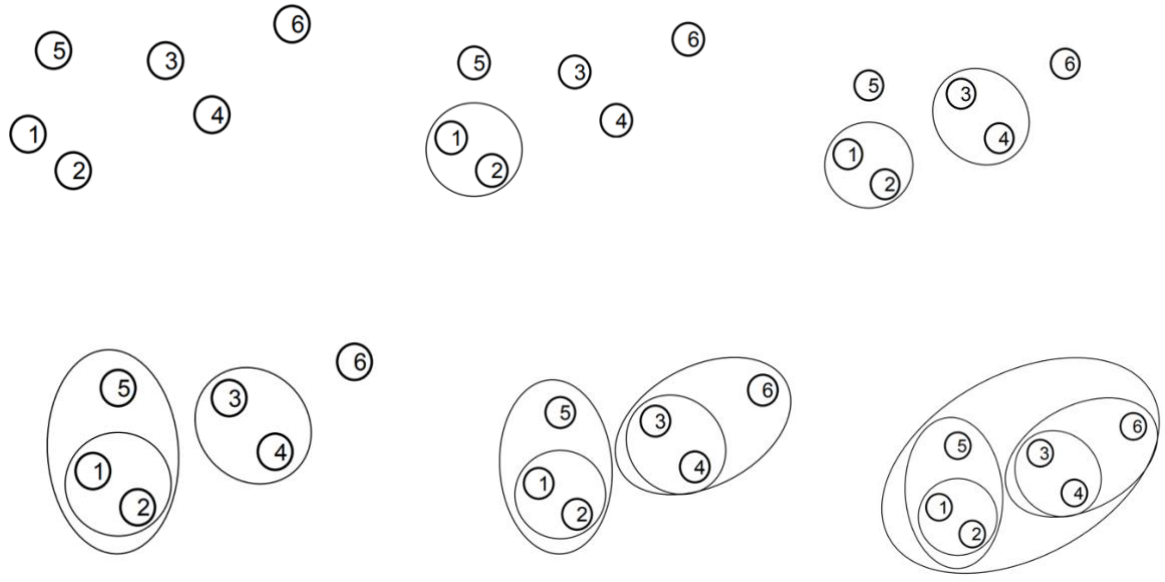
Hierarchical Clustering is divided mainly into 2 types:
Agglomerative (our main focus):
1. Start with every point in its own cluster
2. At each step, merge the two closest clusters
3. Stop when every point is in the same cluster
Divisive:
1. Start with every point in the same cluster
2. At each step, split until every point is in its own cluster

Methods of conduction for Agglomerative Clustering Algorithm
1. Let each point in the dataset be in its own cluster
2. Compute the distance between all pairs of clusters
3. Merge the two closest clusters
4. Repeat 3 & 4 until all points are in the same cluster

Here is an example:



Distance Function:
Distance between points: $d(p_1, p_2)$
Distance between clusters: $D(C_1, C_2)$

$$D_{SL}(C_1, C_2) = \min \{d(p_1, p_2) \mid p_1 \in C_1, p_2 \in C_2\}$$

Average-Link Distance:

$$D_{AL}(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{p_1 \in C_1, p_2 \in C_2} d(p_1, p_2)$$

Centroid Distance is the distance between the centroids of clusters.
$D_c(C_1, C_2) = d(u_1, u_2)$

Ward's Distance is the difference between the spread / variance of points in the merged cluster and the unmerged clusters.

$$D_{WD}(C_1, C_2) = \sum_{p \in C_{12}} d(p, \mu_{12}) - \sum_{p_1 \in C_1} d(p_1, \mu_1) - \sum_{p_2 \in C_2} d(p_2, \mu_2)$$