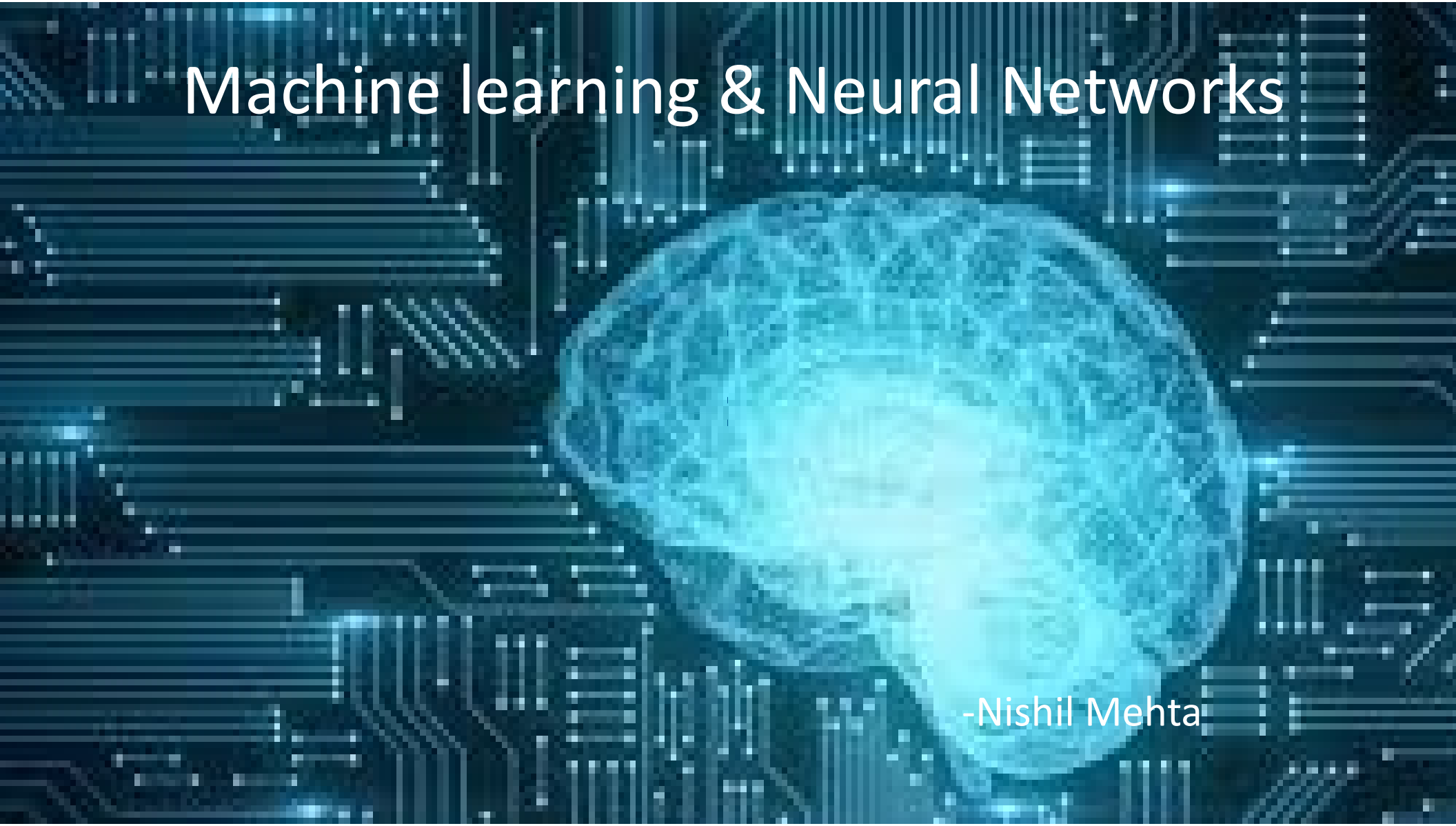

Machine learning & Neural Networks
-Nishil Mehta

Machine learning stems from data science

- Necessary for the effective interpretation of large amounts of data generated.
- Based on principles and techniques from a variety of fields of study, including computer science, statistics, algebra.
- Machine learning and artificial intelligence can be used as a tool for data analysis, pattern recognition, making predictions.

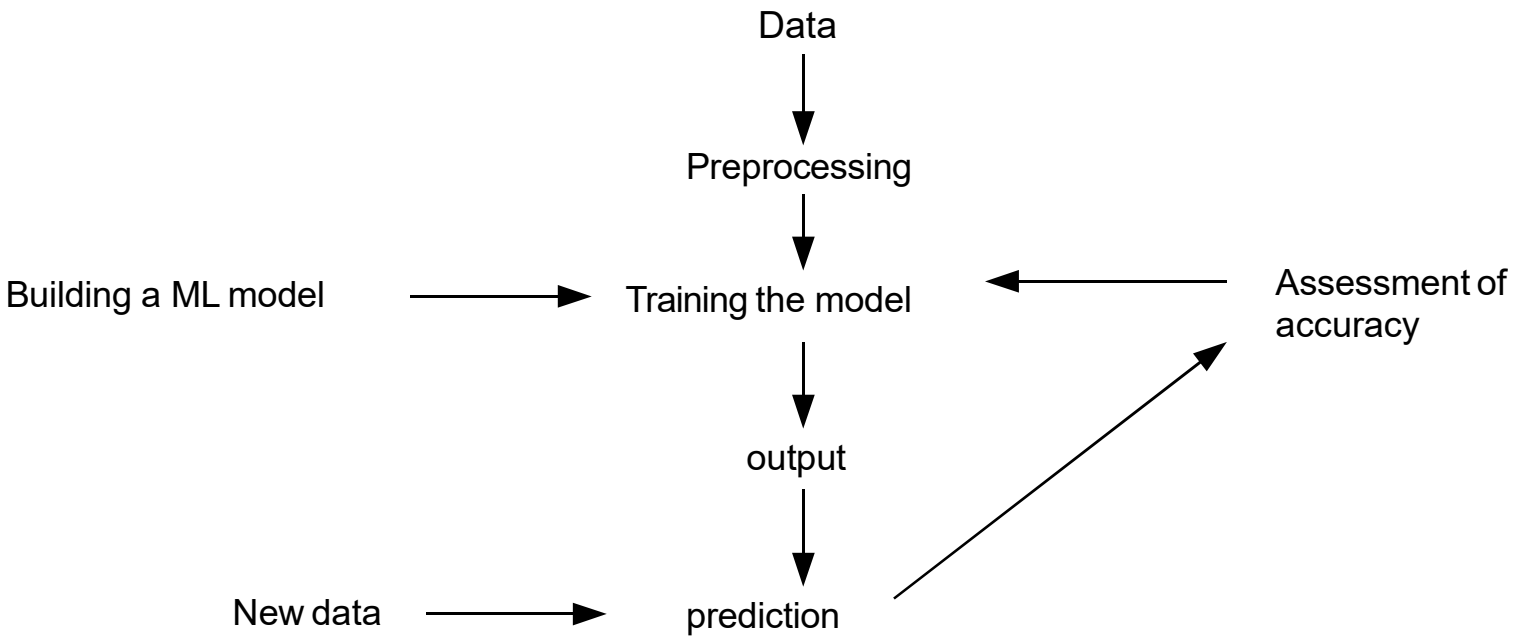
Clearing up some terms

Artificial Intelligence → Techniques enabling machines to mimic human behaviour

Machine Learning → Uses statistical methods to recognize patterns and improve experience

Deep Learning → Involves the computation of tasks by implementing neural networks.

General procedure followed

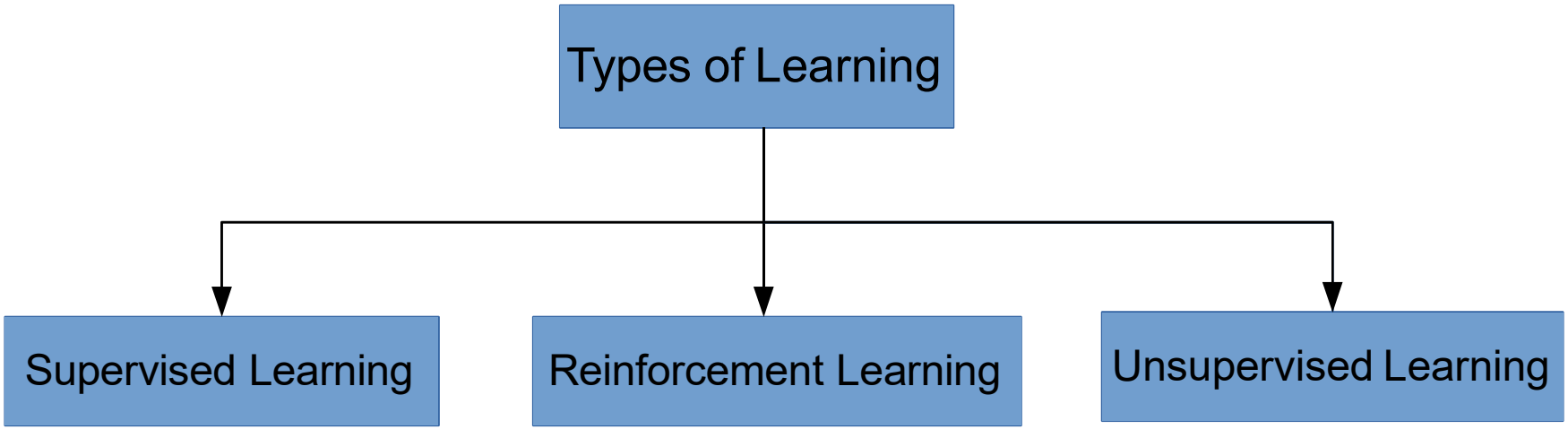


Some common applications

- Commerial Sector: Natural Language Processing, Image/Audio recognition and classification, Trading and Stock Markets, Banking and management sectors, Medical diagnosis.

- Physics: Image recognition in astronomy, prediction of properties of systems in condensed matter physics, simulations and quantum computing, data interpretation and representation in all fields

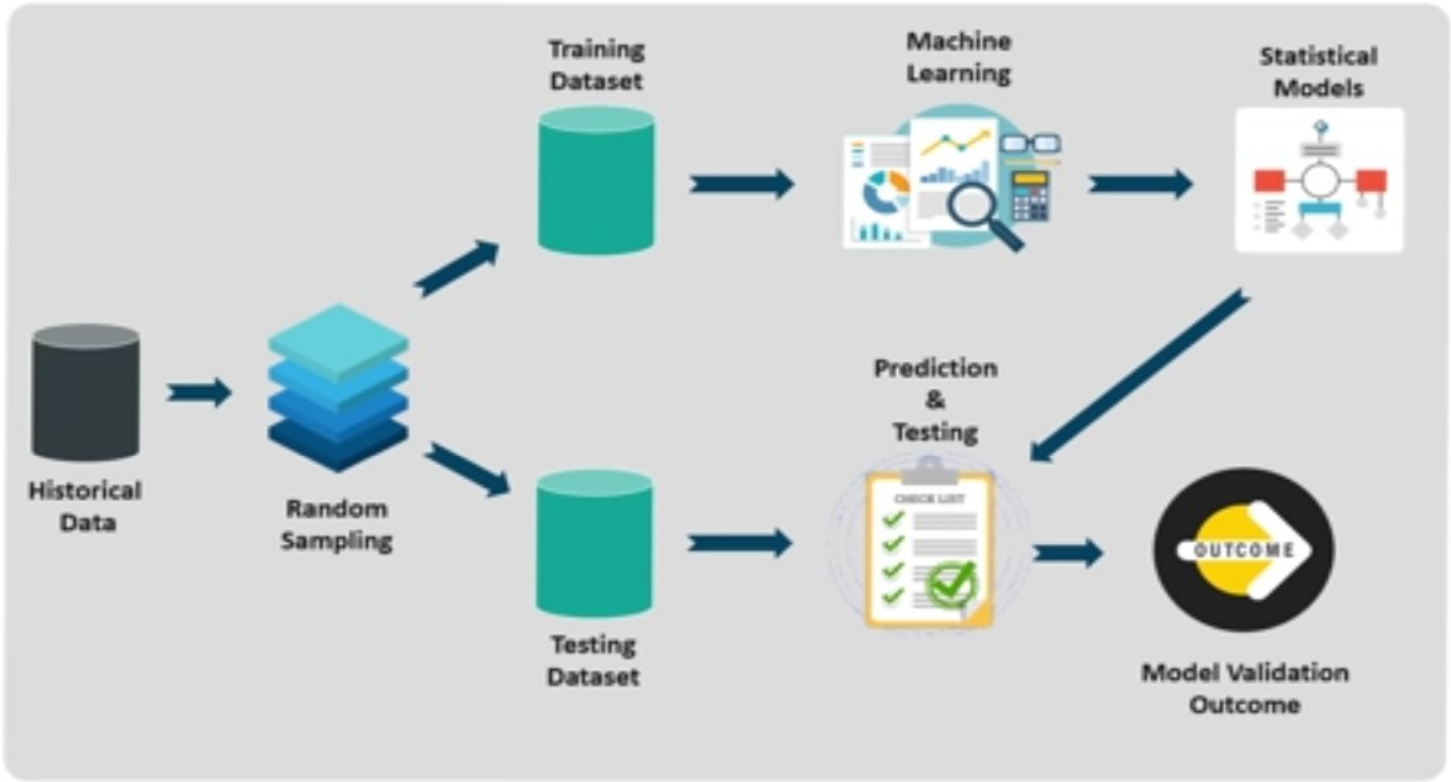
Types of Machine Learning



Supervised Learning

- It has input variable x and output variable y.
- The input is labeled. And the output is known.
- The algorithm processes the data and provides a mapping function from input to output.
- Eg:Facial recognition

- Training and Testing
- Prediction



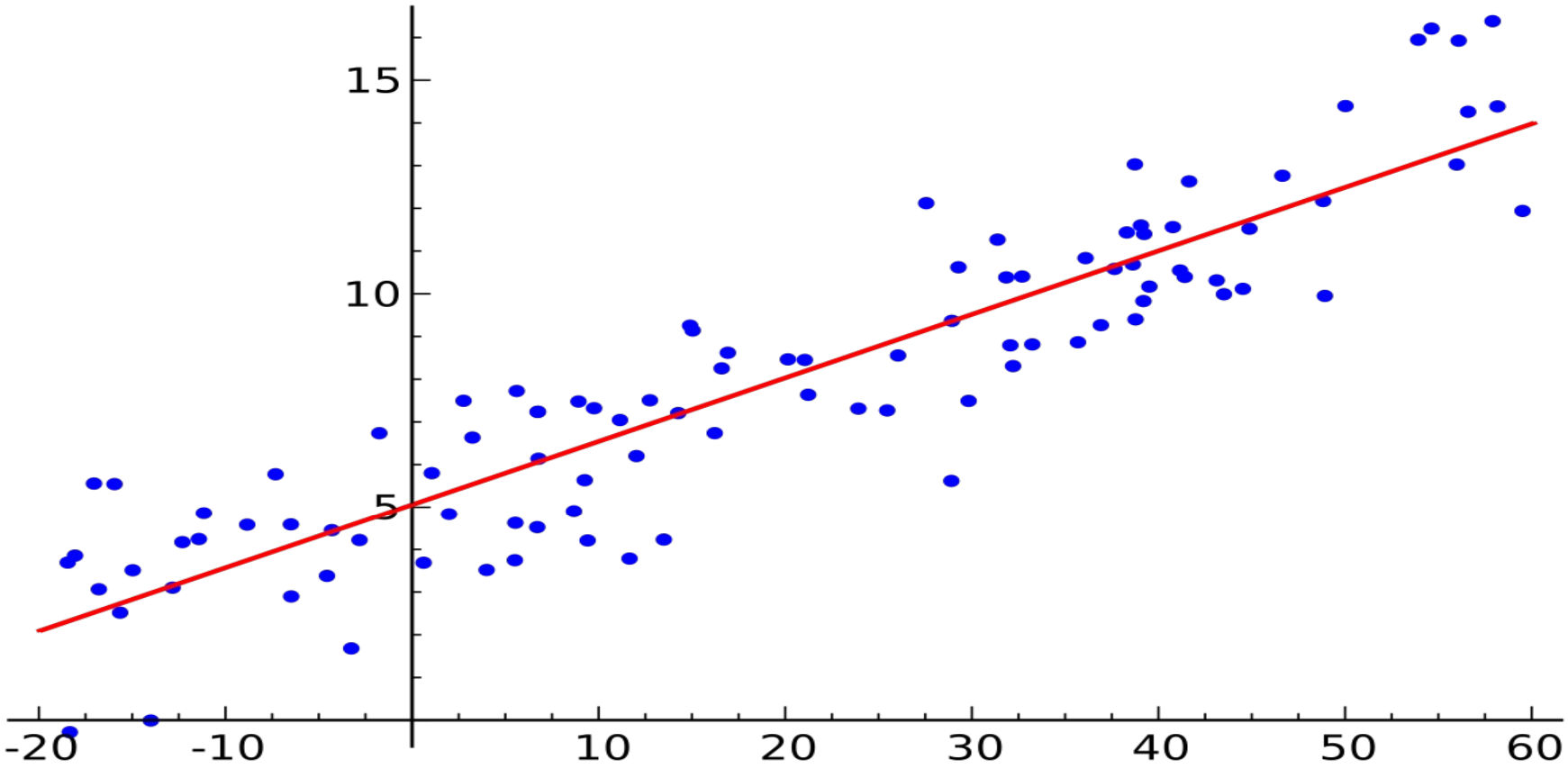
Types of supervised learning

- Linear Regression(real values)
- Logistic Regression(discrete values,boolean,binary)
- K-nearest neighbours
- Decision Tree(categorical and continuous dependent variables)
- Random forest(multiple decision trees)

Naive Bayes Classifier

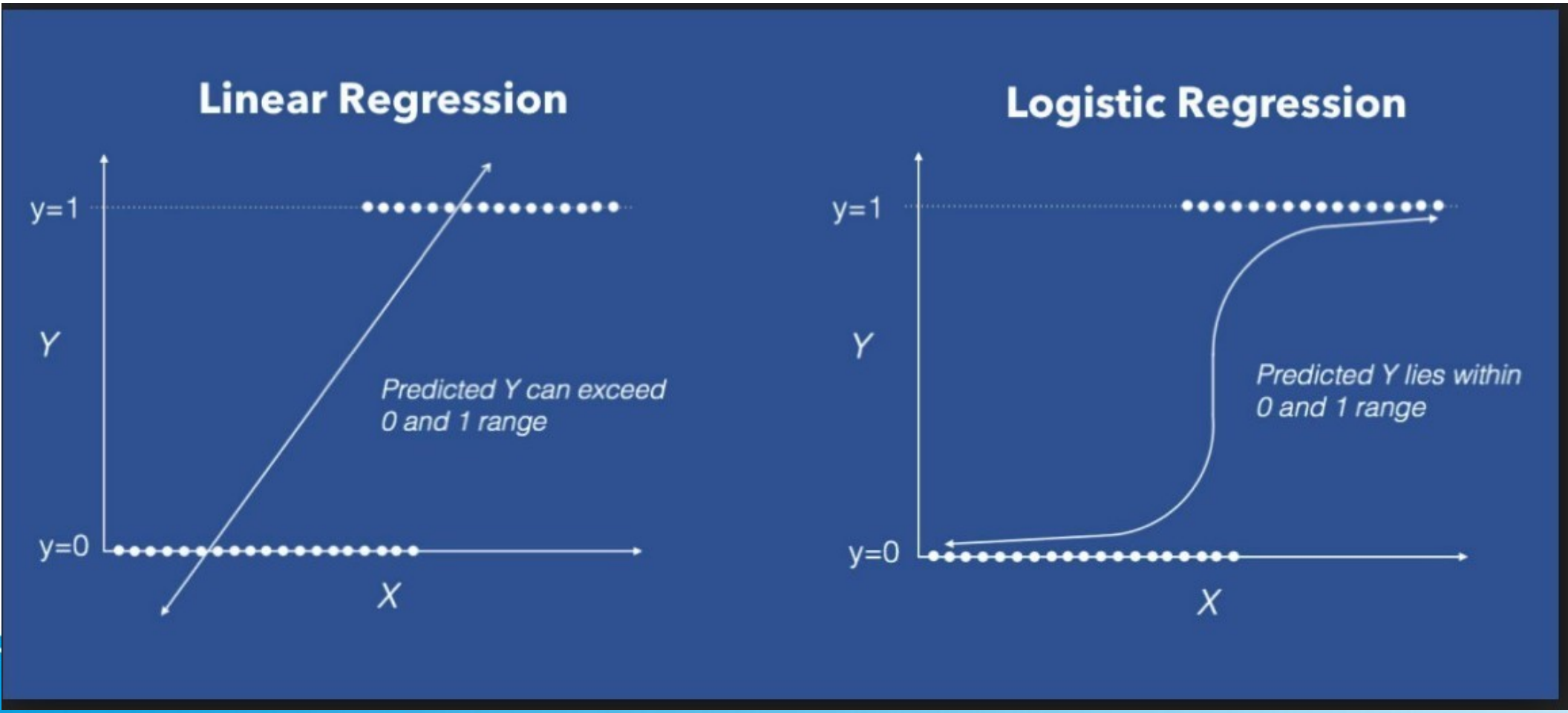
Linear Regression

- Linear regression is a technique used for predicting the unknown value of variable from known value of another variable.
- The algorithm learns the relationship between predictors and outcome so that it can predict future values by model fitting.
- For non linear data, polynomial regression is used.


15
10
5
-20
-10
10
20
30
40
50
60

Logistic Regression

- It usually used for the data where the outcomes is binary,i.e. true or false,win or lose.
- A threshold value is set for the sigmoid curve above which it is 1 or true and below which it is false or 0.

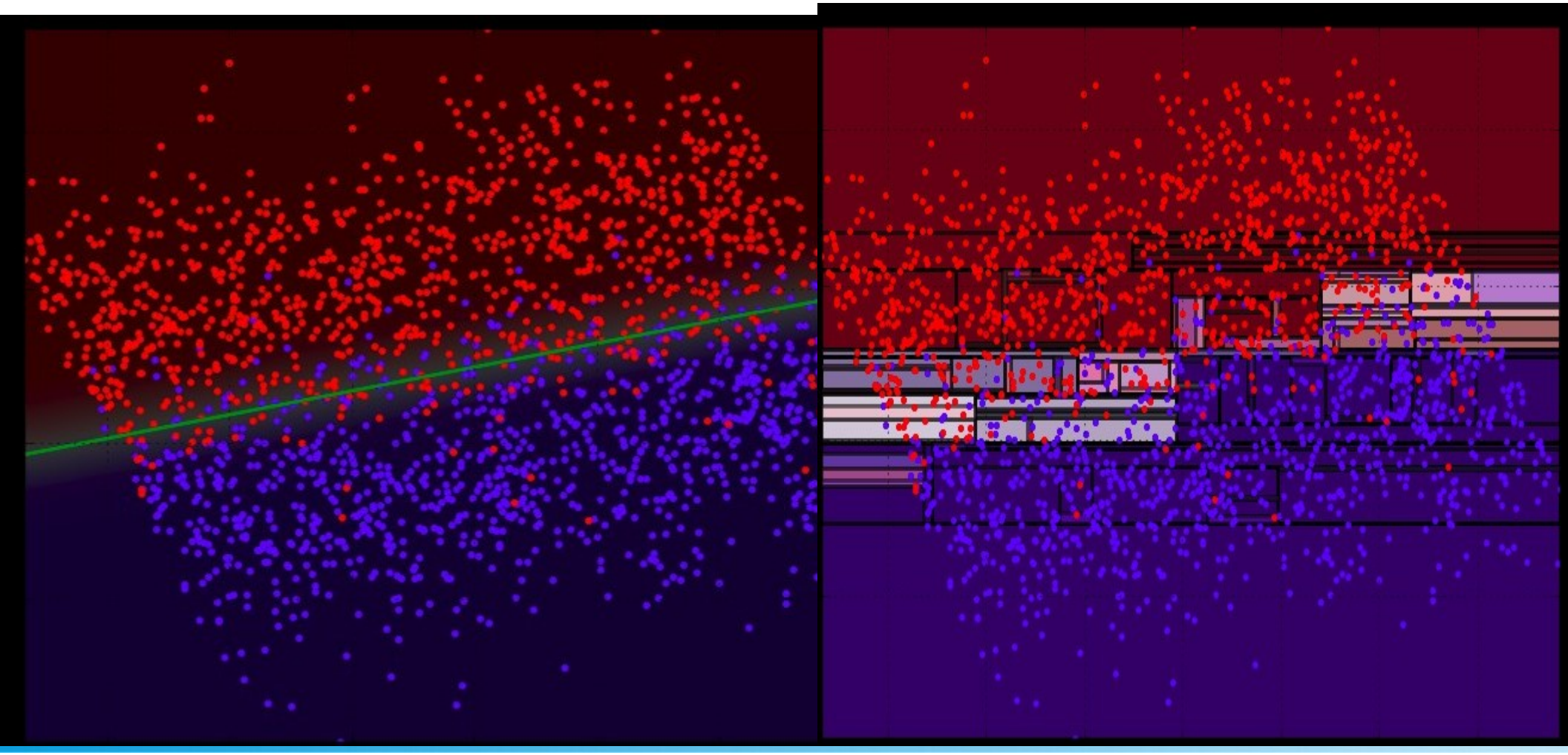

Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed
0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within
0 and 1 range

Decision Tree

Decision Trees bisect the space into smaller and smaller regions, whereas Logistic Regression fits a single line to divide the space exactly into two.

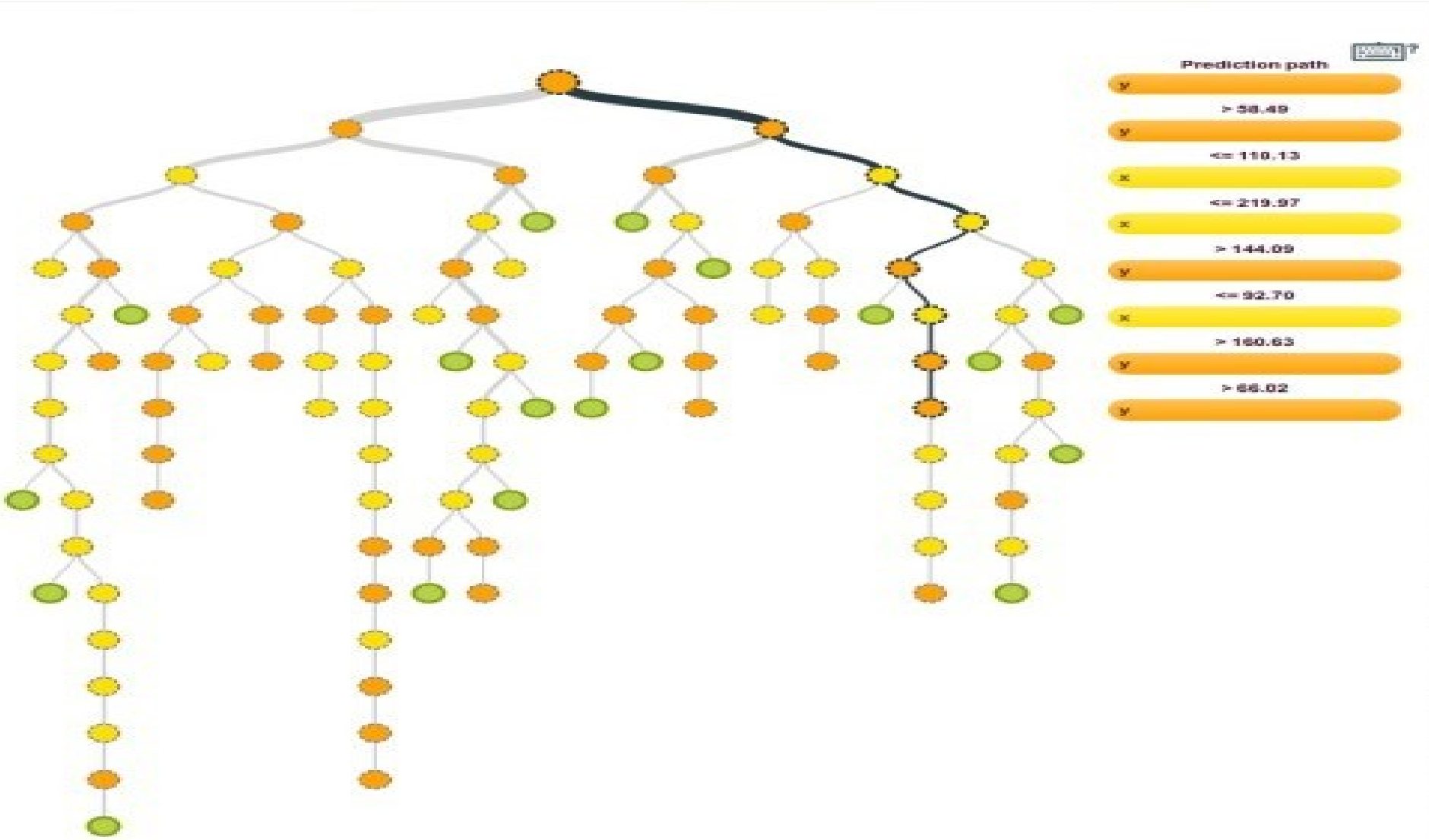
Logistic Regression

Decision Tree



Random Forest

Random Forests is an ensemble classifier which uses many decision tree models to predict the result.


Prediction path
y
> 58.49
y
<= 110.13
x
<= 219.97
x
> 144.09
y
<= 92.70
x
> 160.63
y
> 66.02
y

Naive Bayes Classifier

- It is powerful algorithm based on bayes theorem for predictive modelling.
- It assumes that presence of a particular feature in a class is independent of any other feature.
- All of these properties independently contribute to the probabilty of prediction.
- Eg. Weather prediction

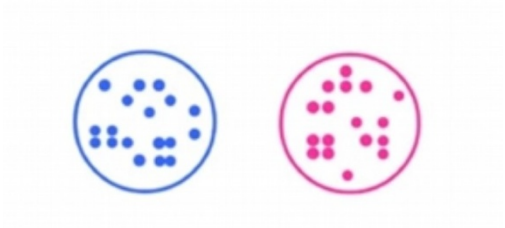
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Unsupervised Learning

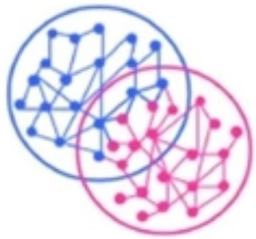
- It is training of model of data which neither classified nor labeled.
- This is used to cluster the data.
- The data of similar types is clustered based on features from the inputs.
- And then it can predict the input belongs to which cluster/group.

Types of Clustering

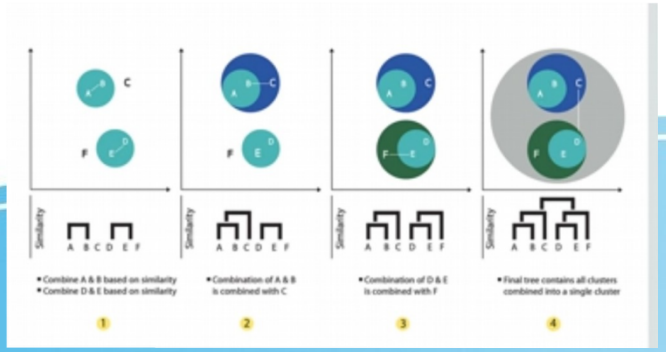
- Exclusive Clustering



- Overlapping Clustering



- Hierarchical Clustering

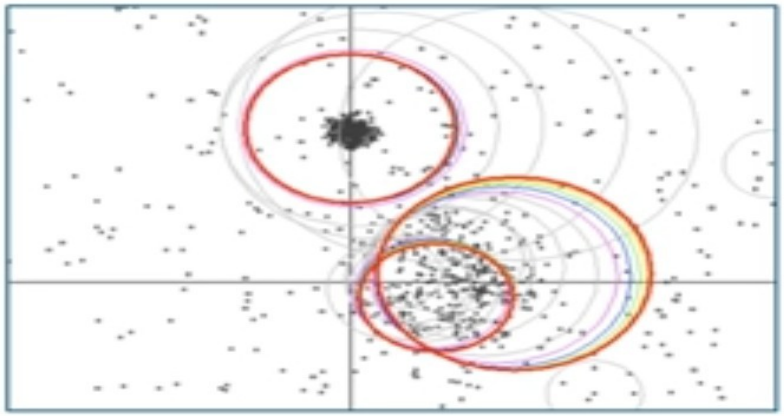


K-Means Clustering(Exclusive)

- Decision of number of clusters is to be made.
- The distance from the randomly chosen centroid to nearest data points is measured.
- The location of centroid is updated to reduce the distance and then again the process is repeated until the centroids are no longer updating the position.
- Eg. Market basket algorithm, Apriori algorithm.

Fuzzy C-means Clustering (Overlapping)

- Instead of measuring the euclidean distance,in this method a circle with certain radius is used.
- A data set can belong to 2 or more cluster and then a degree of membership (0 to 1) is assigned
- Eg. Better representation of nature of behaviour of genes.
- A cut-off membership value is to be assigned.
- Eg. Netflix recommendations.



Reinforcement Learning

- It is learning by interacting with space or environment.
- It selects action by past experience and by new choices. It learns by the consequence of the action.

Reinforcement Learning

- The main aim is to maximize the reward and minimize the cost.
- A reward is instant return from the environment to appraise the last action.
- Cost is usually the computational cost
- Exploitation is get instant rewards.
- Exploration is to get more infomation about the environment.
- The Epsilon Greedy algorithm takes whatever action seems best at the moment.
- Its chooses between exploitation and exploration to maximize the reward.

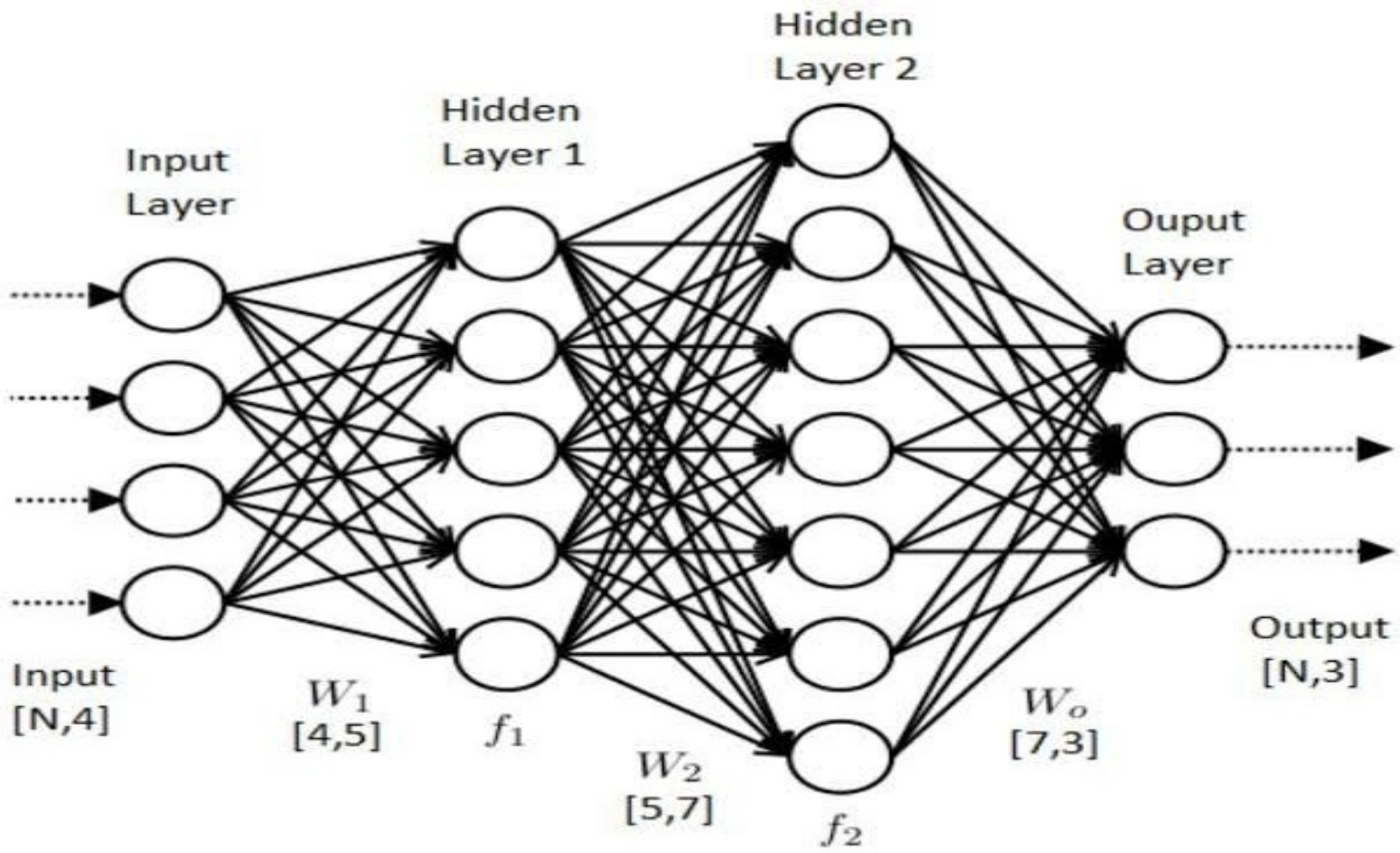
Deep Learning (Neural Network)

The word neural comes from the biological word neruon, which means brain cell. It is so called because the architecture and working of neural network is inspired by the brain. Neural network has a non linear approach to fit data.

Architecture

A neural network consists of the following:

- Neurons (Perceptron)
- Weights
- Biases
- Activation Function


Hidden
Layer 2
Hidden
Layer 1
Input
Layer
Ouput
Layer
Output
[N,3]
Input
[N,4]
W_1
[4,5]
f_1
W_2
[5,7]
f_2
W_o
[7,3]

- **NEURON**

  A neuron can be any variable which stores data or numbers between 0 and 1 (circles in the diagram represent a neuron). Neurons are arranged in two or more layers. The first one is called the input layer (Features). The middle ones are called hidden layers and the last one is called the output layer (Labels).

- **WEIGHTS**

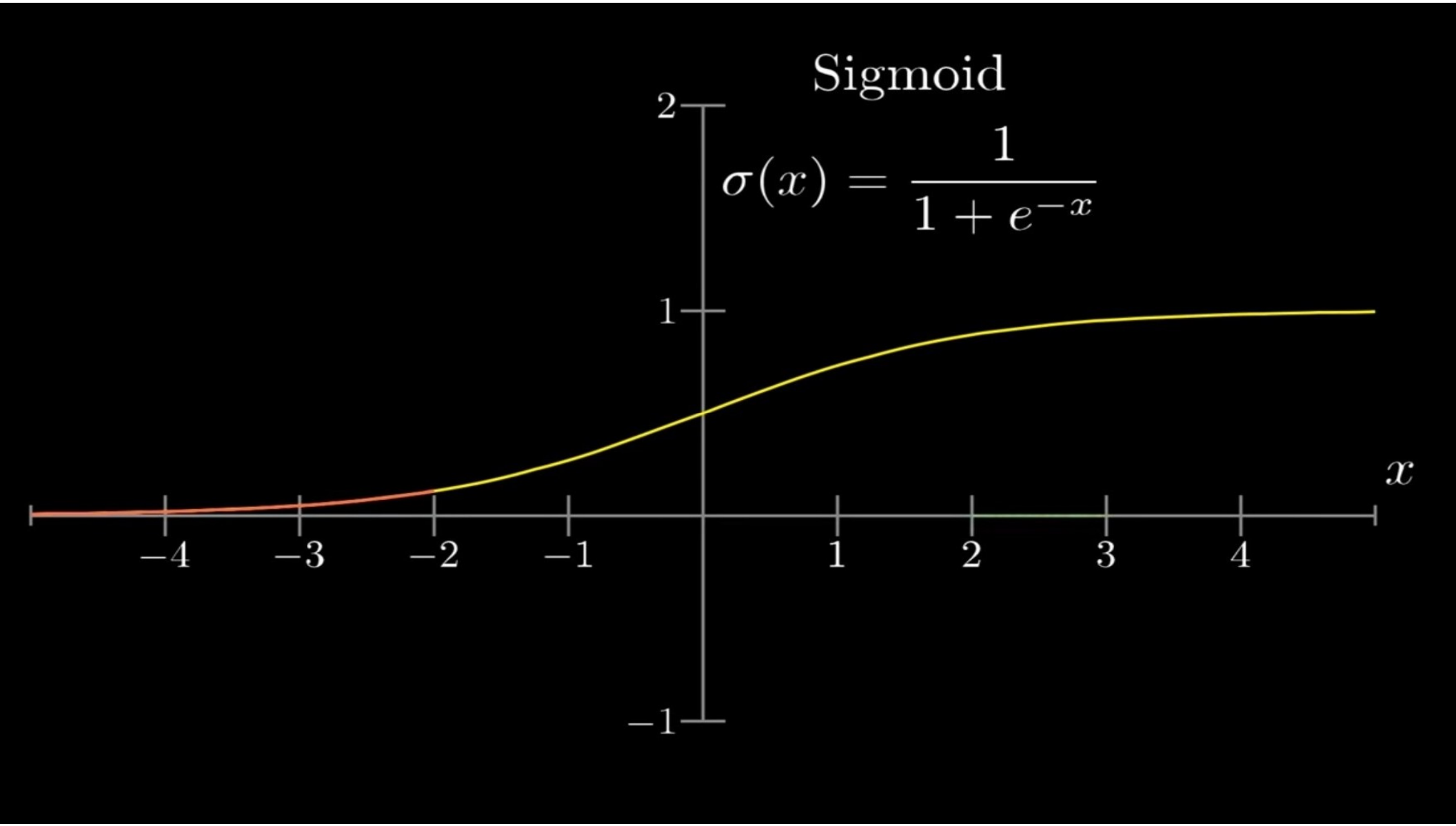
  The neurons are connected via synapses, here weights. Weights is an n dimensional matrix of fractional numbers. They can be understood as the strength of a synapse. They are highly flexible. These weights are multiplied (Matrix multiplication) to the inputs to receive an output.

- **BIASES**

  Bias is a number which acts like a threshold value. It is added to the weighted sum of the neurons, before aplying the activtion function

- **ACTIVATION FUNCTION**

  To the above mentioned output an activation funtion is applied. An activation funtion is primarily responsible for introducing non-linearity in the system. It is the abstraction representing the rate of action potential firing in the cell. It gives an output value between 0 and 1. There are several activation functions: Sigmoid Function, Hyperbolic tangent function.


Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$
2
1
−1
x
−4
−3
−2
−1
1
2
3
4

CALCULATIONS

Let a_n be n number of neurons in the input layer.

Let w_{mn} be n weights associated to the synapses leading to a particular neuron in the next layer having m neurons

Then the weighted sum + Bias (b)= $a_1 w_{m1} + a_2 w_{m2} + a_3 w_{m3} + \ldots\ldots + a_n w_{mn} + b$

Applying the activation (A) function to the same we get,

$$A(a_1 w_{m1} + a_2 w_{m2} + a_3 w_{m3} + \ldots\ldots + a_n w_{mn} + b)$$

This acts as an input to a particular neuron in the next layer.

Matrix Representation:

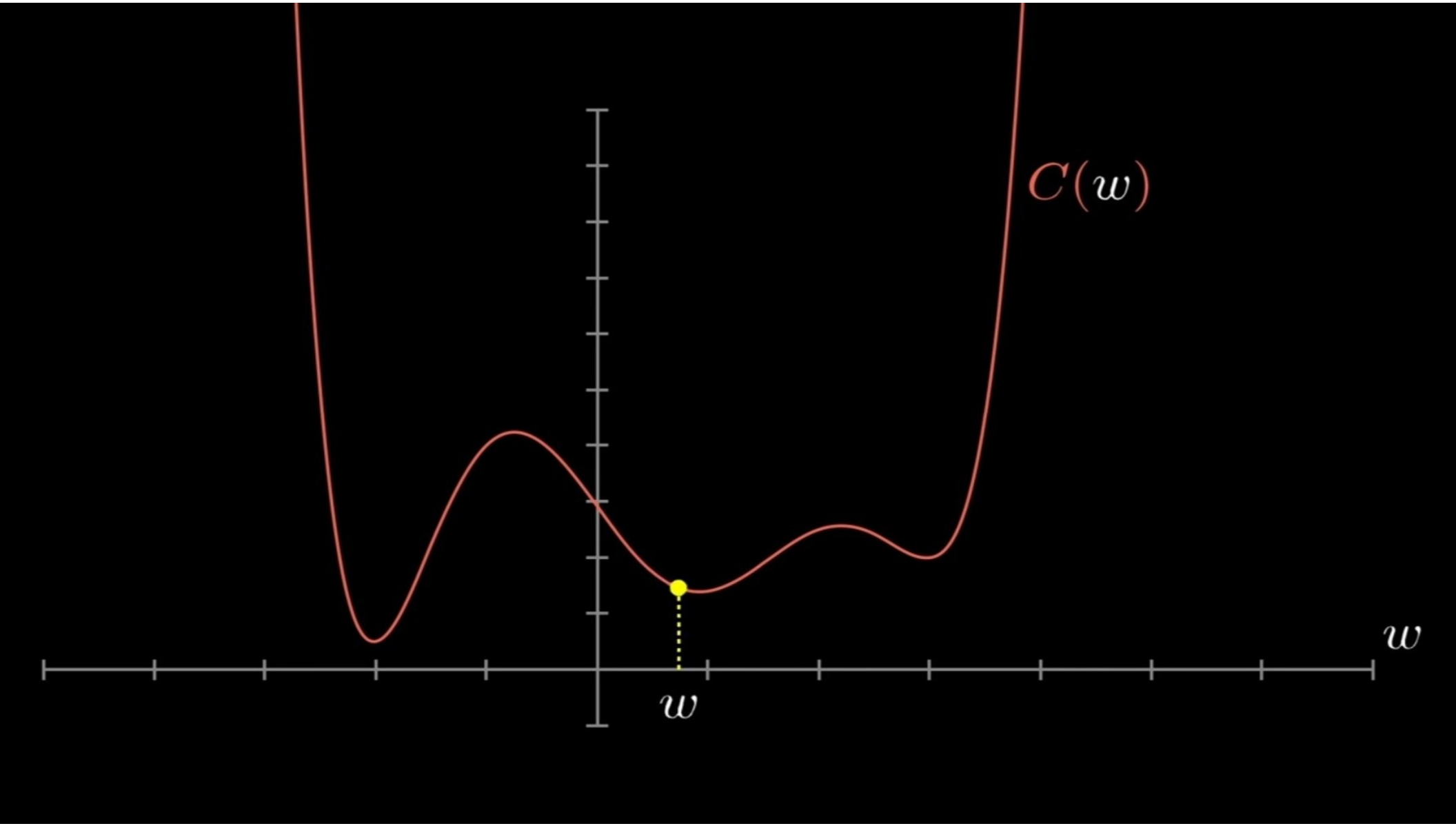
$$A\left(\begin{bmatrix} w_{11} & w_{12} & w_{13} & \ldots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \ldots & w_{2n} \\ \ldots & & & & \\ \ldots & & & & \\ w_{m1} & w_{m2} & w_{m3} & \ldots & w_{mn} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix} \right)$$

The output of this function acts like an input to the next layer.

- Similar matrices are calculated till the output layer is reached.
- Then the error is calculated:

  error = output achieved – expected output
- A function is used to calculate the error, called the **Cost function**
- Squared error is used to compute the cost function.
- Then the primary job of the system is to minimize the cost function. So that we get accurate results from the network.
- For this we need to find the global minimum value of the cost function. To find the global minimum we can find the negative gradient of the cost function: - ∇C and then update the weights accordingly to decrease the error.

  Performing the above process repeatedly is called **Gradient descent.**


$C(w)$
w
w

NOITAGAPORP KCAB

- Back Propogation is an algorithm which adjusts the weights and biases of the network such as to decrease the cost. It not only increases or decreases the weights and biases but also does it such a manner which causes the most rapid decrease in the cost function.

The error is calculated by mean sum squared loss. To figure out which direction to alter our weights found the rate of change of loss (partial derivative). This method is called gradient descent. The difference between predicted and actual output is taken and the derivative of the sigmoid activation is applied to output error. This helps us to figure out the error produced by the previous layer and the process goes on. As the process goes from the output layer to the input layer it is called backward propagation.