Hierarchical Data Analysis

Distance Metrics in Hierarchical Clustering

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

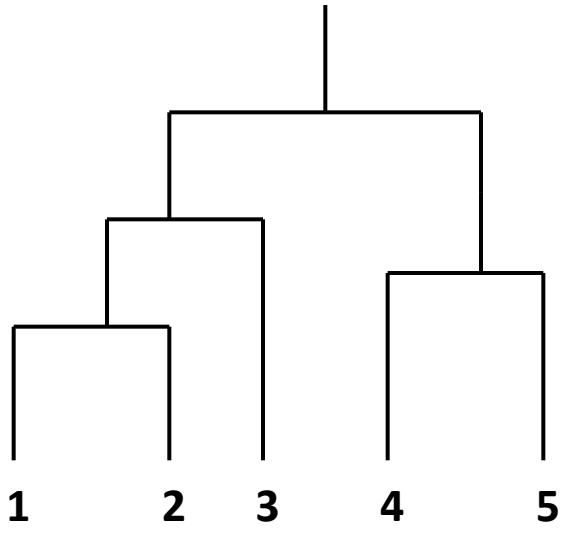
Objective

Apply methods of
hierarchical data analysis

Single-link Clustering: Example

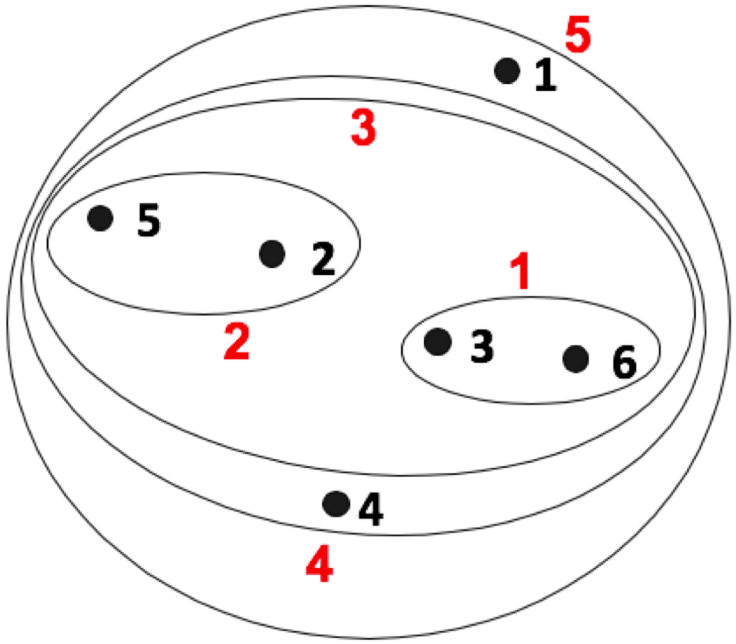
- Determined by one pair of points, i.e., by one link in proximity graph

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

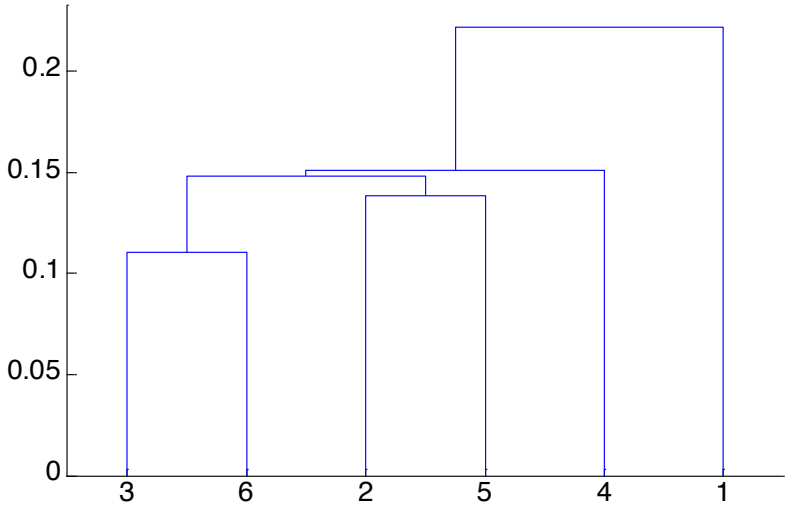


Single-Link Clustering Example

Nested Clusters

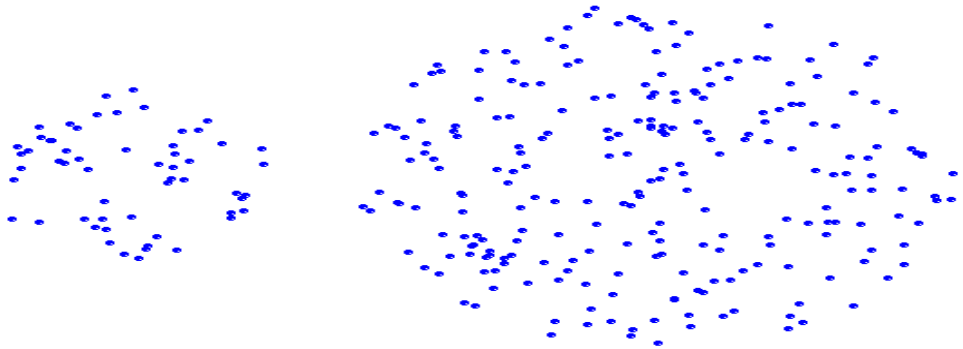


Dendrogram

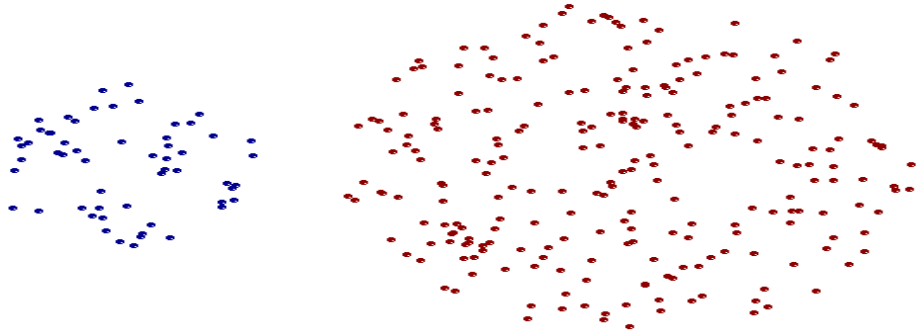


Strengths of Single-Link Clustering

Original Points



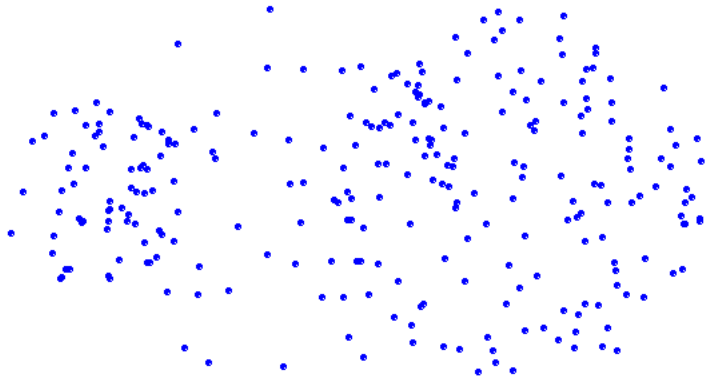
Two Clusters



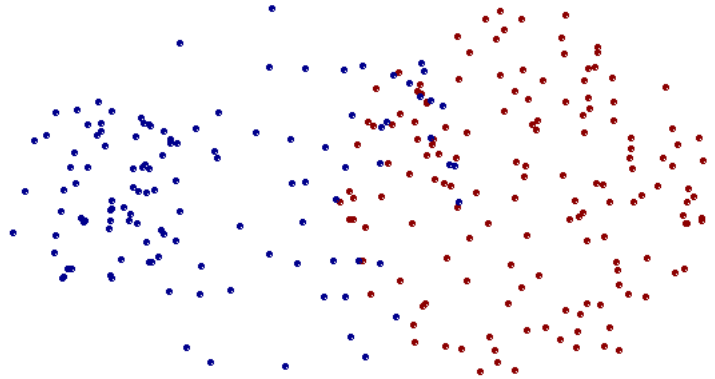
Can handle non-elliptical shapes

Limitations of Single-Link Clustering

Original Points



Two Clusters



- Sensitive to noise and outliers
- It produces long, elongated clusters

Distance Between Two Clusters

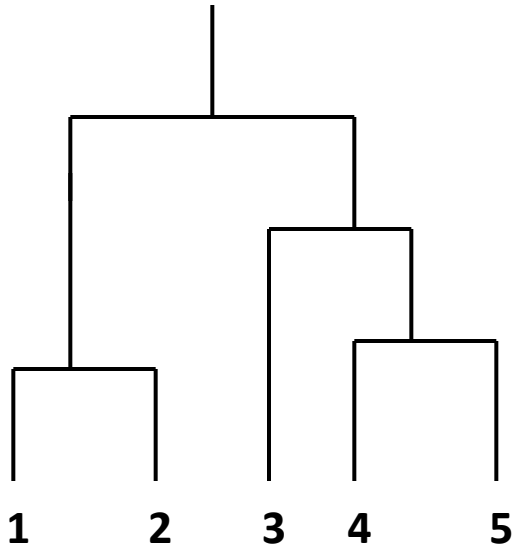
- Complete-link distance between clusters C_i and C_j is the *maximum distance* between any object in C_i and any object in C_j
- The distance is defined by the two most dissimilar objects

$$D_{cl}(C_i, C_j) = \max_{x,y} \{ d(x, y) | x \in C_i, y \in C_j \}$$

Complete-link Clustering: Example

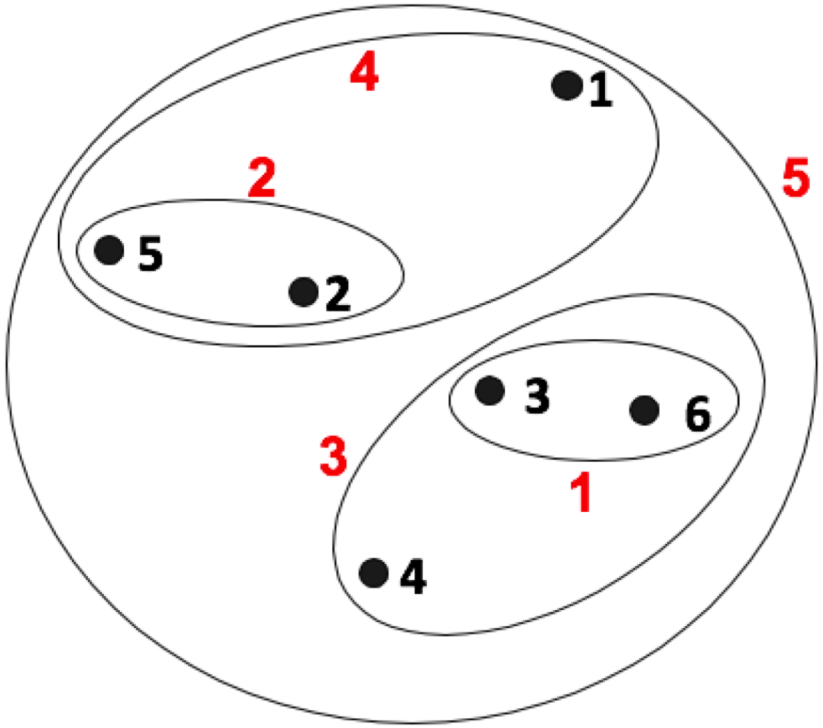
- Distance between clusters is determined by two most distant points in different clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

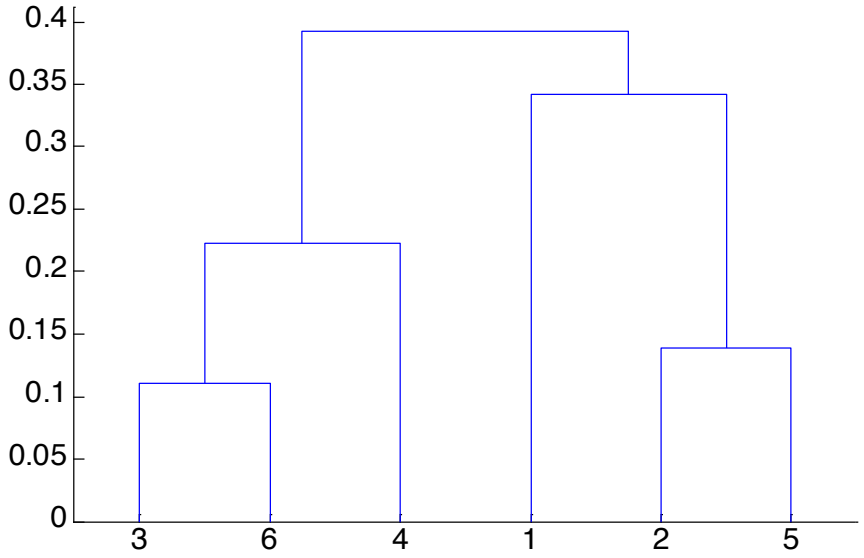


Single-Link Clustering Example

Nested Clusters

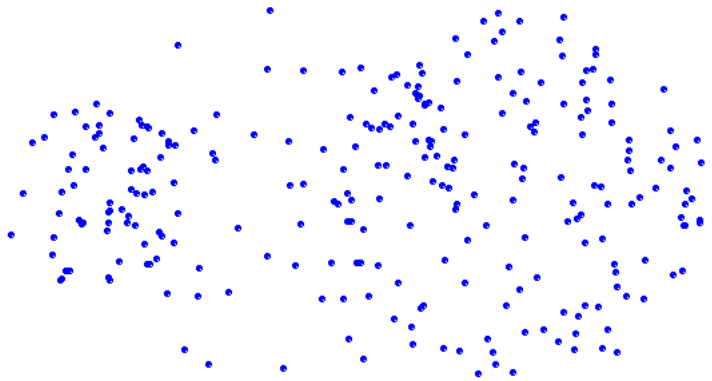


Dendrogram

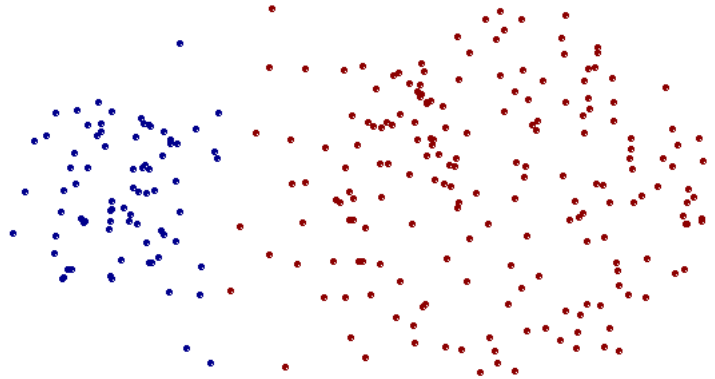


Strengths of Complete-link Clustering

Original Points



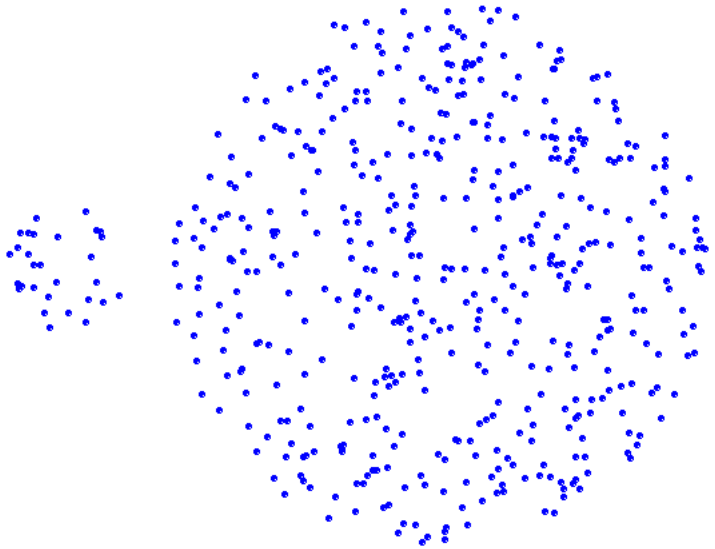
Two Clusters



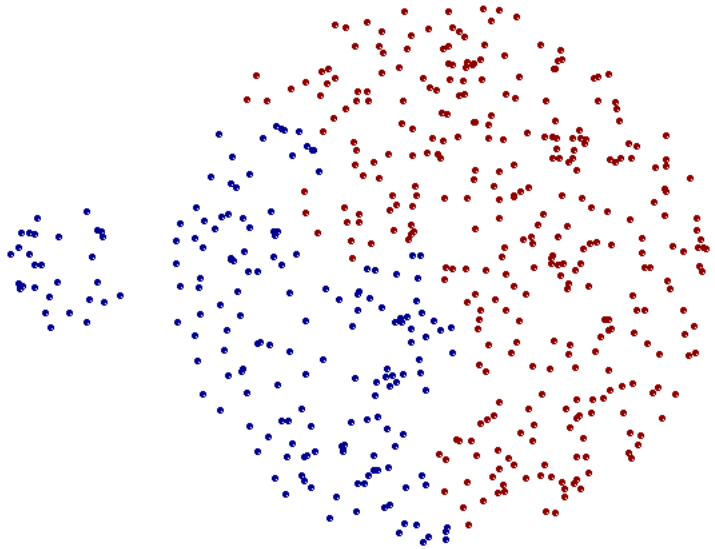
- More balanced clusters (with equal diameter)
- Less susceptible to noise

Limitations of Complete-Link Clustering

Original Points



Two Clusters



- Tends to break large clusters
- All clusters tend to have same diameter – small clusters are merged with larger ones

Distance Between Two Clusters

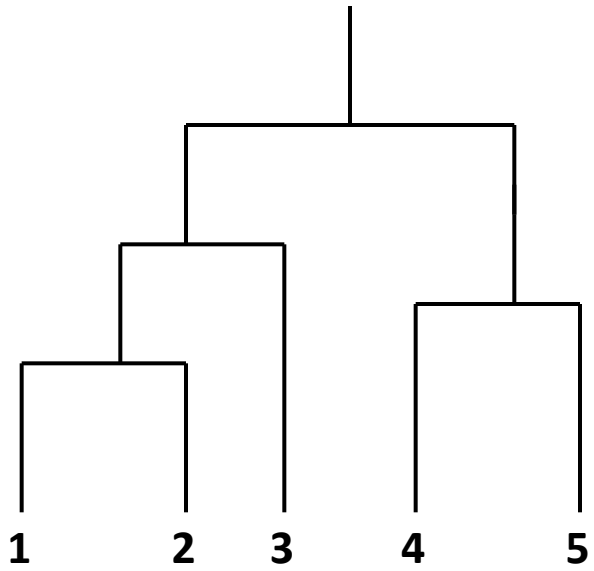
Group average distance between clusters C_i and C_j is the *average distance* between any object in C_i and any object in C_j

$$D_{avg}(C_i, C_j) = \frac{1}{|C_i| \times |C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$$

Average-link Clustering: Example

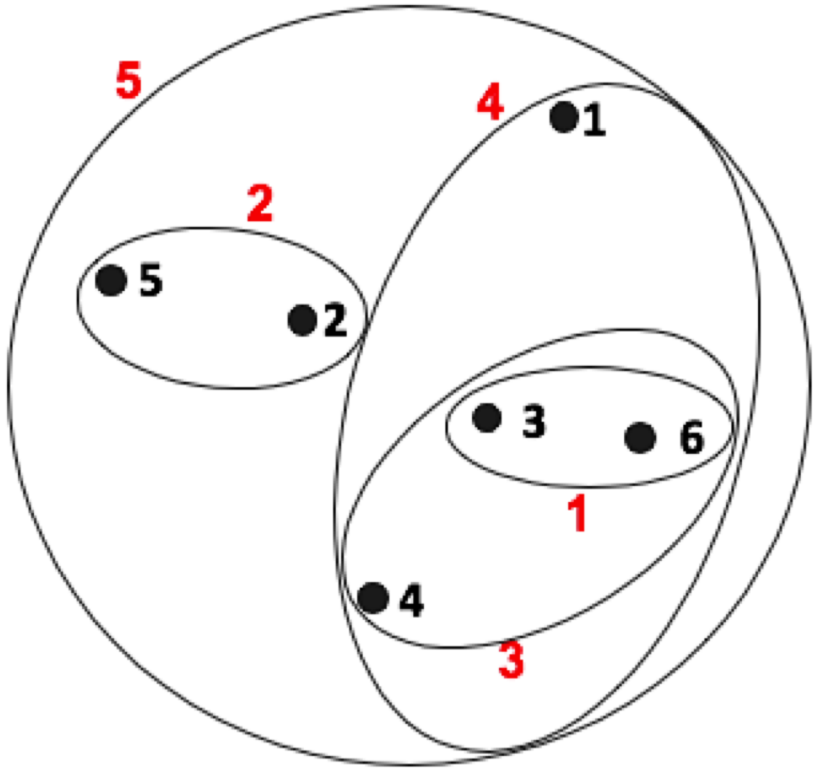
Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

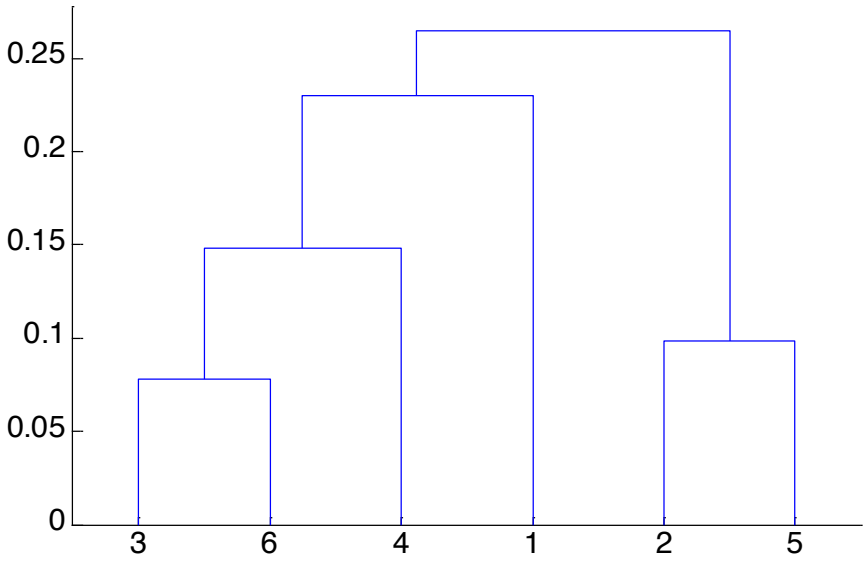


Average-Link Clustering Example

Nested Clusters



Dendrogram



Average-Link Clustering: Discussion

Compromise between Single and Complete Link

Strengths

- Less susceptible to noise and outliers

Limitations

- Biased towards globular clusters

Distance Between Two Clusters

Centroid distance between clusters C_i and C_j is the distance between the centroid r_i of C_i and the centroid r_j of C_j

$$D_{centroids}(C_i, C_j) = d(r_i, r_j)$$

Distance Between Two Clusters

Ward's distance between clusters C_i and C_j is the *difference* between the *total within cluster sum of squares for the two clusters separately*, and the *within cluster sum of squares resulting from merging the two clusters* in cluster C_{ij}

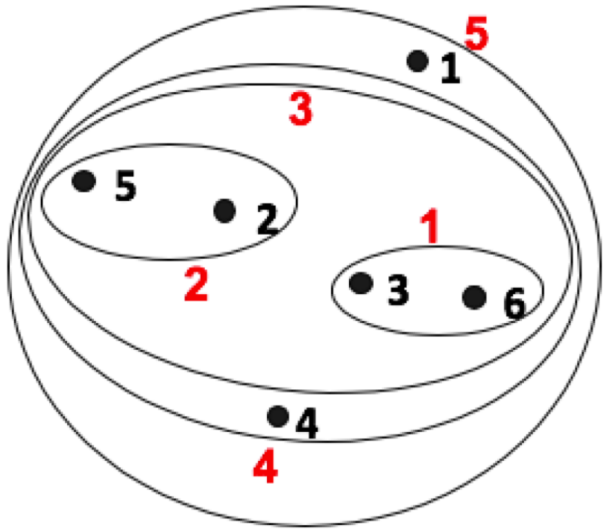
$$D_w(C_i, C_j) = \sum_{x \in C_i} (x - r_i)^2 + \sum_{x \in C_j} (x - r_j)^2 - \sum_{x \in C_{ij}} (x - r_{ij})^2$$

r_i: centroid of C_i
r_j: centroid of C_j
r_{ij}: centroid of C_{ij}

Ward's Distance for Clusters

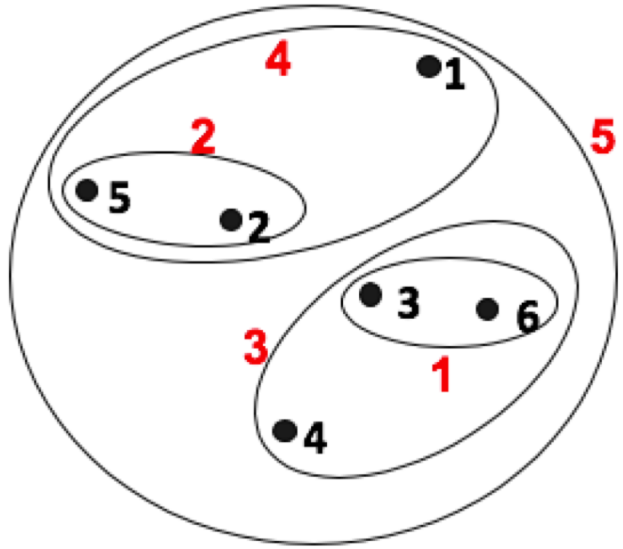
- Similar to group average and centroid distance
- Less susceptible to noise and outliers
- Biased towards globular clusters
- Hierarchical analogue of k-means
  - Can be used to initialize k-means

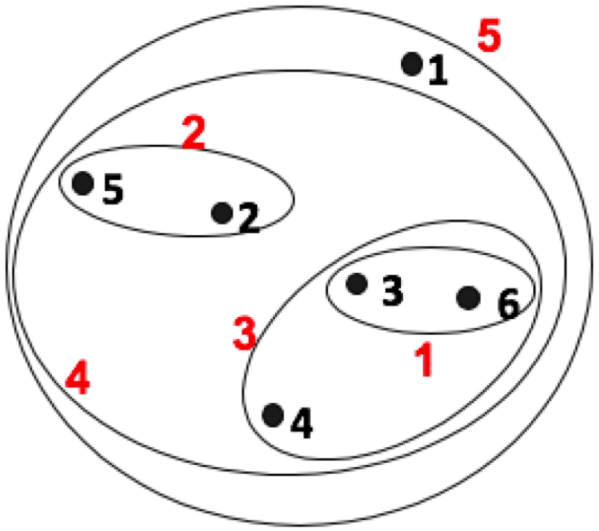
Hierarchical Clustering: Comparison



MIN

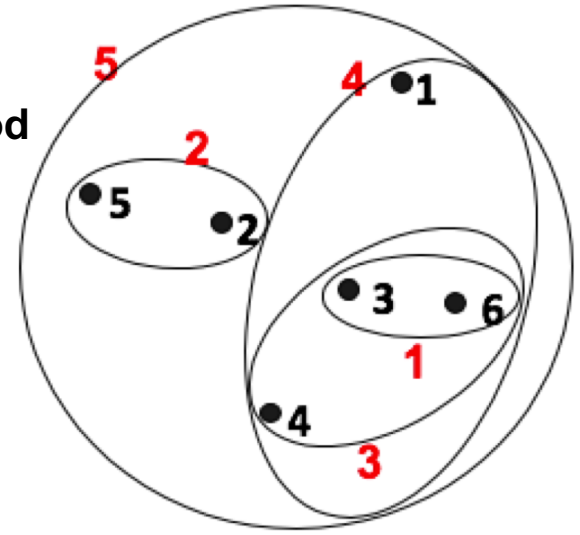
MAX





Group Average

Ward's Method



Hierarchical Clustering: Time and Space Requirements

- For a dataset X consisting of n points
- $O(n^2)$ **space**; it requires storing distance matrix
- $O(n^3)$ **time** in most of the 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 by using appropriate data structures

Hierarchical Clustering Issues

- Distinct clusters are not produced
- Methods for producing distinct clusters but involve specifying somewhat arbitrary cutoff values
- What if data doesn't have a hierarchical structure?
- Is HC appropriate?