# K - Means Clustering Algorithm

## Basic Information

**Clustering algorithms** partition data into distinct groups based on similar items.

**K-Means Clustering Algorithm** is defined as an **unsupervised learning** method, having an iterative process in which the dataset are grouped into k number of predefined non-overlapping clusters or subgroups, making the inner points of the cluster as similar as possible (WCV) while trying to keep the clusters at distinct space (BCV).

Algorithms construct clusters where between-cluster-variation (BCV) should be large in comparison to the within-cluster-variation (WCV) which should be small for good results.

## Machine Learning Distance Metrics

Distance metrics are a key part of several machine learning algorithms. These distance metrics are used in both supervised and unsupervised learning, generally to calculate the similarity between data points.

An effective distance metric improves the performance of our machine learning model, whether that's for classification tasks or clustering.

- Euclidean Distance Metric
  Euclidean Distance represents the shortest distance between two points.

- Manhattan Distance Metric
  Manhattan Distance is the sum of absolute differences between points across all the dimensions.

- Minkowski Distance Metric
  Minkowski Distance is the generalized form of Euclidean and Manhattan Distance.

## Algorithm Process

1. Manually specify K number of clusters to partition the data points into
2. Randomly assign cluster centroids
3. For each data point find the nearest cluster center via distance metric chosen to our cluster centers
4. For k  number of clusters, find cluster centroid and update the location
5. Repeat steps 3 and 4 until convergence (centroids) location do not shift
   (Algorithm Termination)

# K Means Clustering Analysis

Mean Squared Between (MSB)
This metric represents the between cluster variation

Mean Squared Error (MSE)
This metric represents the within cluster variation

## Application

K Means Clustering algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression and more.

The goal usually when we undergo a cluster analysis is either:

- Get a meaningful intuition of the structure of the data we're dealing with.
- Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviors of different subgroups. An example of that is clustering patients into different subgroups and building a model for each subgroup to predict the probability of the risk of having a heart attack.