Clustering

Real-world example

Example: Chapter 2 DM Concepts and Techniques

Table 2.2 A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	*test-1 (nominal)*	*test-2 (ordinal)*	*test-3 (numeric)*
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Similarity and Dissimilarity

Similarity

- **Numerical measure of how alike two data objects are.**
- Is higher when objects are more alike.
- Often falls in the range [0,1]

Dissimilarity

- **Numerical measure of how different are two data objects**
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

Proximity refers to a similarity or dissimilarity

Example: Chapter 2 DM Concepts and Techniques

Table 2.2 A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- *test-1*
  - $d(i, j)$ evaluates to 0 if objects $i = j$, and 1 otherwise
- *test-2*
  - $d(i, j)$ = |i-j| / (n-1), we have 0 to n-1 values
- *test-3*
  - *Normalize (min-max normalization)*
  - *Distance measure (Manhattan or Euclidean distance)*

Example

Table 2.2 A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	*test-1 (nominal)*	*test-2 (ordinal)*	*test-3 (numeric)*
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$\boxed{d(3,1) = \frac{1(1)+1(0.50)+1(0.45)}{3} = 0.65.}$$

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

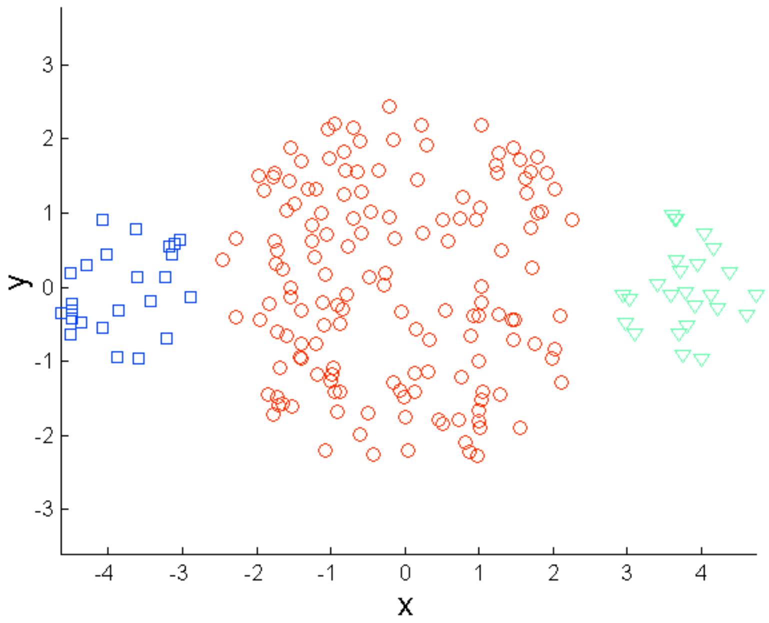
Limitations of K-means

K-means has problems when clusters are of different

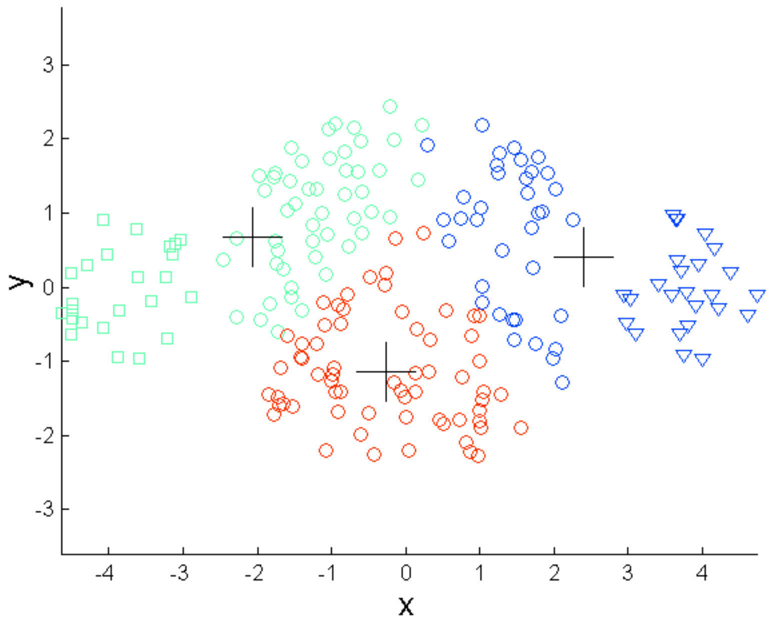
- Sizes
- Densities
- Non-globular shapes

K-means has problems when the data contains outliers.

Limitations of K-means: Clusters with Different Sizes

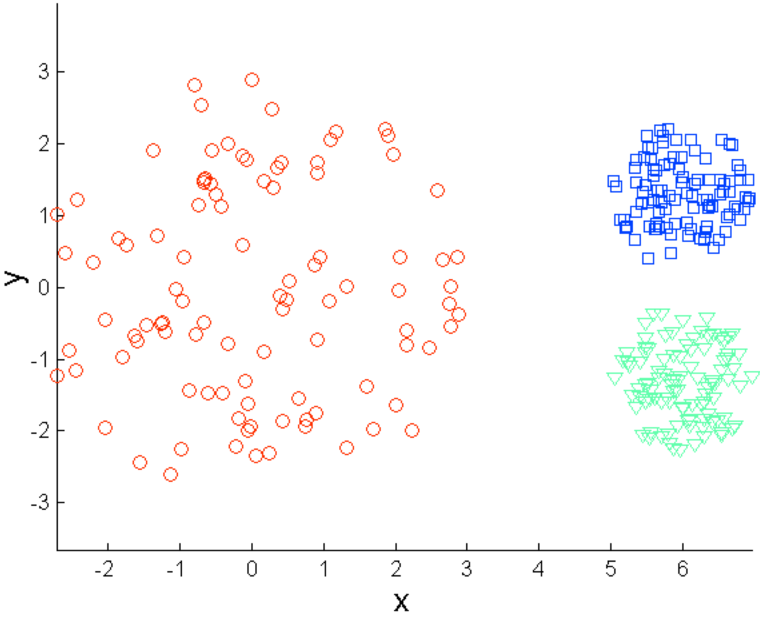


Original Points

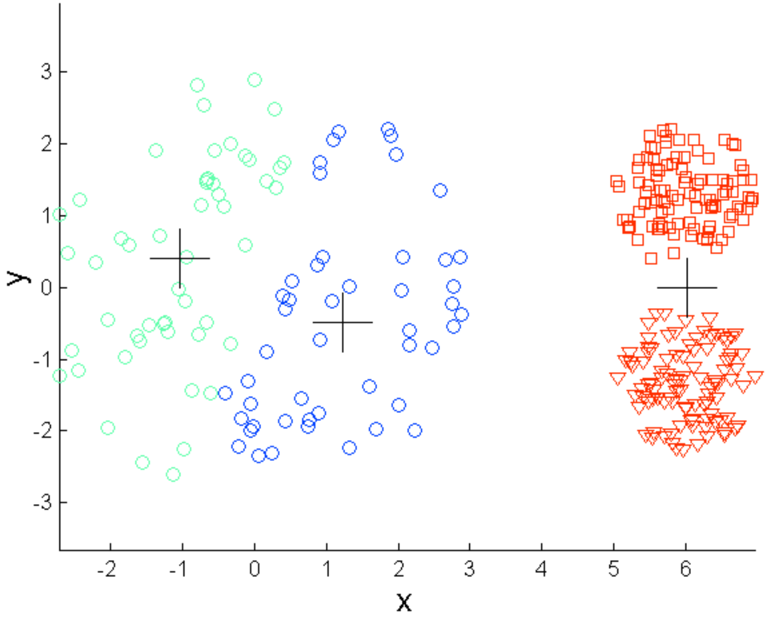


K-means (3 Clusters)

Limitations of K-means: Different Density

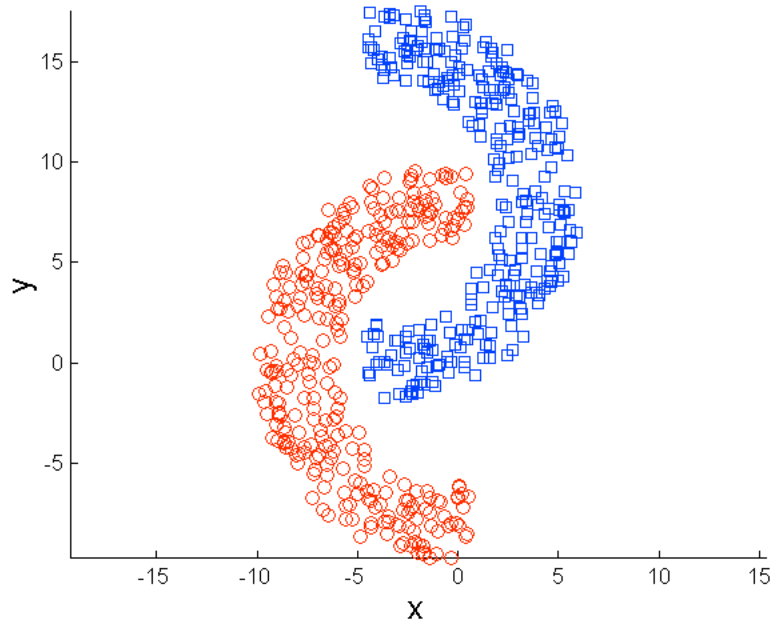


Original Points

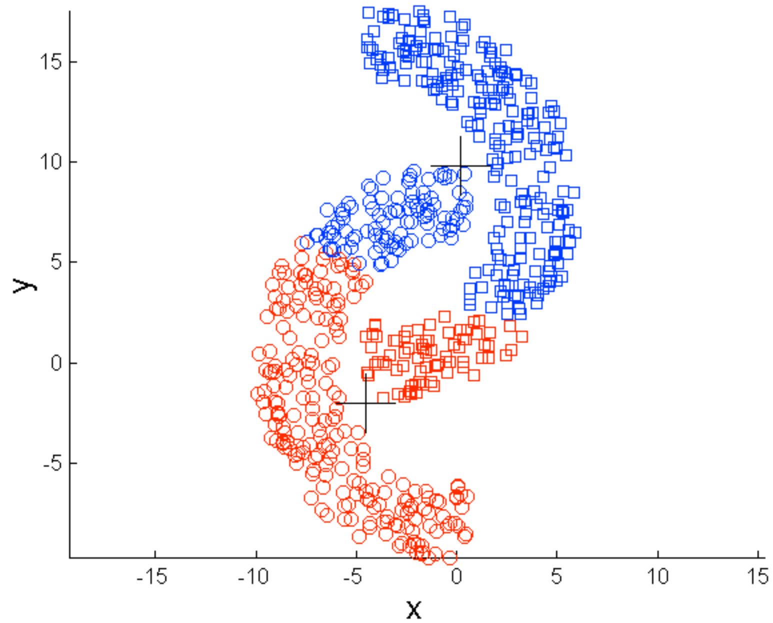


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

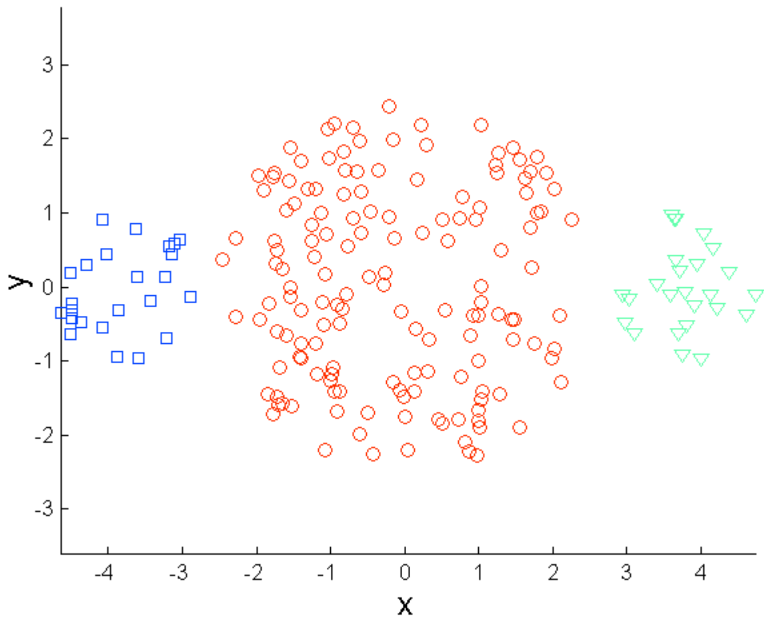


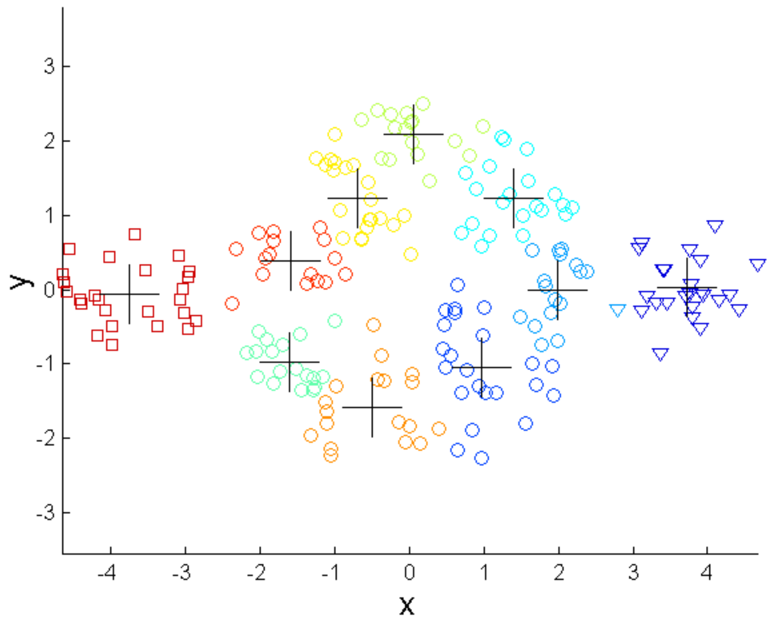
Original Points



K-means (2 Clusters)

Overcoming K-means Limitations



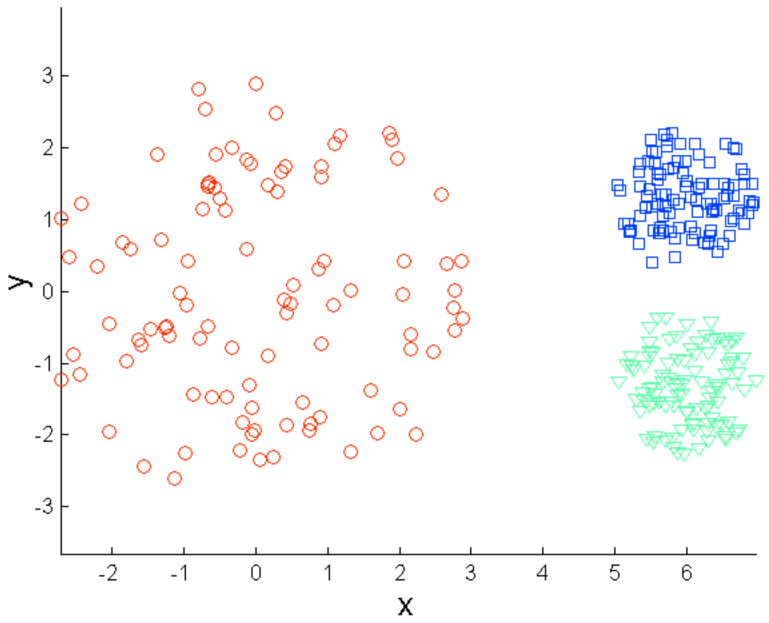


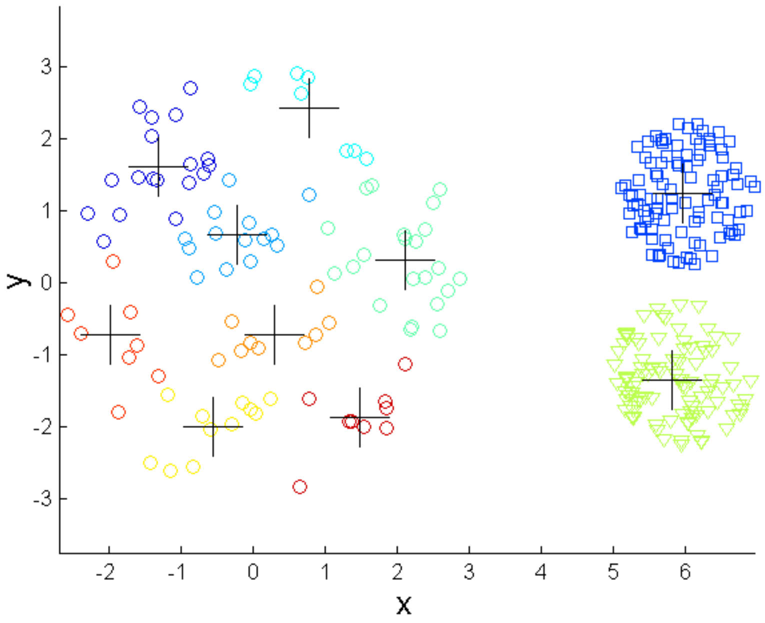
Original Points

K-means Clusters

One solution is to use many clusters.
Find parts of clusters but need to put together.

Overcoming K-means Limitations

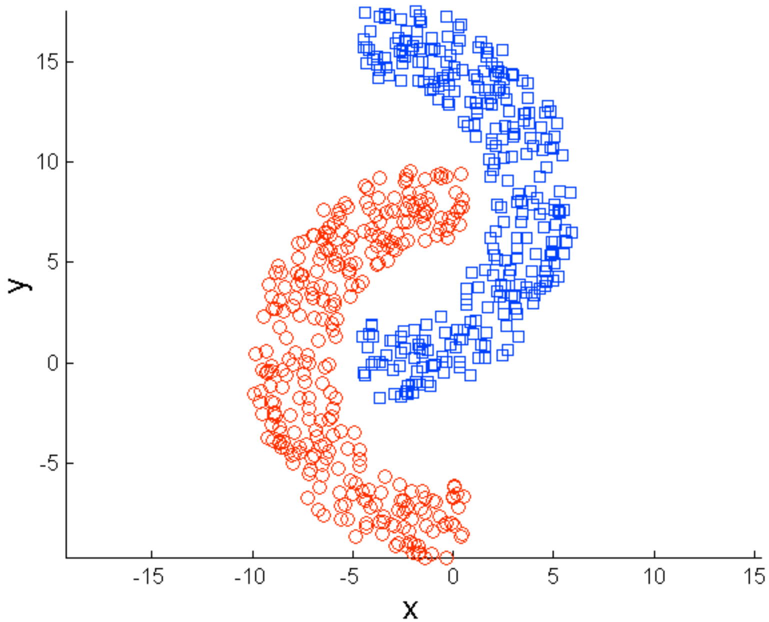


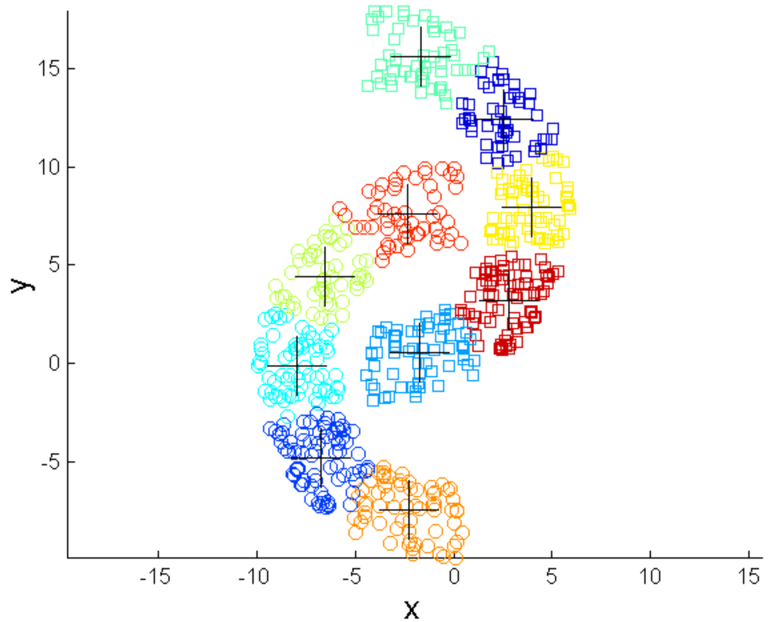


Original Points

K-means Clusters

Overcoming K-means Limitations





Original Points

K-means Clusters

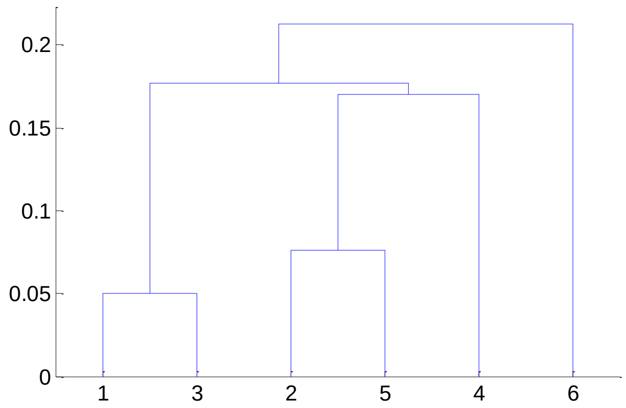
Variations of the *K-Means* Method

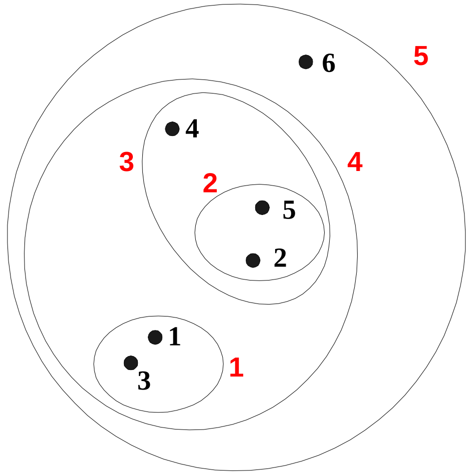
- **A few variants of the *k-means* which differ in**
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- **Handling categorical data: *k-modes* (Huang'98)**
  - Replacing **means** of clusters with **modes**
  - Using new dissimilarity measures to deal with categorical objects
  - Using a **frequency-based method** to update modes of clusters
- **Handling a mixture of categorical and numerical data k-prototype method**

Hierarchical Clustering

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





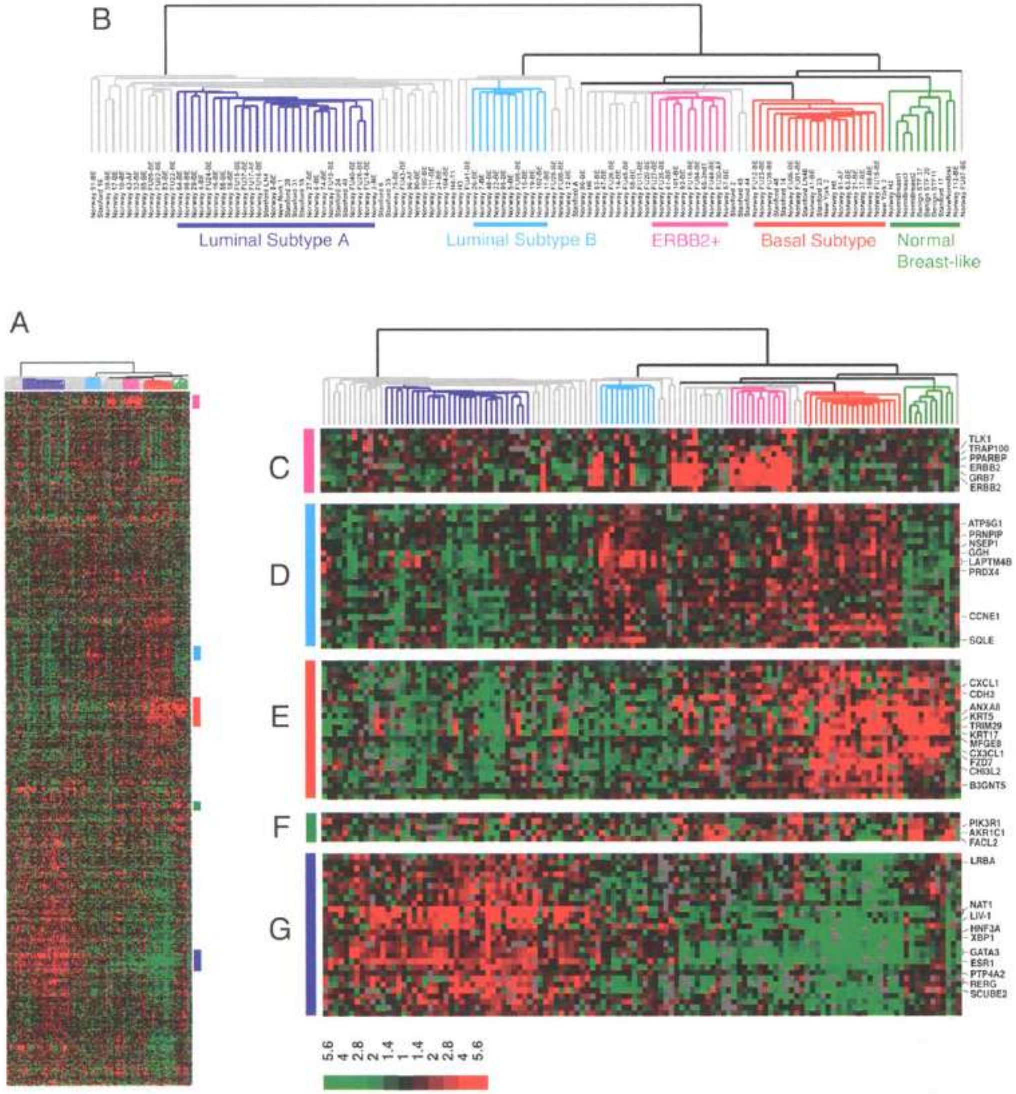
Strengths of Hierarchical Clustering

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

- **Correspond to meaningful taxonomies**
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

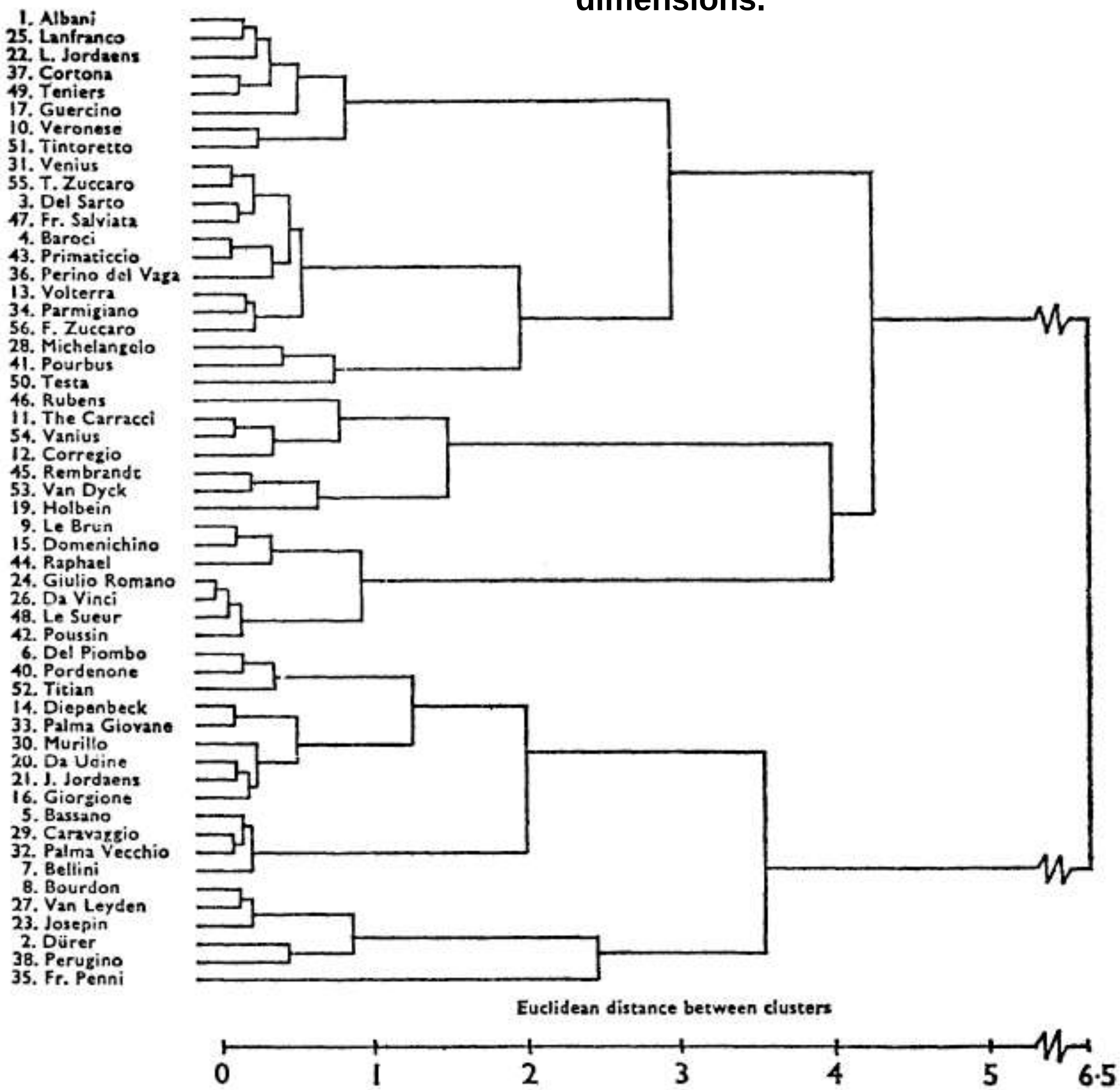
Examples

- Hierarchical clustering of gene expression data lead to new theories
- Later, theories tested in the lab.



"Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets" (Sorlie et al., 2003)

Roger de Piles rated 57 paintings along different dimensions.



Hierarchical vs Kmeans

- Point assignment good when clusters are nice, convex shapes:

- Hierarchical can win when shapes are weird:
  - Note both clusters have essentially the same centroid.

Aside: if you realized you had concentric clusters, you could map points based on distance from center, and turn the problem into a simple, one-dimensional case.

Hierarchical Clustering

Two main types of hierarchical clustering

Agglomerative	Divisive
• Start with the points as individual clusters • At each step, merge the closest pair of clusters until only one cluster (or k clusters) left	• 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 to Merge or split one cluster at a time

Agglomerative Clustering Algorithm

More popular hierarchical clustering technique

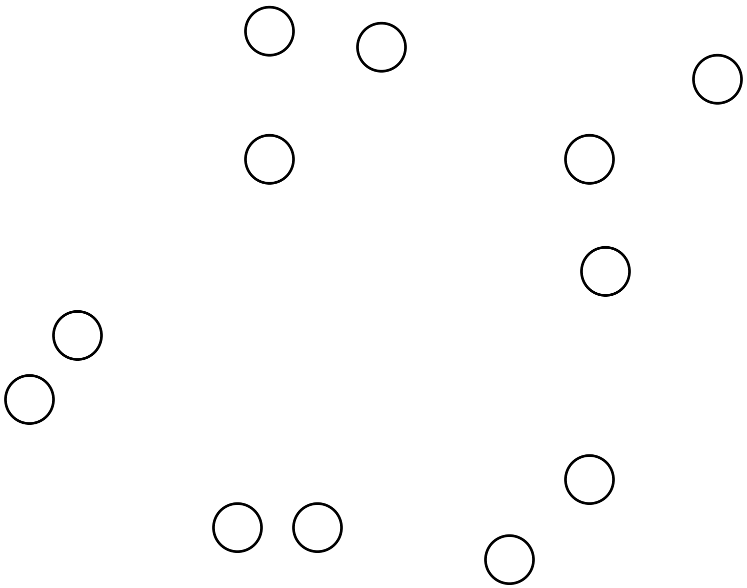
1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

Key operation is the computation of the proximity of two clusters

Different approaches exist to define the distance between clusters

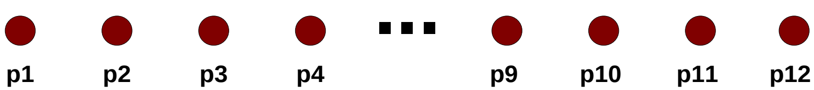
Starting Situation

- Start with clusters of individual points and a proximity matrix



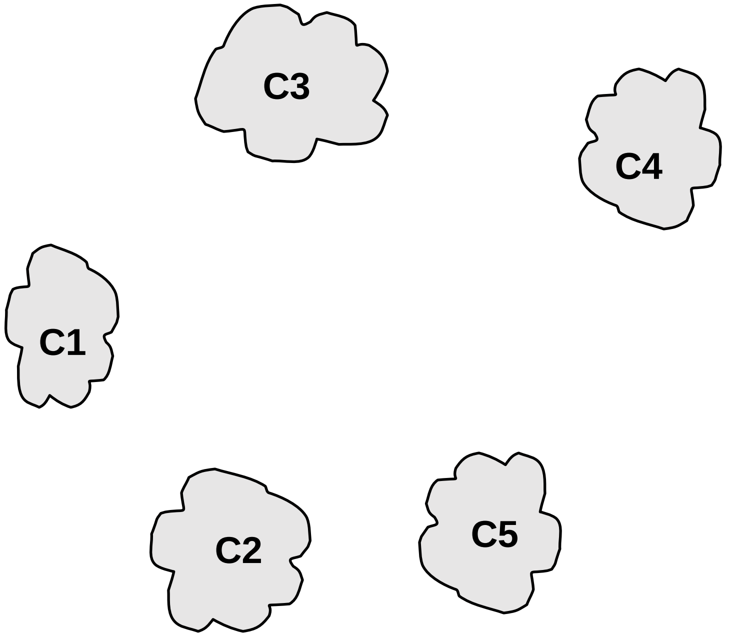
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



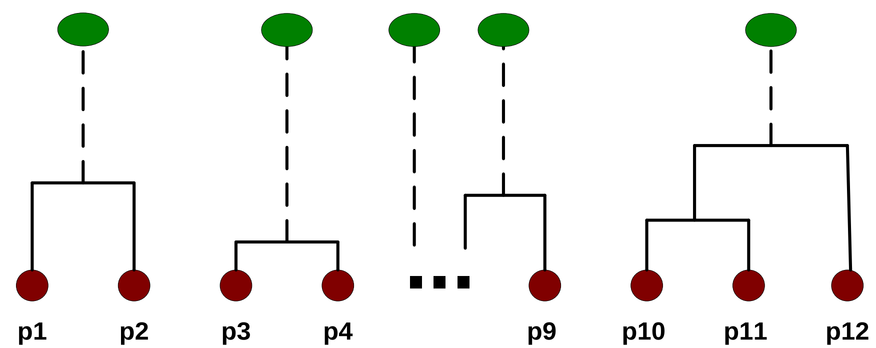
Intermediate Situation

- After some merging steps, we have some clusters



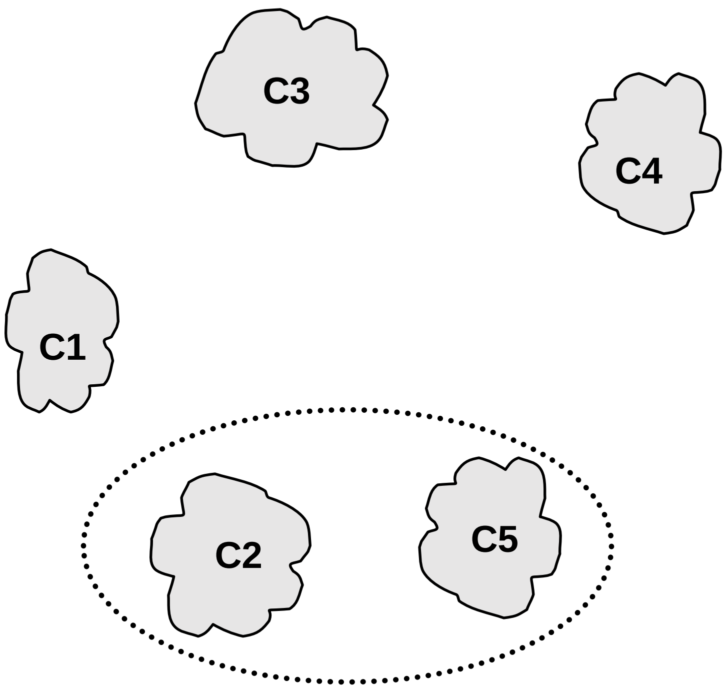
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

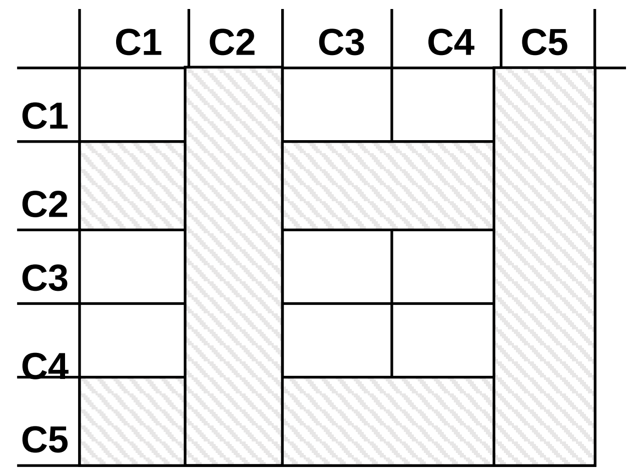
Proximity Matrix



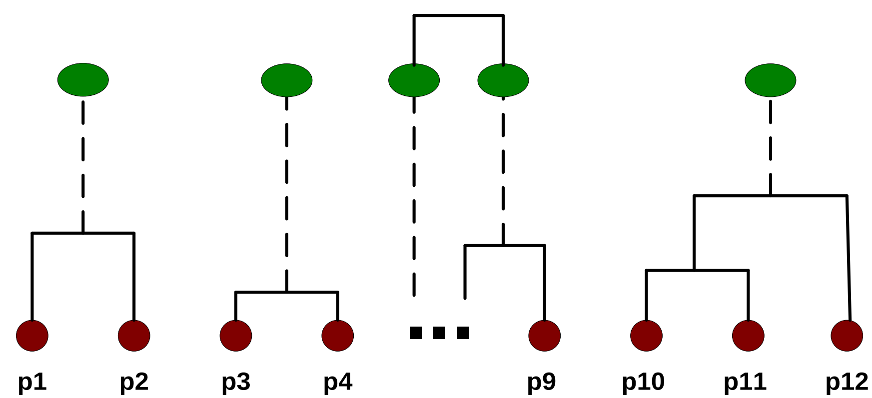
Intermediate Situation

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



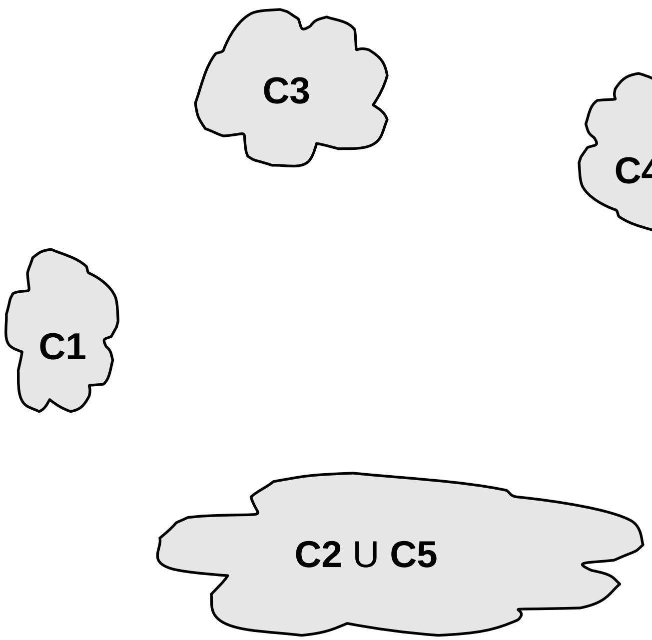


Proximity Matrix



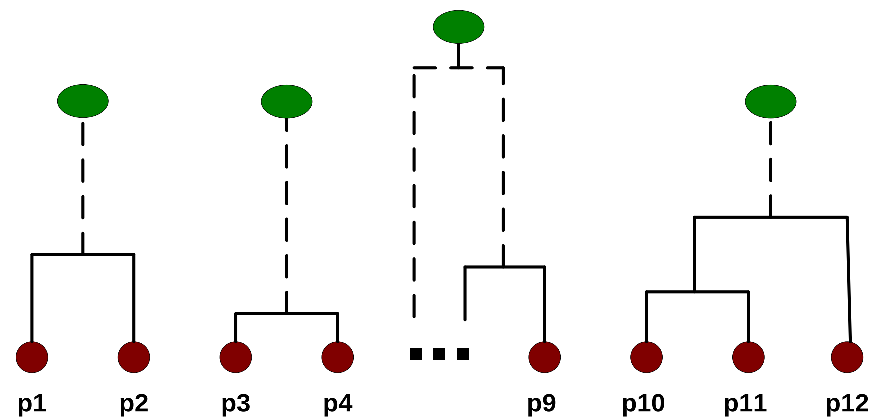
After Merging

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

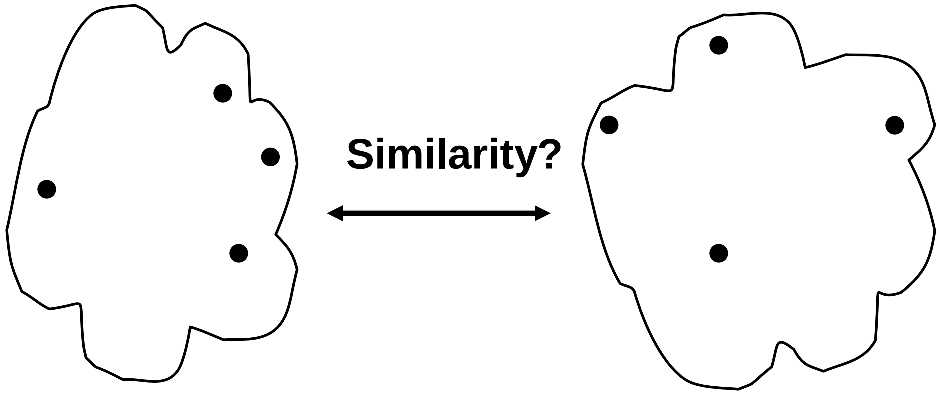


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Similarity

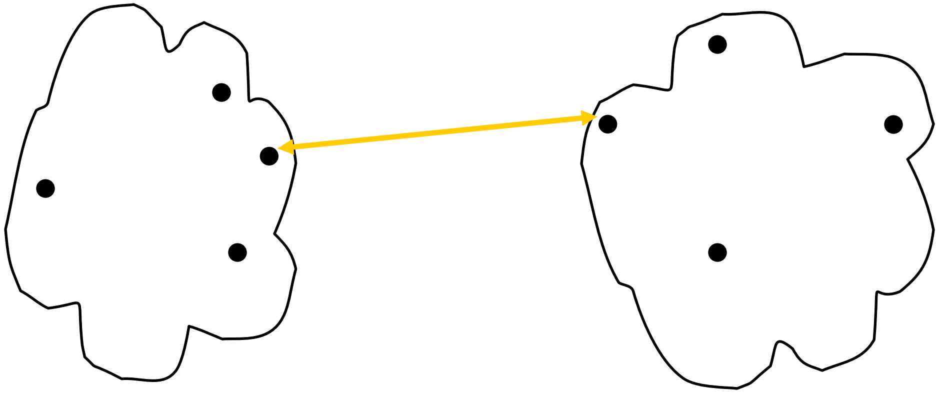


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

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

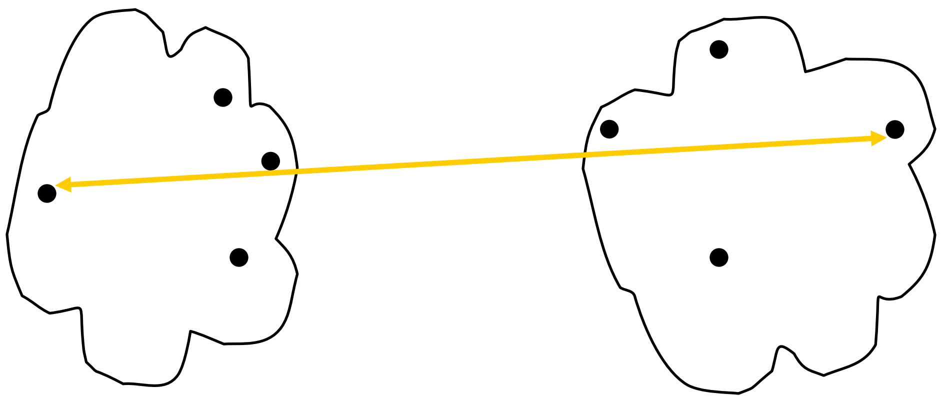


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

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

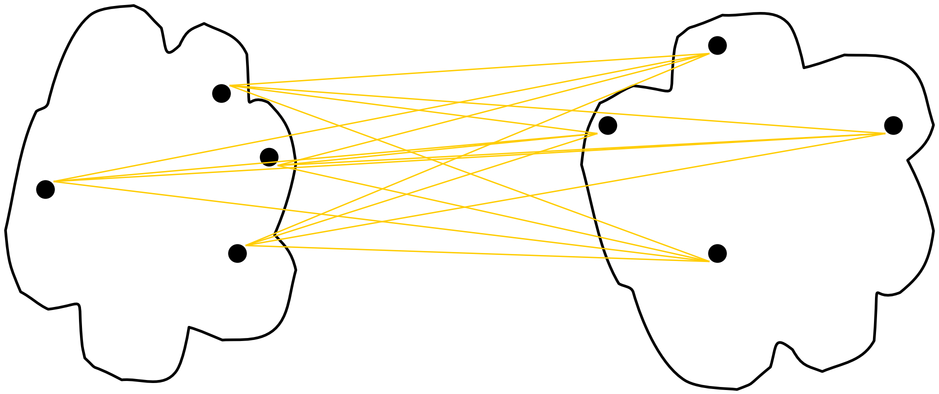


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

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

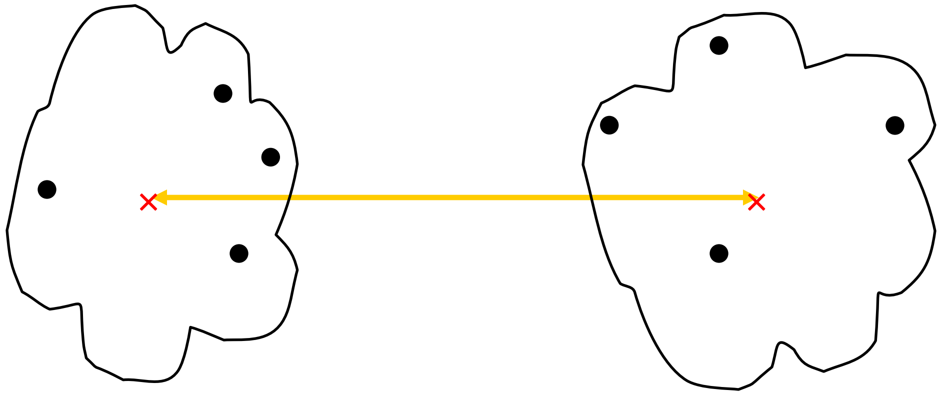


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

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



	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

Three popular choices for Inter-Cluster Similarity

Let G and H are two clusters, $d_{i,j}$ is the distance between two objects and N_G is the number of points in a cluster

Single-linkage: the similarity of two clusters is the similarity of their *most similar* members (closest points)

$$d_{SL}(G,H) = \min_{i \in G, j \in H} (d_{i,j})$$

Complete-linkage: the similarity of two clusters is the similarity of their *most dissimilar* members (farthest points)

$$d_{CL}(G,H) = \max_{i \in G, j \in H} (d_{i,j})$$

Group Average: the average similarity between clusters

$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d_{i,j}$$

Example from Book Principle of Data Mining

- Single Link

	a	b	c	d	e	f
a	0	12	6	3	25	4
b		0	19	8	14	15
c			0	12	5	18
d				0	11	9
e					0	7
f						0

	ad	b	c	e	f
ad	0	8	6	11	4
b		0	19	14	15
c			0	5	18
e				0	7
f					0

Example: Single Link

- Distance Matrix after two Mergers

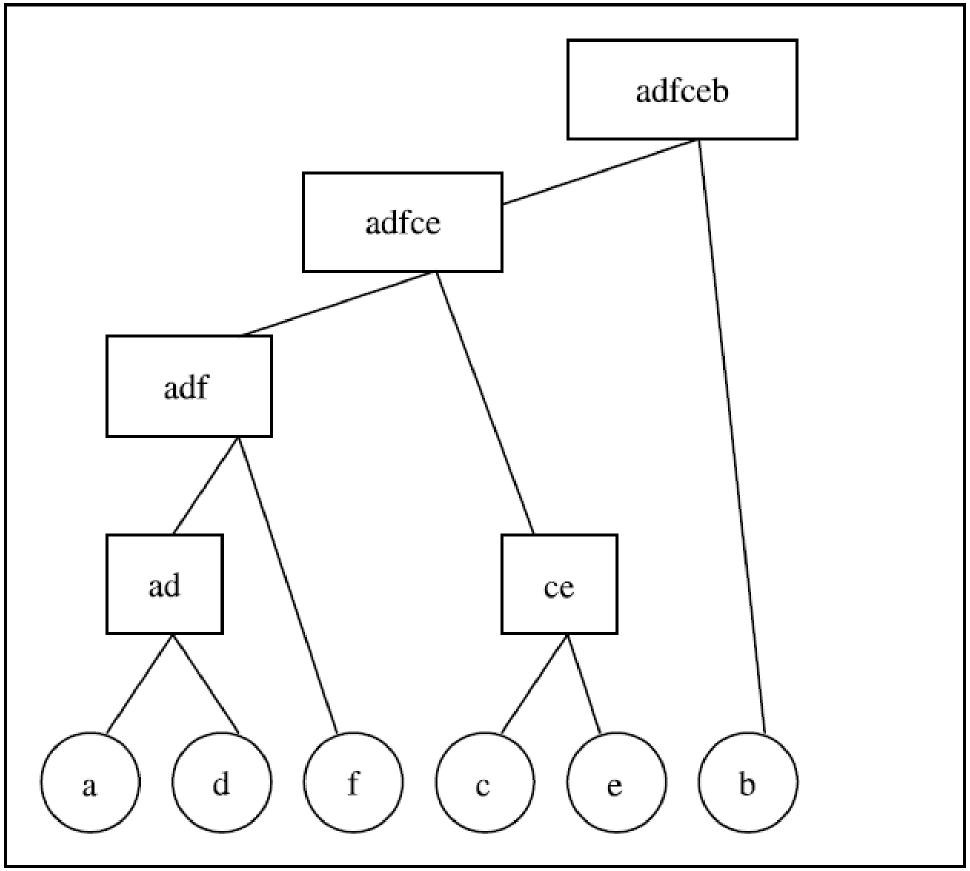
	ad	b	c	e	f
ad	0	8	6	11	4
b		0	19	14	15
c			0	5	18
e				0	7
f					0

	adf	b	c	e
adf				
b		0	19	14
c			0	5
e				0

	adf	b	ce
adf	0	8	6
b		0	14
ce			0

	$adfce$	b
$adfce$	0	8
b	8	0

Example: Single Link



Example from Book Principle of Data Mining

- Complete Link min(max distances)

	a	b	c	d	e	f
a	0	12	6	3	25	4
b		0	19	8	14	15
c			0	12	5	18
d				0	11	9
e					0	7
f						0

$$d_{CL}(G,H) = \max_{i \in G, j \in H} (d_{i,j})$$

	ad	b	c	e	f
ad	0	12	12	25	9
b		0	19	14	15
c			0	5	18
e				0	7
f					

Example Complete Link min(max distances)

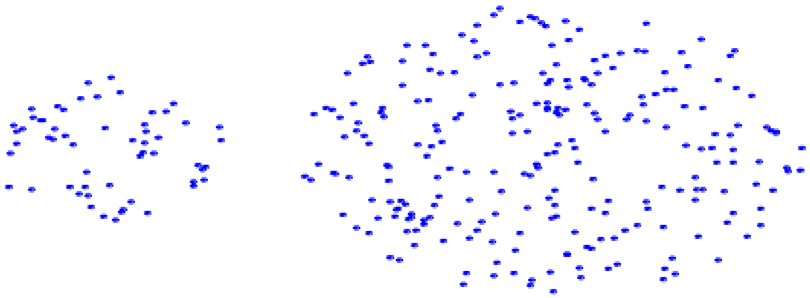
	ad	b	c	e	f
ad	0	12	12	25	9
b		0	19	14	15
c			0	5	18
e					7
f					

	ad	b	ce	f
ad	0	12	25	9
b		0	19	15
ce				18
f				

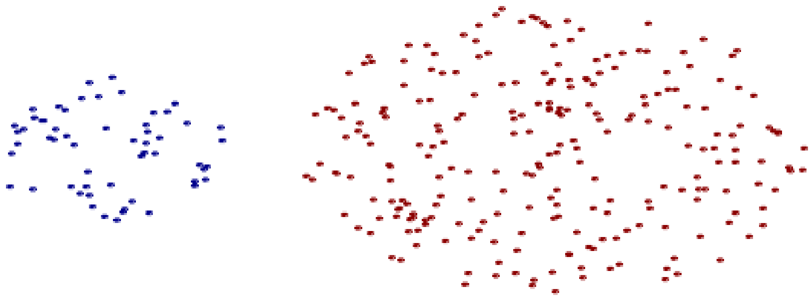
	adf	b	ce
adf		15	25
b			19
ce			

Strength of MIN

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph.



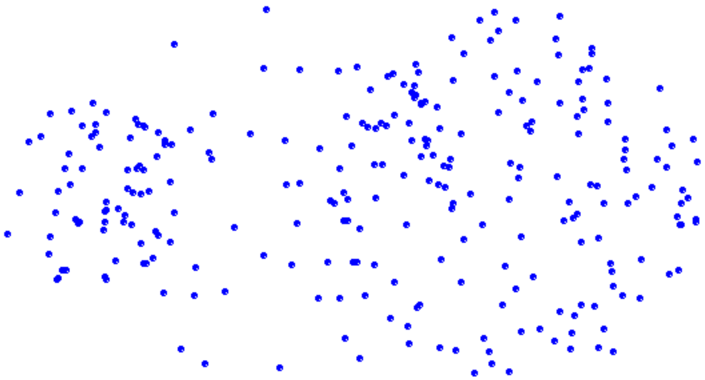
Original Points



Two Clusters

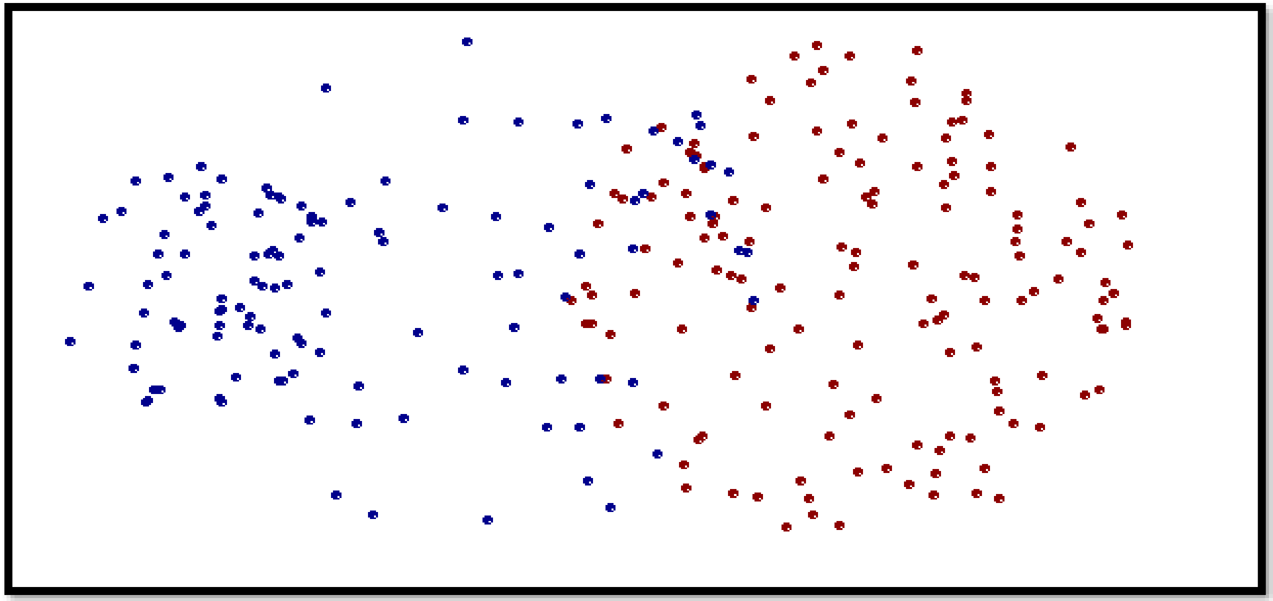
Can handle non-elliptical shapes

Limitations of MIN



Original Points

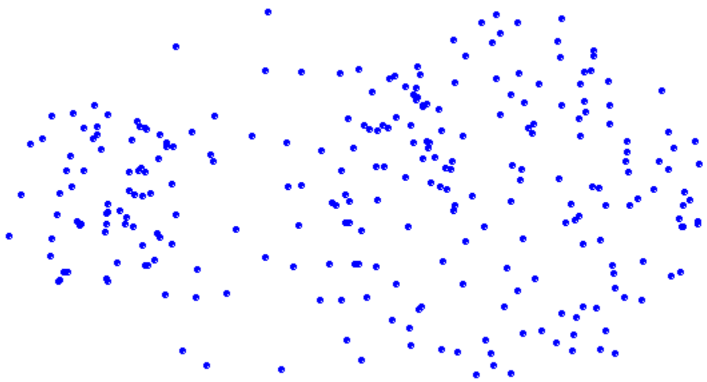
- Sensitive to noise and outliers



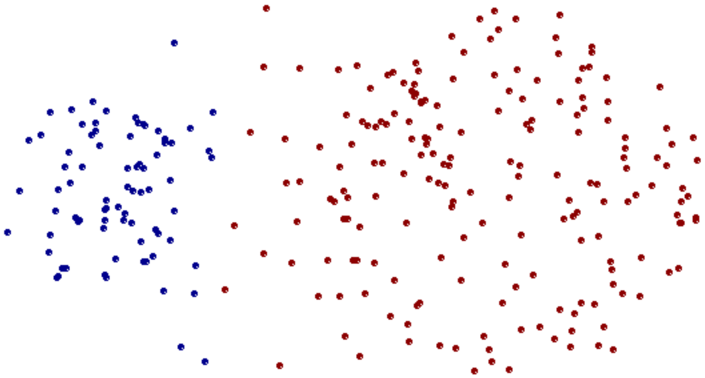
Two Clusters

Strength of MAX

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters



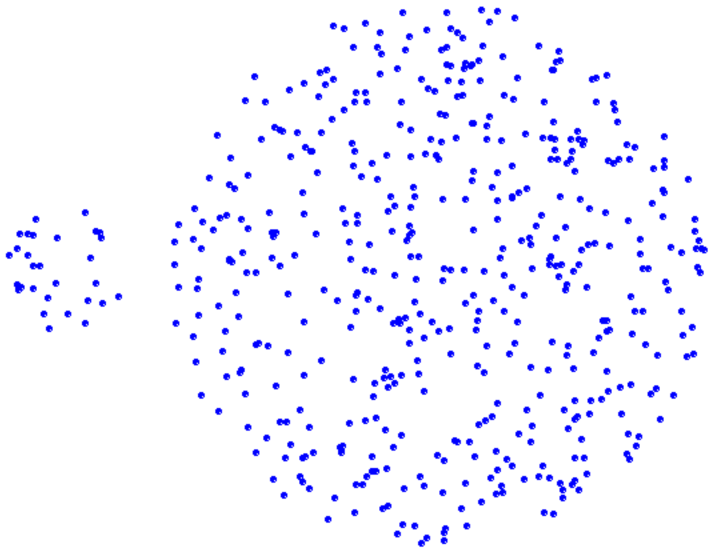
Original Points



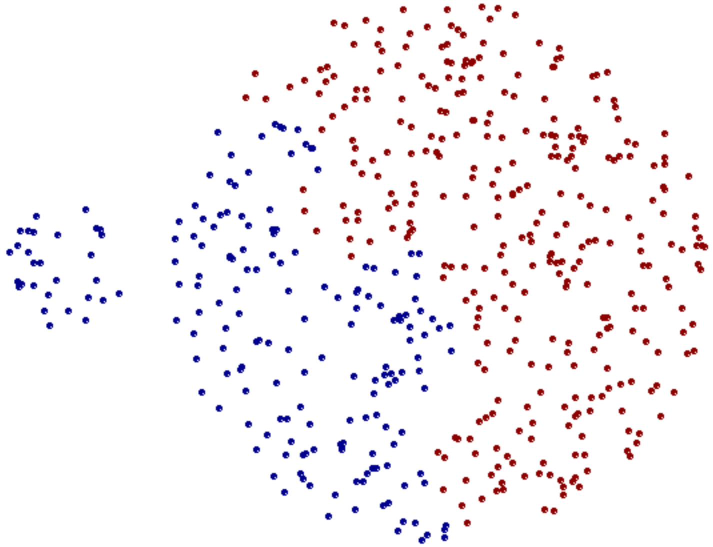
Two Clusters

Less susceptible to noise and outliers

Limitations of MAX



Original Points



Two Clusters

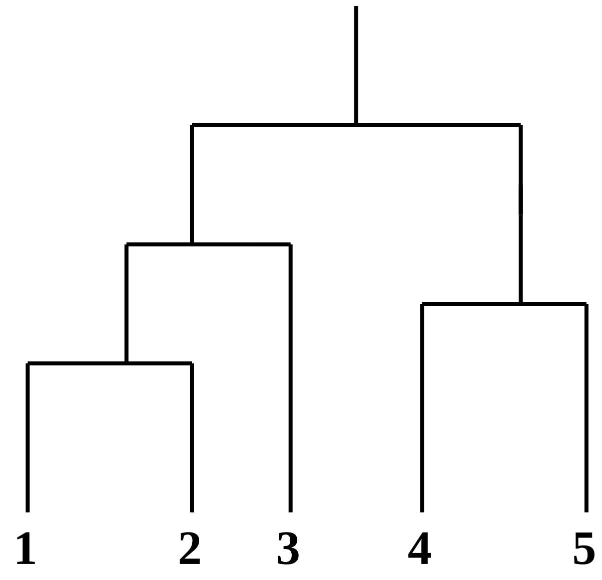
- Tends to break large clusters
- Biased towards globular clusters

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



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

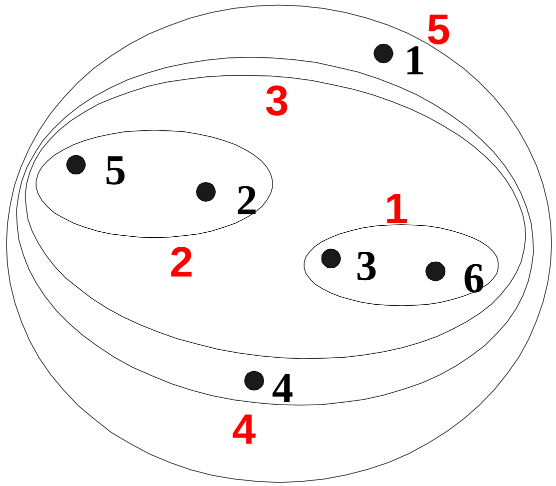
- Similarity of two clusters is based on the increase in squared error when two clusters are merged
  - Similar to group average if distance between points is distance squared

- Less susceptible to noise and outliers

- Biased towards globular clusters

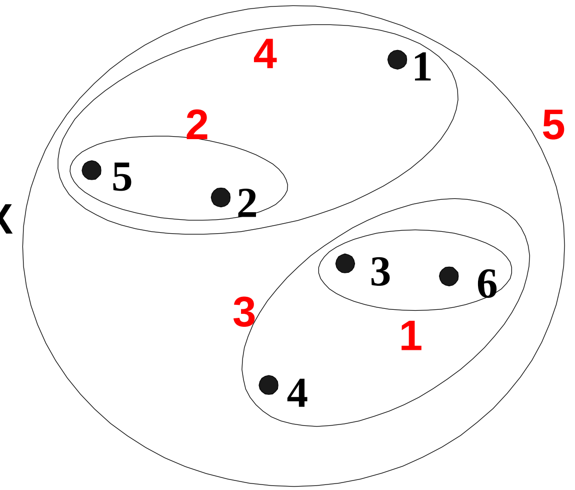
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

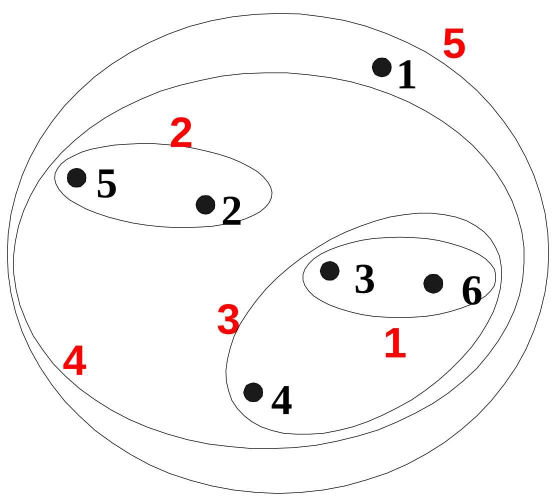
Hierarchical Clustering: Comparison



MIN

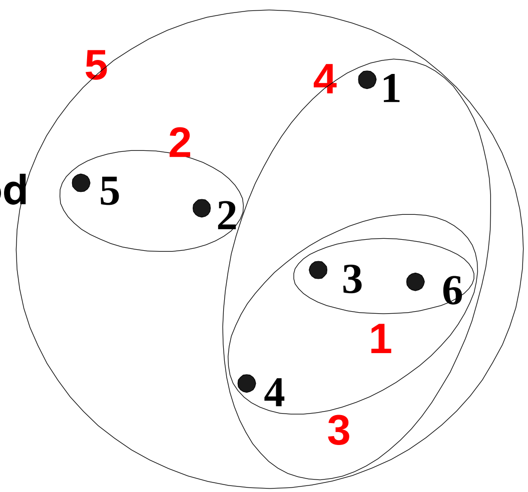
MAX





Group Average

Ward's Method



Properties of intergroup similarity

- Single linkage
  - can produce “chaining,” where a sequence of close observations in different groups cause early merges of those groups
- Complete linkage has the opposite problem.
  - It might not merge close groups because of outlier members that are far apart.
- Group average represents a natural compromise,
  - but depends on the scale of the similarities. Applying a monotone transformation to the similarities can change the results.

Hierarchical Clustering: Time and Space

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

- TIME
  - $O(N^3)$ 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 if we use a special structure like a heap or sorted lists

Hierarchical Clustering: Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- Do not scale well: time complexity of at least $O(N^2 \log N)$, where *n* is the number of total objects
- 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