# DATA MINING LECTURE NOTES-2

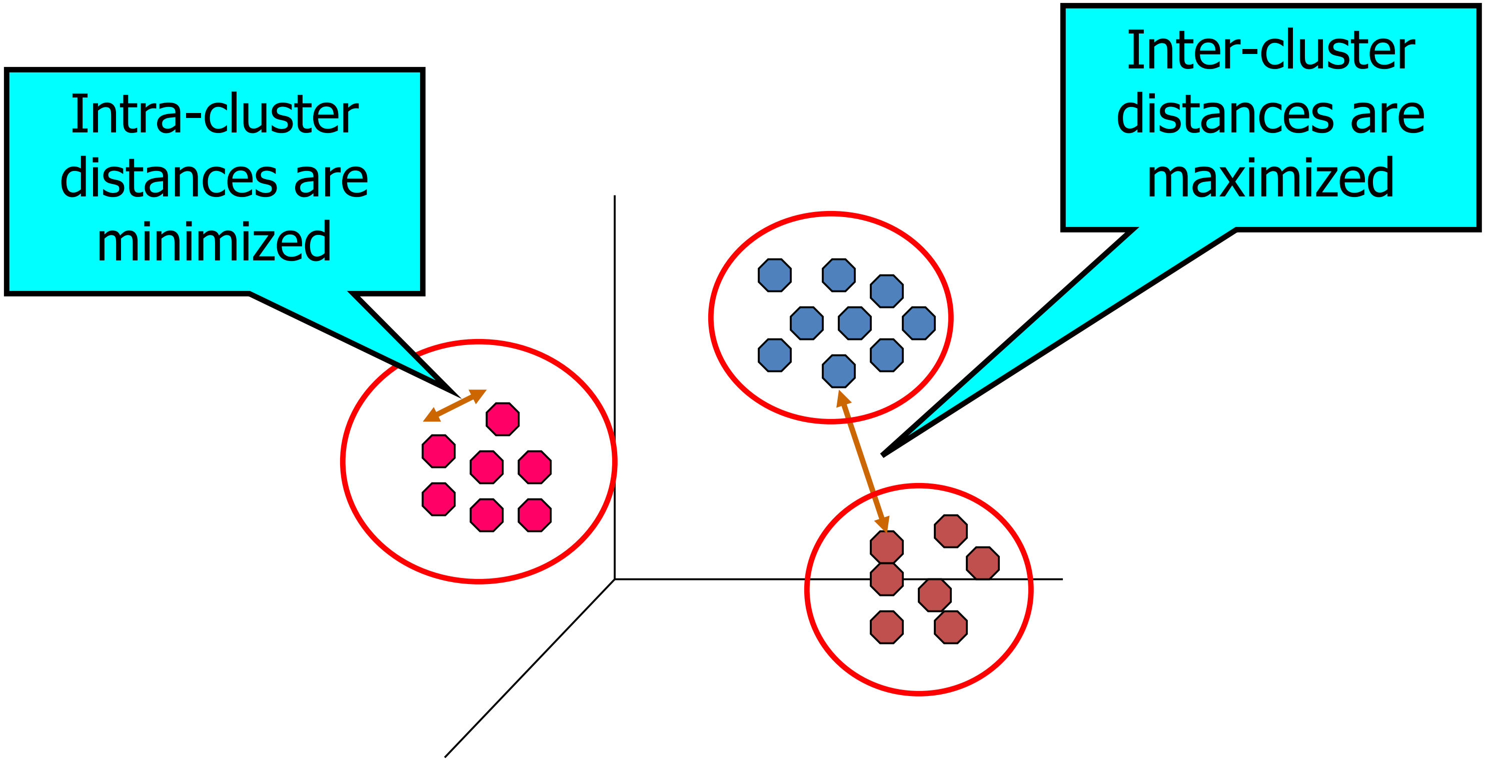
BSc.(H) Computer Science: VI Semester

Teacher: Ms. Sonal Linda

# CLUSTERING

# What is a Clustering?

- In general a grouping of objects such that the objects in a group (cluster) are similar (or related) to one another and different from (or unrelated to) the objects in other groups



# Clustering Algorithms

- K-means and its variants
- Hierarchical clustering
- DBSCAN

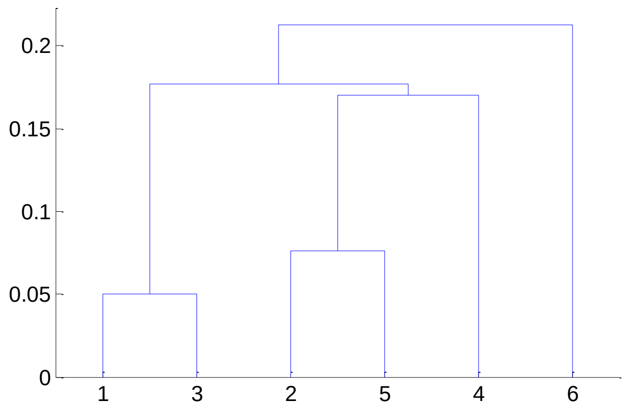
# HIERARCHICAL CLUSTERING

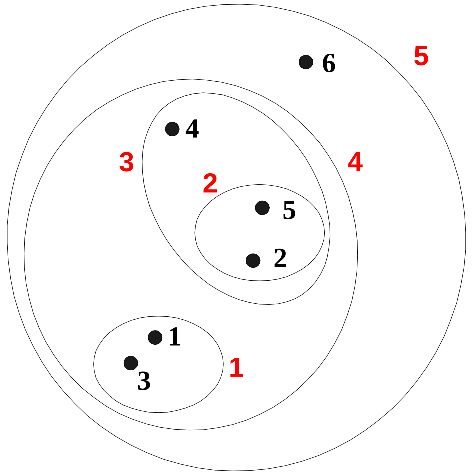
# Hierarchical Clustering

- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

# Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits



# Strengths of Hierarchical Clustering

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

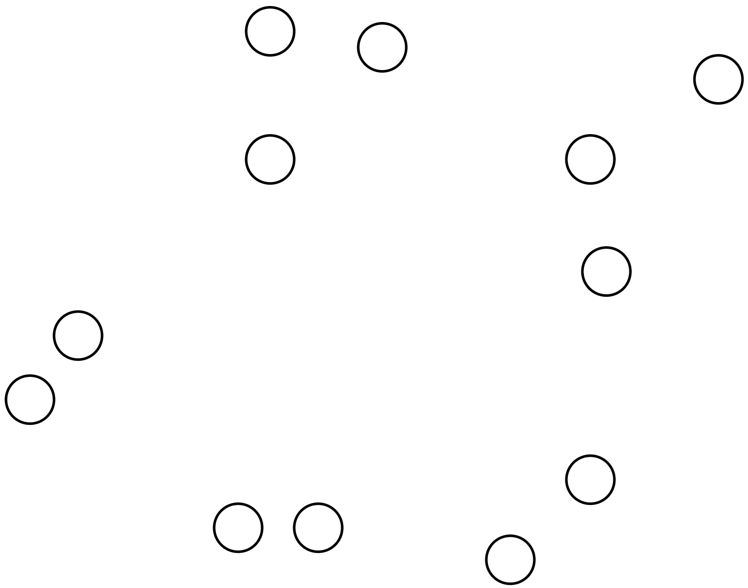
- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

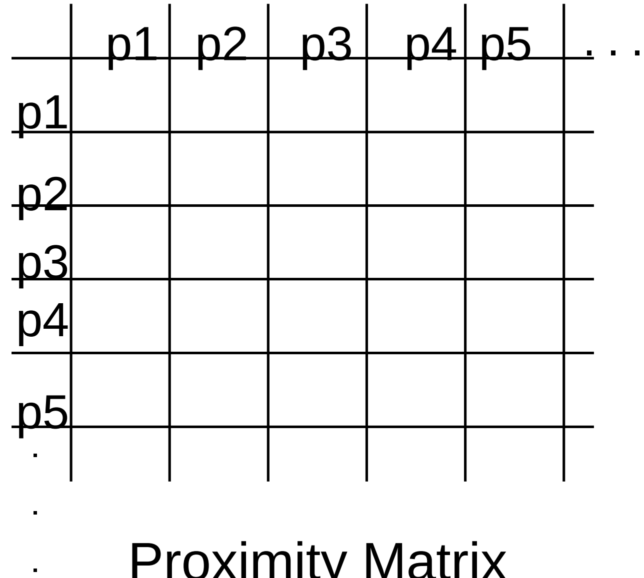
# Agglomerative Clustering Algorithm

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

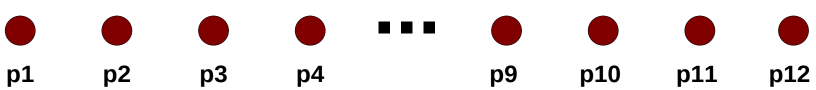
# Starting Situation

- Start with clusters of individual points and a proximity matrix



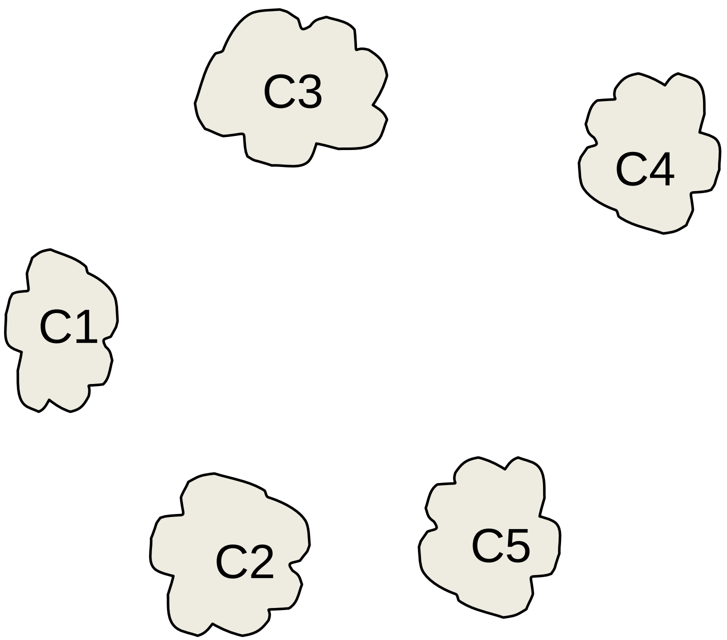



Proximity Matrix



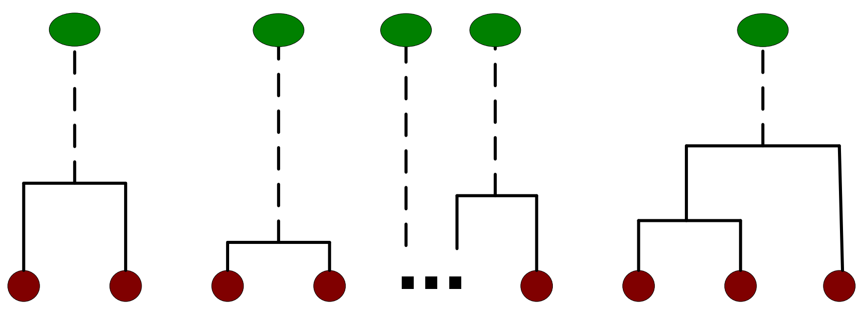
# Intermediate Situation

- After some merging steps, we have some clusters



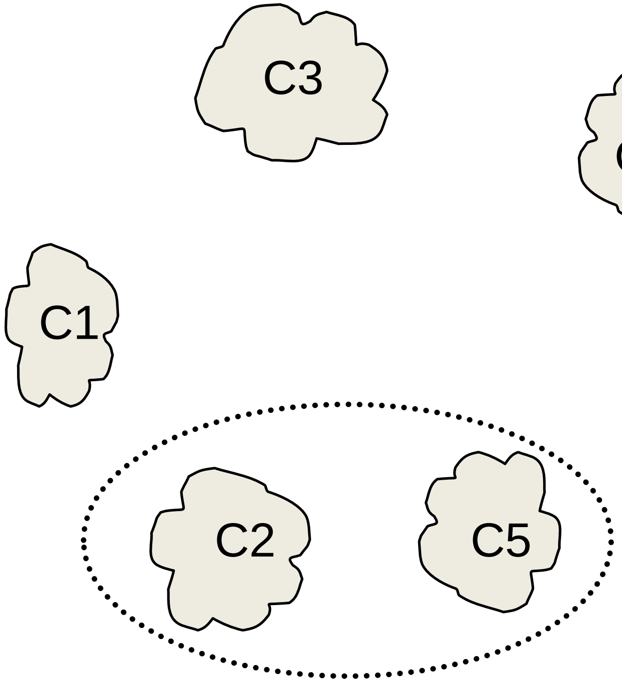

| | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| C1 | | | | | |
| C2 | | | | | |
| C3 | | | | | |
| C4 | | | | | |
| C5 | | | | | |

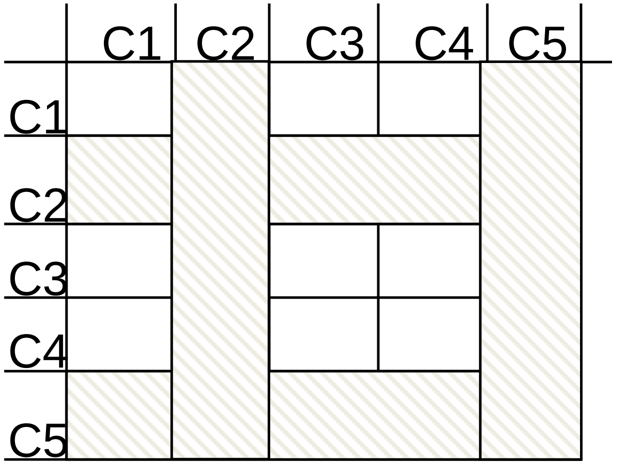
Proximity Matrix



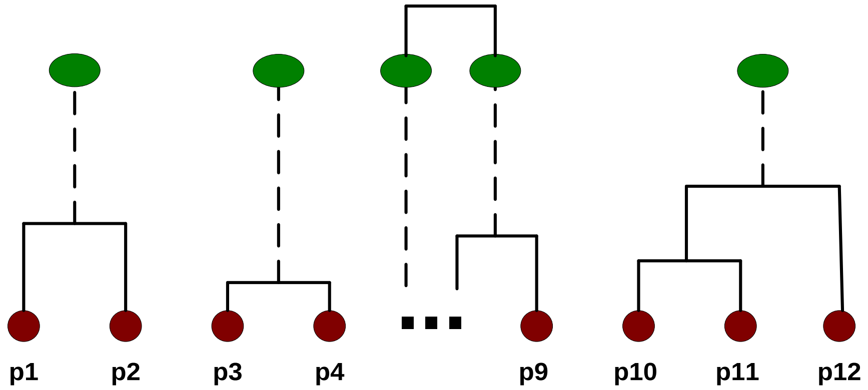
# Intermediate Situation

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



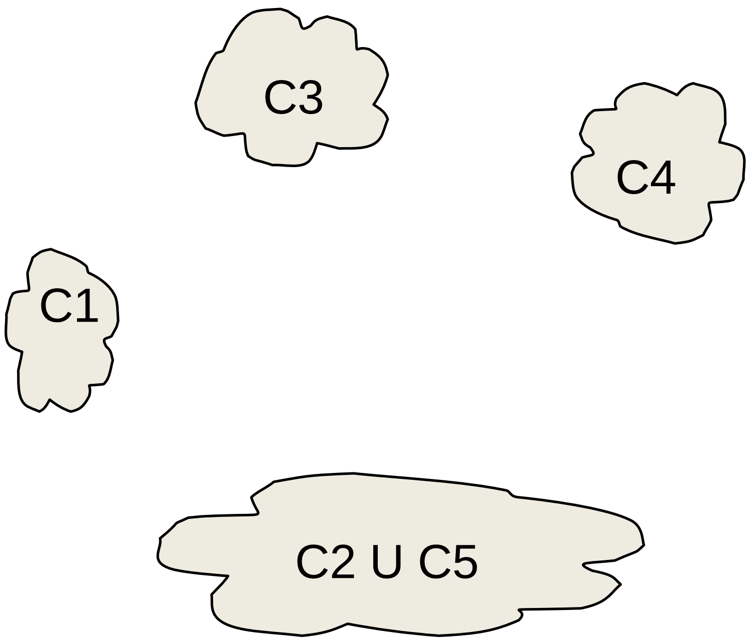




Proximity Matrix



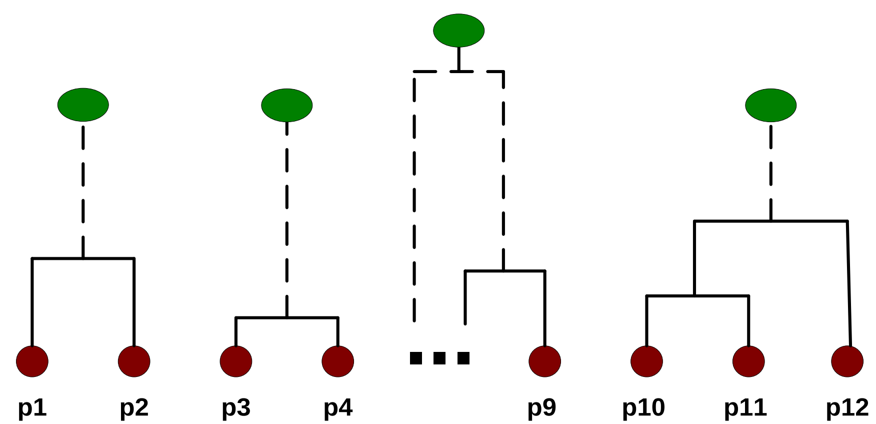
# After Merging

- The question is "How do we update the proximity matrix?"

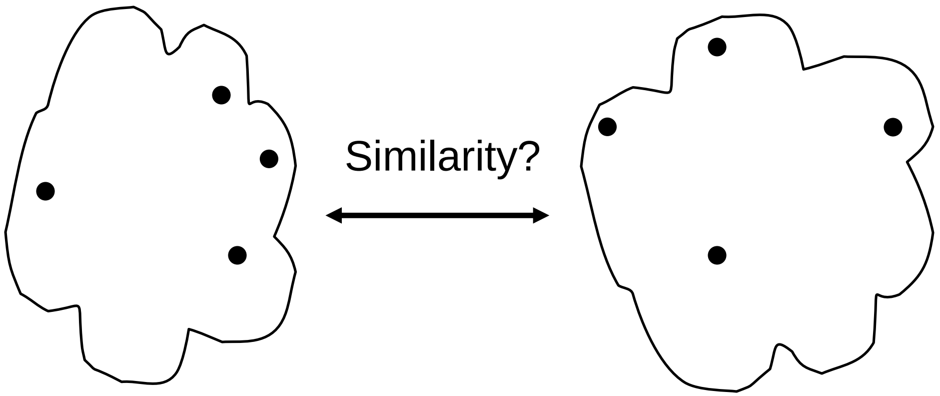



| | C1 | C2 U C5 | C3 | C4 |
|---|---|---|---|---|
| C1 | | ? | | |
| C2 U C5 | ? | ? | ? | ? |
| C3 | | ? | | |
| C4 | | ? | | |

Proximity Matrix



# How to Define Inter-Cluster Similarity

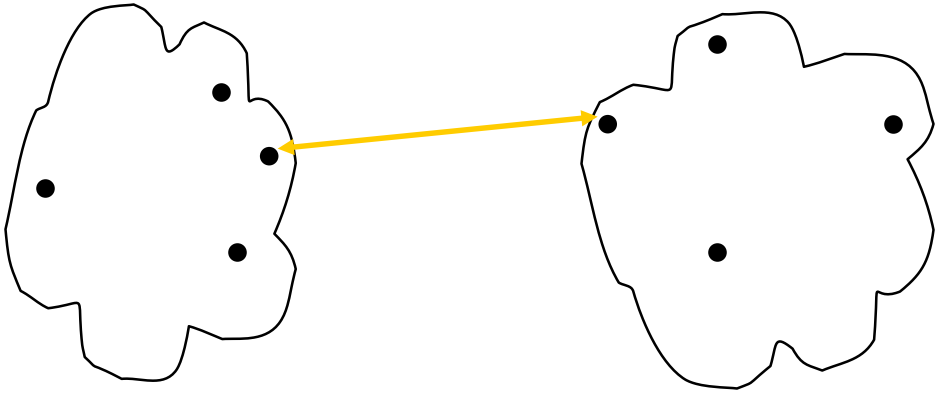



| | p1 | p2 | p3 | p4 | p5 | . . . |
|---|---|---|---|---|---|---|
| p1 | | | | | | |
| p2 | | | | | | |
| p3 | | | | | | |
| p4 | | | | | | |
| p5 | | | | | | |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

# How to Define Inter-Cluster Similarity

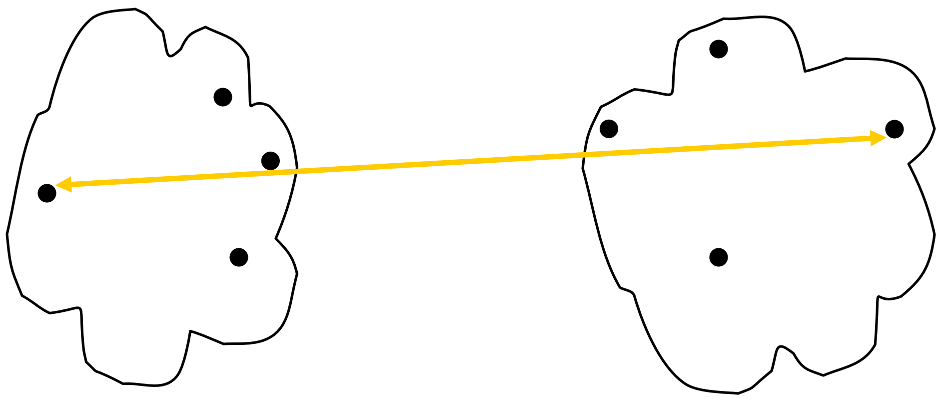


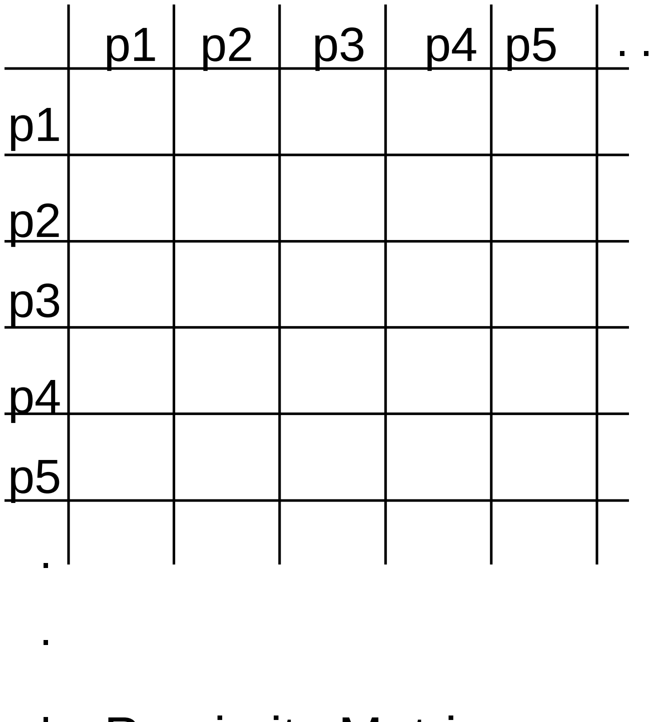
| | p1 | p2 | p3 | p4 | p5 | . . . |
|---|---|---|---|---|---|---|
| p1 | | | | | | |
| p2 | | | | | | |
| p3 | | | | | | |
| p4 | | | | | | |
| p5 | | | | | | |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

# How to Define Inter-Cluster Similarity

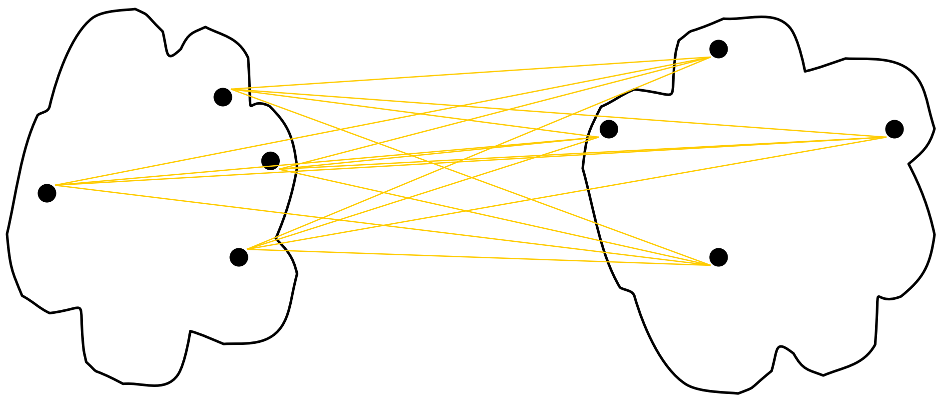





Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  – Ward's Method uses squared error

# How to Define Inter-Cluster Similarity

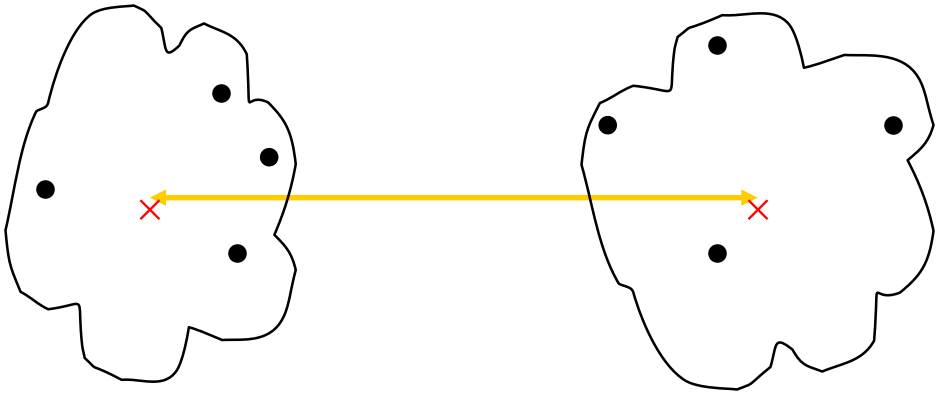


| | p1 | p2 | p3 | p4 | p5 | . . . |
|---|---|---|---|---|---|---|
| p1 | | | | | | |
| p2 | | | | | | |
| p3 | | | | | | |
| p4 | | | | | | |
| p5 | | | | | | |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward’s Method uses squared error

# How to Define Inter-Cluster Similarity



| | p1 | p2 | p3 | p4 | p5 | . . . |
|---|---|---|---|---|---|---|
| p1 | | | | | | |
| p2 | | | | | | |
| p3 | | | | | | |
| p4 | | | | | | |
| p5 | | | | | | |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |

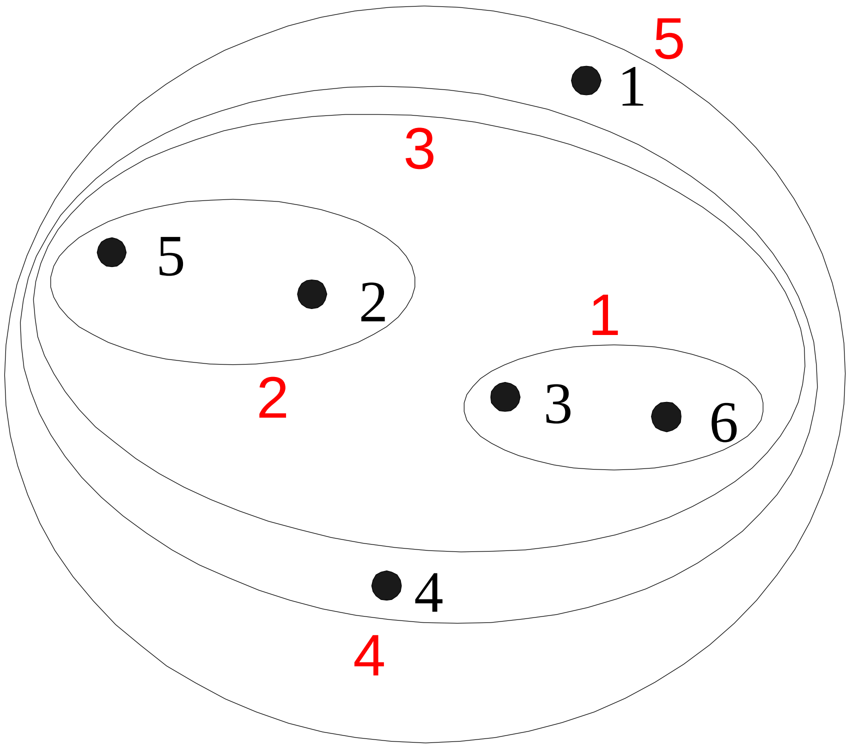
Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  – Ward's Method uses squared error

# Single Link – Complete Link

- Another way to view the processing of the hierarchical algorithm is that we create links between their elements in order of increasing distance
  - The MIN – Single Link, will merge two clusters when a single pair of elements is linked
  - The MAX – Complete Linkage will merge two clusters when all pairs of elements have been linked.

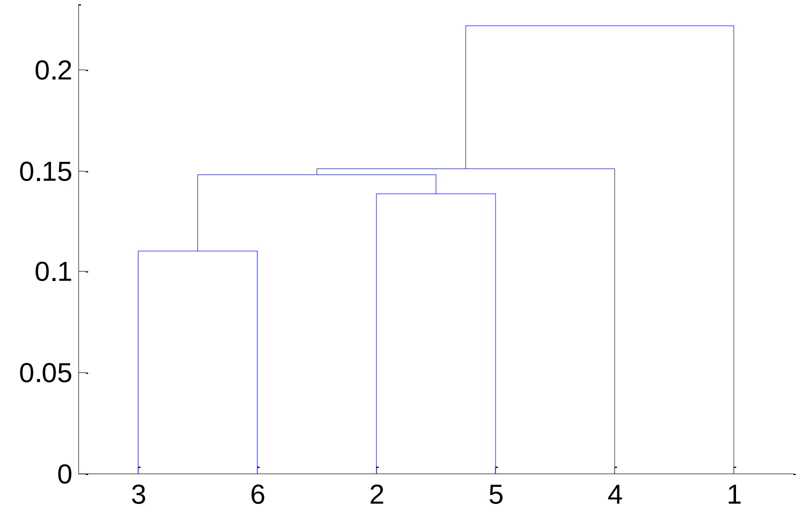
# Hierarchical Clustering: MIN



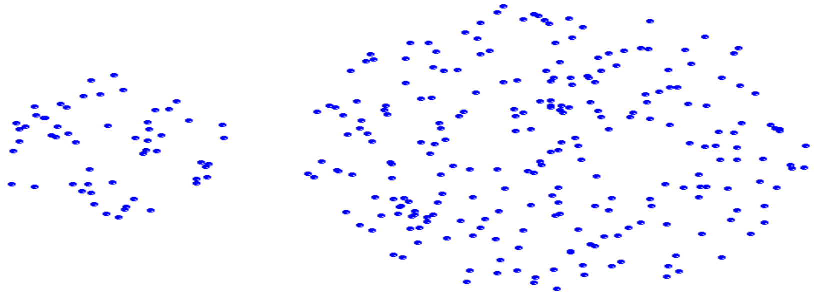

Nested Clusters

| | 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 |

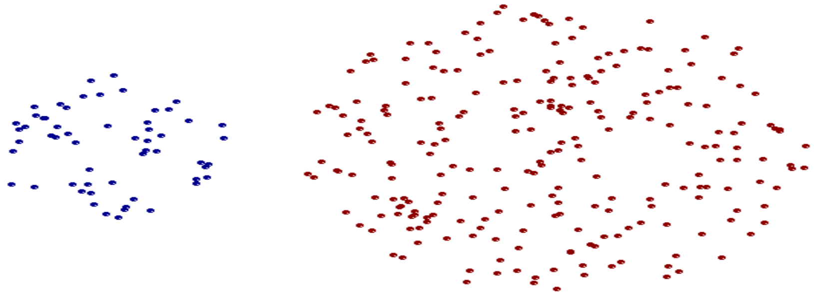
Dendrogram



# Strength of MIN



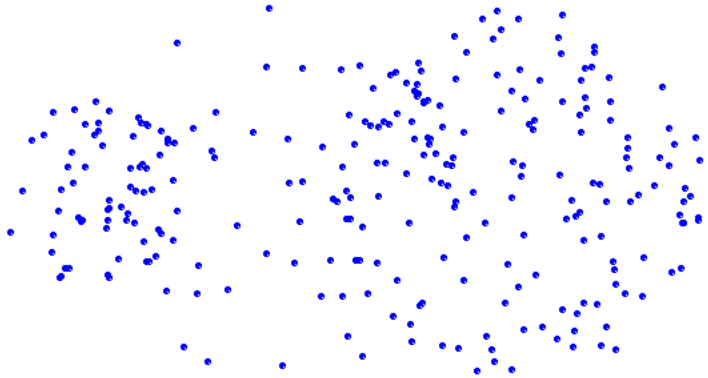
Original Points



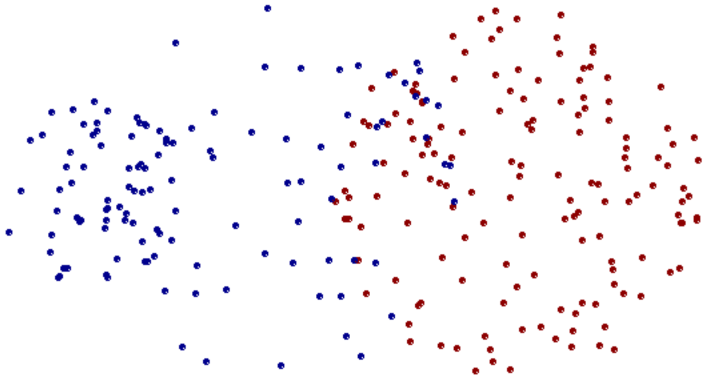
Two Clusters

- Can handle non-elliptical shapes

# Limitations of MIN



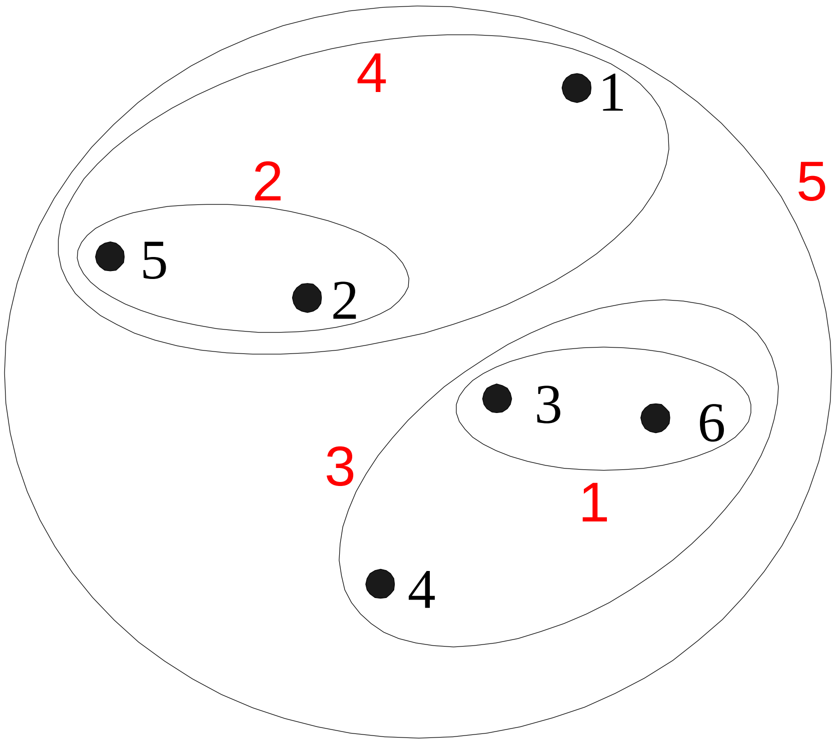
Original Points



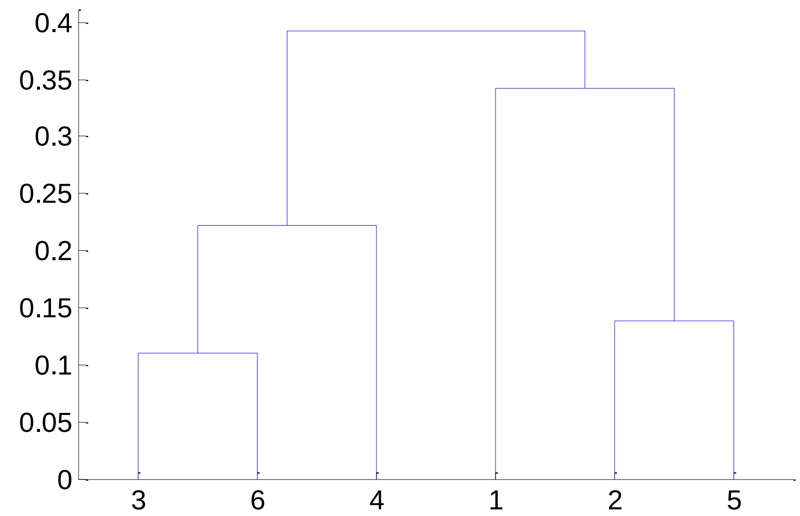
Two Clusters

- Sensitive to noise and outliers

# Hierarchical Clustering: MAX



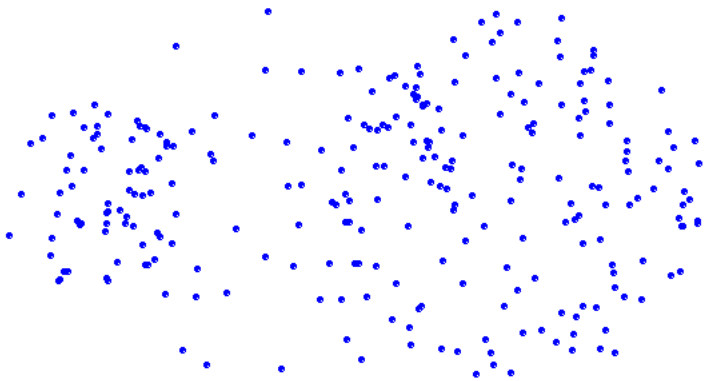

|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | .24 | .22 | .37 | .34 | .23 |
| 2 | .24 | 0 | .15 | .20 | .14 | .25 |
| 3 | .22 | .15 | 0 | .15 | .28 | .11 |
| 4 | .37 | .20 | .15 | 0 | .29 | .22 |
| 5 | .34 | .14 | .28 | .29 | 0 | .39 |
| 6 | .23 | .25 | .11 | .22 | .39 | 0 |



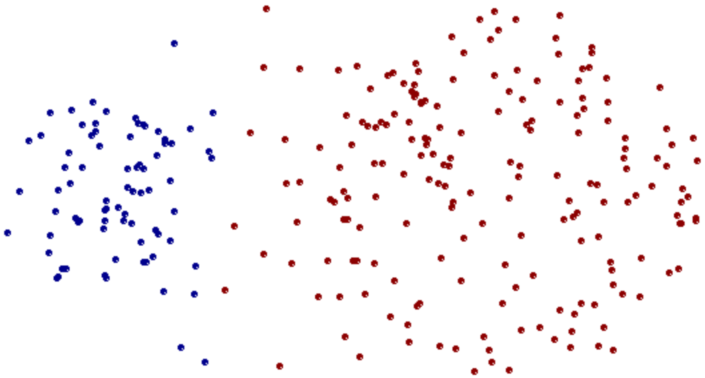

Nested Clusters

Dendrogram

# Strength of MAX



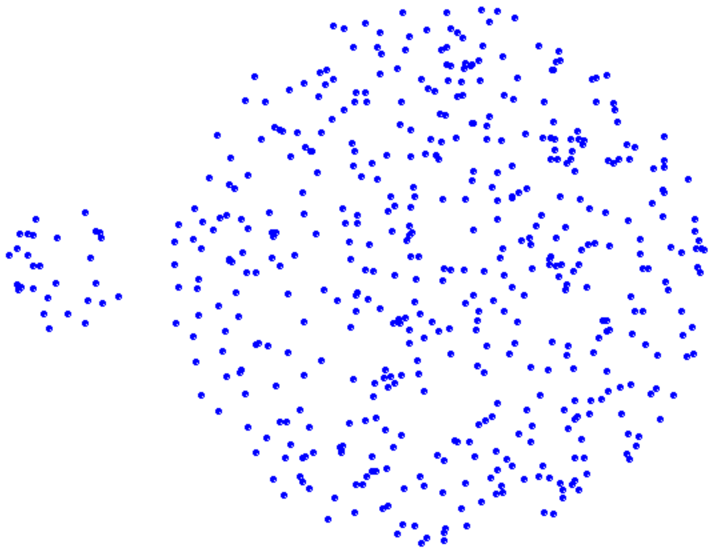
Original Points



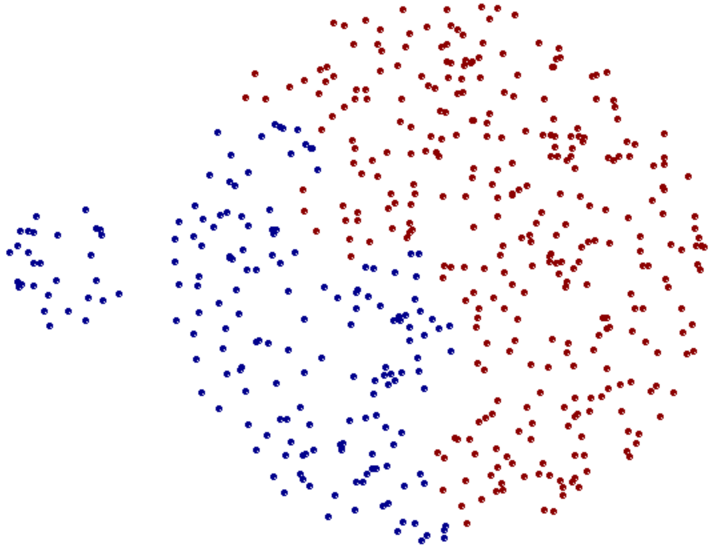
Two Clusters

- Less susceptible to noise and outliers

# Limitations of MAX



Original Points



Two Clusters

- Tends to break large clusters
- Biased towards globular clusters

# Cluster Similarity: Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

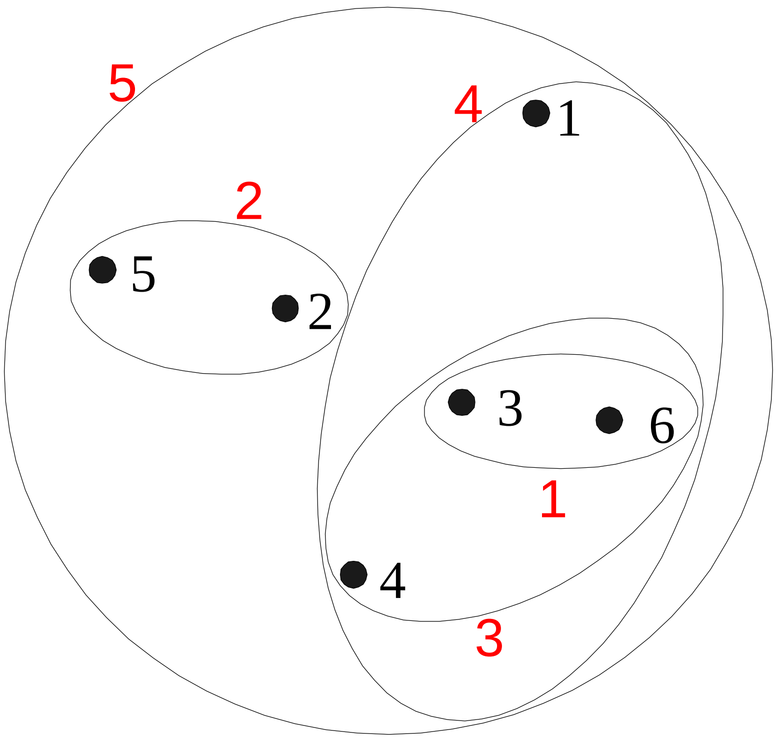
$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum\limits_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

- Need to use average connectivity for scalability since total proximity favors large clusters

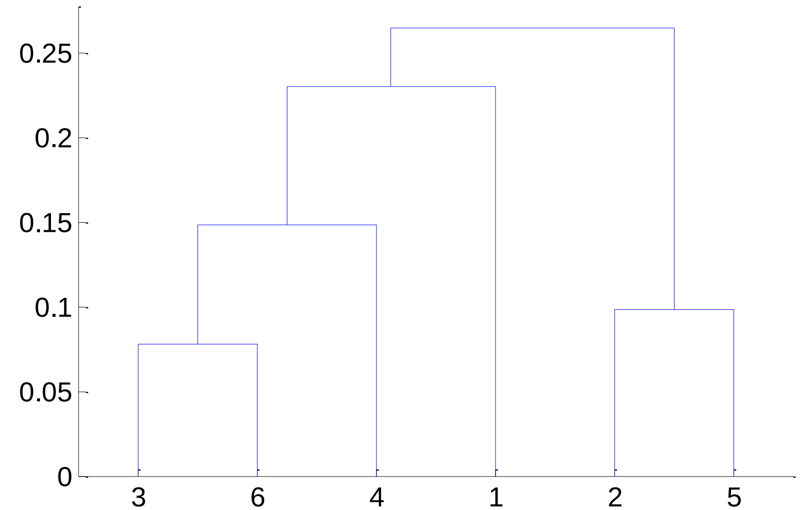
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | .24 | .22 | .37 | .34 | .23 |
| 2 | .24 | 0 | .15 | .20 | .14 | .25 |
| 3 | .22 | .15 | 0 | .15 | .28 | .11 |
| 4 | .37 | .20 | .15 | 0 | .29 | .22 |
| 5 | .34 | .14 | .28 | .29 | 0 | .39 |
| 6 | .23 | .25 | .11 | .22 | .39 | 0 |

# Hierarchical Clustering: Group Average

| | 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 |




Nested Clusters




Dendrogram

# Hierarchical Clustering: Group Average

- Compromise between Single and Complete Link

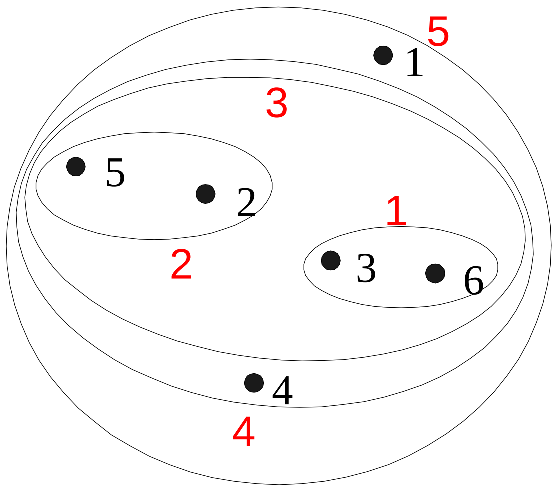
- Strengths
  - Less susceptible to noise and outliers

- Limitations
  - Biased towards globular clusters

# Cluster Similarity: Ward's Method

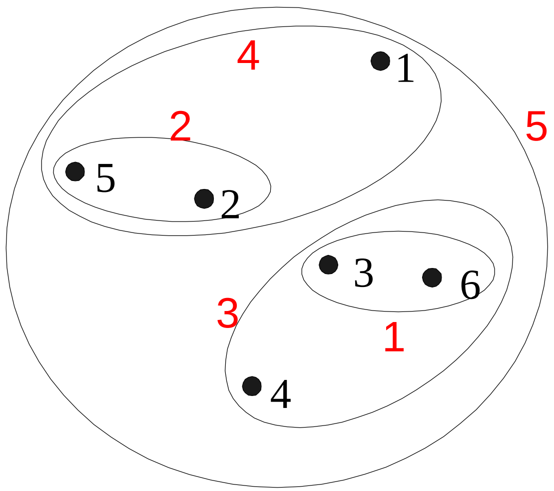
- Similarity of two clusters is based on the increase in squared error (SSE) when two clusters are merged
  - Similar to group average if distance between points is distance squared
- Less susceptible to noise and outliers
- Biased towards globular clusters
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

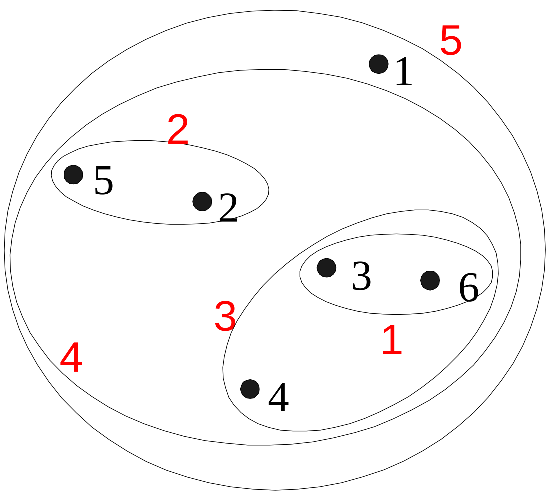
# Hierarchical Clustering: Comparison




MIN

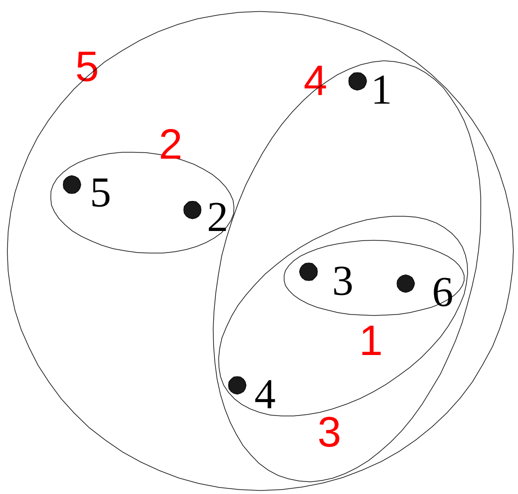
MAX




Group Average

Ward's Method



# Hierarchical Clustering: Time and Space requirements

- $O(N^2)$ space since it uses the proximity matrix.
  - N is the number of points.

- $O(N^3)$ time in many cases
  - There are N steps and at each step the size, $N^2$, proximity matrix must be updated and searched
  - Complexity can be reduced to $O(N^2 \log(N))$ time for some approaches

# Hierarchical Clustering: Problems and Limitations

- Computational complexity in time and space
- 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 convex shapes
  - Breaking large clusters