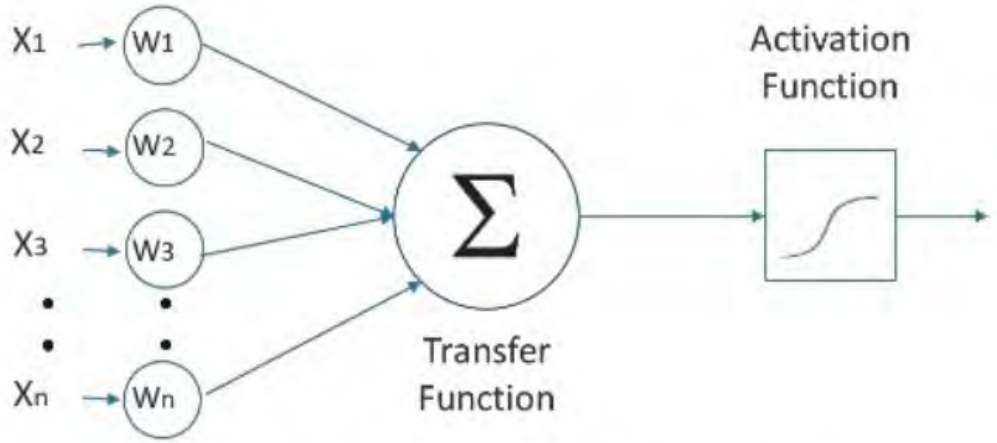
How Neuron Works?

Deep learning is a form of machine learning that uses a model of computing that's very much inspired by the structure of the brain, so lets understand that first.



A Perceptron

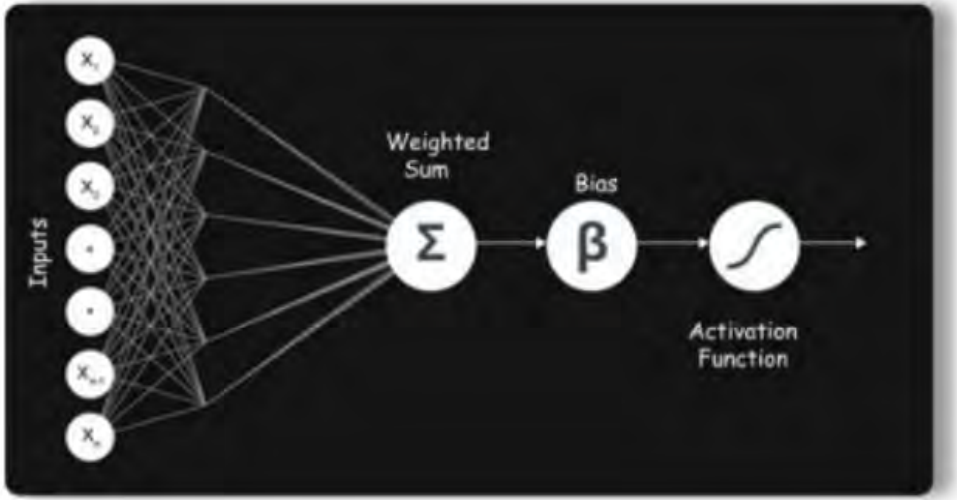
Each neuron has a set of inputs, each of which is given a specific weight. The neuron computes some function on these weighted inputs and gives the output.



Schematic for a neuron in a neural net

Role of Weights and Bias

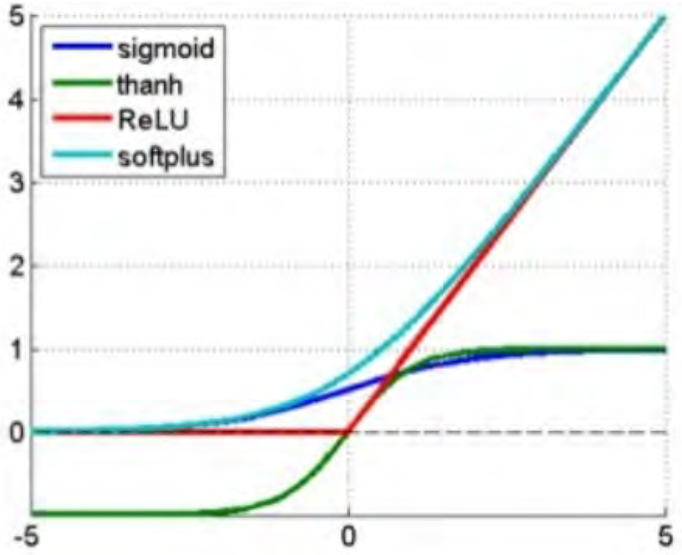
- For a perceptron, there can be one more input called bias
- While the weights determine the slope of the classifier line, bias allows us to shift the line towards left or right



Activation Functions

- Activation function translates the inputs into outputs
- It uses a threshold to produce an output

1. Linear or Identity
2. Unit or Binary Step
3. Sigmoid or Logistic
4. Tanh
5. ReLU
6. SoftMax



What are Tensors?

Tensors are the standard way of representing data in deep learning

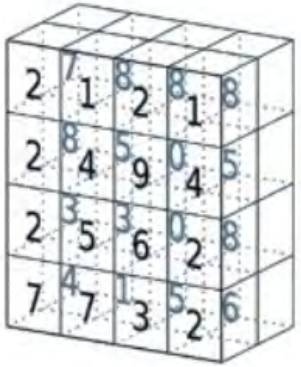
Tensors are just multidimensional arrays, an extension of 2-dimensional tables (matrices) to data with higher dimension.

't'
'e'
'n'
's'
'o'
'r'

Tensor 0f Dimension -6

3	1	4	1
5	9	2	6
5	3	5	8
9	7	9	3
2	3	8	4
6	2	6	4

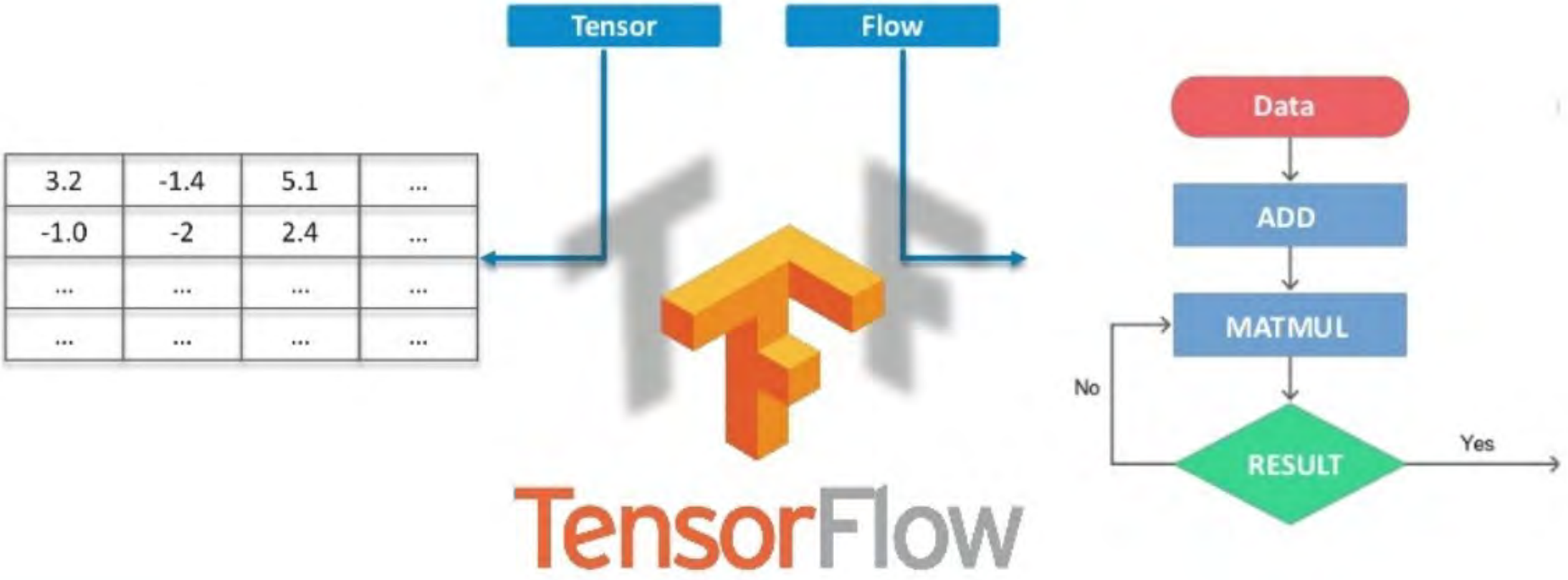
Tensor 0f Dimension –[6,4]



Tensor 0f Dimension –[6,4,2]

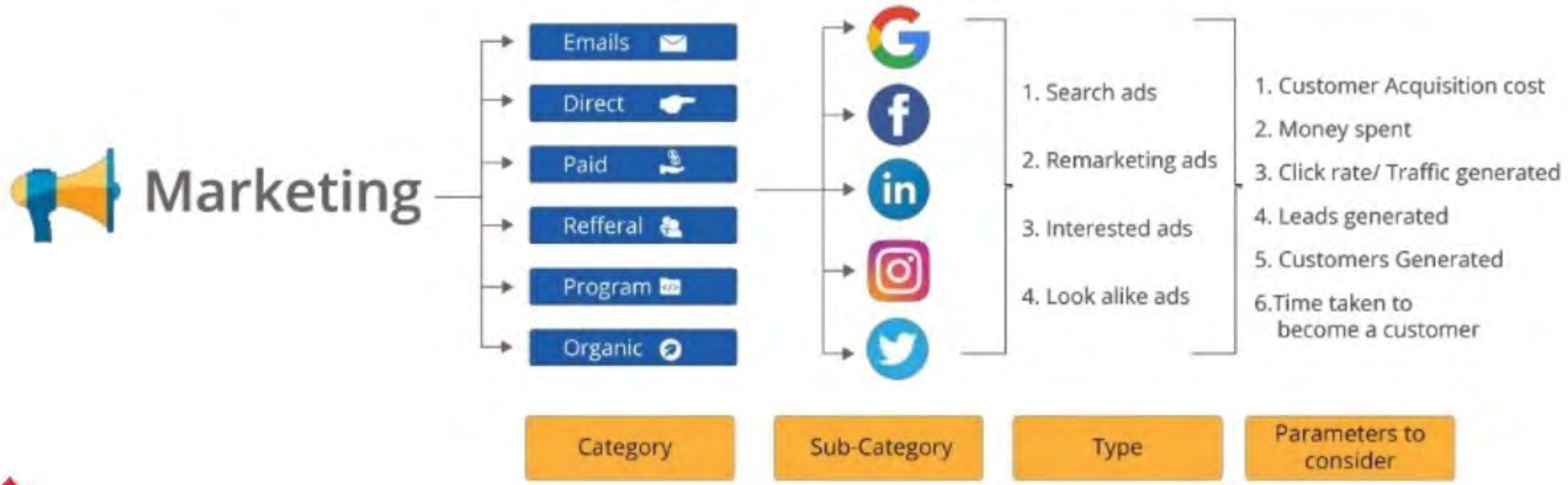
What is TensorFlow?

In Tensorflow, computation is approached as a dataflow graph

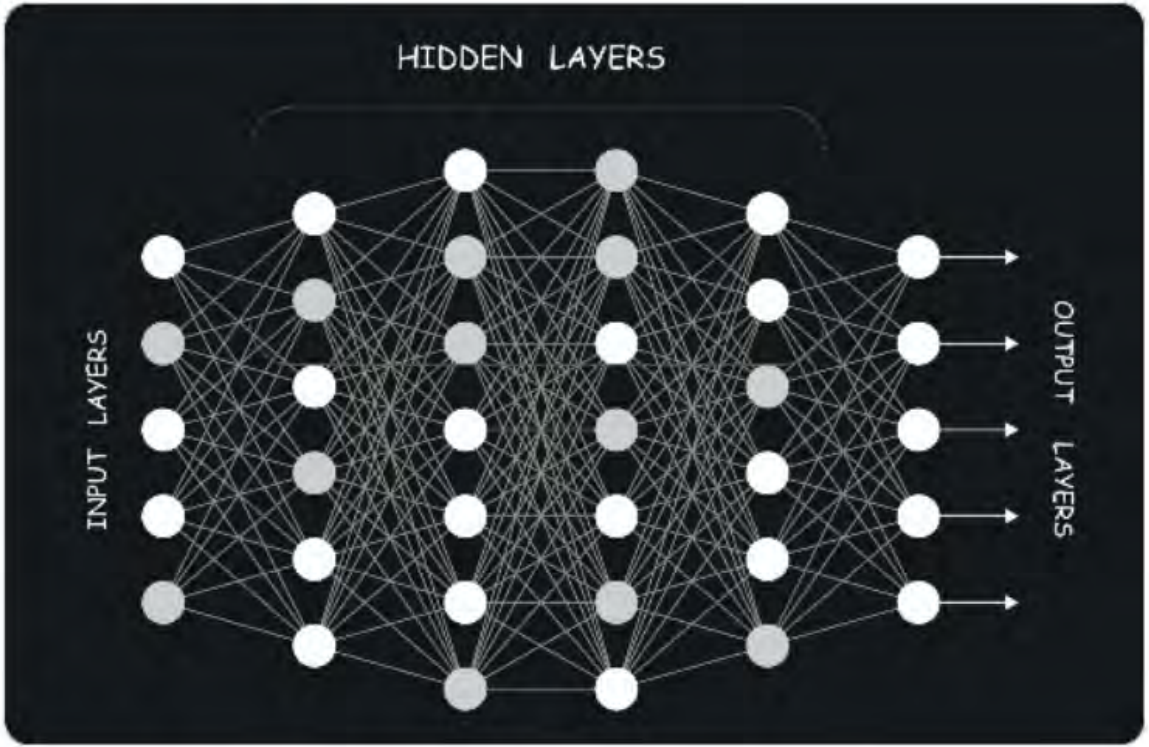


Perceptron Problems

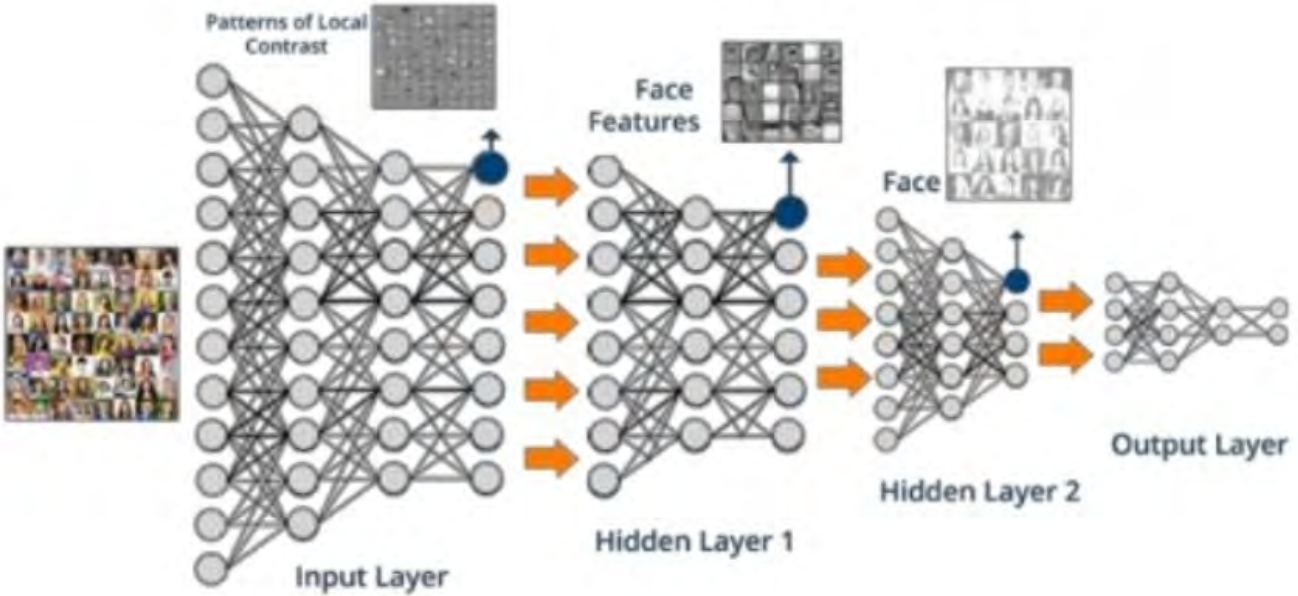
- Single-Layer Perceptrons **cannot classify non-linearly separable data points.**
- Complex problems, that involve **a lot of parameters** cannot be solved by Single-Layer Perceptrons.



Deep Neural Network



Deep Neural Network



Training Network Weights

- We can estimate the weight values for our training data using 'stochastic gradient descent' optimizer.
- Stochastic gradient descent requires two parameters:
- **Learning Rate**: Used to limit the amount each weight is corrected each time it is updated.
- **Epochs**: The number of times to run through the training data while updating the weight.
- These, along with the training data will be the arguments to the function.

MNIST Dataset

Deep Neural Network