Clustering
Lecture 3: Hierarchical Methods

Jing Gao

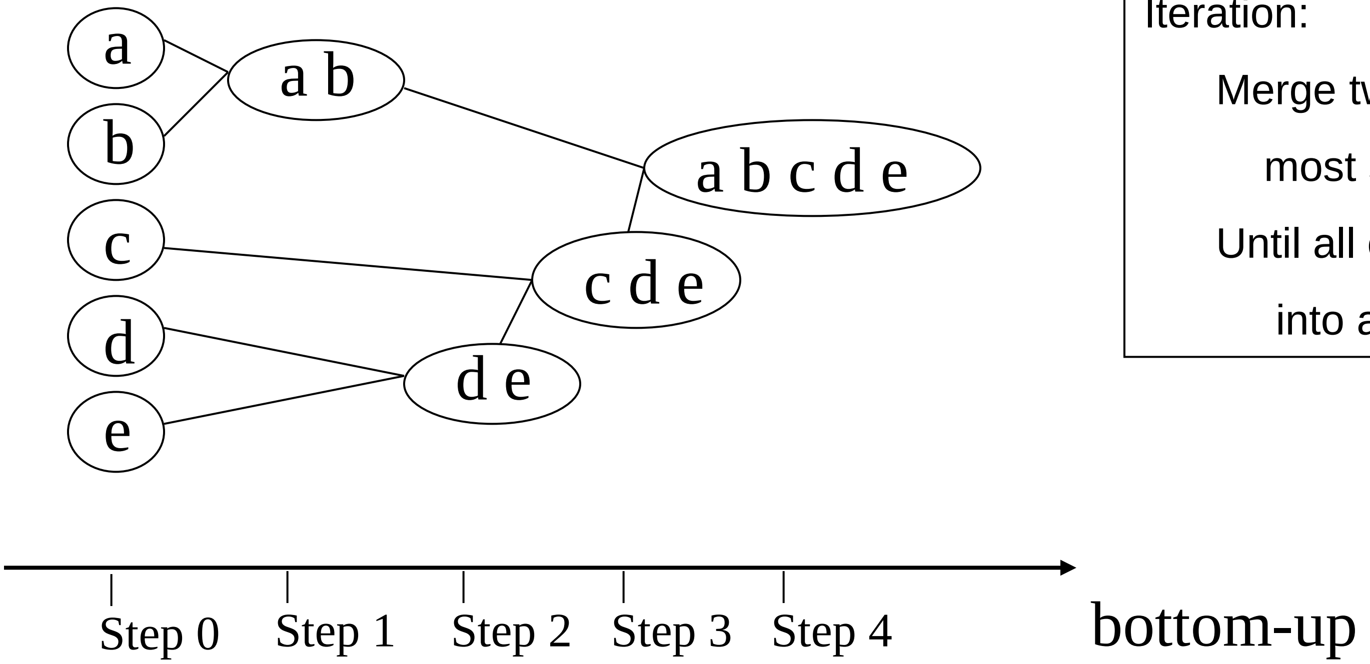
SUNY Buffalo

Outline

- **Basics**
  - Motivation, definition, evaluation
- **Methods**
  - Partitional
  - Hierarchical
  - Density-based
  - Mixture model
  - Spectral methods
- **Advanced topics**
  - Clustering ensemble
  - Clustering in MapReduce
  - Semi-supervised clustering, subspace clustering, co-clustering, etc.

Hierarchical Clustering

- **Agglomerative approach**



Initialization:

Each object is a cluster

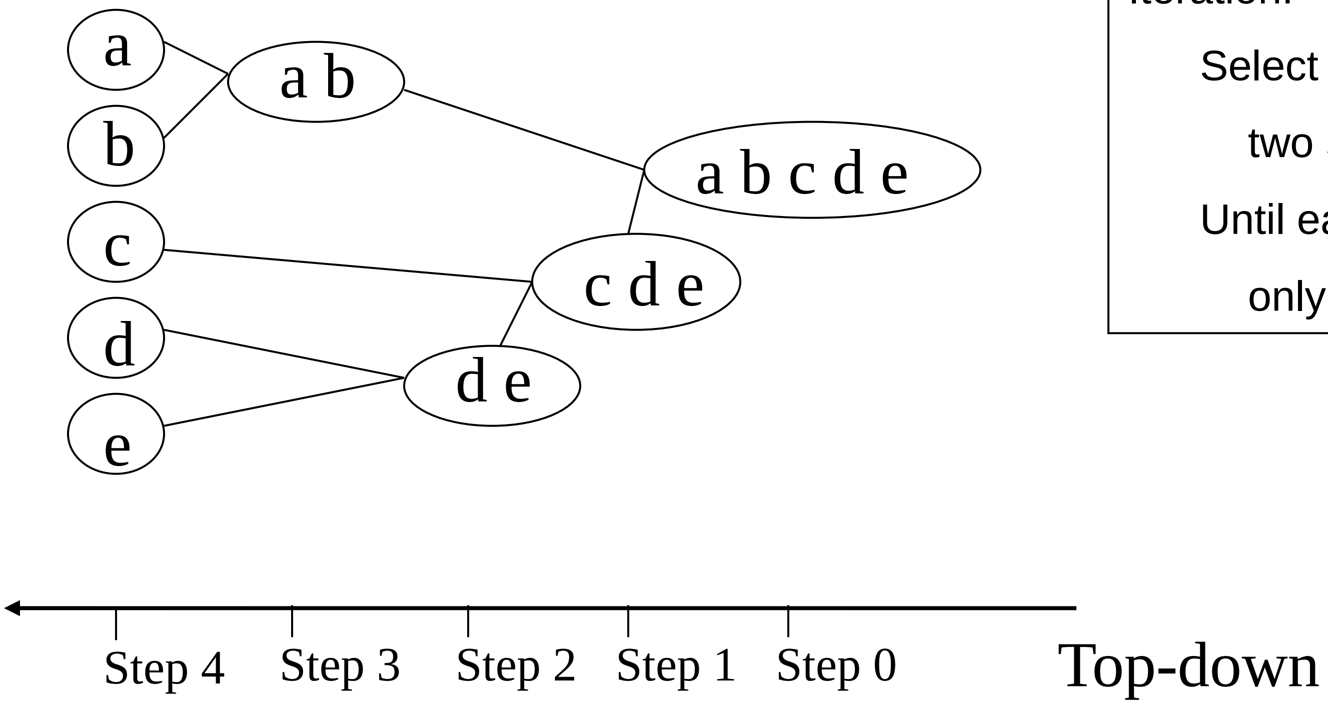
Iteration:

Merge two clusters which are most similar to each other;

Until all objects are merged into a single cluster

Hierarchical Clustering

- **Divisive Approaches**



Initialization:

All objects stay in one cluster

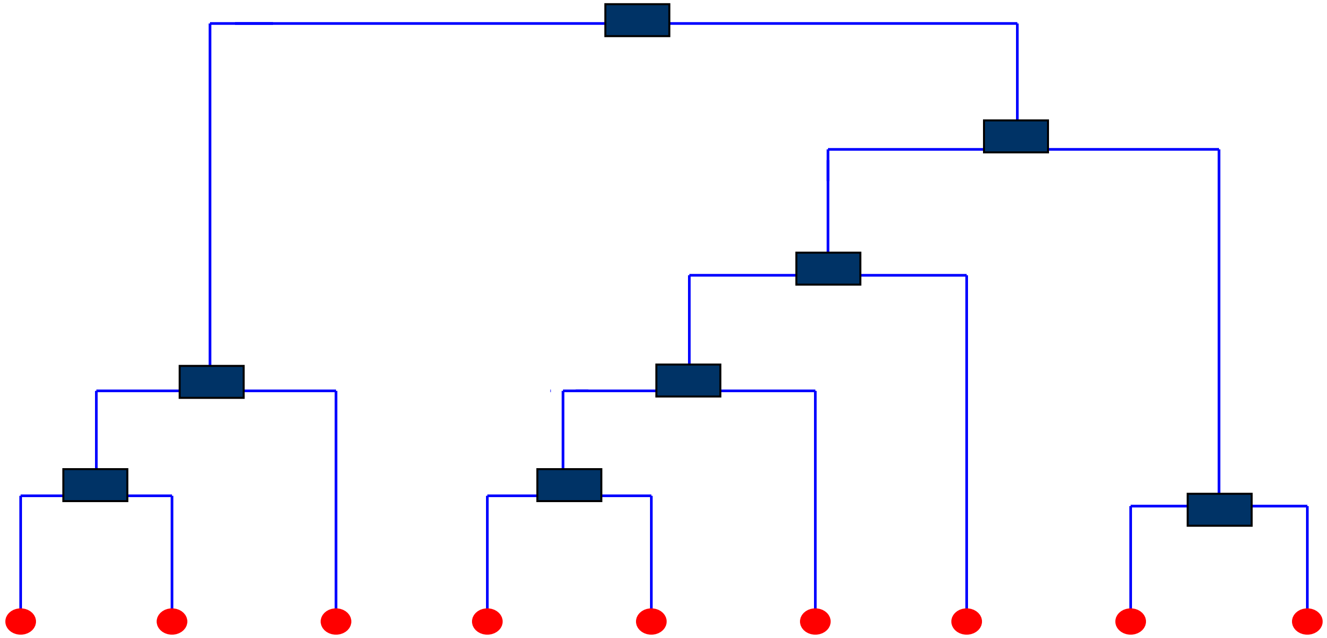
Iteration:

Select a cluster and split it into two sub clusters

Until each leaf cluster contains only one object

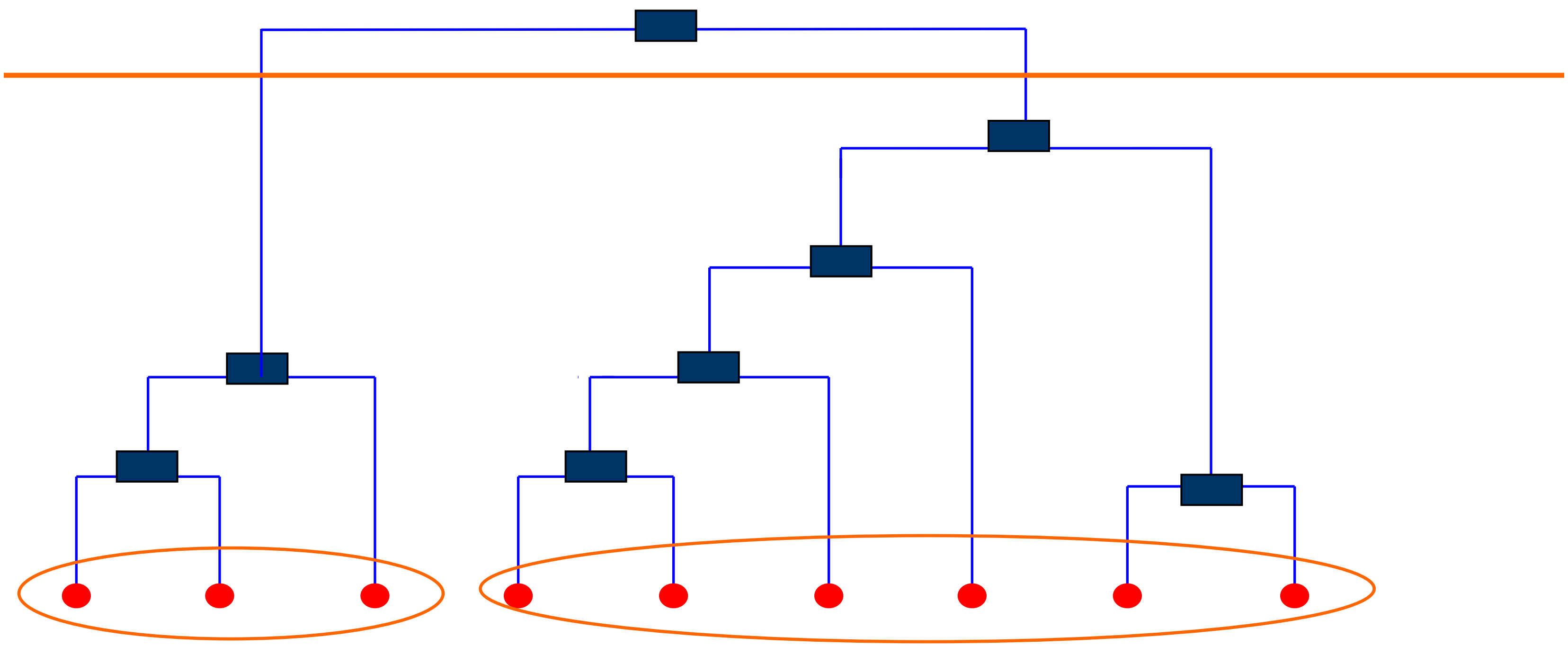
Dendrogram

- A tree that shows how clusters are merged/split hierarchically
- Each node on the tree is a cluster; each leaf node is a singleton cluster



Dendrogram

- A clustering of the data objects is obtained by cutting the *dendrogram* at the desired level, then each connected component forms a cluster

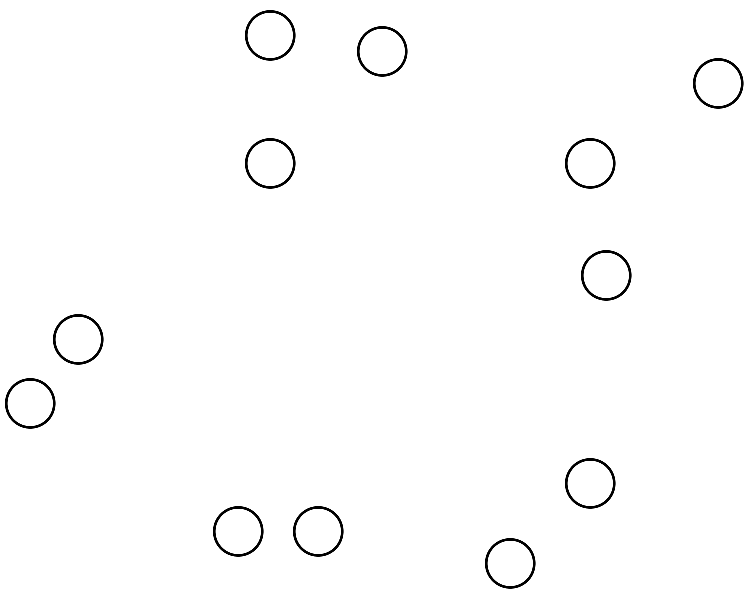


Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique
- Basic algorithm is straightforward
  1. Compute the distance matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the distance matrix
  6. **Until** only a single cluster remains
- Key operation is the computation of the distance between two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

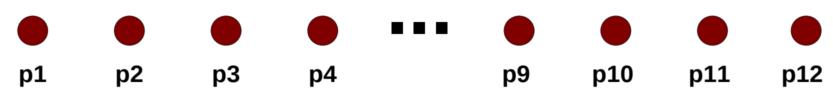
Starting Situation

- Start with clusters of individual points and a distance matrix



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Distance Matrix

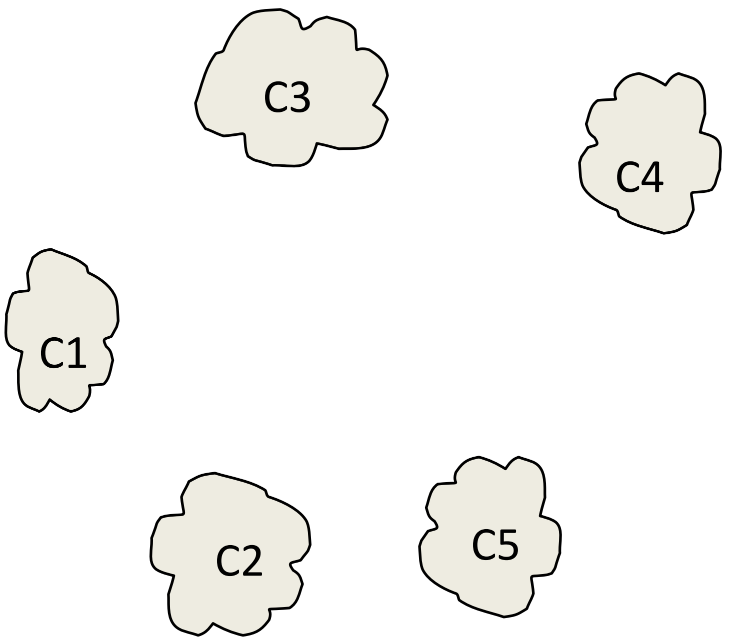


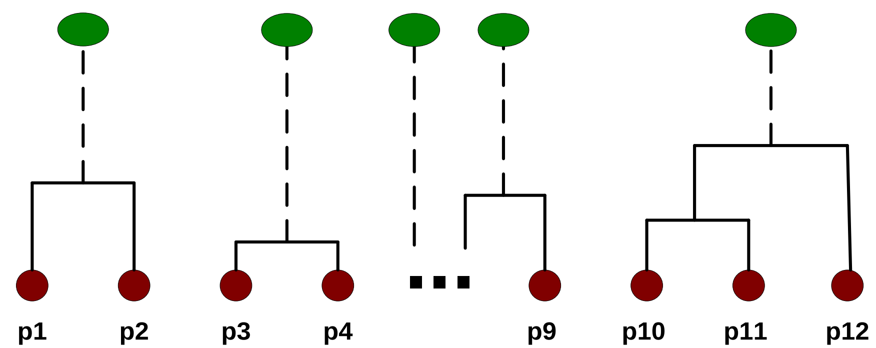
Intermediate Situation

- After some merging steps, we have some clusters
- Choose two clusters that has the smallest distance (largest similarity) to merge

	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

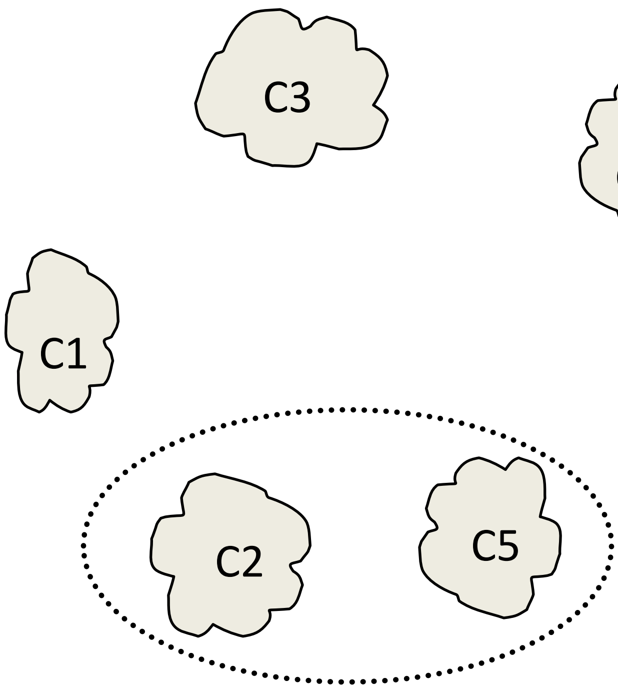
Distance Matrix

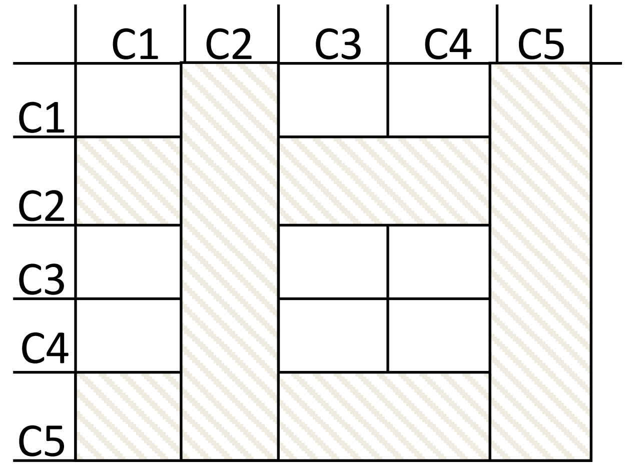




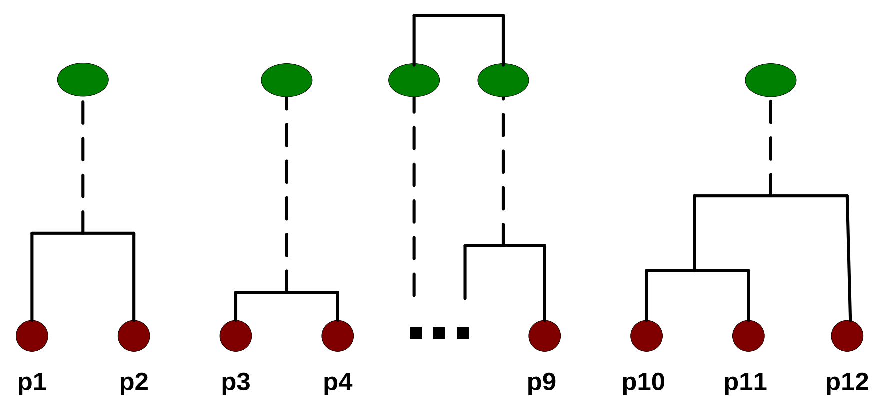
Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the distance matrix.



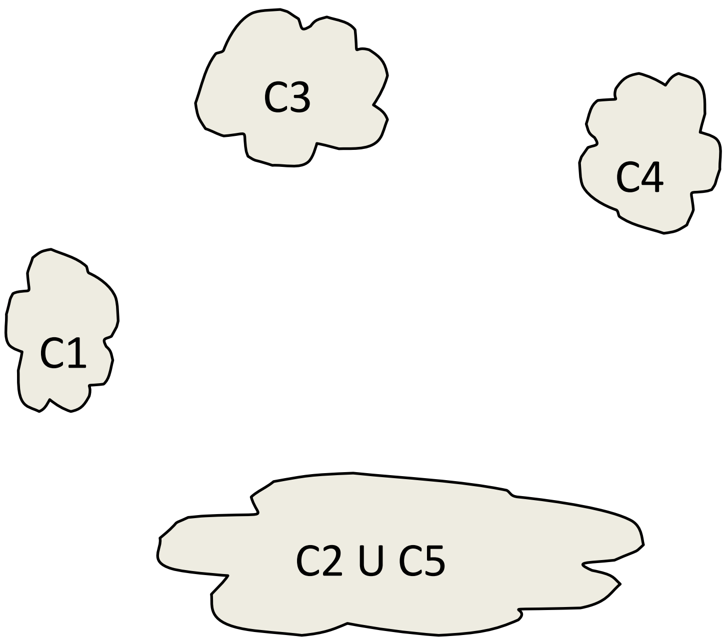


Distance Matrix



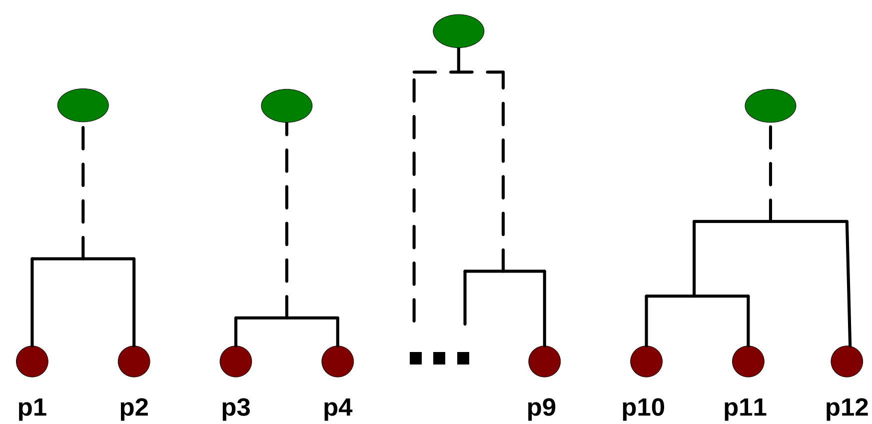
After Merging

- The question is “How do we update the distance matrix?”

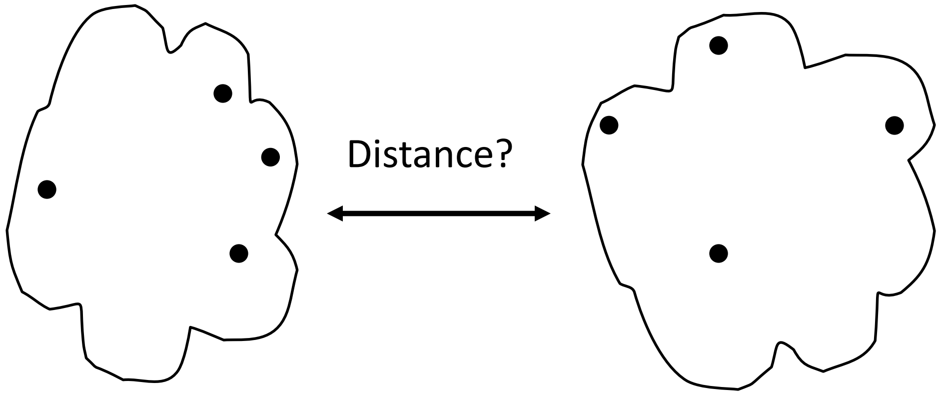


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Distance Matrix



How to Define Inter-Cluster Distance



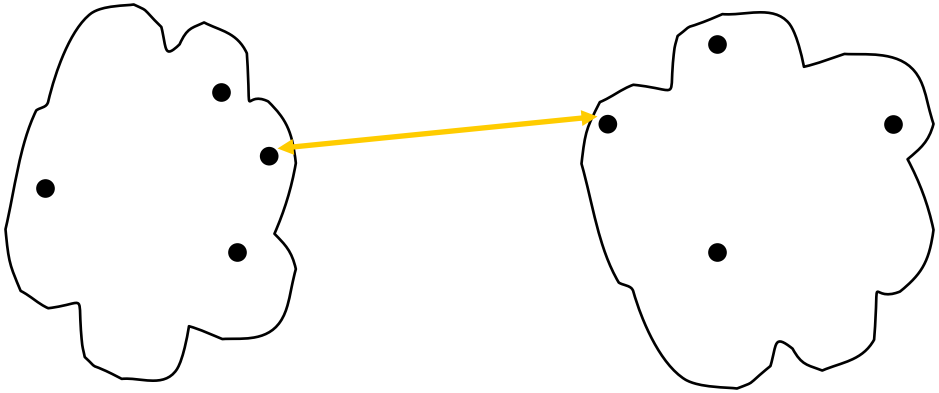
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Distance Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- ……

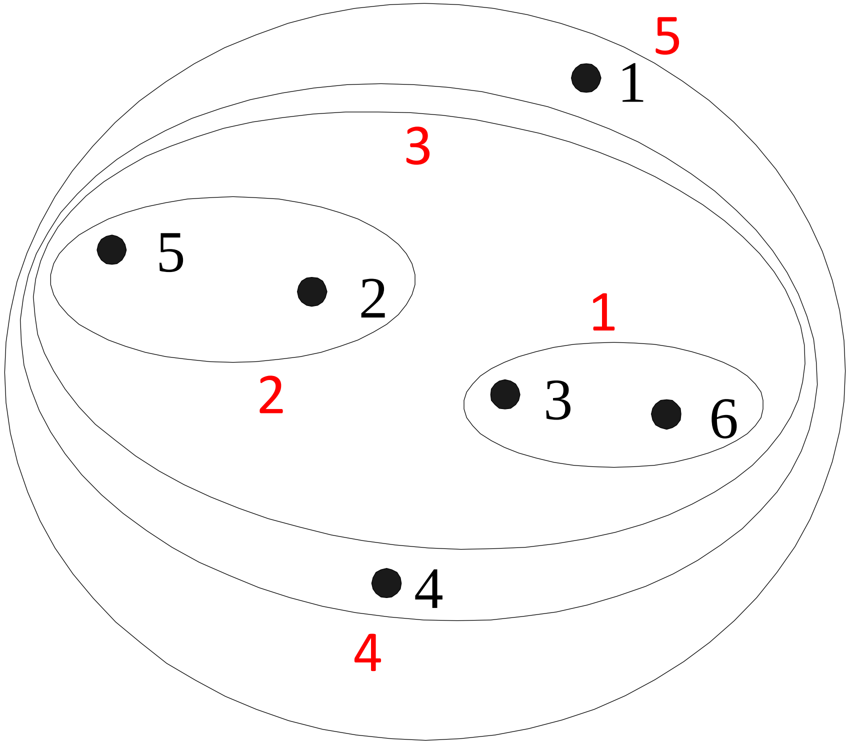
MIN or Single Link

- **Inter-cluster distance**
  - The distance between two clusters is represented by the distance of the *closest pair of data objects* belonging to different clusters.
  - Determined by one pair of points, i.e., by one link in the proximity graph

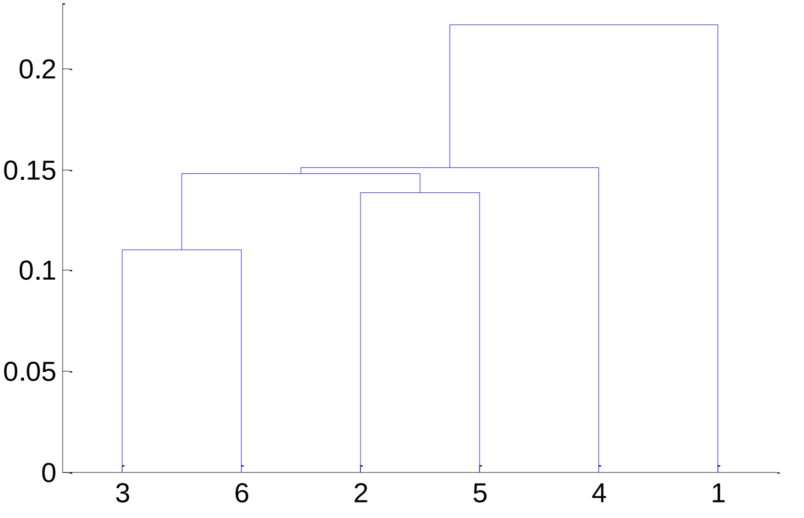


$$d_{\min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p, q)$$

MIN

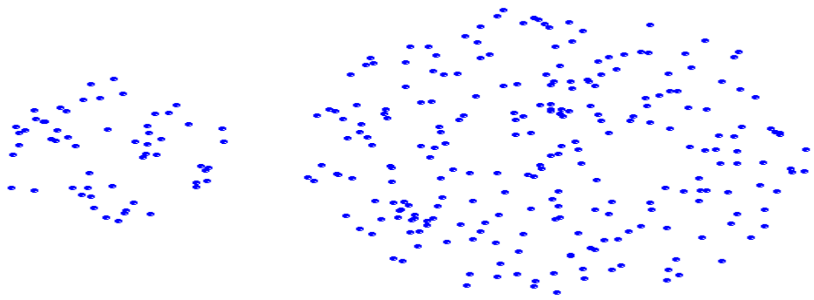


Nested Clusters

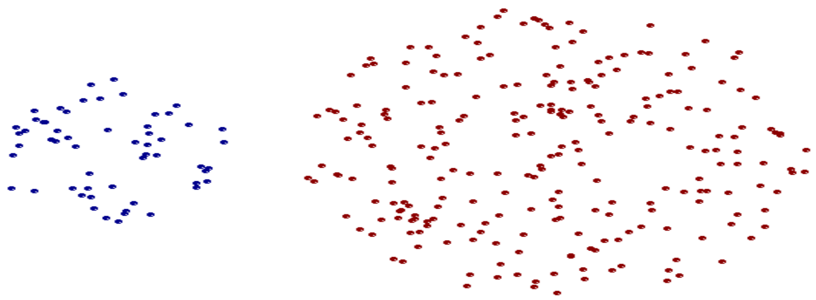


Dendrogram

Strength of MIN



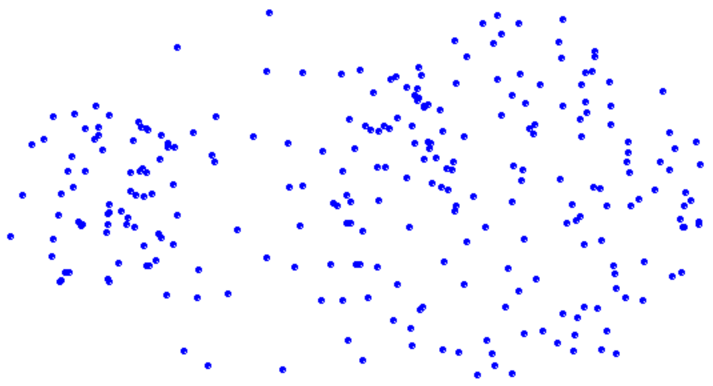
Original Points



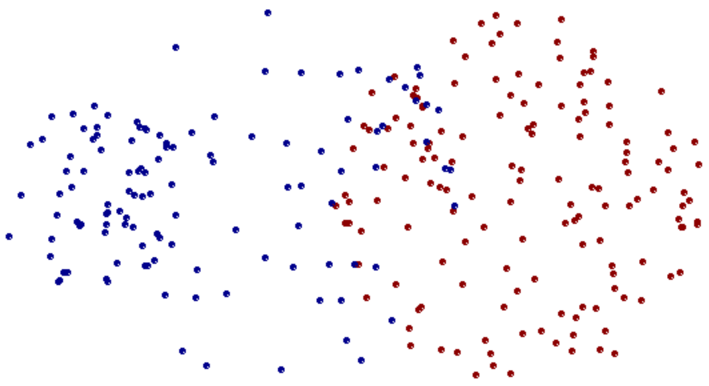
Two Clusters

- Can handle non-elliptical shapes

Limitations of MIN



Original Points

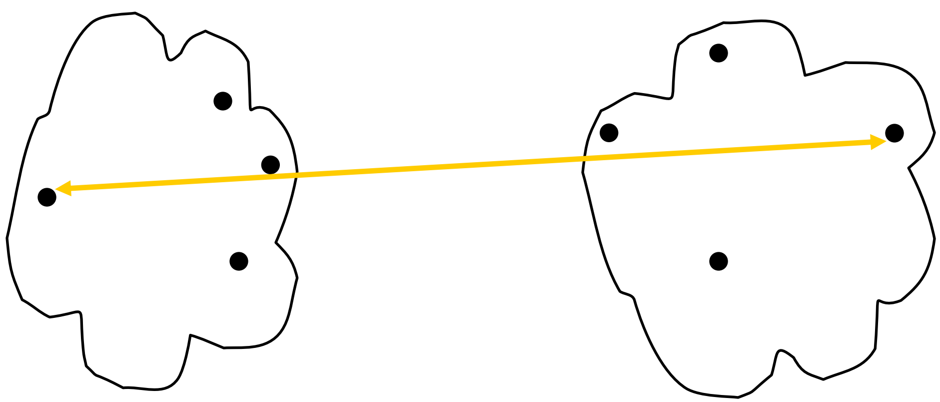


Two Clusters

- Sensitive to noise and outliers

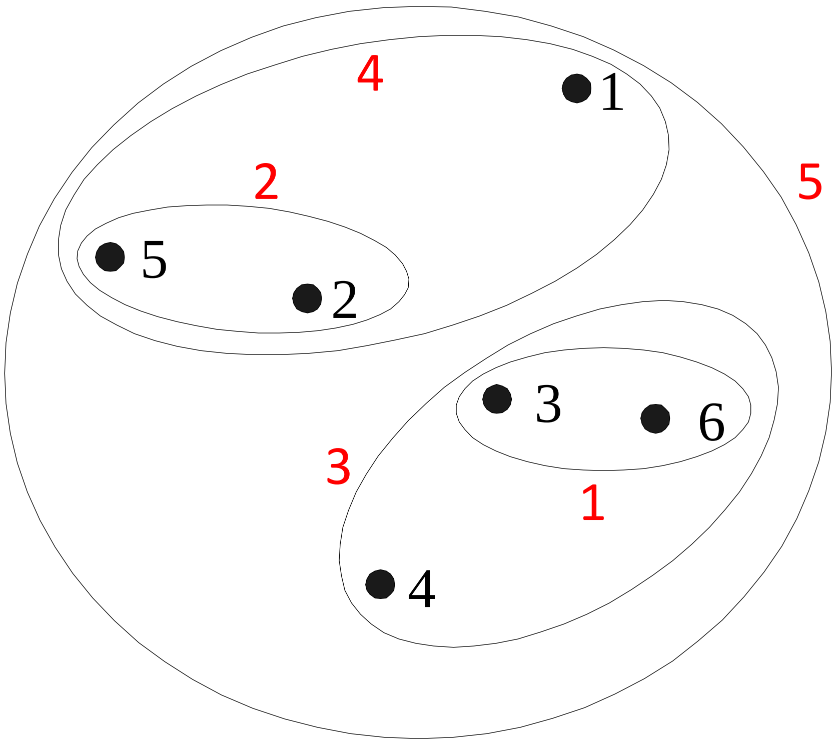
MAX or Complete Link

- **Inter-cluster distance**
  - The distance between two clusters is represented by the distance of the *farthest pair of data objects* belonging to different clusters

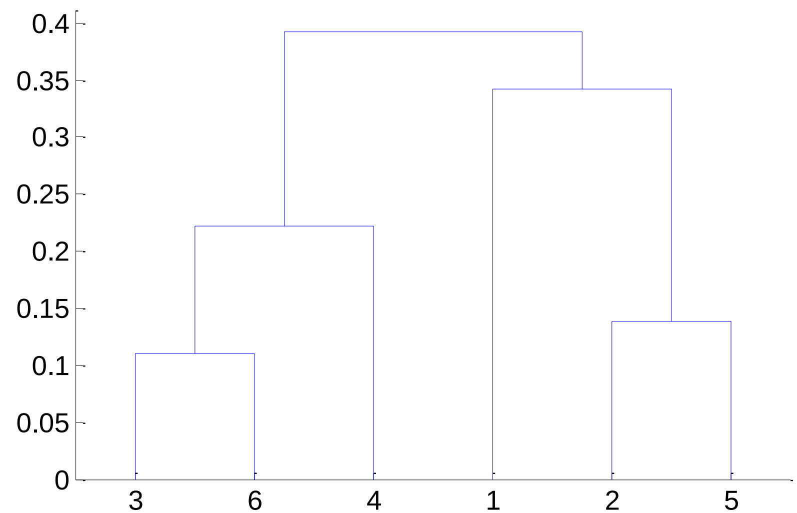


$$d_{\min}(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p, q)$$

MAX

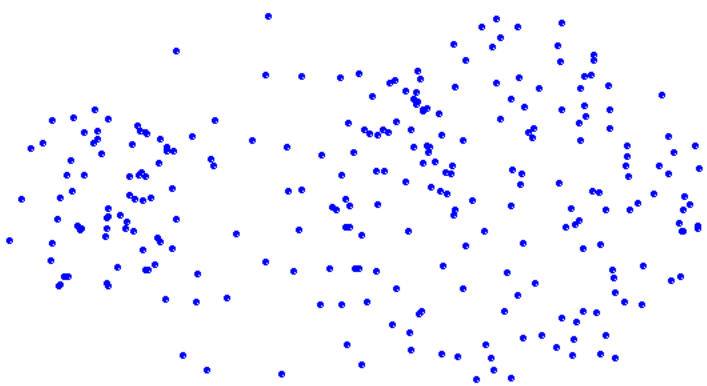


Nested Clusters

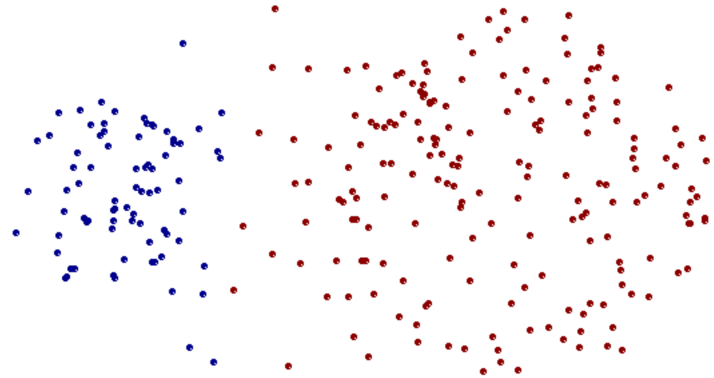


Dendrogram

Strength of MAX



Original Points

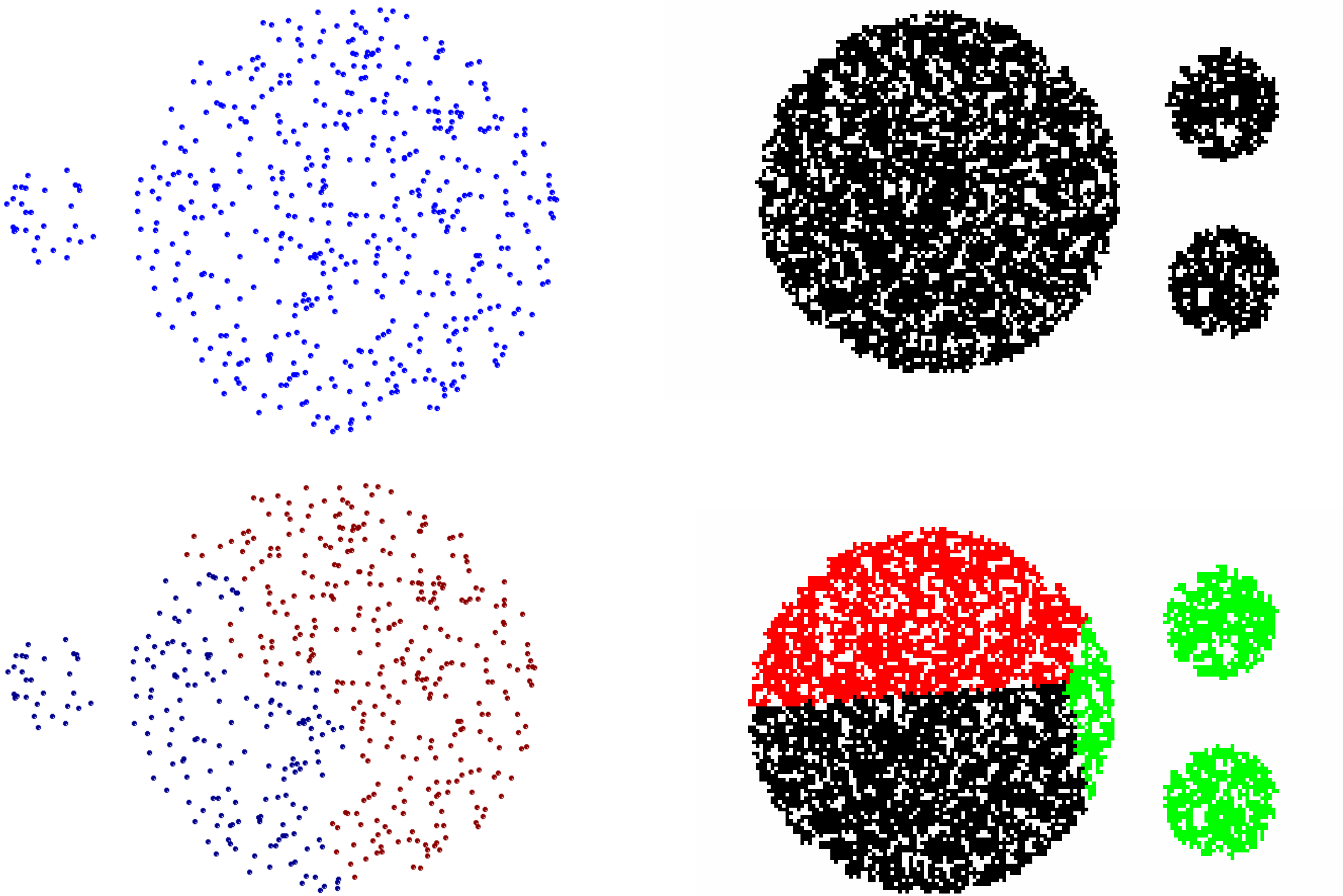


Two Clusters

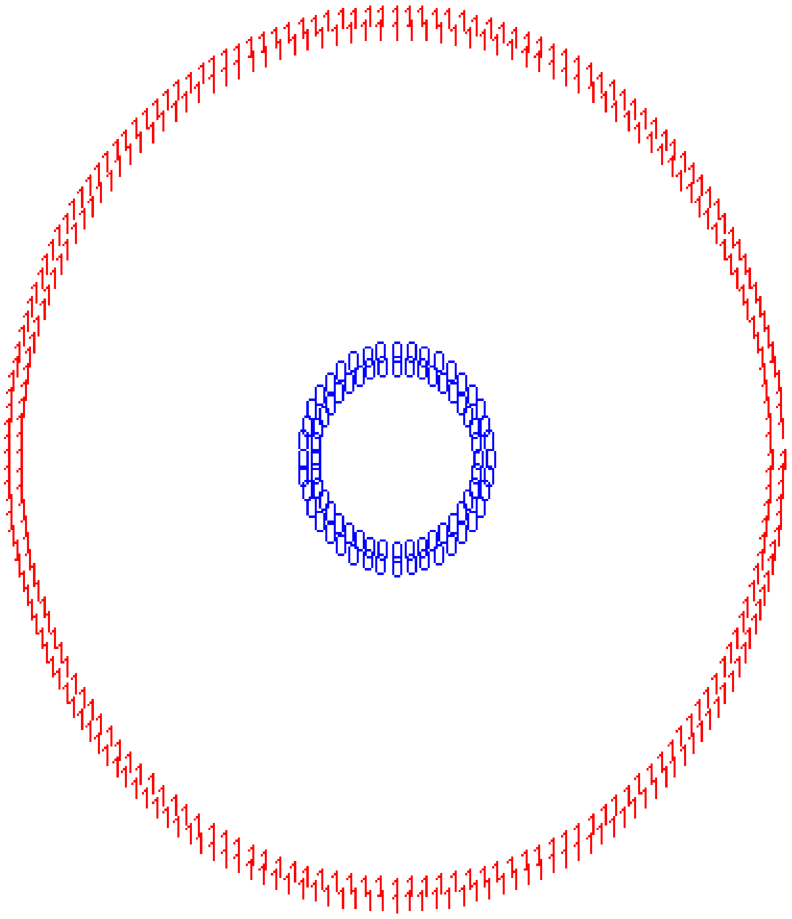
- Less susceptible to noise and outliers

Limitations of MAX

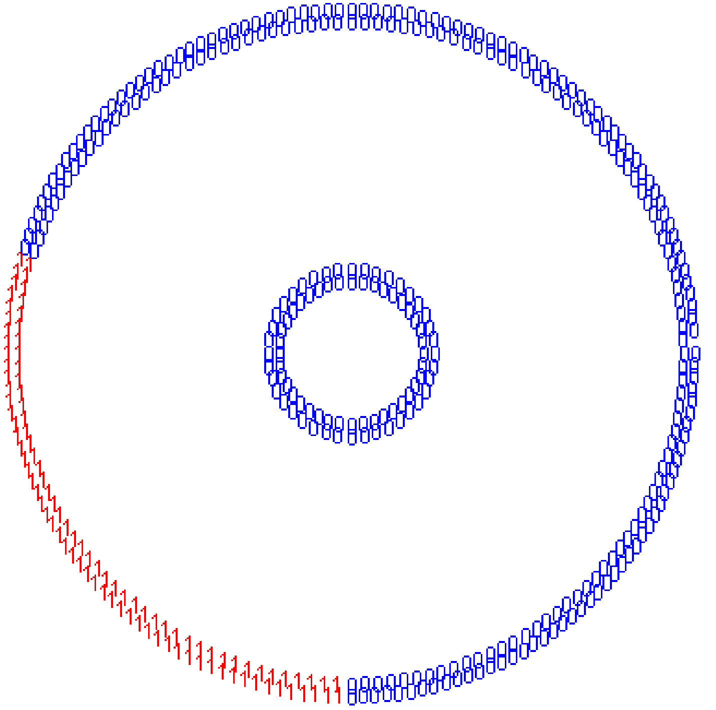
- Tends to break large clusters



Limitations of MAX



MIN (2 clusters)

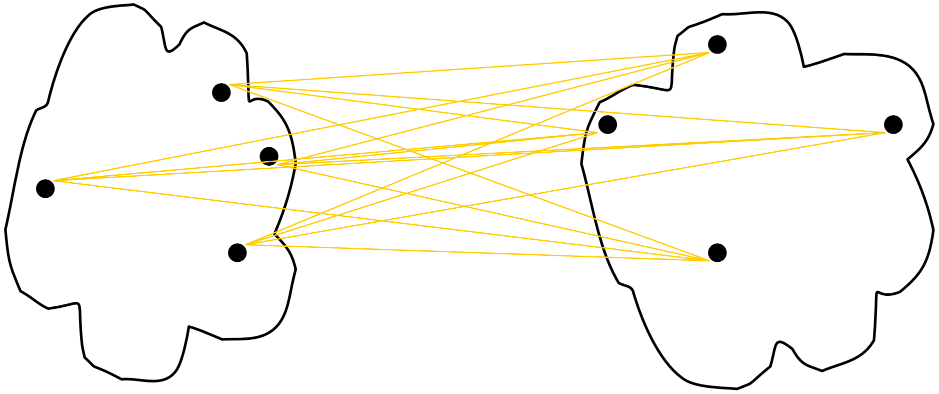


MAX (2 clusters)

- Biased towards globular clusters

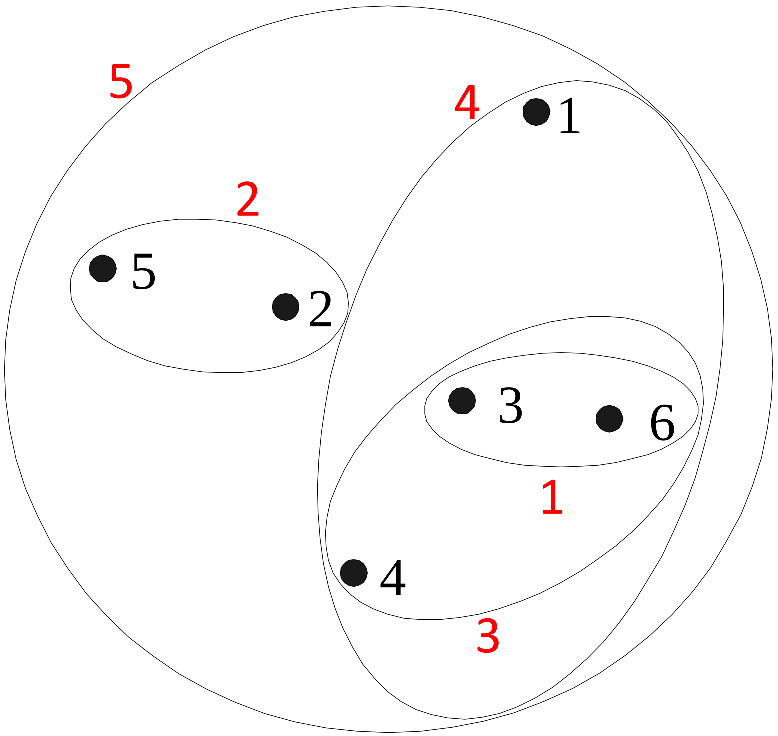
Group Average or Average Link

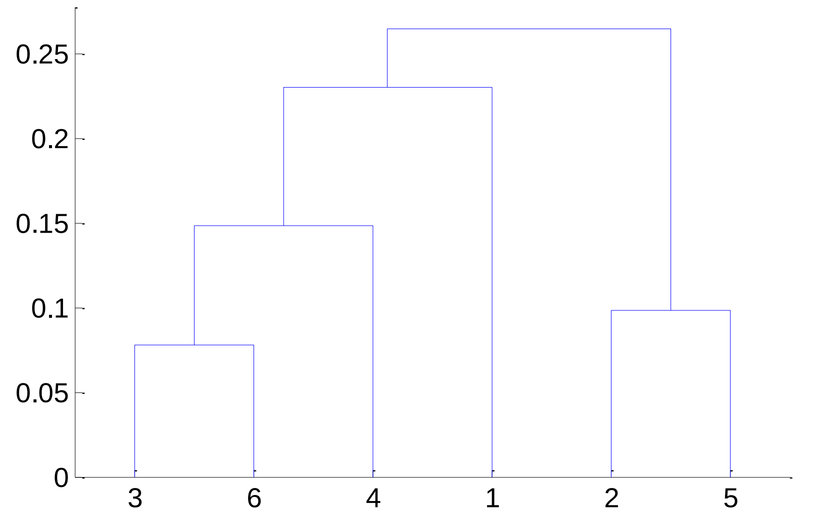
- **Inter-cluster distance**
  - The distance between two clusters is represented by the *average* distance of *all pairs of data objects* belonging to different clusters
  - Determined by all pairs of points in the two clusters



$$d_{\min}(C_i, C_j) = \underset{p \in C_i, q \in C_j}{avg} d(p, q)$$

Group Average





Nested Clusters

Dendrogram

Group Average

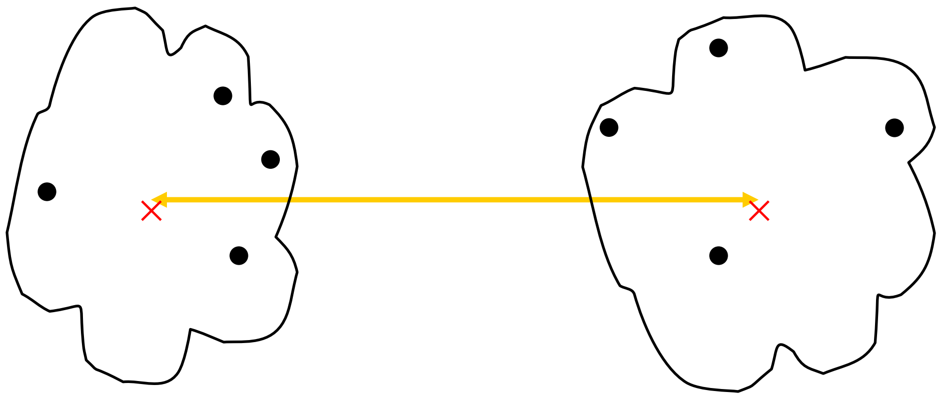
- Compromise between Single and Complete Link

- Strengths
  - Less susceptible to noise and outliers

- Limitations
  - Biased towards globular clusters

Centroid Distance

- **Inter-cluster distance**
  - The distance between two clusters is represented by the distance between *the centers of the clusters*
  - Determined by cluster centroids



$$d_{mean}(C_i, C_j) = d(m_i, m_j)$$

Ward's Method

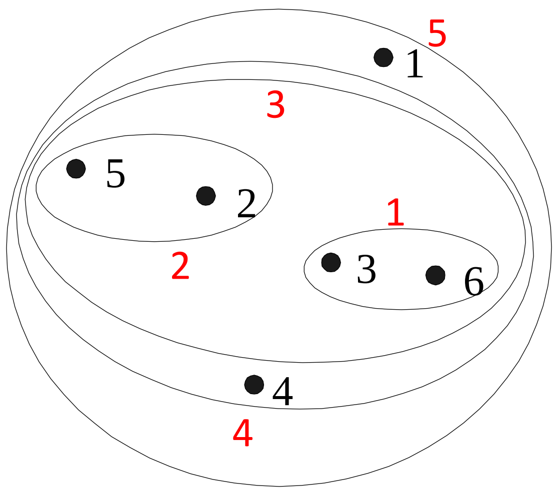
- Similarity of two clusters is based on the increase in squared error when two clusters are merged
  - Similar to group average if distance between points is squared distance

- Less susceptible to noise and outliers

- Biased towards globular clusters

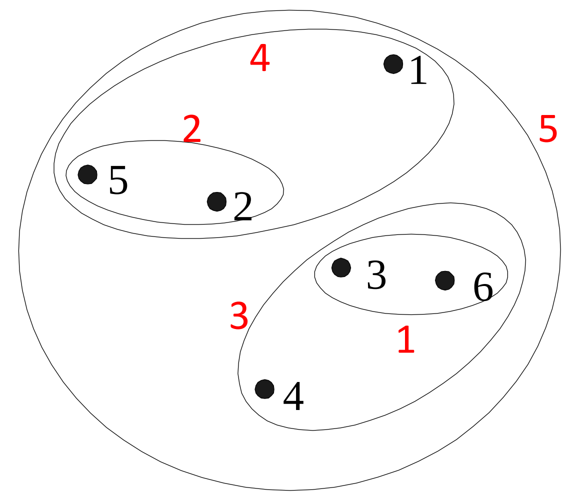
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

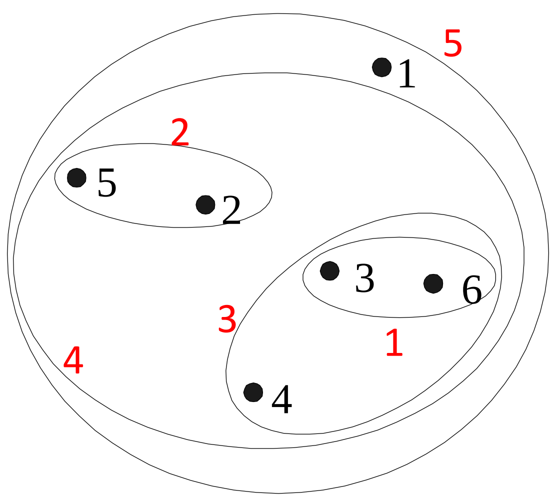
Comparison



MIN

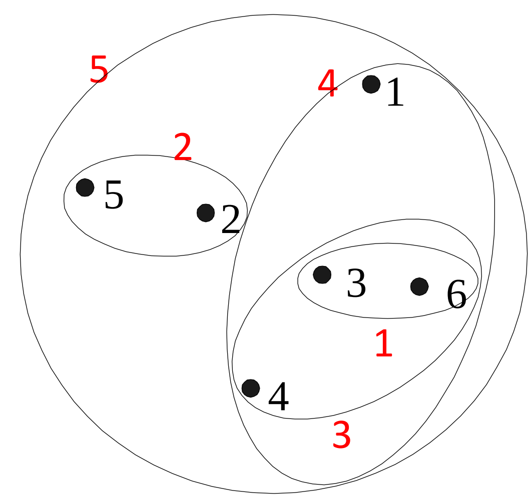
MAX





Group Average

Ward's Method



Time and Space Requirements

- O(N^2) space since it uses the distance matrix
  - N is the number of points
- O(N^3) time in many cases
  - There are N steps and at each step the size, N^2, distance matrix must be updated and searched
  - Complexity can be reduced to O(N^2 log(N)) time for some approaches

Strengths

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level

- They may correspond to meaningful taxonomies
  - e.g., shopping websites—electronics (computer, camera, ..), furniture, groceries

Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and irregular shapes
  - Breaking large clusters

Take-away Message

- Agglomerative and divisive hierarchical clustering
- Several ways of defining inter-cluster distance
- The properties of clusters outputted by different approaches based on different inter-cluster distance definition
- Pros and cons of hierarchical clustering