What Is Data Science?

What is Data Science?

“Torture the data, and it will confess to anything.”
~ Ronald Coase, Economics, Nobel Prize

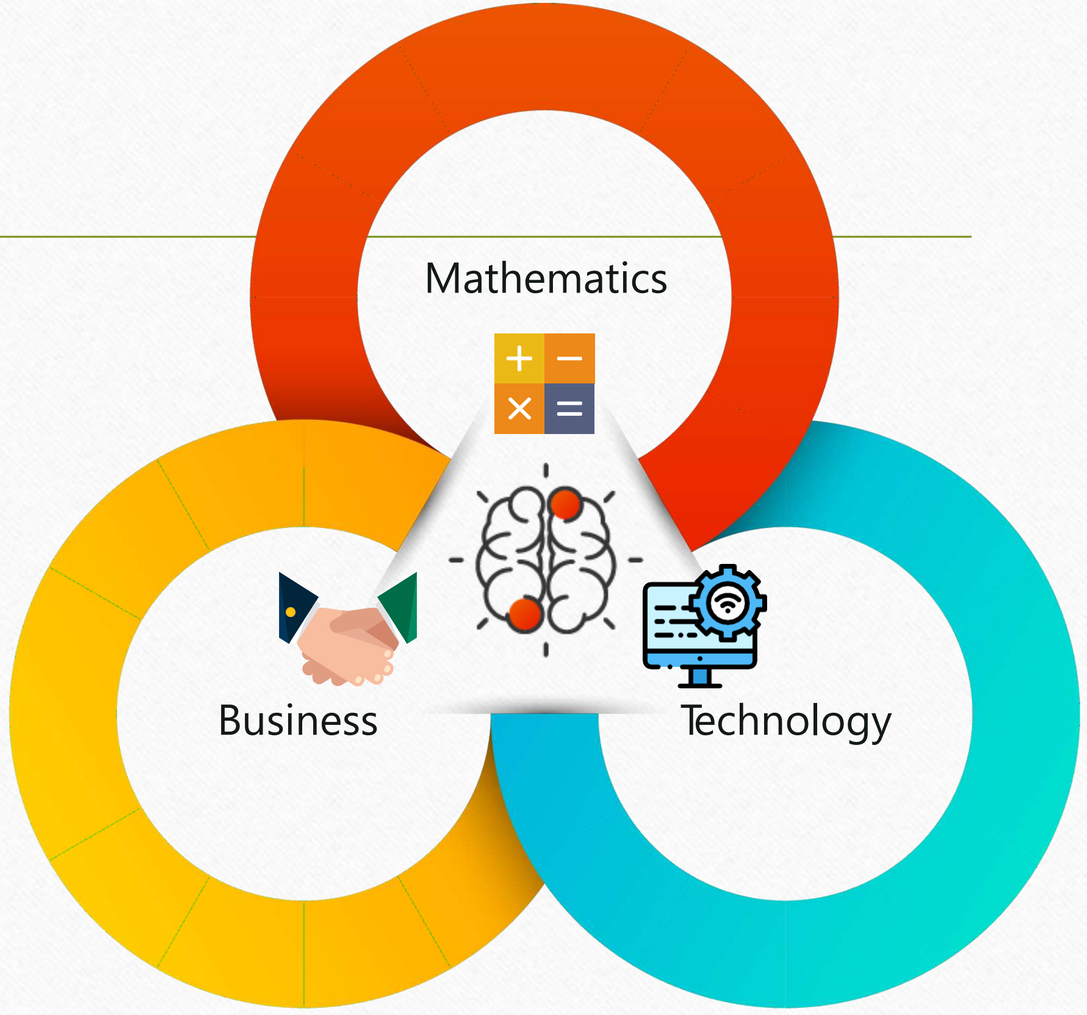
Data Science is the process of extracting knowledge and insights from data by using *scientific methods*.

Scientific methods:

Programming + **Statistics** + **Business**

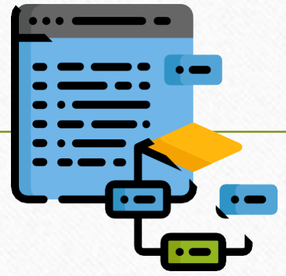
Who Is A Data Scientist?

Who Is A Data Scientist?

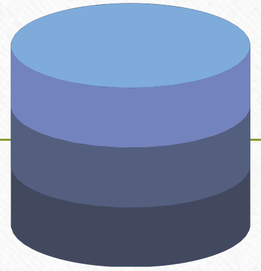


Data Science – Skill Set

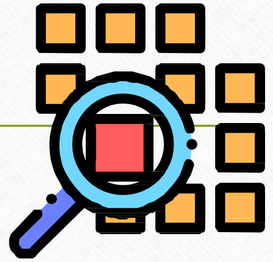
Statistics



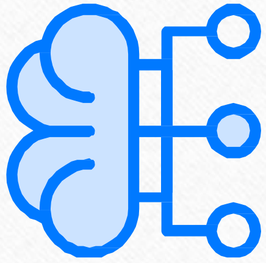
Programming languages



Data extraction & processing



Data wrangling & exploration



Machine Learning

Big Data processing frameworks

Data visualisation

Data Science Job Roles

Data Science Job Roles

Data Scientist

Data Analyst

Data Architect

Data Engineer

Statistician

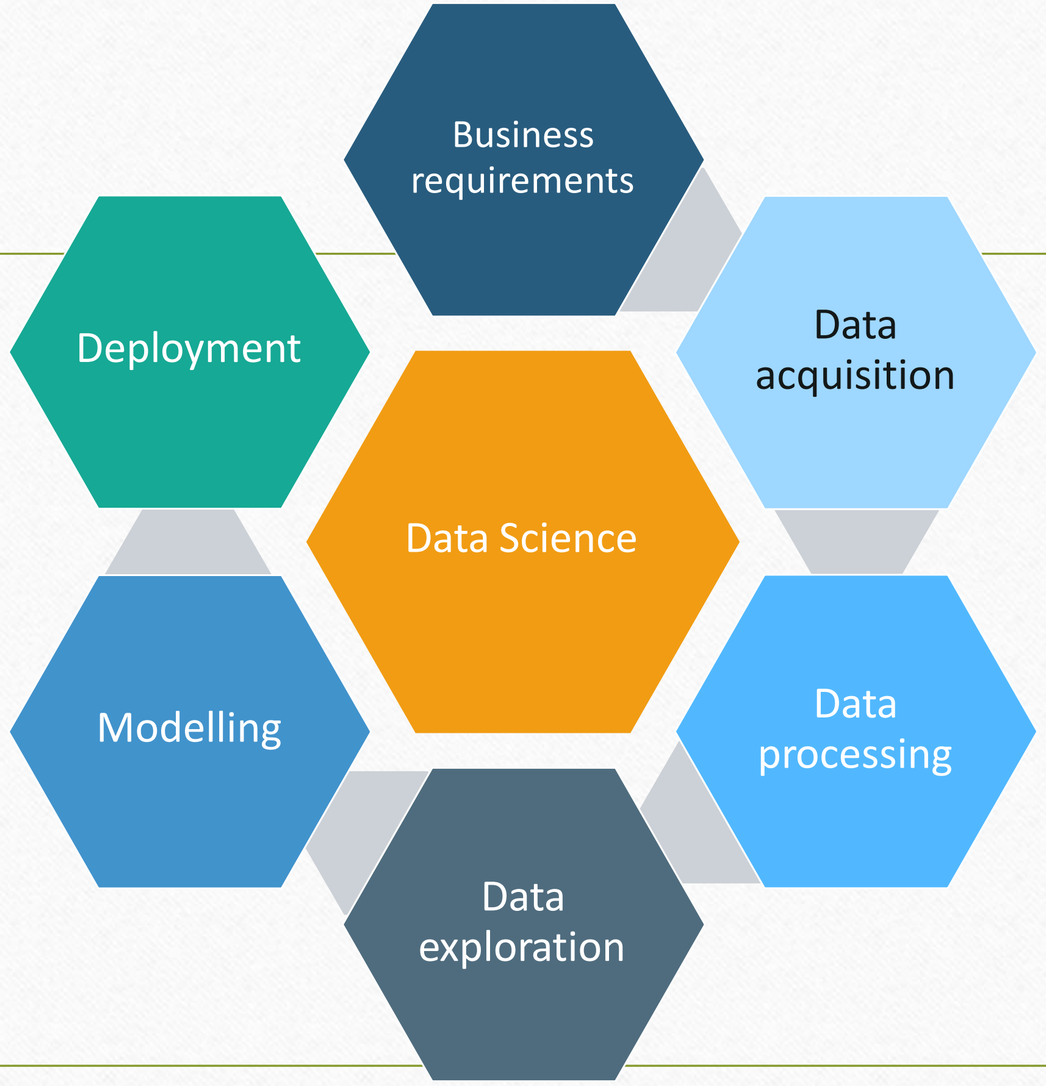
Database Administrator

Business Analyst

Data & Analytics Manager

Data Science Life Cycle

Data Life Cycle



Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Understand the problem

Identify central objectives

Identify variables that need to be predicted

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

What data do I need for my project?

What are the data sources?

How can I obtain the data?

What is the most efficient way to store and access all of it?

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Transform data into desired format

Data cleaning

- Missing values
- Corrupted data
- Remove unnecessary data

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

understand the patterns in the data

Retrieve useful insight

form hypotheses

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Determine optimal data features for the machine-learning model

Create a model that predicts the target most accurately

Evaluate & test the efficiency of the model

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Check the deployment environment for dependency issues

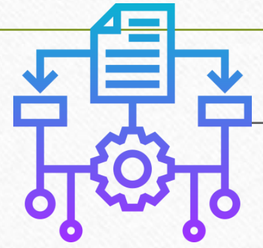
Deploy the model in a pre-production/ test environment

Monitor the performance

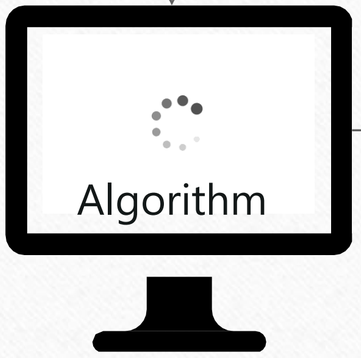
Introduction To Machine Learning

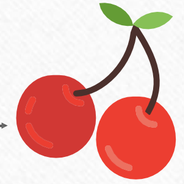
What Is Machine Learning?

Machine learning is a subset of artificial intelligence (AI) which provides machines the ability to learn automatically & improve from experience without being explicitly programmed.



Data





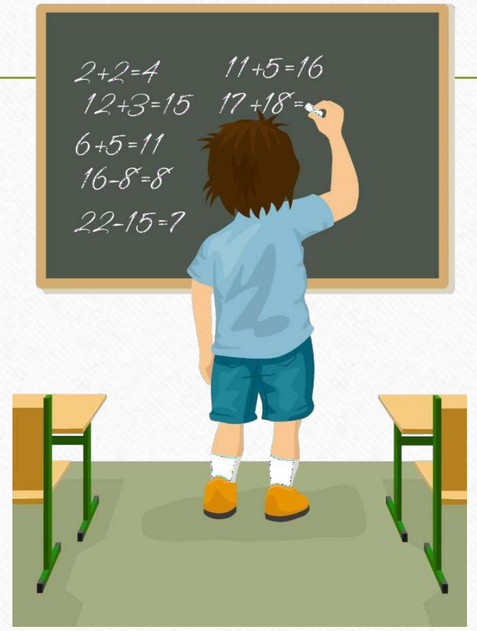
Cherry

Apple

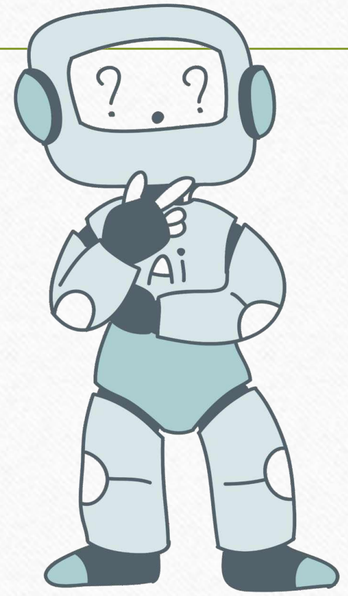
Orange

Types Of Machine Learning

Supervised Learning



Unsupervised Learning



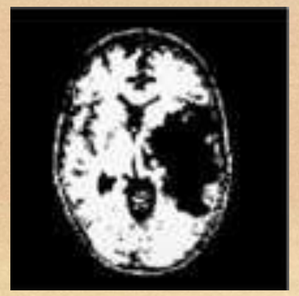
Reinforcement Learning

K – Means Use Case

Brain Tumour Detection Using K - means

K-Means clustering is an *unsupervised learning* algorithm used to partition *a dataset* into *k* clusters in which each data point belongs to the cluster with the nearest mean.

Brain tumour segmentation deals with the implementation of the k-means algorithm for detection of range and shape of tumour in brain MR images.



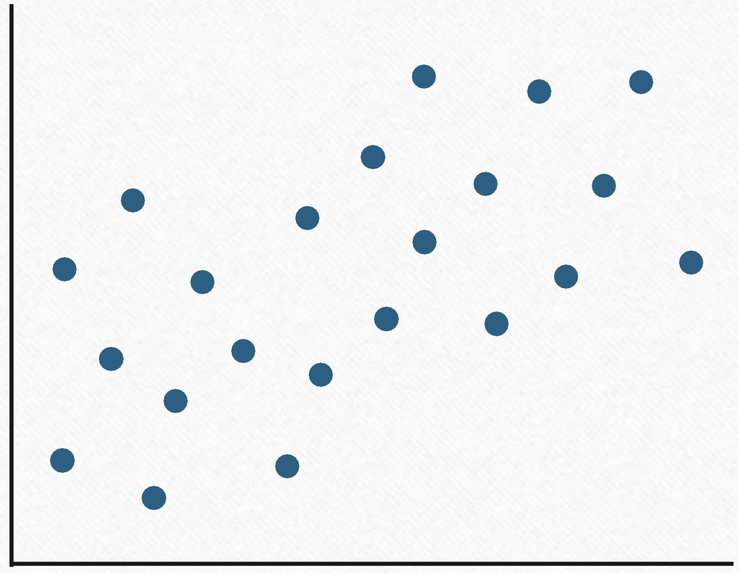
K – Means Algorithm

K – Means Algorithm

- Initialization
- Cluster assignment
- Move centroid
- Optimization
- Convergence

➢Randomly initialize *k* points called the cluster centroids. Here, k = 2

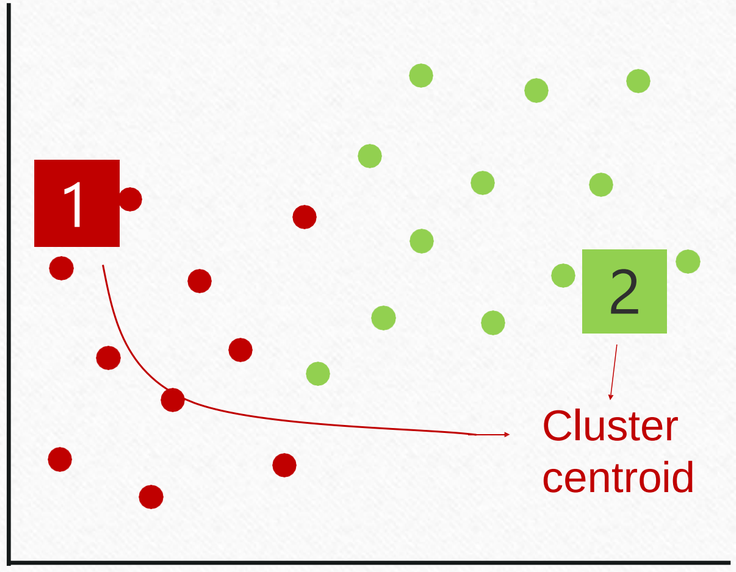
➢Value of k(number of clusters) can be determined by the elbow curve.



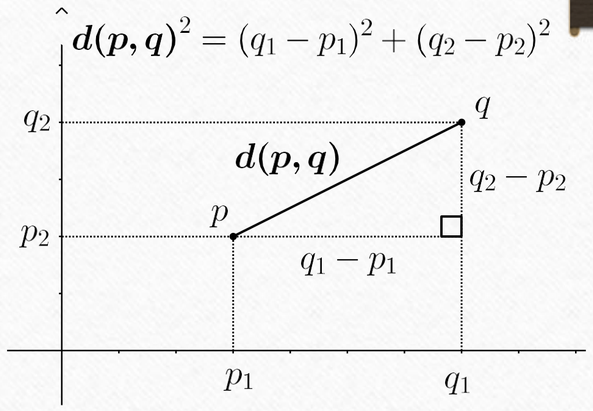
K – Means Algorithm

- Initialization
- Cluster assignment
- Move centroid
- Optimization
- Convergence

- Compute the distance between the data points and the cluster centroid initialized.
- Depending upon the minimum distance, data points are divided into two groups.



$$d(p,q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2$$

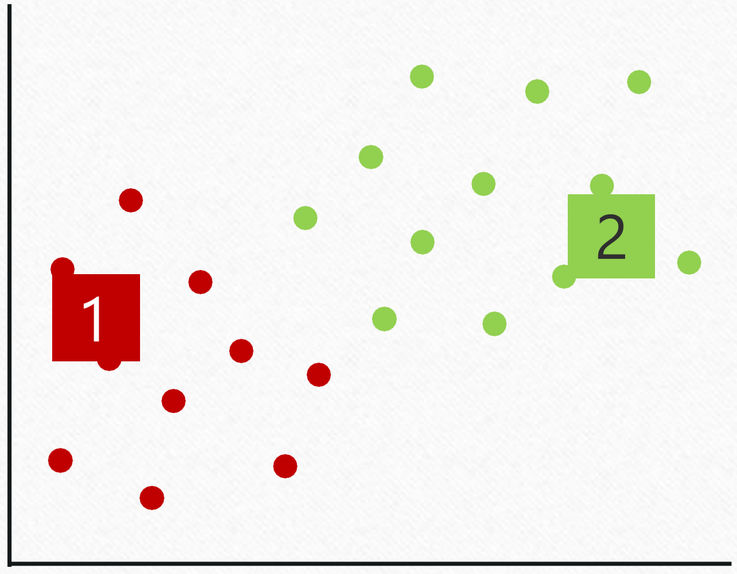


Euclidean distance

K – Means Algorithm

- Initialization
- Cluster assignment
- Move centroid
- Optimization
- Convergence

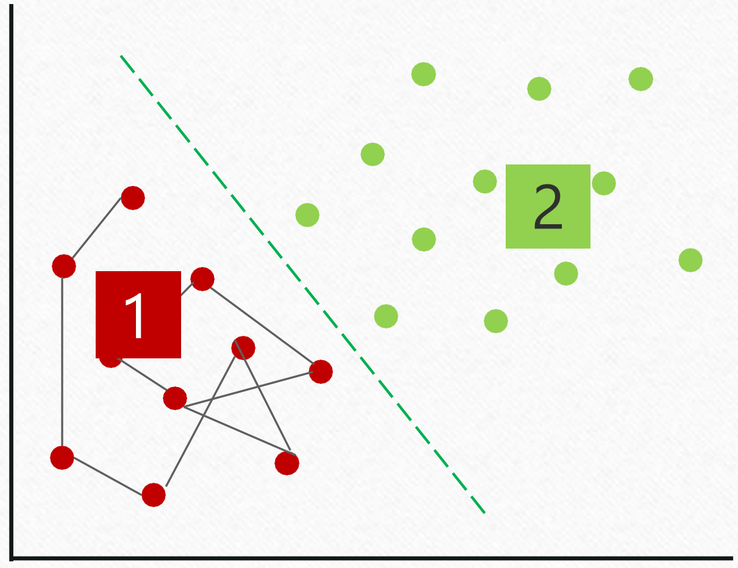
- Compute mean of red dots & reposition red cluster centroid to this mean
- Compute mean of green dots & reposition green cluster centroid to this mean.



K – Means Algorithm

- Initialization
- Cluster assignment
- Move centroid
- Optimization
- Convergence

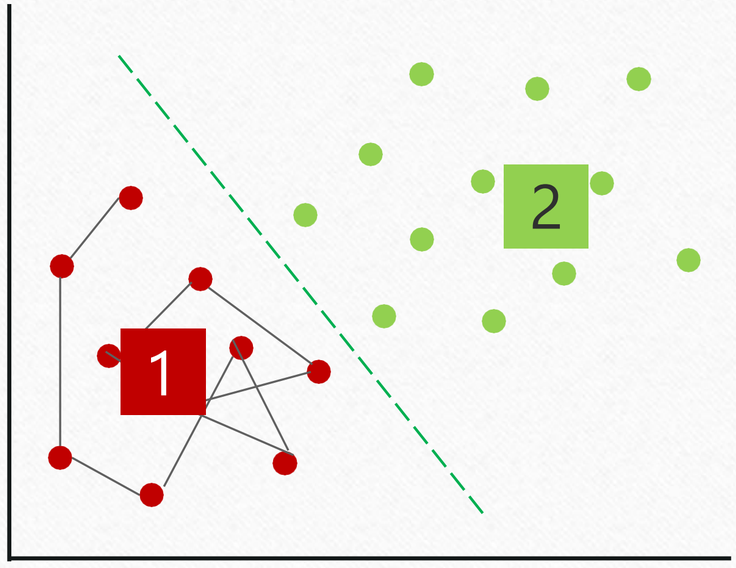
➢Repeat previous two steps iteratively till the cluster centroids stop changing their positions.



K – Means Algorithm

- Initialization
- Cluster assignment
- Move centroid
- Optimization
- Convergence

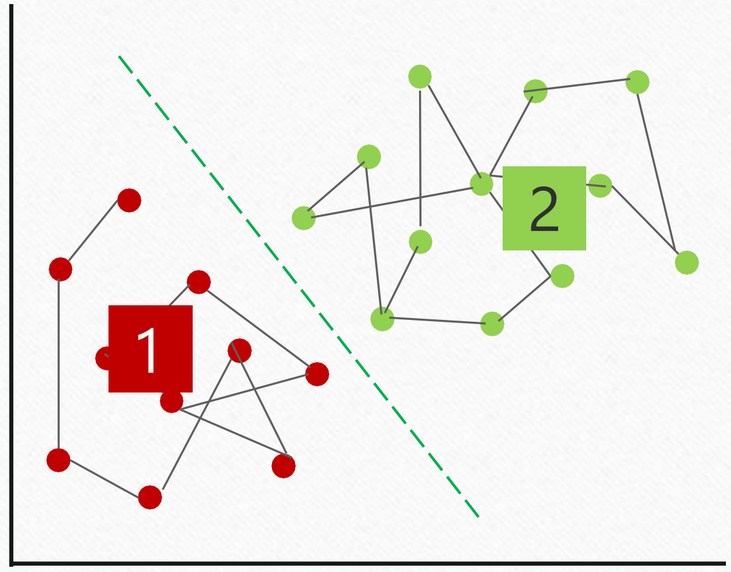
➢Repeat previous two steps iteratively till the cluster centroids stop changing their positions.



K – Means Algorithm

- Initialization
- Cluster assignment
- Move centroid
- Optimization
- Convergence

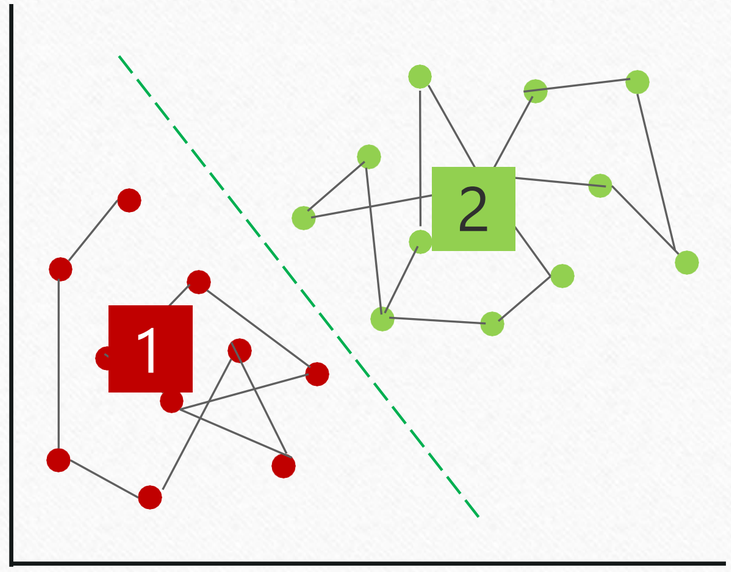
➢Repeat previous two steps iteratively till the cluster centroids stop changing their positions.



K – Means Algorithm

- Initialization
- Cluster assignment
- Move centroid
- Optimization
- Convergence

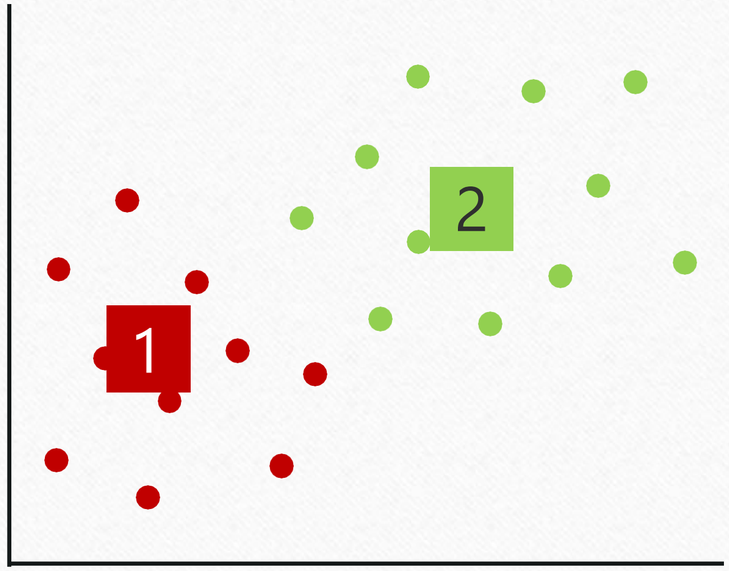
➢Repeat previous two steps iteratively till the cluster centroids stop changing their positions.



K – Means Algorithm

- Initialization
- Cluster assignment
- Move centroid
- Optimization
- Convergence

- Finally, k-means clustering algorithm converges.
- Divides the data points into two clusters clearly visible in red and green.



Data Analytics vs Data Analysis

Data Analytics vs Data Analysis

IIM Gwalior

9

• Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight. Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed business decisions.

• Tools used for data analysis are Open Refine, Rapid Miner, KNIME, Google Fusion Tables, Node XL, Wolfram Alpha, Tableau Public, etc. Tools used in Data analytics are Python, Tableau Public, SAS, Apache Spark, Excel, etc.