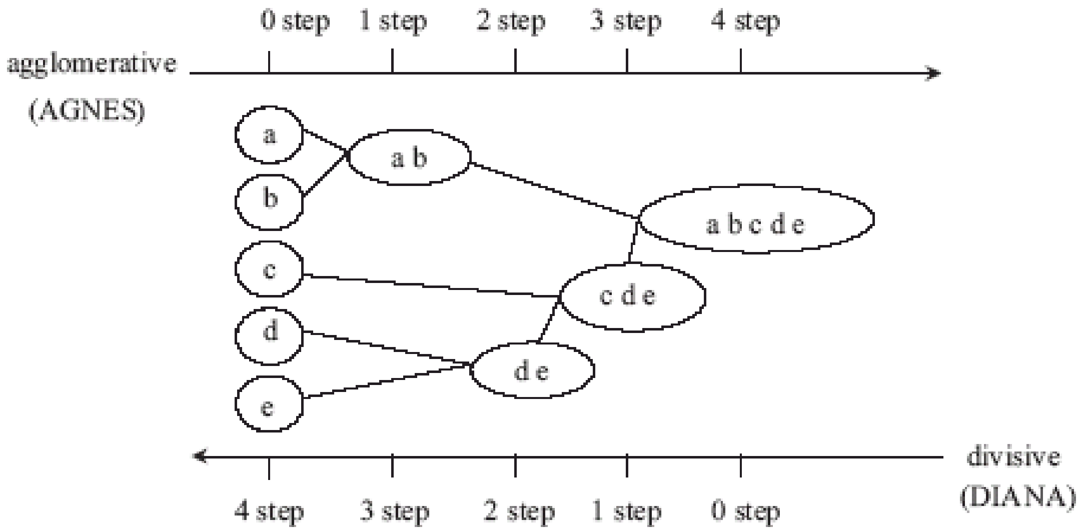
Hierarchical Clustering

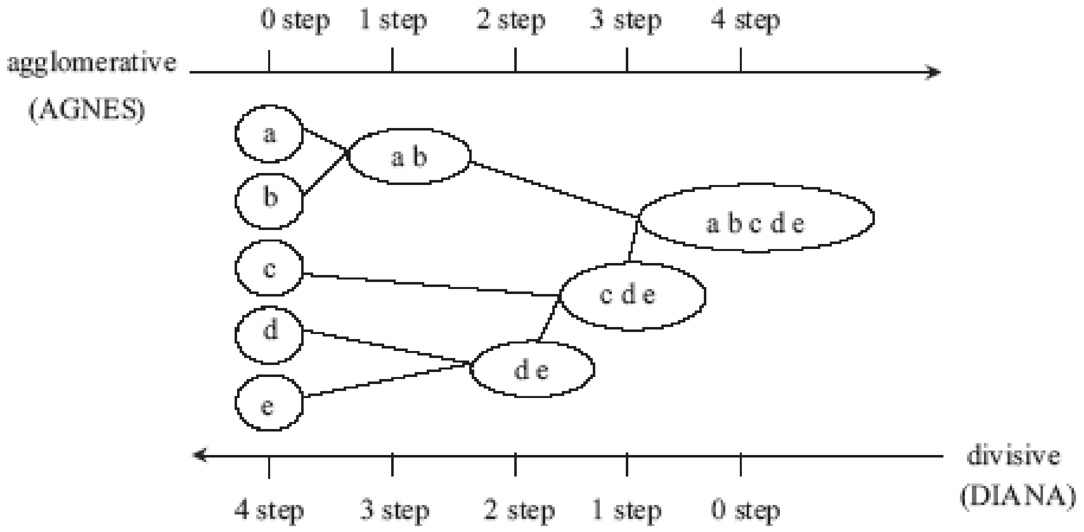
- **They work by grouping data objects into a tree of clusters**



Hierarchical Clustering

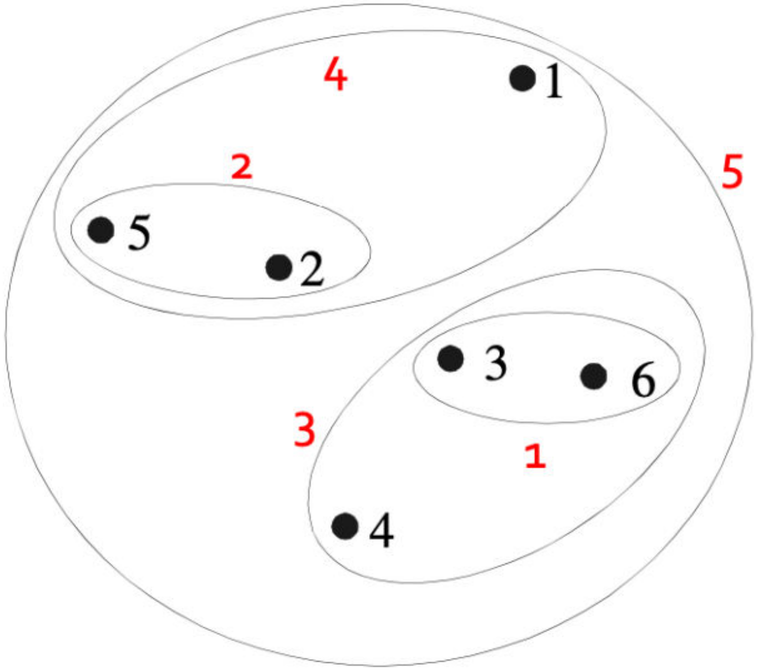
Agglomerative

- It is a bottom-up strategy
- It first places each object in its own cluster
- Then merges these clusters into larger and larger clusters



Hierarchical Clustering

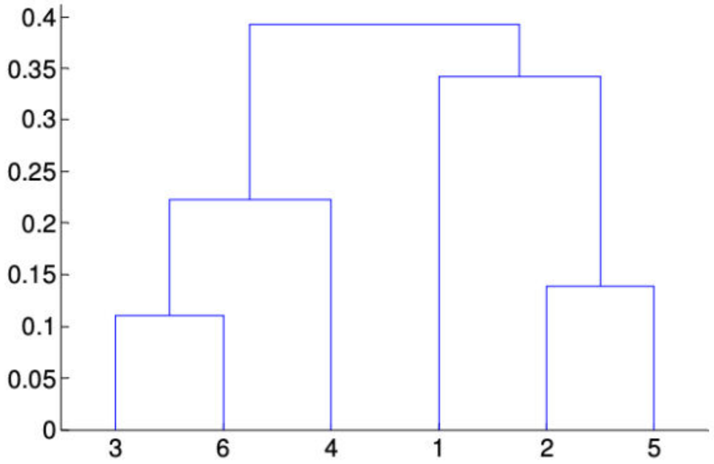
Agglomerative



Nested Clusters

Dendrogram

	1	2	3	4	5	6
1	0	.24	.22	.37	.34	.23
2	.24	0	.15	.20	.14	.25
3	.22	.15	0	.15	.28	.11
4	.37	.20	.15	0	.29	.22
5	.34	.14	.28	.29	0	.39
6	.23	.25	.11	.22	.39	0



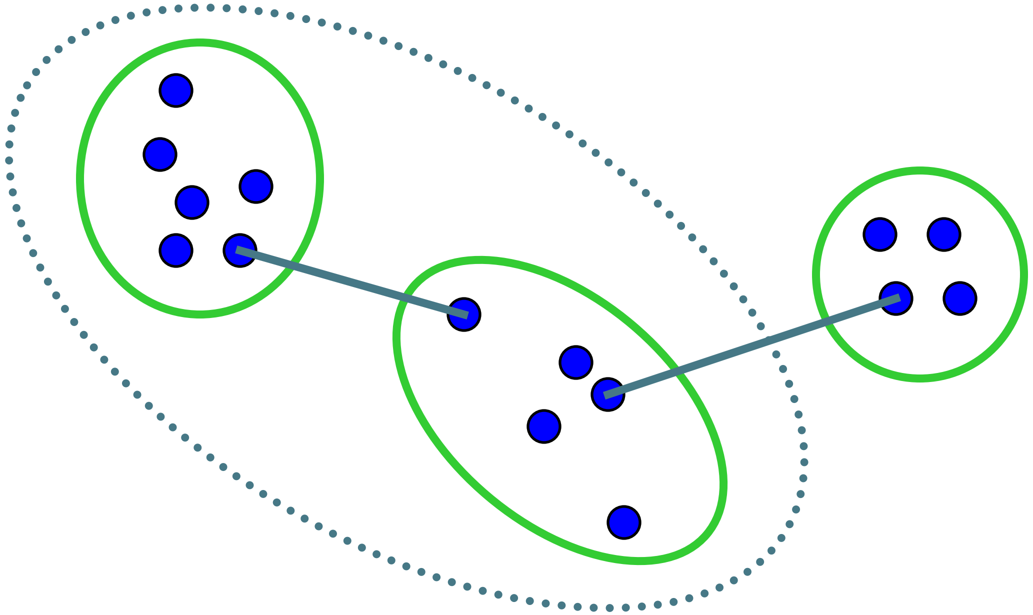
Hierarchical Clustering

- **The splitting and merging of clusters is based on distance measures**

- **Four widely used distance measures are**

Single Link

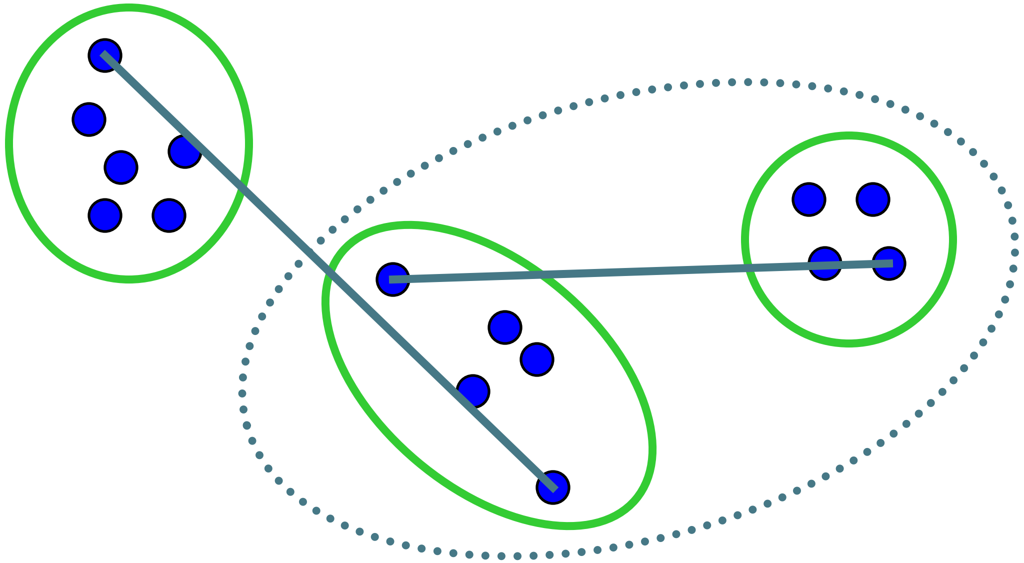
- cluster similarity = similarity of two most similar members



Minimum distance: $d_{min}(C_i, C_j) = min_{p \in Ci, p' \in C_j} |p - p'|$

Complete Link

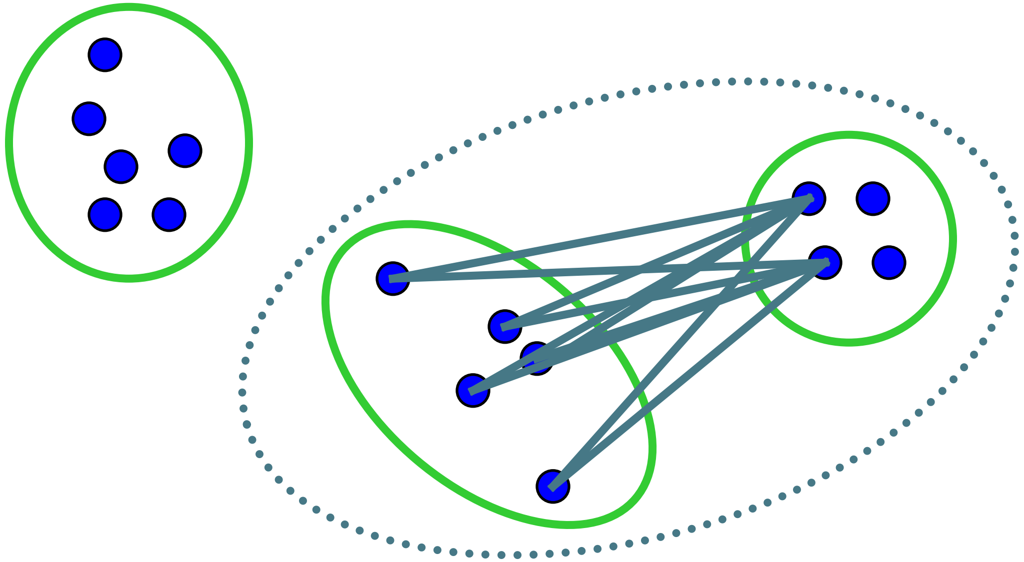
- cluster similarity = similarity of two least similar members



Maximum distance: $d_{max}(C_i, C_j) = max_{p \in Ci, p' \in C_j} |p - p'|$

Group Average

- cluster similarity = average similarity of all pairs



Average distance: $d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \Sigma_{p \in C_i} \Sigma_{p' \in C_j} |p - p'|$

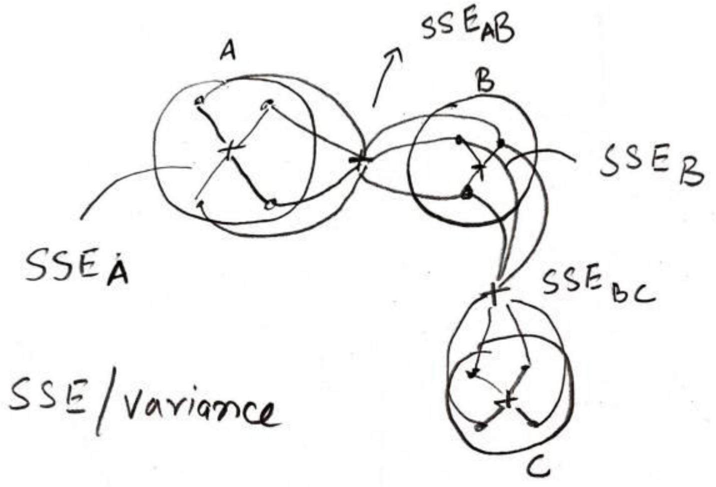
Hierarchical Clustering

Mean distance: $d_{mean}(C_i, C_j) = |m_i - m_j|$

where m_i is the mean data object of cluster C_i

WARD

- ‘ward’ minimizes the variance of the clusters being merged.



$$\text{distance}(A,B) = SSE_{AB} - (SSE_A + SSE_B)$$

$$\text{distance}(B,C) = SSE_{BC} - (SSE_B + SSE_C)$$

It promotes intracluster similarity and intercluster dissimilarity.

Hierarchical Clustering

Example: Let us consider five data samples: A, B, C, D, and E

Let the inter-point distances between these examples be given by the following distance matrix.

$$\mathbf{DM}(1) = \begin{array}{cccccc} & A & B & C & D & E \\ A & 0 & & & & \\ B & 1 & 0 & & & \\ C & 5 & 3 & 0 & & \\ D & 6 & 8 & 4 & 0 & \\ E & 8 & 7 & 6 & 2 & 0 \end{array}$$

Hierarchical Clustering

Using the *nearest neighbor* measure, also known as the *single linkage* measure, we merge *A* and *B* to form a cluster, since they are closest.
Next we compute the distances between this cluster and the remaining examples. We can get these distances from the above distance matrix.

The values for these are as follows:

d(AB)C = min{dAC, dBC} = dBC = 3
d(AB)D = min{dAD, dBD} = dAD = 6
d(AB)E = min{dAE, dBE} = dBE = 7

Hierarchical Clustering

Updated distance matrix

$$\mathbf{DM}(2) = \begin{array}{c|cccc} & AB & C & D & E \\ \hline AB & 0 & & & \\ C & 3 & 0 & & \\ D & 6 & 4 & 0 & \\ E & 7 & 6 & 2 & 0 \end{array}$$

Since the smallest entry in above distance matrix is 2, examples *D* and *E* are merged to form another cluster

Updated distance matrix

$$\mathbf{DM}(3) = \begin{array}{c|ccc} & AB & C & DE \\ \hline AB & 0 & & \\ C & 3 & 0 & \\ DE & 6 & 4 & 0 \end{array}$$

This matrix indicates that *C* should be merged with *A* and *B*.
At this stage we have only two clusters left that are joined to form a single cluster of five examples.

Hierarchical Clustering

- **Termination Condition**

    - The user can specify the desired number of clusters as a termination condition

    - The quality of clusters can also be a termination condition

How to Choose the Number of Clusters:
Lifetime Method

The *lifetime* of a cluster is defined as the absolute value of the difference between the dendrogram level at which it is created and the level at which it is absorbed into a larger cluster.

Using lifetime as a criterion, a user can search for cluster that have a large lifetime.

Hierarchical Clustering

- The merge (or split) decisions are critical in these types of algorithms
- Once a merging (or splitting) of two clusters is made, it cannot be undone and furthering clustering proceeds on the basis of this decision