# Clustering

* Popular unsupervised learning model. Aims to find similar groups in a data.
* Similarity of data items is measured using **distance metrics**. As the distance is small, data items are similar. Else, dissimilar.
* **Standardization** is recommended so that all attributes can have equal impact on computational distance.
* Quality of clusters can be measured using **compactness** (intra cluster cohesion-should be low) and **isolation** (inter-cluster cohesion should be high).

# Hierarchical clustering

**Agglomerative** clustering: (bottom up)

* Unlike K-Means we don’t define K-first (in simple we do not know what’s K value is).
* Builds clusters from bottom level and merges most similar pair of clusters. First assume every point as a cluster, then slowly merge them until you get optimal number of K clusters.
* Optimal is K is defined by elbow curve and cross-validation (model performance)

**Divisive(top -down)**

* Builds clusters from top level (root, everything as 1 cluster) and splits root into set of child clusters, each child cluster is recursively divided further.
* Stops when each cluster with single data point.
* Optimal is K is defined by elbow curve and cross-validation (model performance)

**Both approaches use distance matrix:** Euclidean mostly, Manhattan, Chebychev etc

**Do not require number of clusters ‘k”:** is determined using model performance/ elbow curve

**Need termination condition.**

# K-Means

Classifies given data set into ‘K’ number of clusters. Each cluster has cluster centre called centroid. This clustering is done in such a way that variation with in the cluster should be less and between the clusters should be high. Here we specify K (we know what K value is prior to building model)

* Randomly choose ‘K’ data points to be initial centroid.
* Assign each data point to the closest centroid.
* Re-compute centroids using current cluster membership.
* Repeat these steps till the convergence condition met

Convergence/ stopping condition:

* Until no/ minimum number of re-arrangements of data points to different clusters.
* Until no/minimum change of centroid.
* Minimum decrease in sum of square distance within cluster.

**Model Performance:**

* If the ground truth labels are not known, evaluation must be performed using the model itself. The Silhouette Coefficient is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters (0 to 1 , 1 is better).
* If the ground truth is available then, we split the data into train and test (70-30) and we measure **“homogeneity\_score” and “completeness\_score”**