Clustering

* Clustering is a method of “**Unsupervised Learning”**
* Machine learning technique that involves the grouping of data points.
* Given set of data points, we can use a clustering algorithm to classify each data point into a specific group, so that the data points in the same group are more similar to the other data points in the group.

# Unsupervised Learning

Algorithms belonging to the family of Unsupervised Learning have no variable to predict tied to the data, the data only has an input which would be multiple variables that describe the data.

# Types of Clustering

1. **Hard Clustering:**

* Each data point either belongs to a cluster completely or not (ex. All red or all green)
* Also called as **“Exclusive Clustering”**
* **Example: K-Means Clustering Algorithm**

1. **Soft Clustering :** 
   * A probability or likelihood of that data point to be in those clusters is assigned.
   * Also called as **“Overlapping Clustering”**
   * **Example: C-Means Clustering Algorithm**

# Types of Clustering methods

# K-Means Clustering Algorithm

* K-means clustering is a type of **unsupervised learning**, which is used when you have **unlabeled data** (i.e., data without defined categories or groups)
* The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.
* The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided.
* Data points are clustered based on feature similarity.
* Result of the K-means clustering are:

1. **The centroid of the K clusters,** which can be used to label new data
2. **Labels for the training data** (each data point is assigned to a single cluster)

# Steps in K-Means Clustering

1. Specify the desired number of clusters(K)
2. Randomly assign each data point to a cluster
3. Calculate cluster centroid
4. Re- assign each data point to the closest cluster centroid
5. Re calculate cluster centroid
6. Repeat 4 and 5 until it reaches to Global Optima (there is no further switching of data points between two clusters for two successive repeats)

# Hierarchical Clustering

* This algorithm builds hierarchy of clusters.
* This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into one cluster.in the end, this algorithm terminates when there is only a single cluster left.
* The results of the hierarchical clustering can be shown using dendrogram.
* Bottom up approach and top-down approach