Learning Analytics

Dr. Bowen Hui
Computer Science
University of British Columbia Okanagan

Last Class

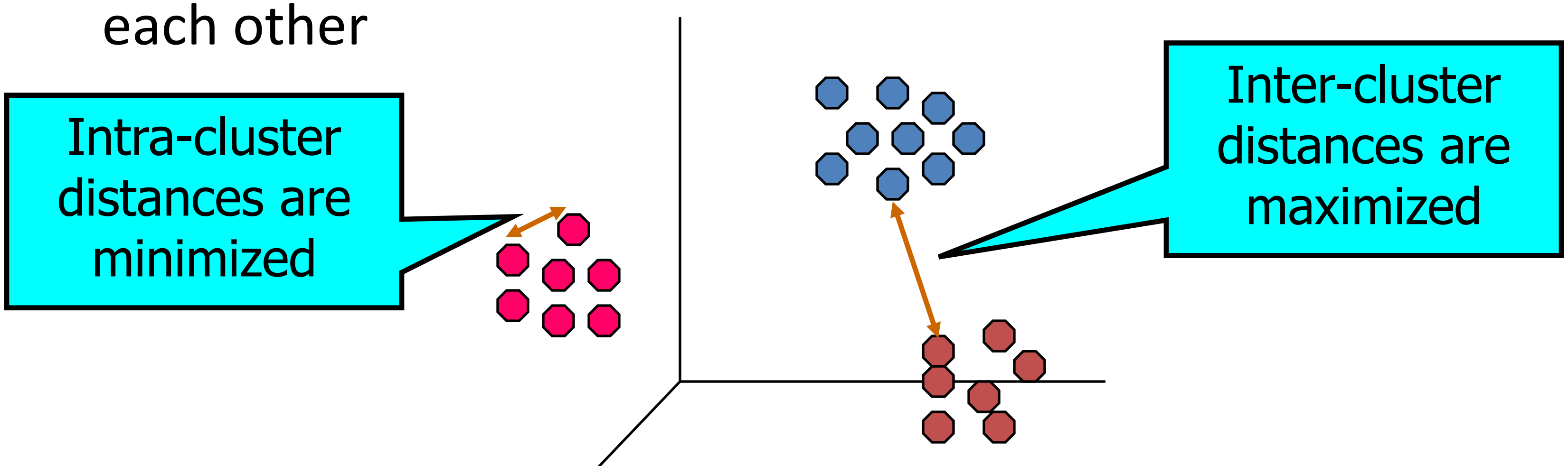
- Introduced difference between categorization vs decision making
- One specific classification technique: decision trees
  - Given labeled dataset
  - Learn the underlying decision tree (model)
  - Predict label for unseen data
- This class:
  - How it works with unlabeled data
  - Specific algorithms:
    - Agglomerative hierarchical clustering
    - K-means

Common Applications

- Marketing research: discover target customer segments
- Biology: categorize genes, derive animal/plant taxonomies
- Climate: find patterns in atmospheric pressure and ocean temperature
- Health: identify different types of depression, detect patterns in spatial/temporal distribution of a disease
- Compression: uses prototype to generate lossy data (vector quantization)

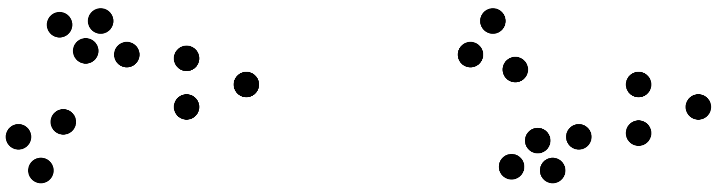
Overview

- Cluster analysis is an analysis technique to group data into meaningful clusters
  - Clusters should capture natural structure of the data
  - Goal: data within a cluster should be similar to each other and data in different clusters should be dissimilar from each other



Find clusters to minimize or maximize an objective function

Ambiguous Notion of "Cluster"



How many clusters?

Ambiguous Notion of "Cluster"

How many clusters?

Six Clusters

Two Clusters

Four Clusters

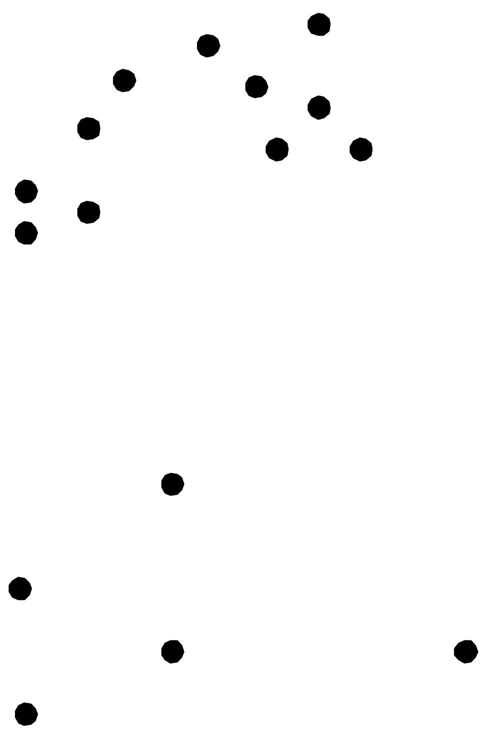
Types of Clustering

- Major distinctions
  - Hierarchical (nested) vs. Partitional (unnested)
  - Exclusive vs. Overlapping vs. Fuzzy
  - Complete vs. Partial

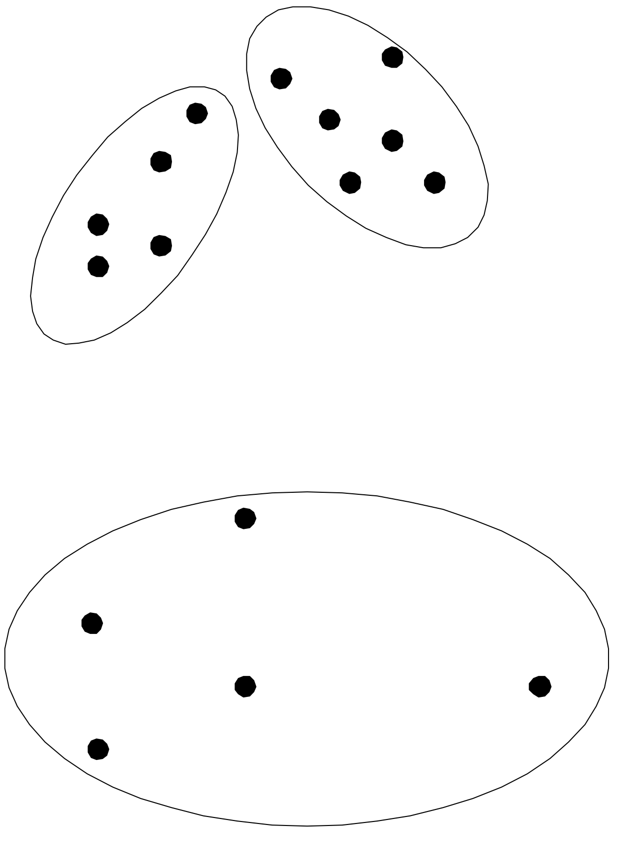
Most Common Distinction

- Hierarchical (nested) vs. Partitional (unnested)
  - Partitional: clusters are non-overlapping
  - Hierarchical: nested clusters organized as tree
    - Clusters may have subclusters

Visualizing Partitional Clustering

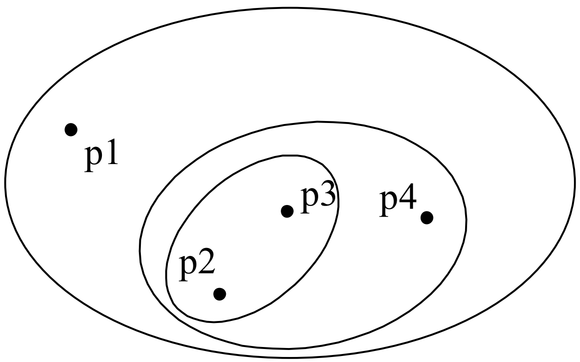


Original Points

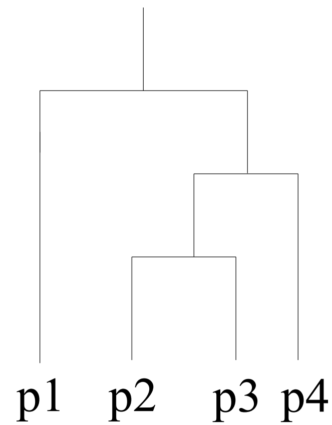


A Partitional Clustering

Visualizing Hierarchical Clustering



Traditional Hierarchical Clustering



Traditional Dendrogram

Object Membership

- Exclusive vs. Overlapping vs. Fuzzy
  - Exclusive: each point belongs to a single cluster
  - Overlapping: an point can simultaneously belong to multiple clusters
    - E.g. a person may be a student and an employee
  - Fuzzy: every point belongs to every cluster with a weight in [0,1]
    - Clusters are treated as fuzzy sets
    - Typically impose constraint that for each point, its sum of the weights = 1.0
    - Doesn’t truly address case when a point belongs to multiple clusters (because the weight would exceed 1.0)

Object Membership

- Complete vs. Partial
  - Complete: every point is assigned to a cluster
  - Partial: not every point is assigned a cluster
    - Motivation: sometimes data has noise or outliers
    - E.g. clustering news articles – may only be interested in articles that are tightly related to common themes, then ignore articles that are "generic"

Agglomerative Hierarchical Clustering

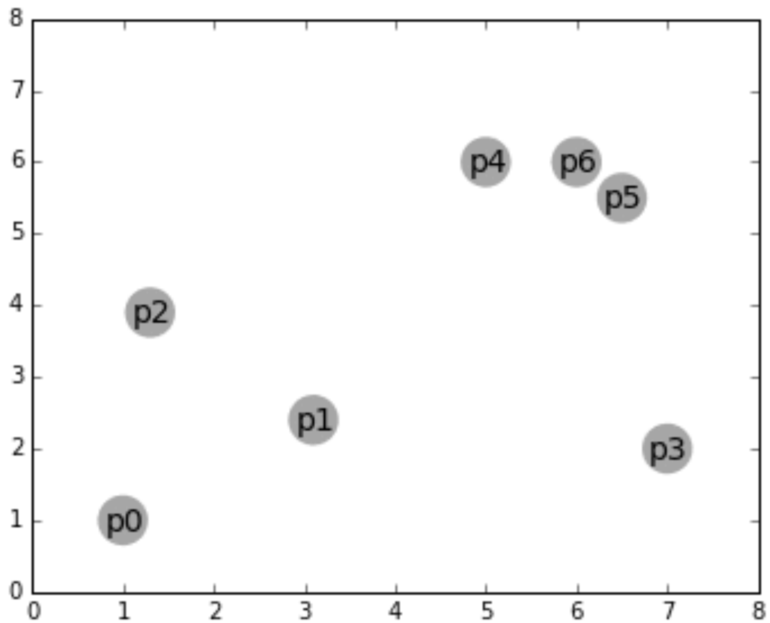
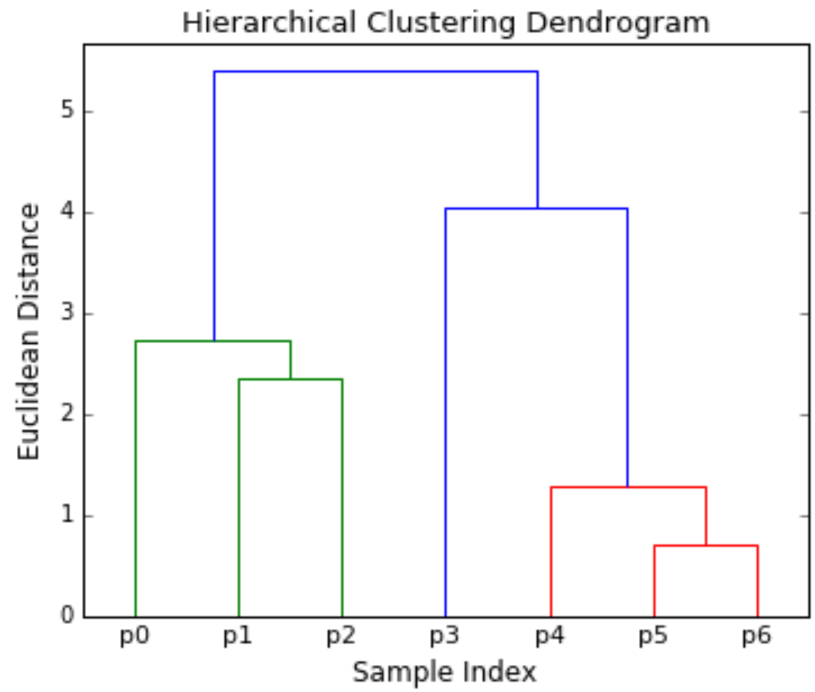


Image taken from towardsdatascience.com



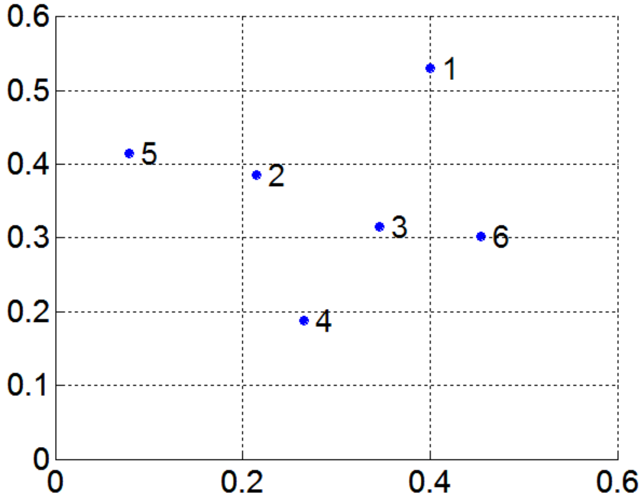
- General algorithm:
  - Start with each point being an individual cluster
  - Repeat: merge “closest” pair of clusters
  - Stop: when one cluster remains

Basic Agglomerative Hierarchical Clustering Algorithm

- Compute the proximity matrix between all pairs of points
- Every point starts as its own cluster

Proximity Matrix

- Example:



Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

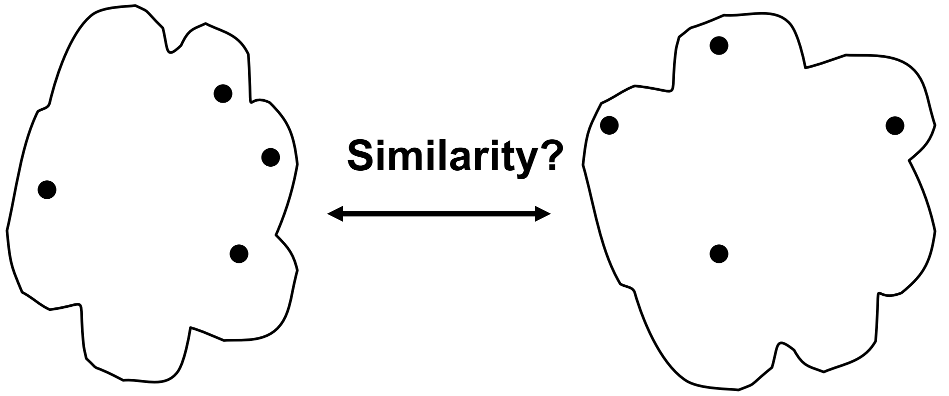
- Given six 2D points, compute pairwise distances using Euclidean distance (L2 distance)
- Recall: $d(\mathbf{p},\mathbf{q})^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2$

Basic Agglomerative Hierarchical Clustering Algorithm

- Compute the proximity matrix between all pairs of points
- Every point starts as its own cluster
- Repeat
  - Merge the two closest clusters
  - Update proximity matrix based on new clusters
- Until one cluster remains

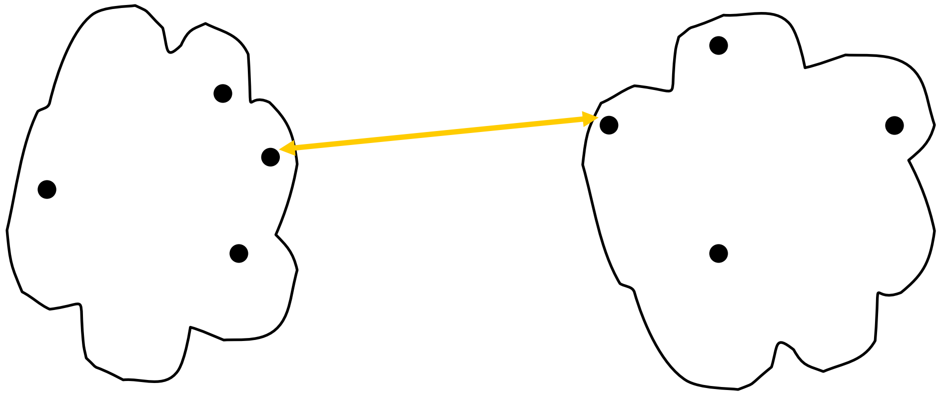
- Key operation: distance computation
  - Specific choice is that distinguishes between specific agglomerative hierarchical techniques

Cluster Proximity



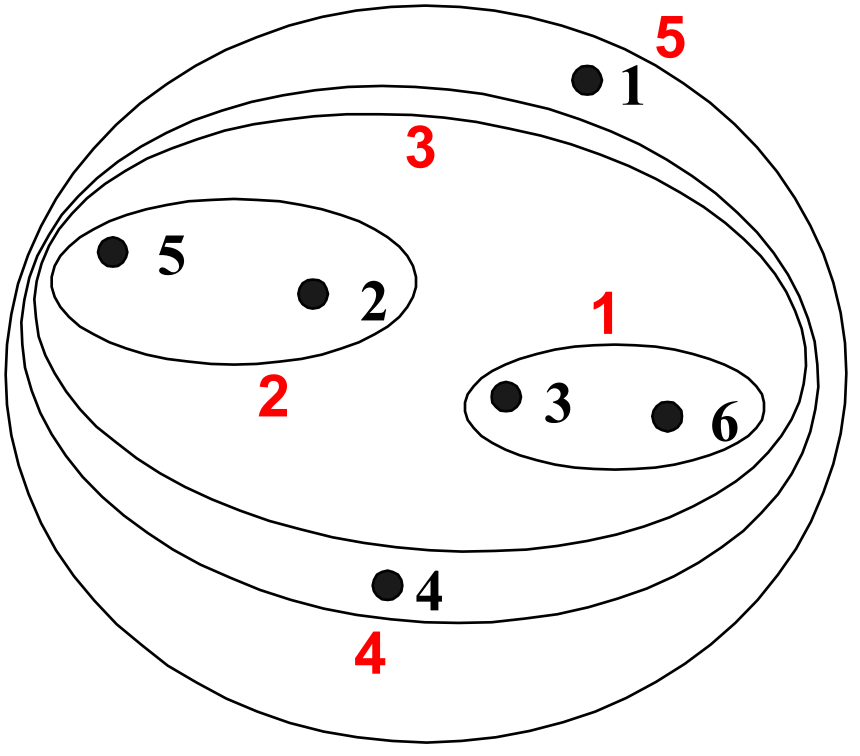
- Most common:
  - MIN
  - MAX
  - Group average
  - Distance between centroids

Cluster Proximity

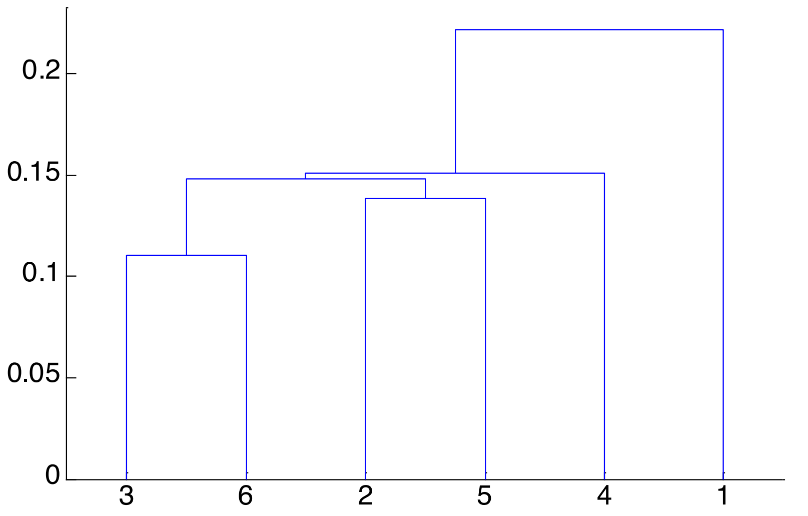


- Most common:
  - MIN
  - MAX
  - Group average
  - Distance between centroids

Example with MIN



Nested Clusters

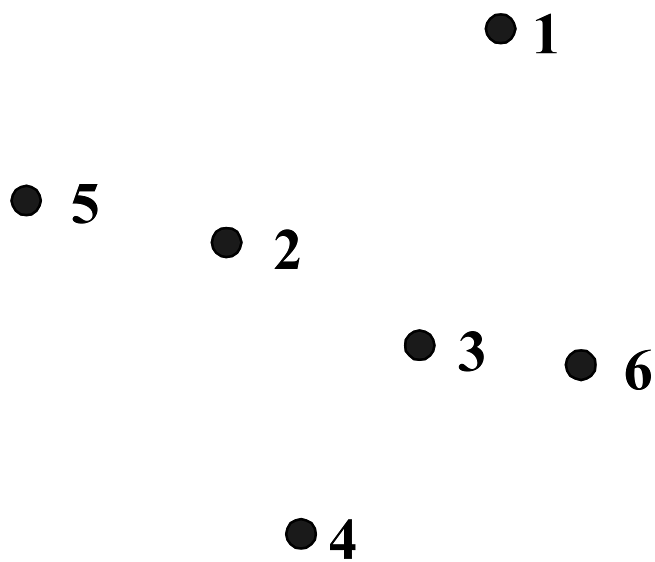


Dendrogram

Promixity of two clusters is based on 2 closest points in different clusters

- Determined by one pair of points

Example with MIN



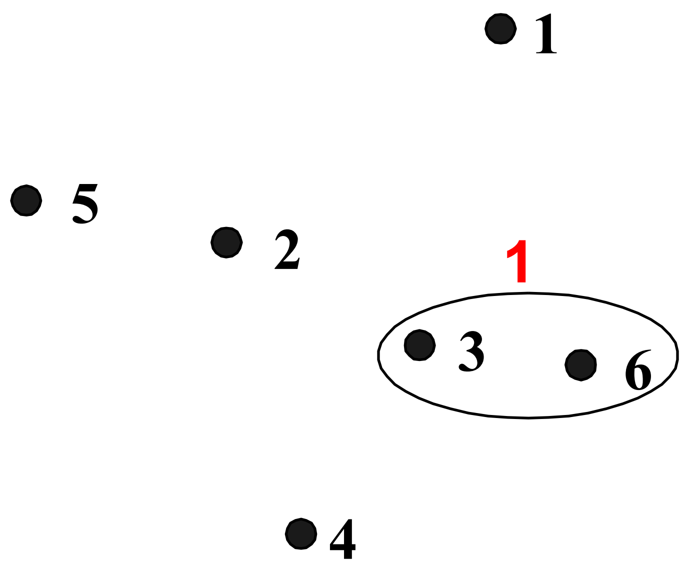
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Nested Clusters **Distance Matrix**

Promixity of two clusters is based on 2 closest points in different clusters

- Determined by one pair of points

Example with MIN



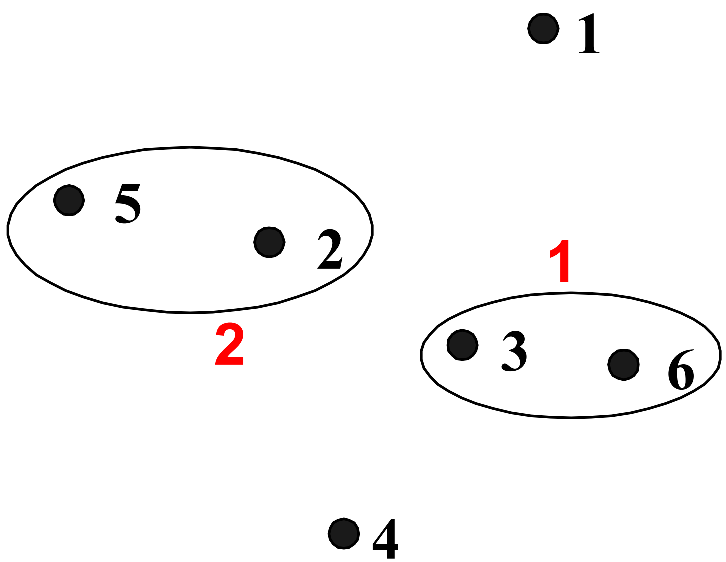
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25		0.22	0.39	0.00

Nested Clusters **Distance Matrix**

Promixity of two clusters is based on 2 closest points in different clusters

- Determined by one pair of points

Example with MIN



Nested Clusters

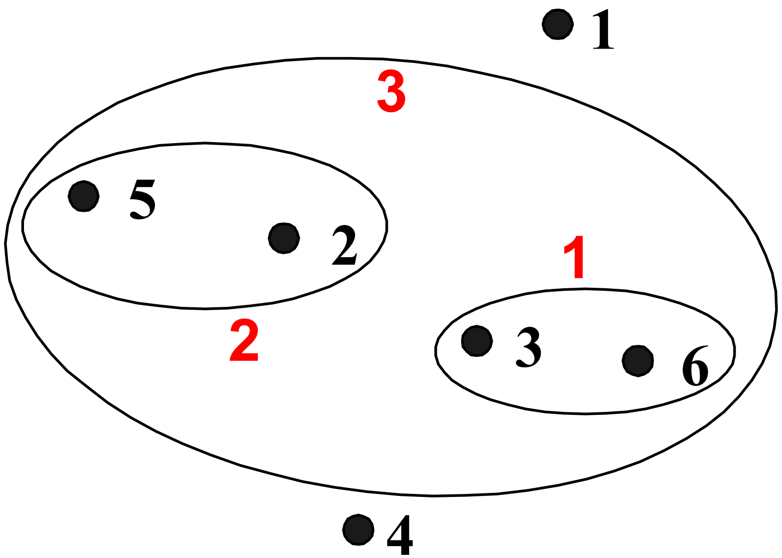
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	[illegible]	0.28	0.29	0.00	0.39
p6	0.23	0.25	[illegible]	0.22	0.39	0.00

Distance Matrix

Promixity of two clusters is based on 2 closest points in different clusters

- Determined by one pair of points

Example with MIN



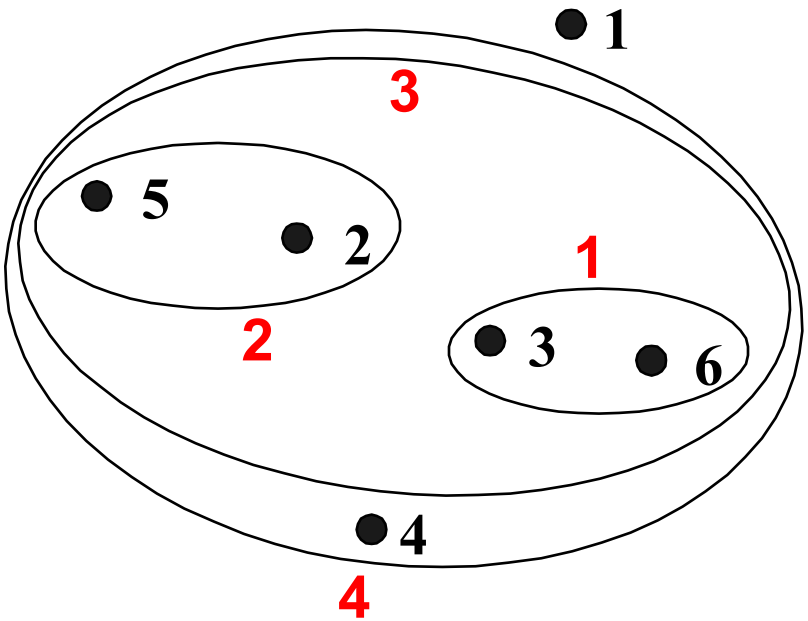
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22		0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34		0.28	0.29	0.00	0.39
p6	0.23	0.25		0.22	0.39	0.00

Nested Clusters **Distance Matrix**

Promixity of two clusters is based on 2 closest points in different clusters

- Determined by one pair of points

Example with MIN



Nested Clusters

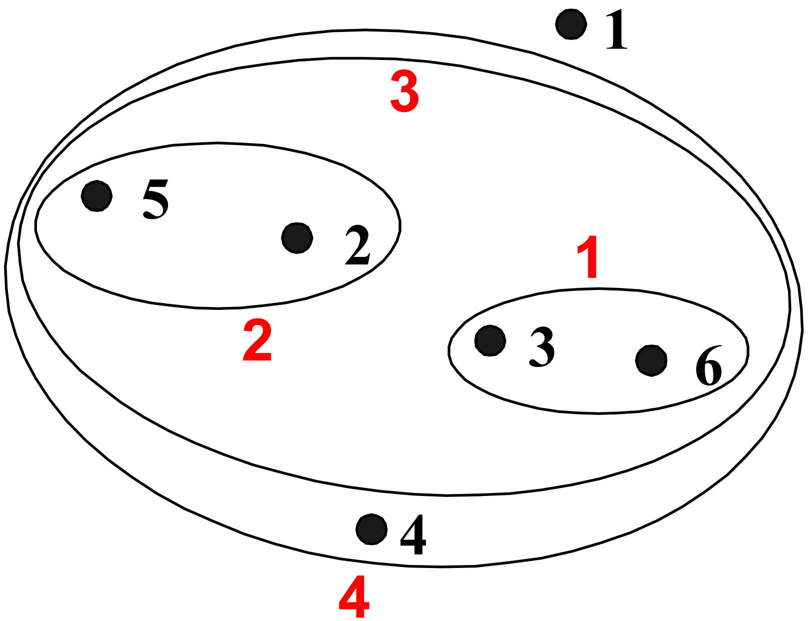
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22		0.00	0.15	0.28	0.11
p4	0.37	0.20		0.00	0.29	0.22
p5	0.34		0.28	0.29	0.00	0.39
p6	0.23	0.25		0.22	0.39	0.00

Distance Matrix

Promixity of two clusters is based on 2 closest points in different clusters

- Determined by one pair of points

Example with MIN



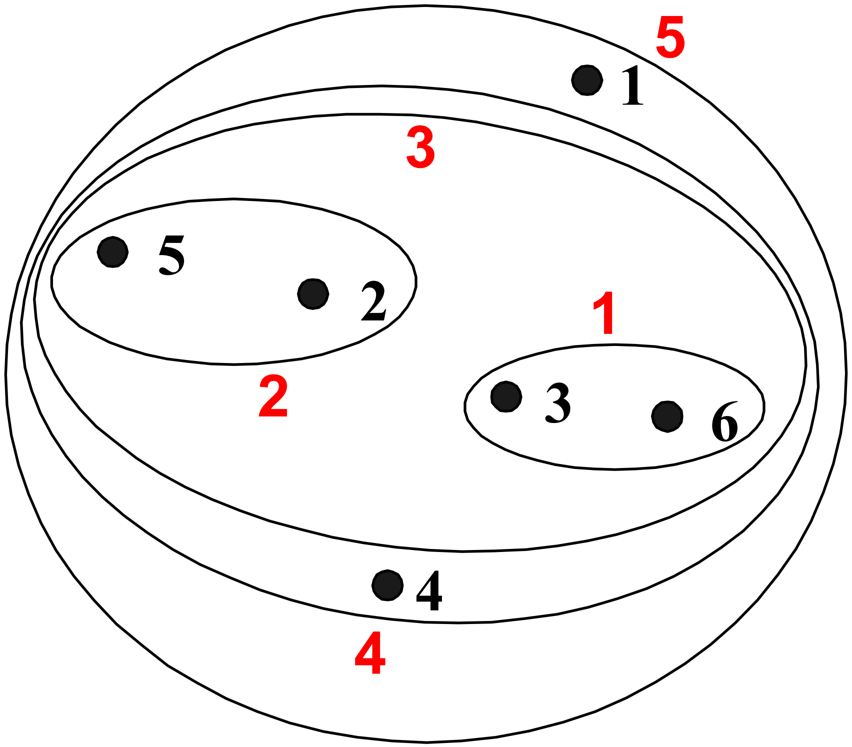
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22		0.00	0.15	0.28	0.11
p4	0.37			0.00	0.29	0.22
p5	0.34		0.28	0.29	0.00	0.39
p6	0.23	0.25		0.22	0.39	0.00

Nested Clusters **Distance Matrix**

Promixity of two clusters is based on 2 closest points in different clusters

- Determined by one pair of points

Example with MIN



Nested Clusters

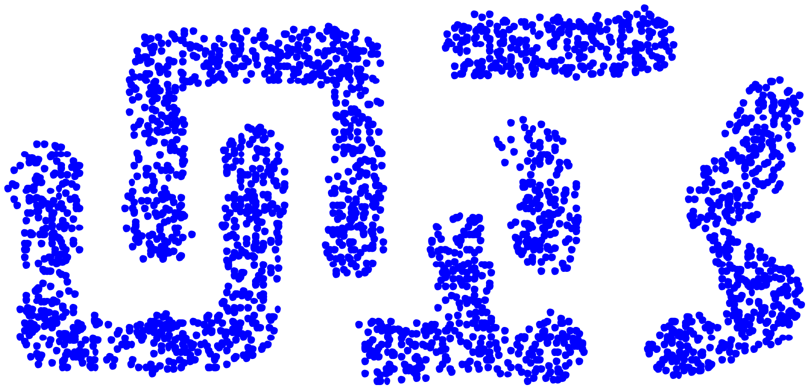
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	[illegible]	[illegible]	0.00	0.15	0.28	0.11
p4	0.37	[illegible]	[illegible]	0.00	0.29	0.22
p5	0.34	[illegible]	0.28	0.29	0.00	0.39
p6	0.23	0.25	[illegible]	0.22	0.39	0.00

Distance Matrix

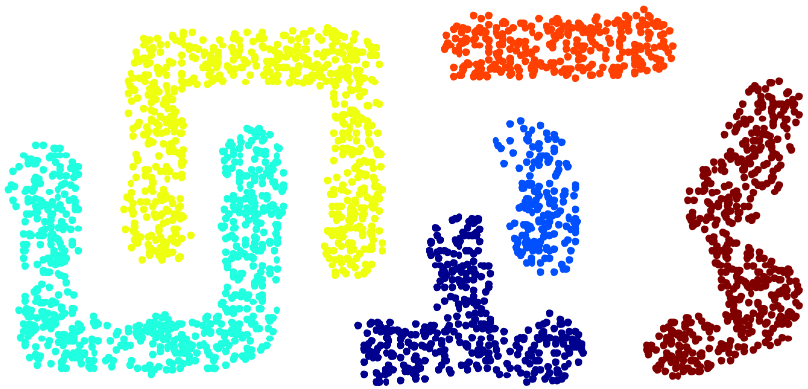
Promixity of two clusters is based on 2 closest points in different clusters

- Determined by one pair of points

Strength of MIN



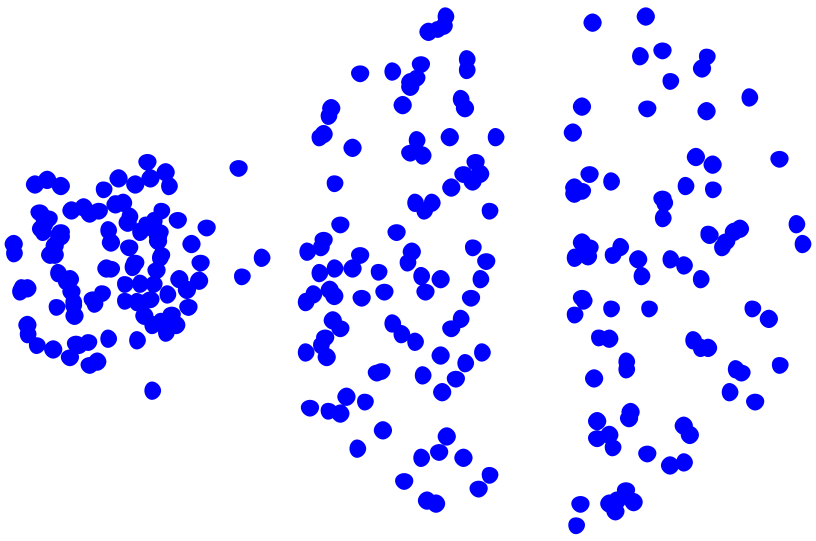
Original Points



Six Clusters

Handles non-elliptical shapes

Limitations of MIN

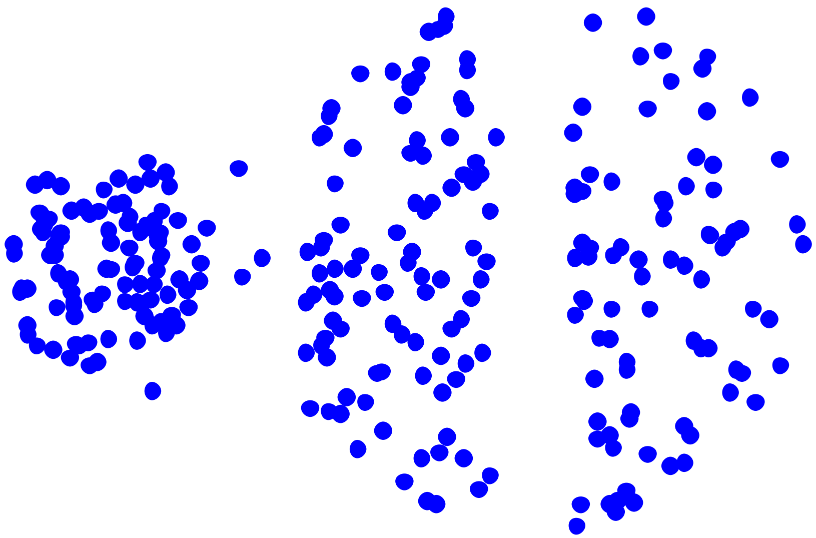


Original Points

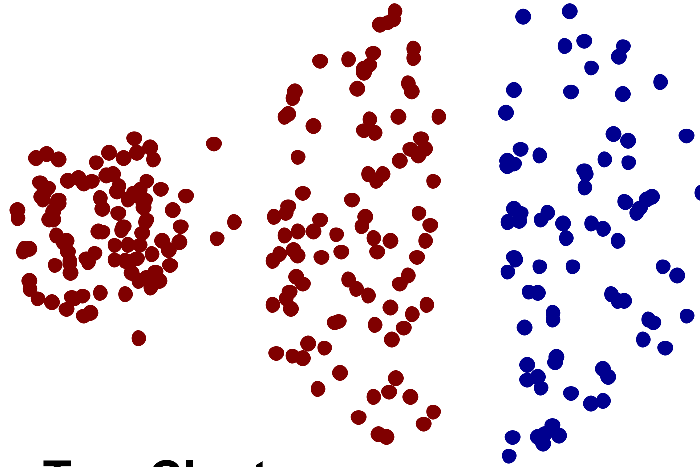
Two Clusters?

Three Clusters?

Limitations of MIN



Original Points

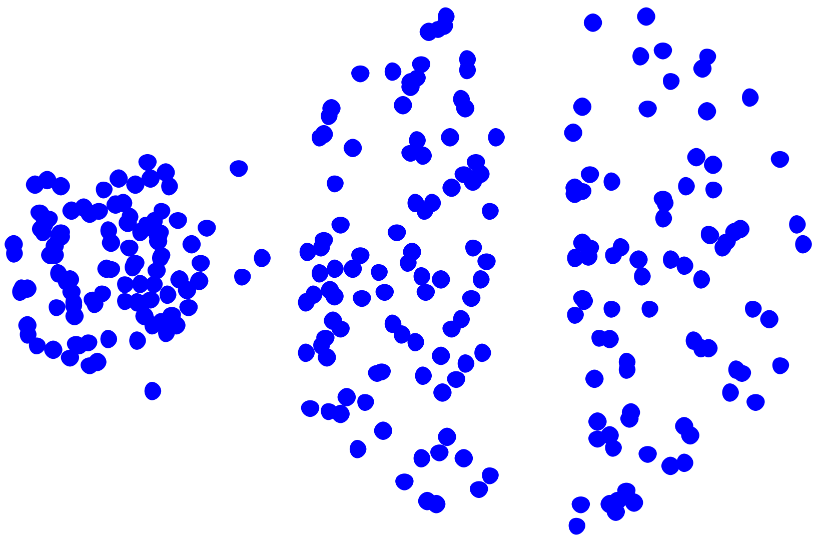


Two Clusters

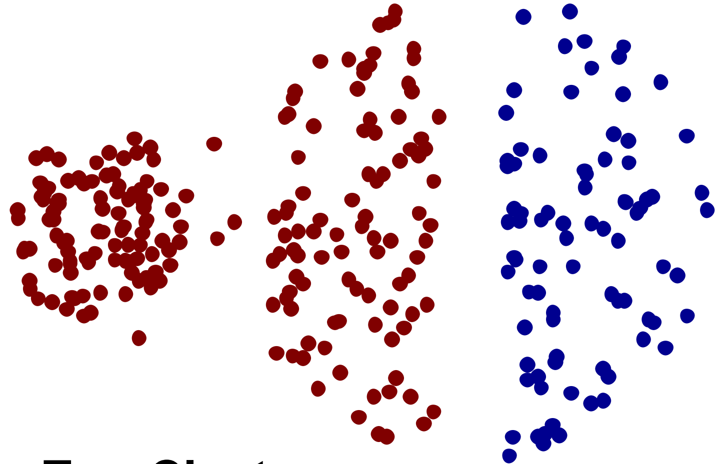
Sensitive to noise and outliers

Three Clusters?

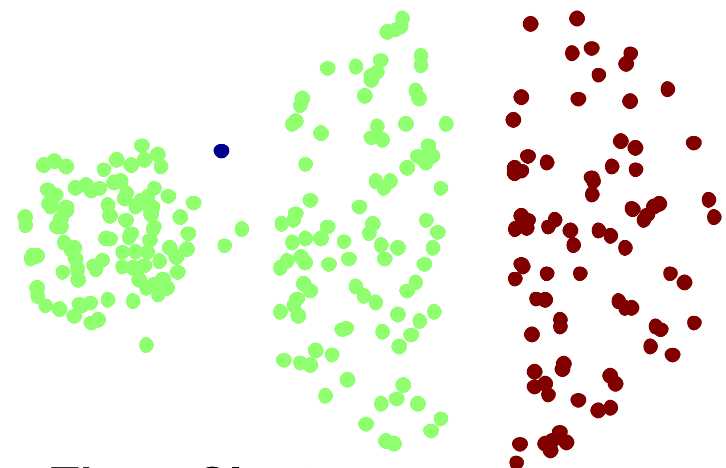
Limitations of MIN



Original Points



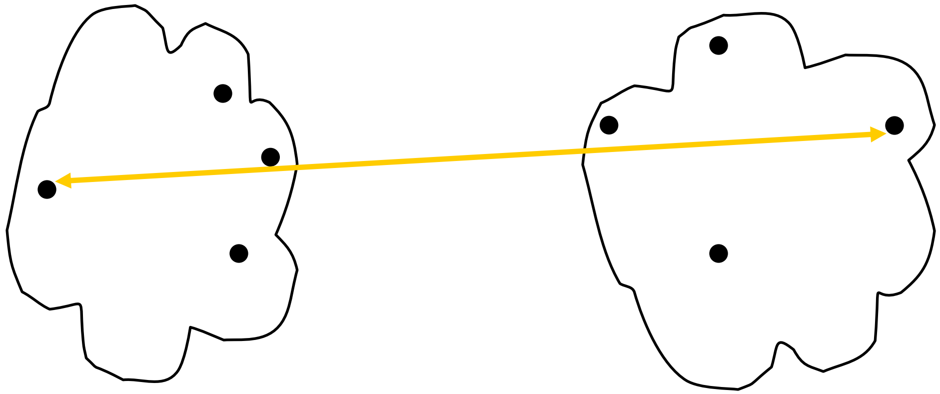
Two Clusters



Three Clusters

Sensitive to noise and outliers

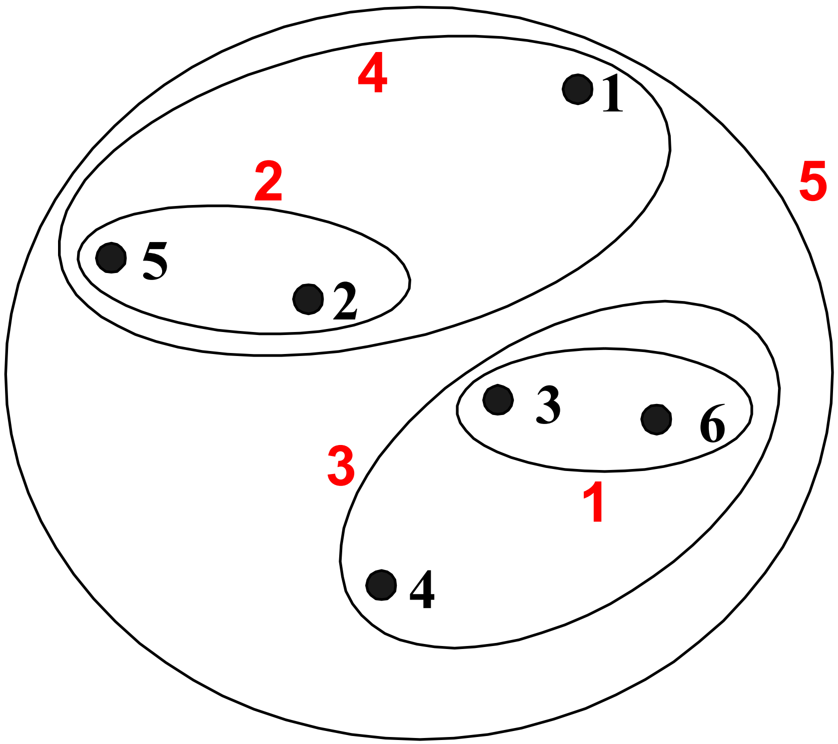
Cluster Proximity



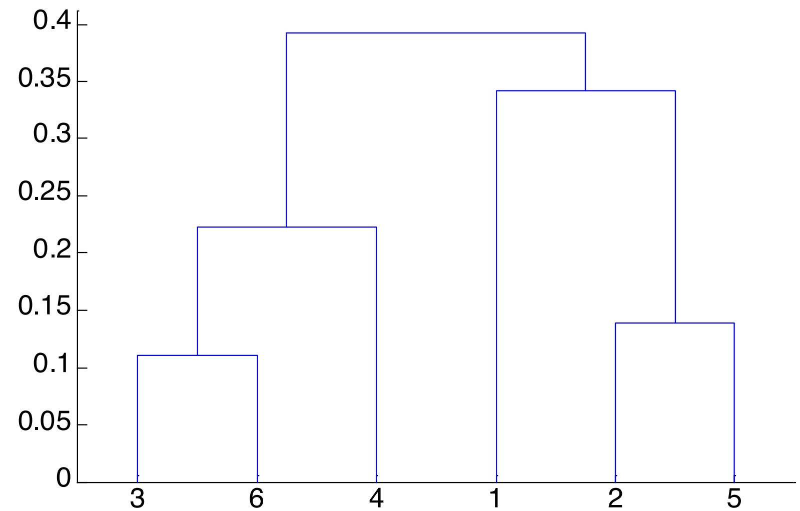
- Most common:
  - MIN
  - MAX
  - Group average
  - Distance between centroids

Example with MAX

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00



Nested Clusters

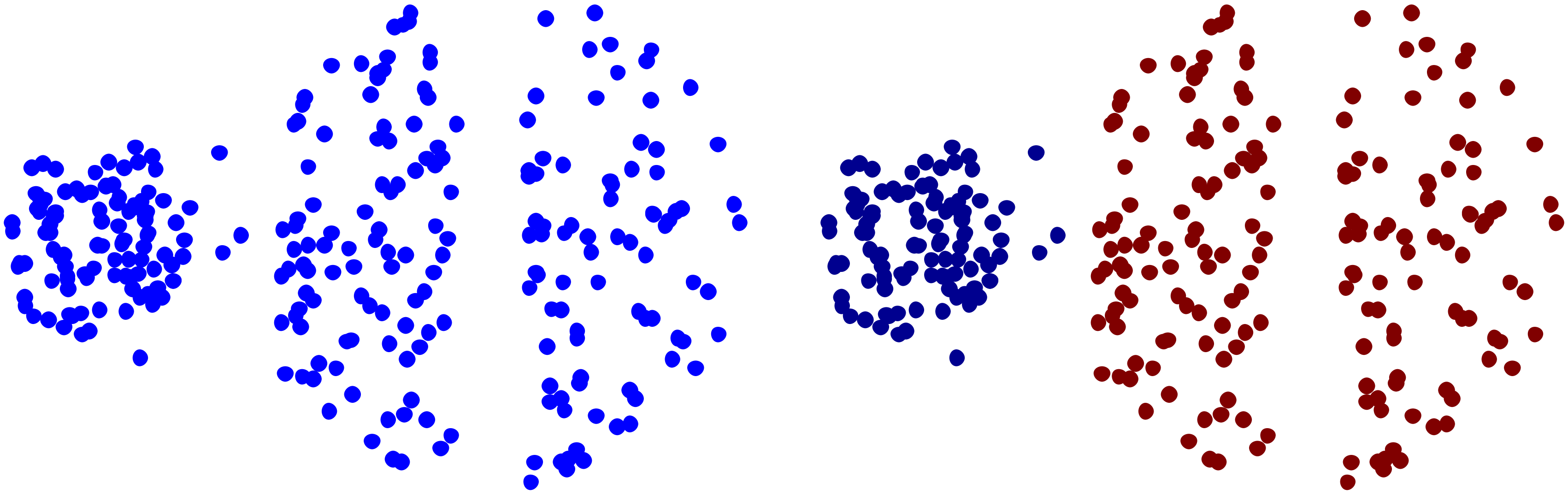


Promixity of two clusters is based on 2 furthest points in different clusters

- Determined by all pairs of points

Still merge based on two closest clusters

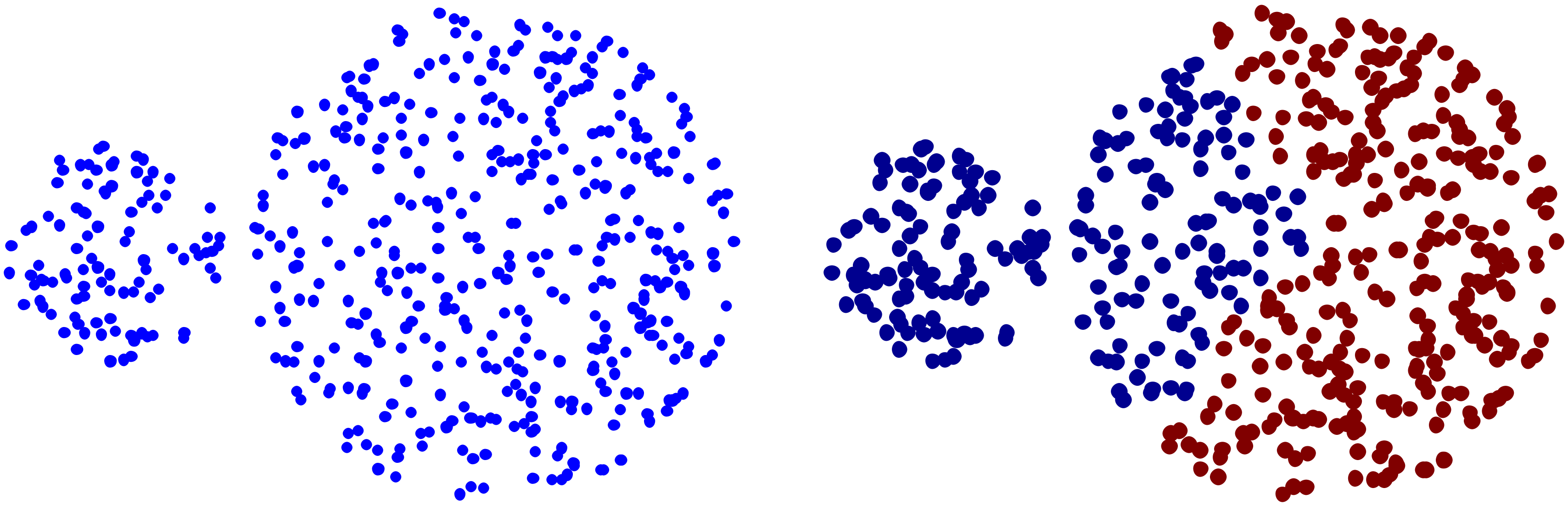
Strength of MAX



Original Points **Two Clusters**

Less susceptible to noise and outliers

Limitations of MAX

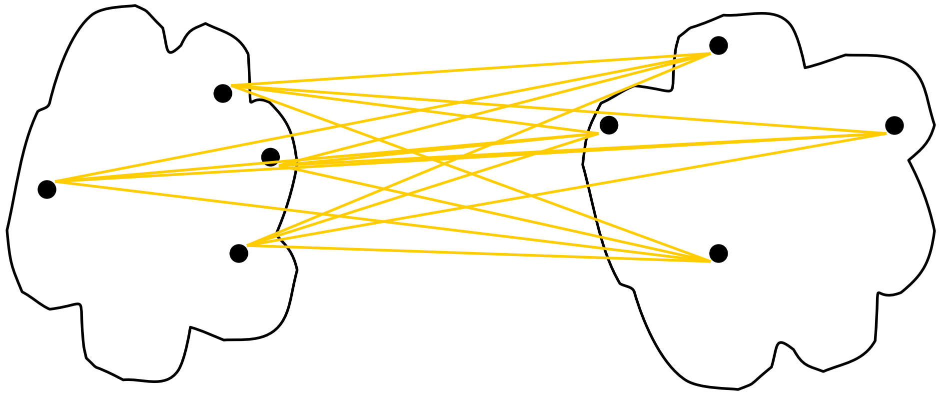


Original Points   Two Clusters

Tends to break large clusters

Biased towards globular clusters

Cluster Proximity

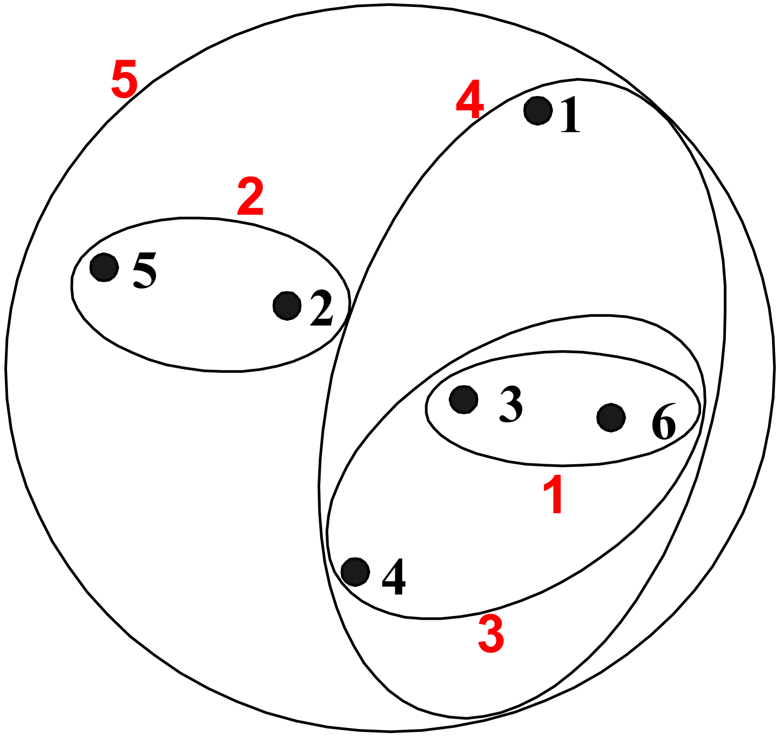


- Most common:
  - MIN
  - MAX
  - Group average
  - Distance between centroids

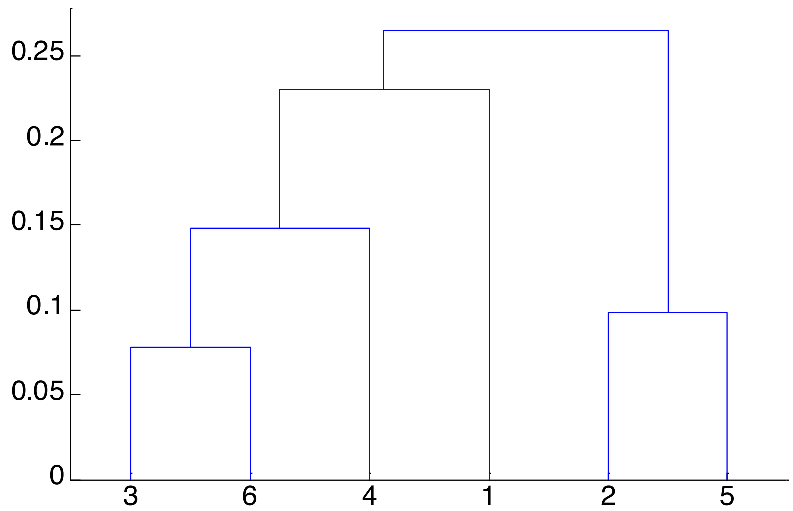
$$proximity(C_i, C_j) = \frac{\sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_j}} proximity(\mathbf{x}, \mathbf{y})}{m_i \times m_j}$$

where $m_i = |C_i|$ and $m_j = |C_j|$

Example with Group Average



Nested Clusters

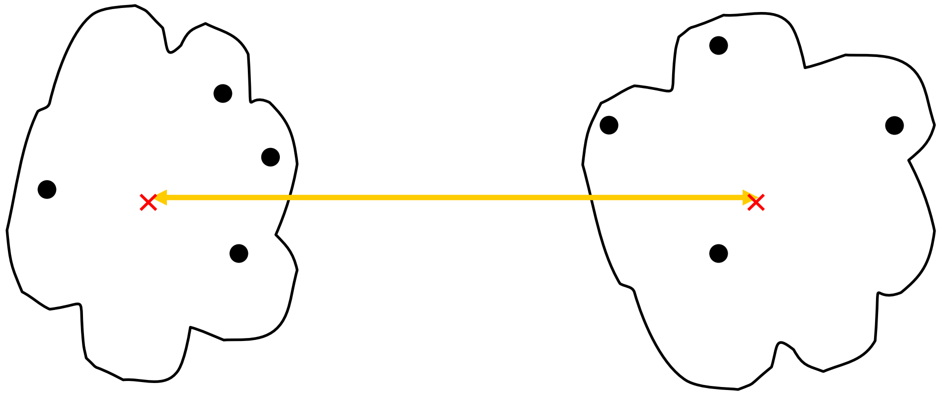


Dendrogram

Proximity of two clusters is based on average pairwise distances between all points in clusters

Still merge based on two closest clusters

Cluster Proximity



- Most common:
  - MIN
  - MAX
  - Group average
  - Distance between centroids

Ward's method is hierarchical analog to k-means

Key Issues in Hierarchical Clustering

- Once clusters are merged, cannot be undone
- Lack of global objective function
  - Merge decisions based on local objectives
- Different schemes have one/more problem:
  - Sensitive to noise and outliers
  - Difficulty with clusters of diff sizes and non-elliptical shapes
  - Breaks up large clusters
- Most often used for creating taxonomy

Key Ideas

- Clustering
  - Forms groups with unlabeled data
  - Hierarchical vs. partitional clustering
- Algorithm:
  - Agglomerative hierarchical clustering
  - Compute and update proximity matrix
  - Main operation is to calculate cluster distance
- Specific Methods
  - MIN
  - MAX
  - Group Average