

Classical Data Analysis



Master in Big Data Solutions 2017-2018

Francisco Gutierrez

francisco.gutierrez@bts.tech

Sara Hajian

sara.hajian@bts.tech

Session 2.2.6 - Clustering

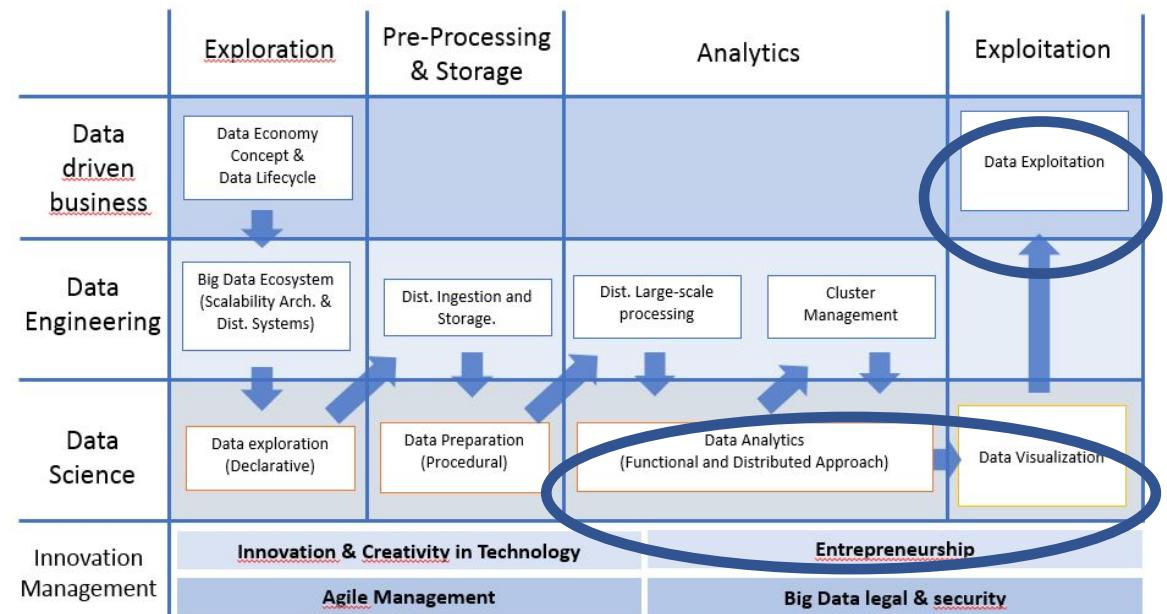
Hierarchical clustering

Sara Hajian

What we will learn

Session1: Clustering

Hierarchical clustering

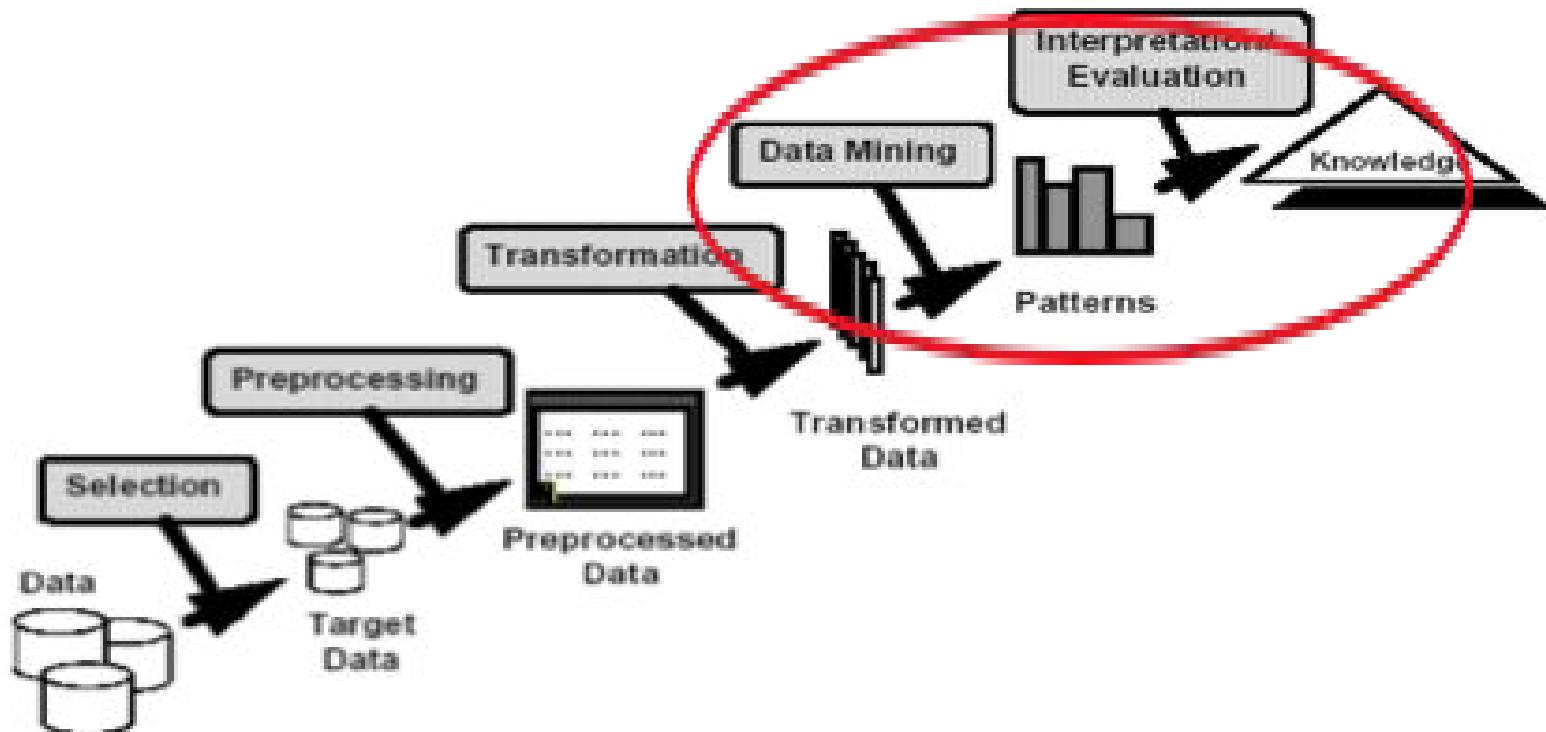


We will learn how to:

- description of hierarchical clustering
- different types of hierarchical clustering
- agglomerative hierarchical clustering algorithm
- measures for cluster proximity
- limitations and strengths

What is data mining

Non-trivial extraction of implicit, previously unknown and potentially useful information from data (i.e., **discovery of meaningful patterns**)



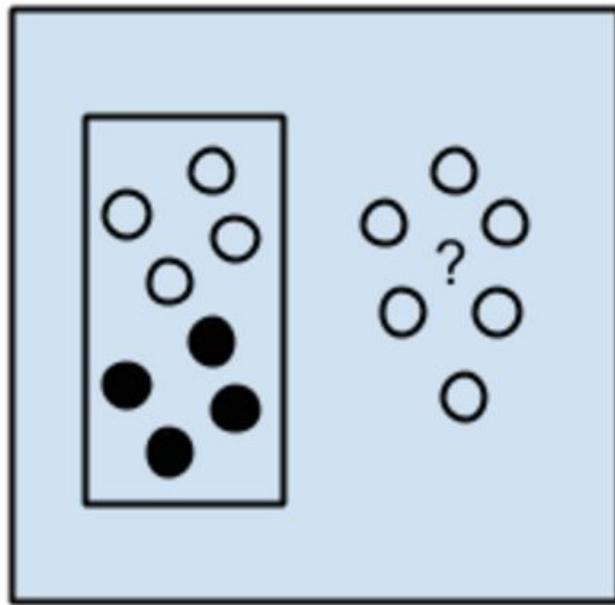
Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

Data mining algorithms

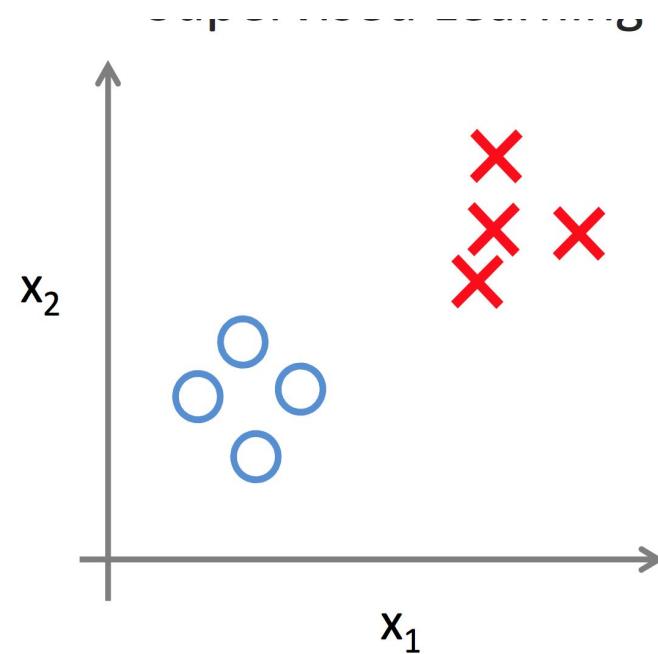
Supervised: Classification and regression

Unsupervised: Clustering

Data mining algorithms

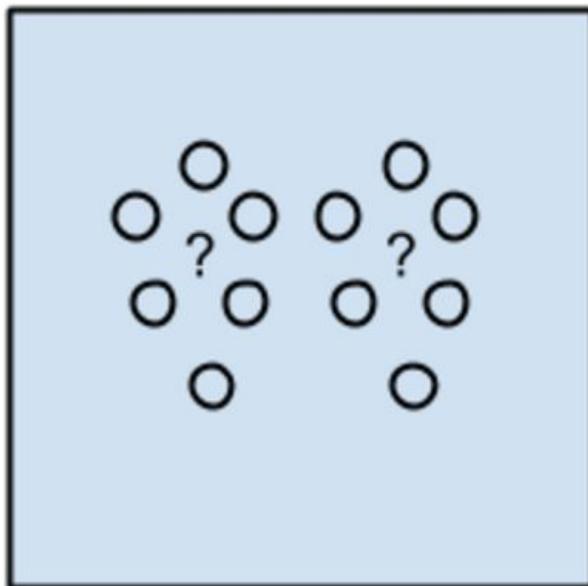


Supervised Learning
Algorithms

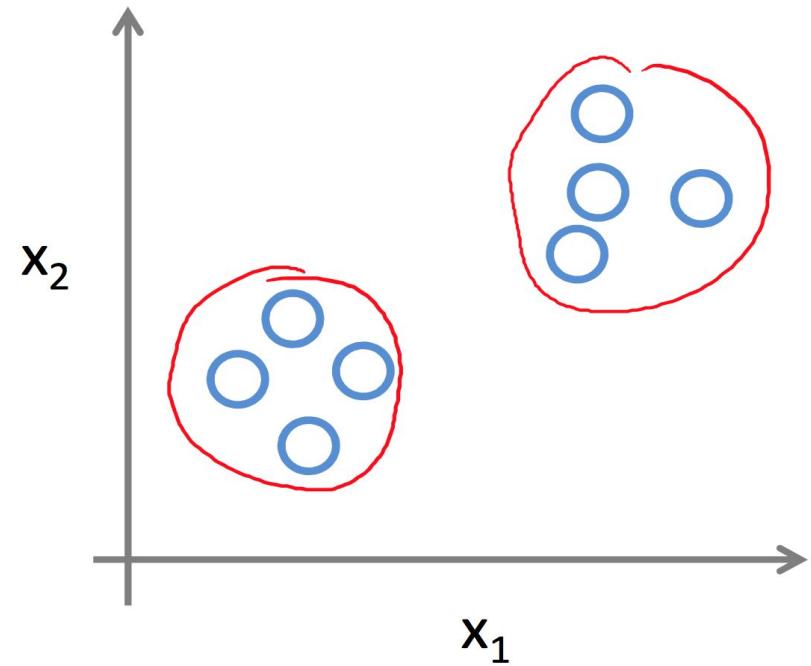


Examples: regression and classification

Data mining algorithms



Unsupervised Learning
Algorithms



Examples: clustering

Classical Data analysis: Curriculum

Part I: Regression

Linear regression

Logistic regression

Part II: Classification

Support vector machines (SVM)

Decision trees

Ensemble methods

Clustering

K-means

Hierarchical clustering

Supervised learning algorithms

Classical Data analysis: Curriculum

Part I: Regression

Linear regression

Logistic regression

Part II: Classification

Support vector machines (SVM)

Decision trees

Ensemble methods

Clustering

K-means

Hierarchical clustering

Supervised learning algorithms

Classical Data analysis: Curriculum

Part I: Regression

Linear regression

Logistic regression

Part II: Classification

Support vector machines (SVM)

Decision trees

Ensemble methods

Clustering

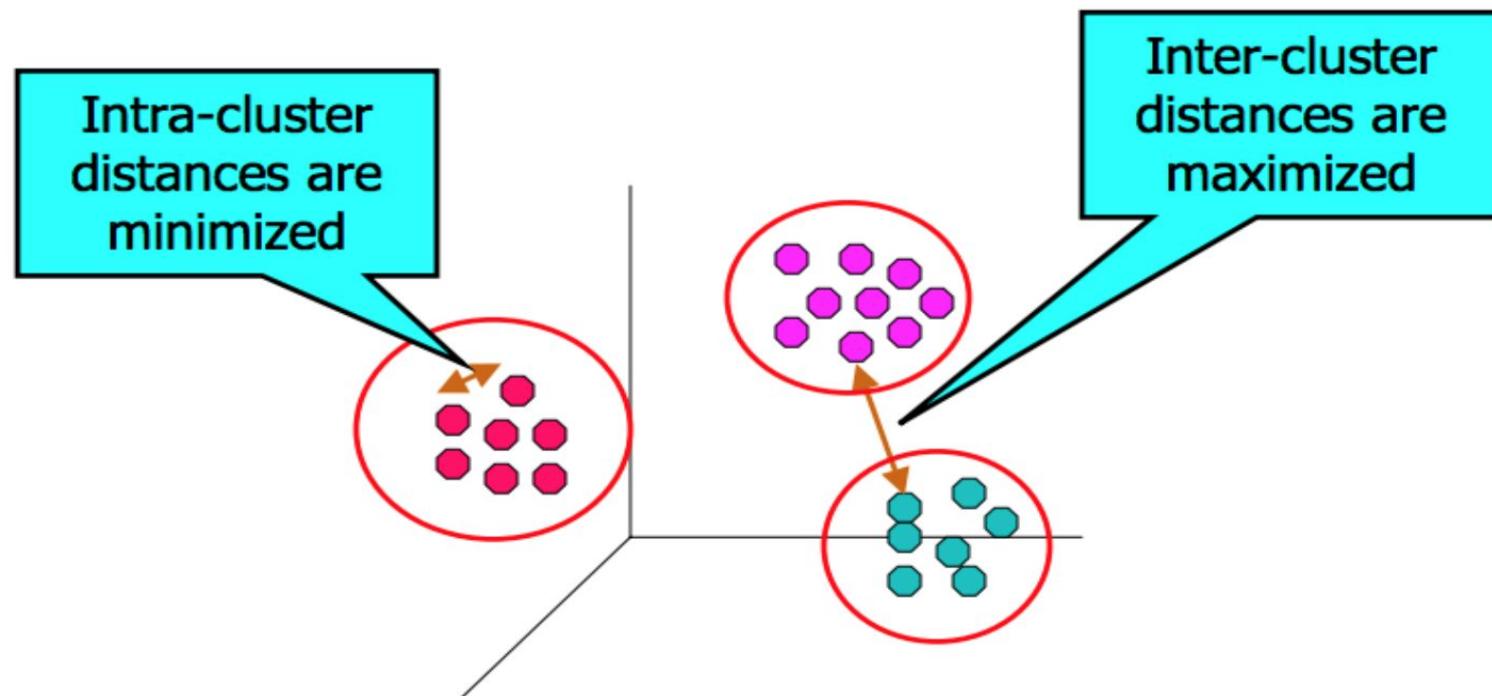
K means

Hierarchical clustering

Unsupervised learning
algorithms

Clustering analysis: definition

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Clustering analysis: applications

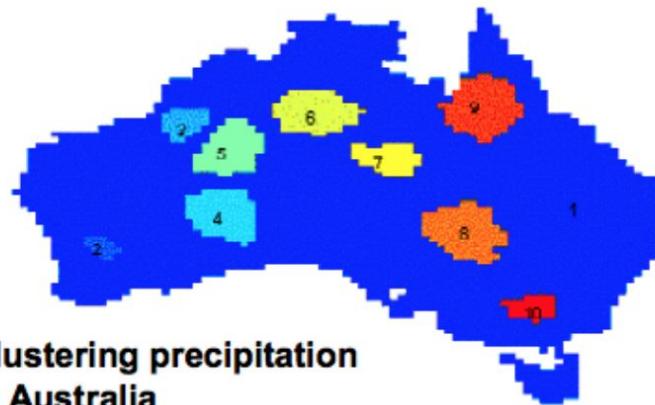
Understanding

- Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

	<i>Discovered Clusters</i>	<i>Industry Group</i>
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Summarization

- Reduce the size of large data sets



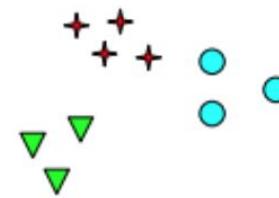
Clustering analysis: applications

- A key intermediate step for other data mining tasks
 - Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, etc.
 - Outlier detection: Outliers—those “far away” from any cluster
- Data summarization, compression, and reduction
 - Ex. Image processing: Vector quantization
- Collaborative filtering, recommendation systems, or customer segmentation
 - Find like-minded users or similar products
- Dynamic trend detection
 - Clustering stream data and detecting trends and patterns
- Multimedia data analysis, biological data analysis and social network analysis
 - Ex. Clustering images or video/audio clips, gene/protein sequences, etc.

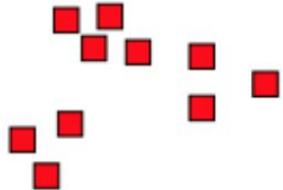
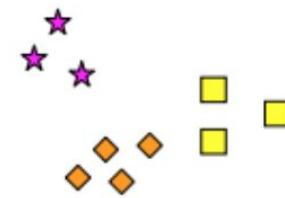
What is a cluster?



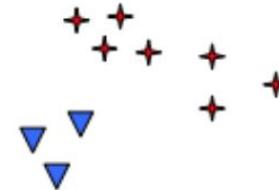
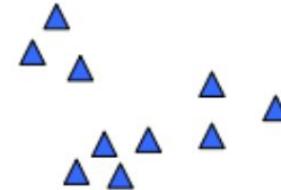
How many clusters?



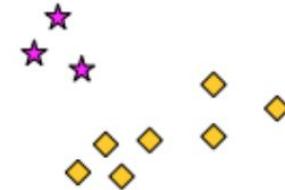
Six Clusters



Two Clusters



Four Clusters



Types of clusterings

A **clustering** is a set of clusters

Important distinction between **hierarchical** and **partitional** sets of clusters

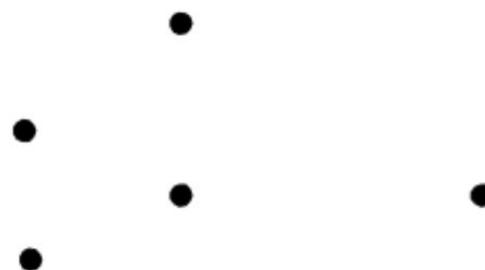
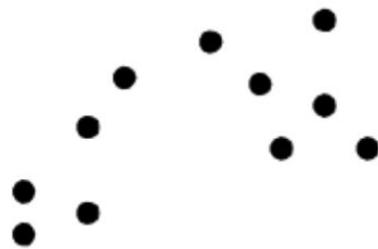
Partitional Clustering

- A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

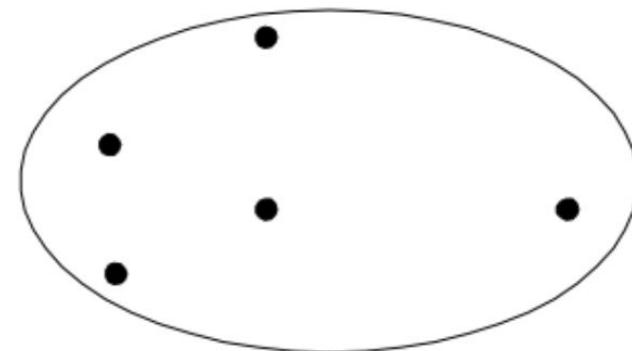
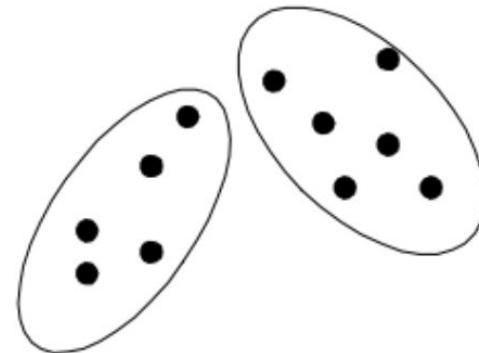
Hierarchical clustering

- A set of nested clusters organized as a hierarchical tree

Partitional clustering

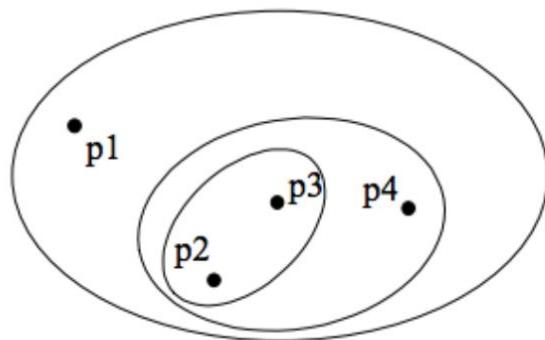


Original Points

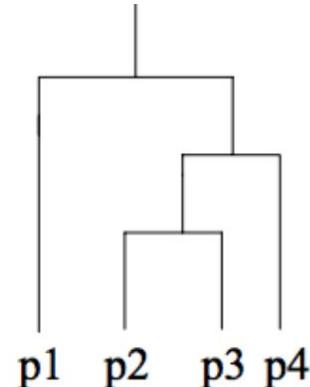


A Partitional Clustering

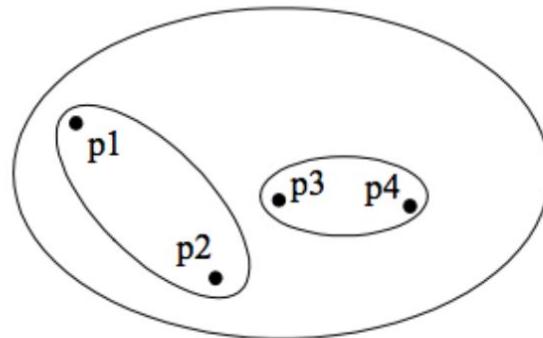
Hierarchical clustering



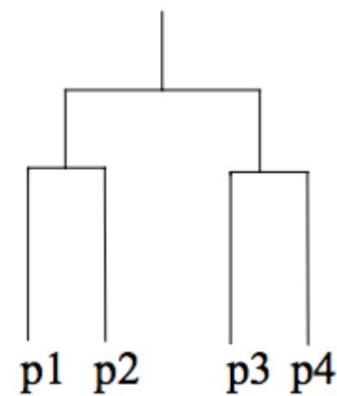
Traditional Hierarchical Clustering



Traditional Dendrogram



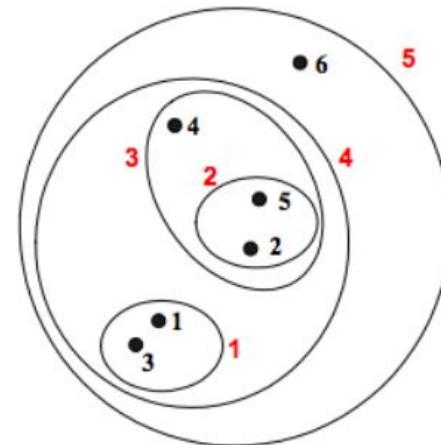
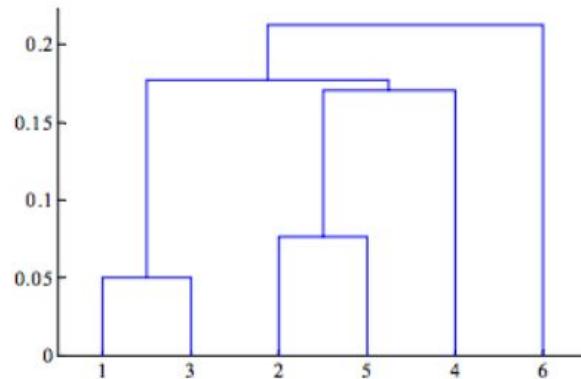
Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

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

Hierarchical clustering

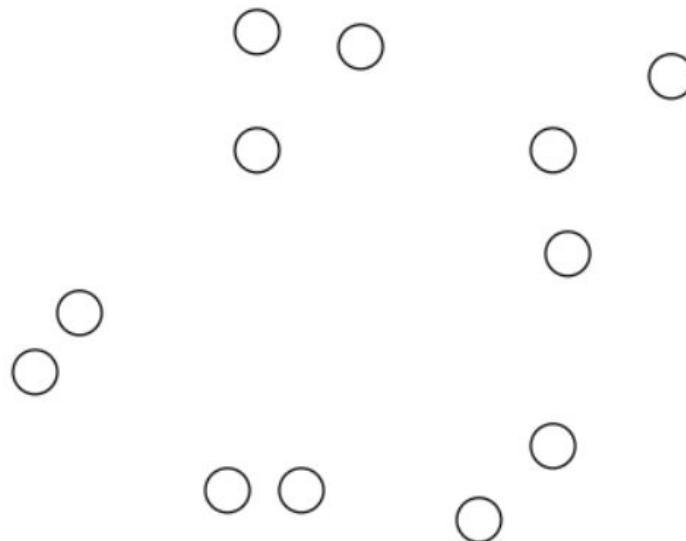
- Two main types of hierarchical clustering
 - Agglomerative:
 - ◆ Start with the points as individual clusters
 - ◆ At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
 - Divisive:
 - ◆ Start with one, all-inclusive cluster
 - ◆ At each step, split a cluster 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

Agglomerative hierarchical clustering

- 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 point

- Start with clusters of individual points and a proximity matrix



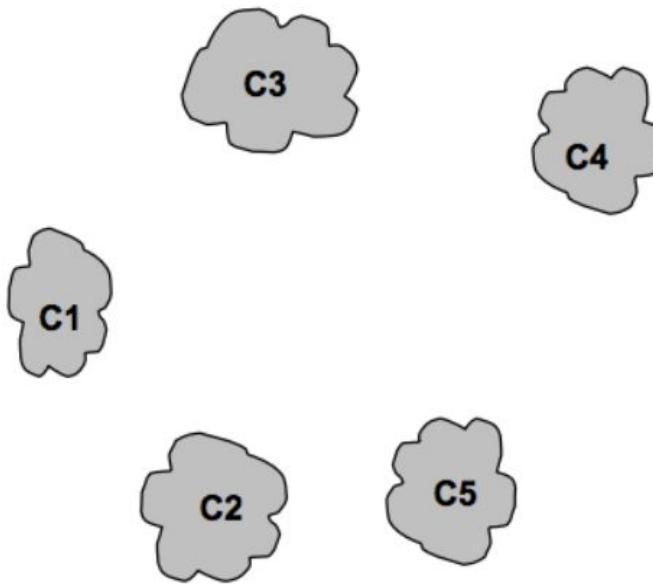
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						

Proximity Matrix

 p1  p2  p3  p4 ...  p9  p10  p11  p12

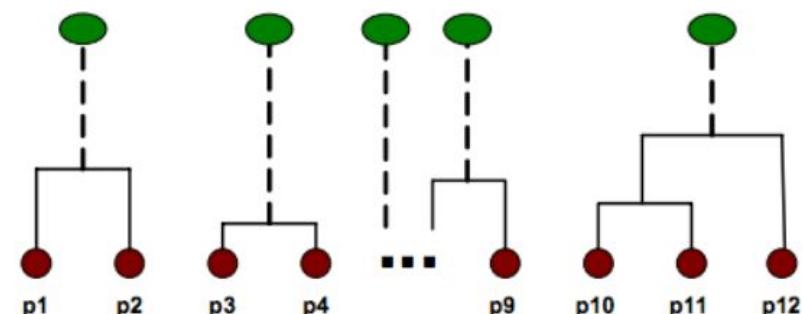
Intermediate point

- After some merging steps, we have some clusters



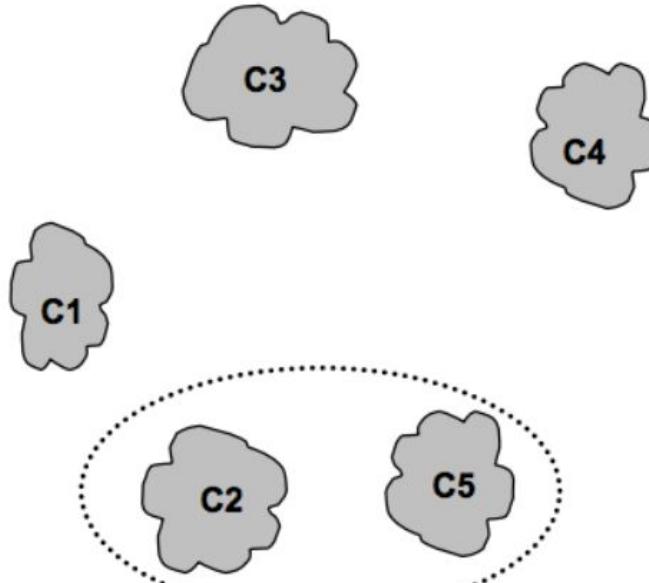
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



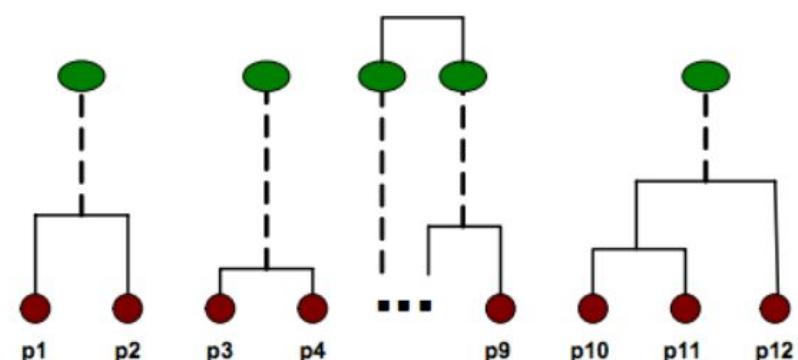
Intermediate point

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



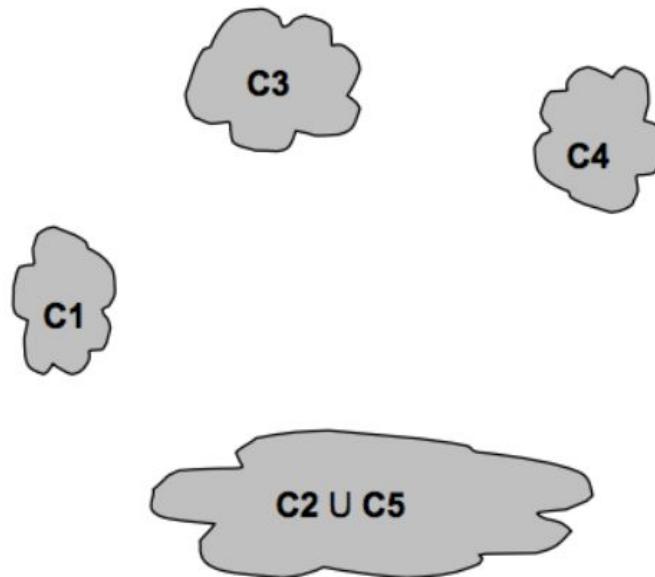
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



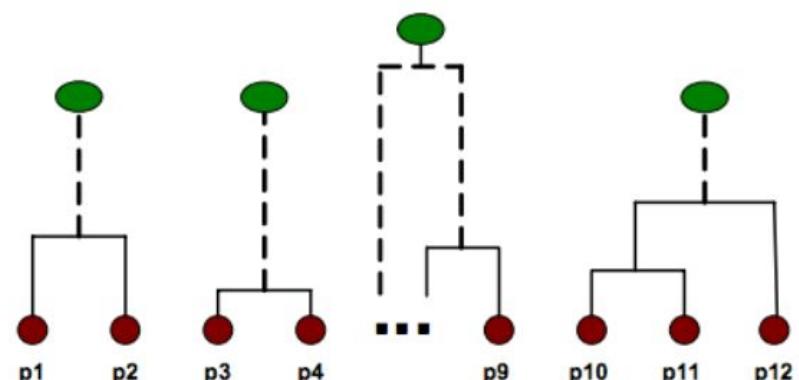
After merging

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

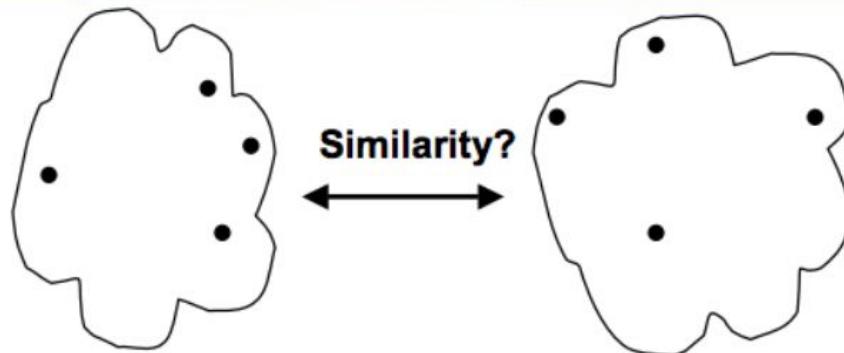


		C ₂ U C ₅			
		C ₁	C ₅	C ₃	C ₄
C ₁	C ₁	?			
	C ₂ ∪ C ₅	?	?	?	?
C ₃		?			
C ₄		?			

Proximity Matrix



How to define inter-cluster similarity

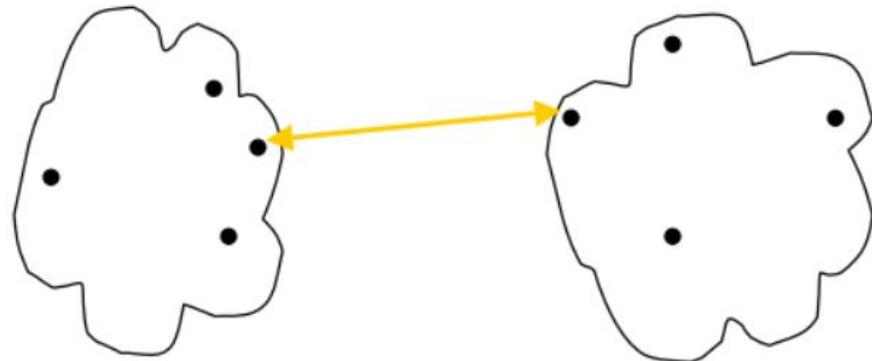


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - Ward's Method uses squared error

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

Proximity Matrix

How to define inter-cluster similarity

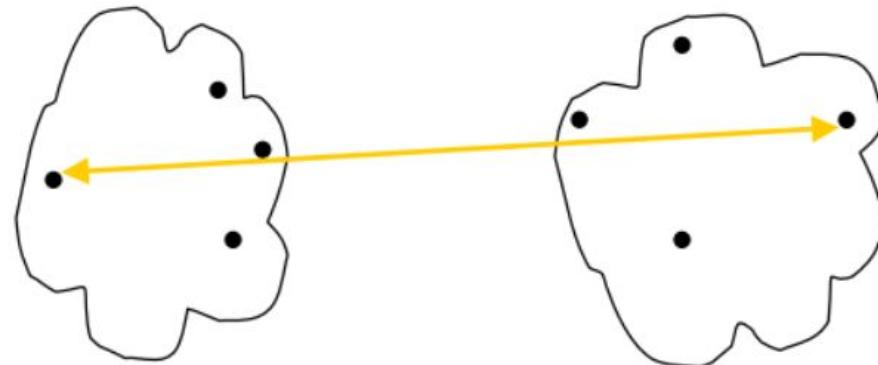


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - Ward's Method uses squared error

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

● **Proximity Matrix**

How to define inter-cluster similarity

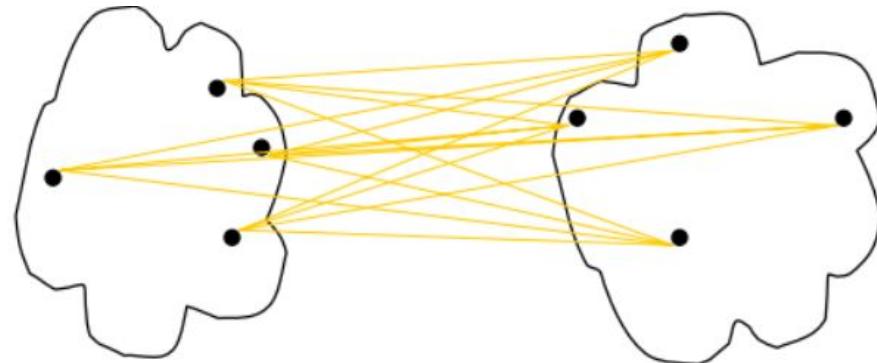


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - Ward's Method uses squared error

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

● **Proximity Matrix**

How to define inter-cluster similarity

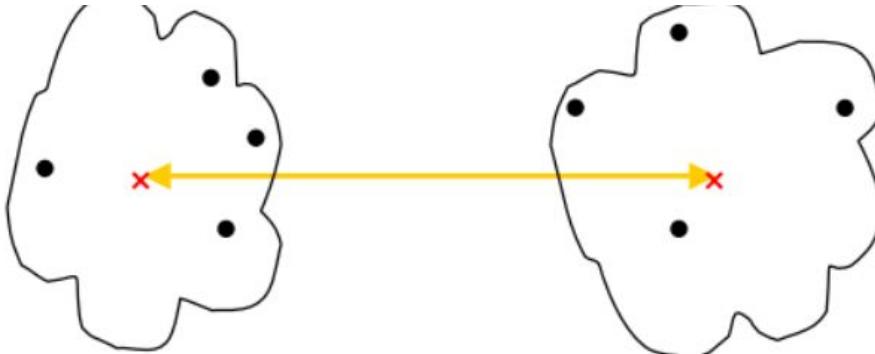


- MIN
- MAX
- **Group Average**
- Distance Between Centroids
- Other methods driven by an objective function
 - Ward's Method uses squared error

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

Proximity Matrix

How to define inter-cluster similarity

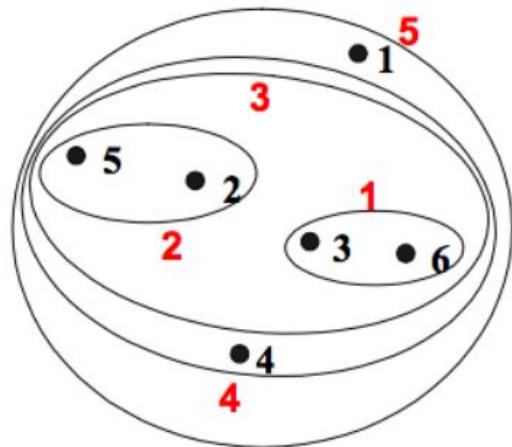


- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
 - Ward's Method uses squared error

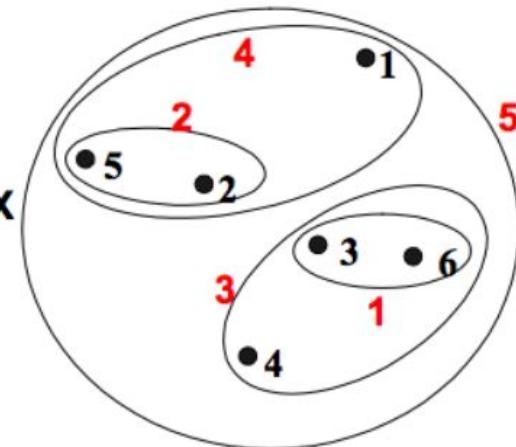
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

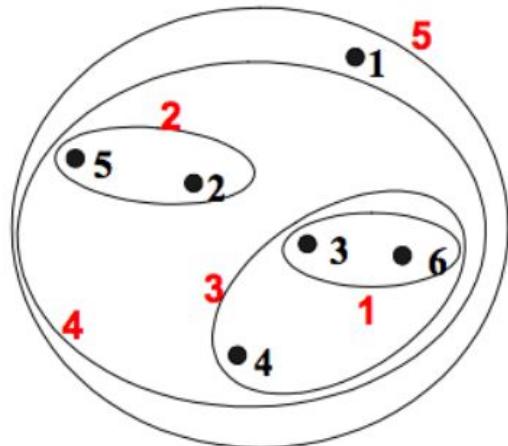
Hierarchical clustering: comparison



MIN

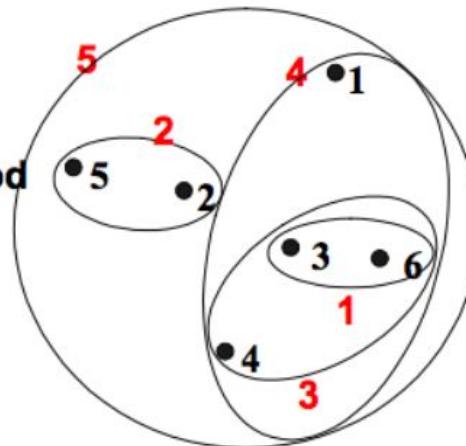


MAX



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

- 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

Cluster validity

- For supervised classification we have a variety of measures to evaluate how good our model is
 - Accuracy, precision, recall
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?
- But “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
 - To avoid finding patterns in noise
 - To compare clustering algorithms
 - To compare two sets of clusters
 - To compare two clusters

Different aspects of cluster validation

1. Determining the **clustering tendency** of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.
 - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the ‘correct’ number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

Measures of cluster validity

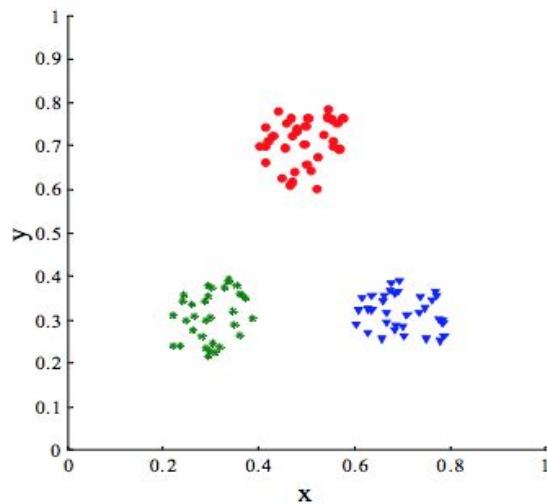
- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
 - **External Index:** Used to measure the extent to which cluster labels match externally supplied class labels.
 - ◆ Entropy
 - **Internal Index:** Used to measure the goodness of a clustering structure *without* respect to external information.
 - ◆ Sum of Squared Error (SSE)
 - **Relative Index:** Used to compare two different clusterings or clusters.
 - ◆ Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as **criteria** instead of **indices**
 - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Measuring cluster validity via correlation

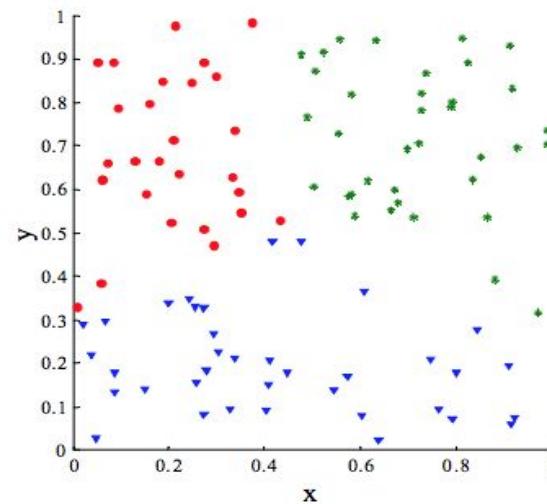
- Two matrices
 - Proximity Matrix
 - “Incidence” Matrix
 - ◆ One row and one column for each data point
 - ◆ An entry is 1 if the associated pair of points belong to the same cluster
 - ◆ An entry is 0 if the associated pair of points belongs to different clusters
- Compute the correlation between the two matrices
 - Since the matrices are symmetric, only the correlation between $n(n-1) / 2$ entries needs to be calculated.
- High correlation indicates that points that belong to the same cluster are close to each other.
- Not a good measure for some density or contiguity based clusters.

Measuring cluster validity via correlation

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.



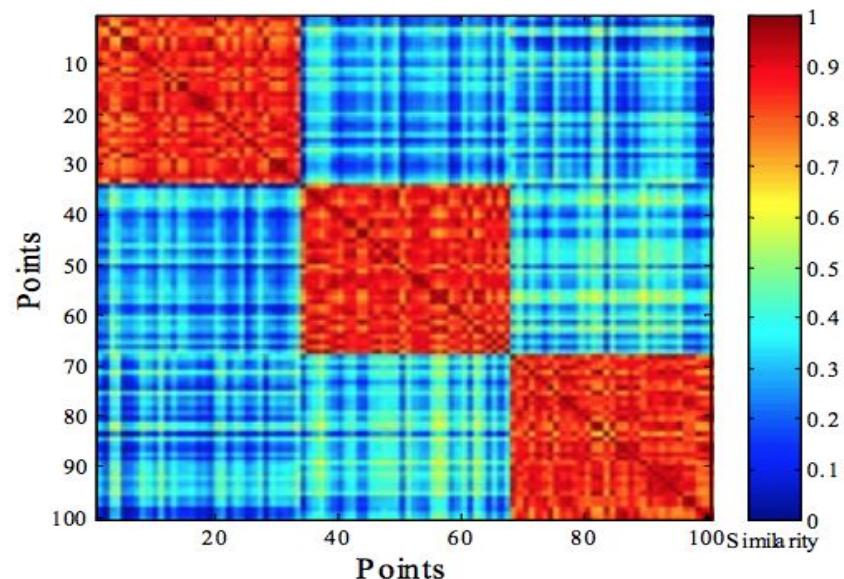
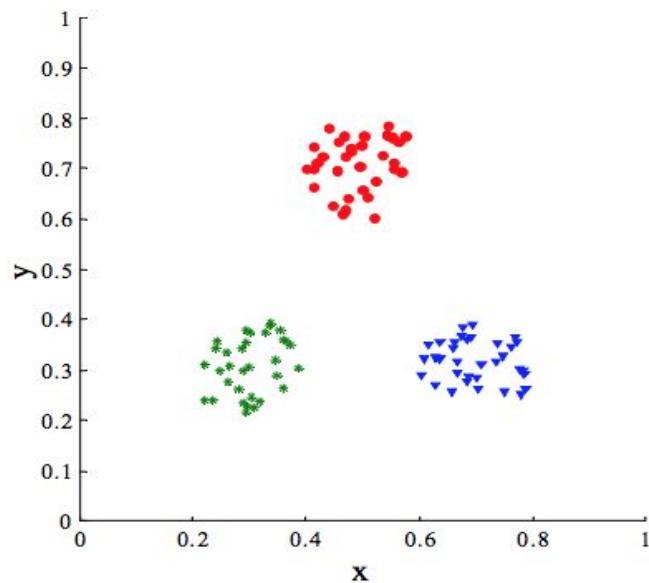
Corr = -0.9235



Corr = -0.5810

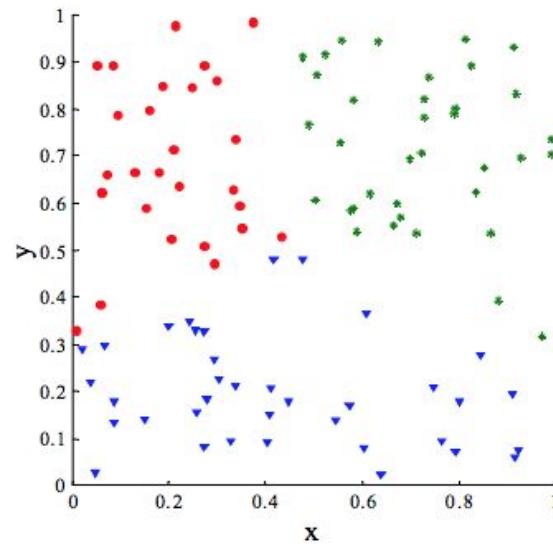
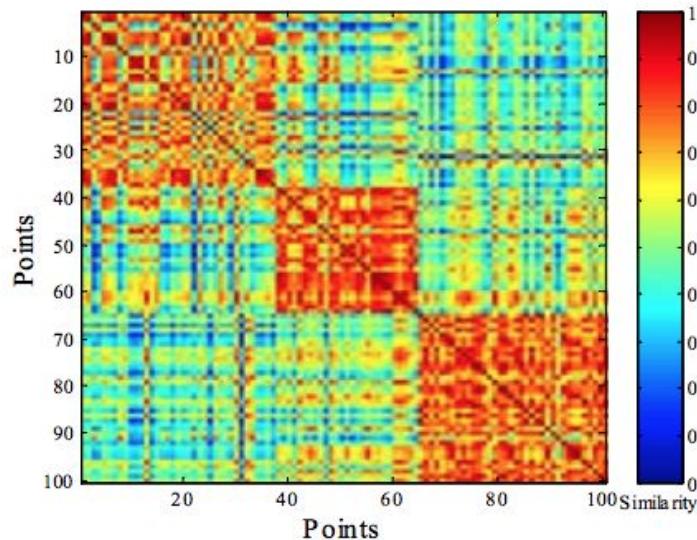
Using similarity matrix for cluster validation

- Order the similarity matrix with respect to cluster labels and inspect visually.



Using similarity matrix for cluster validation

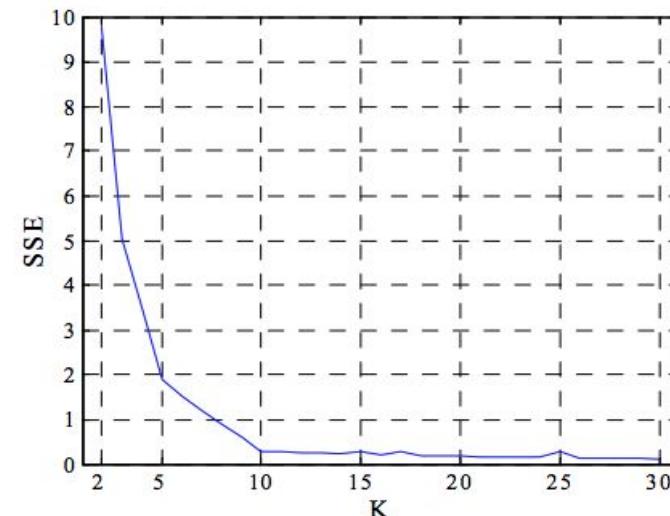
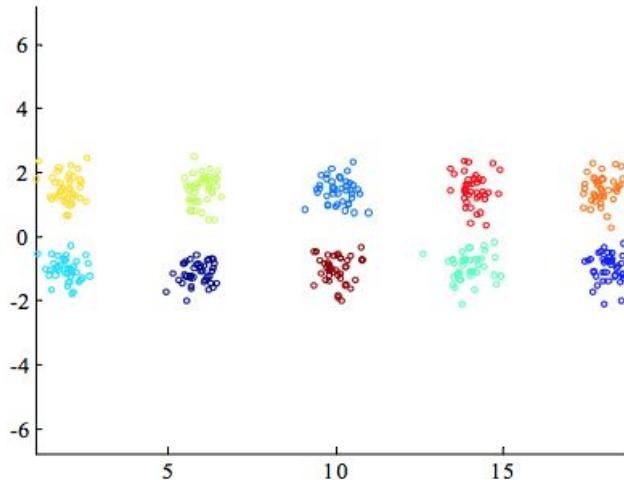
- Clusters in random data are not so crisp



K-means

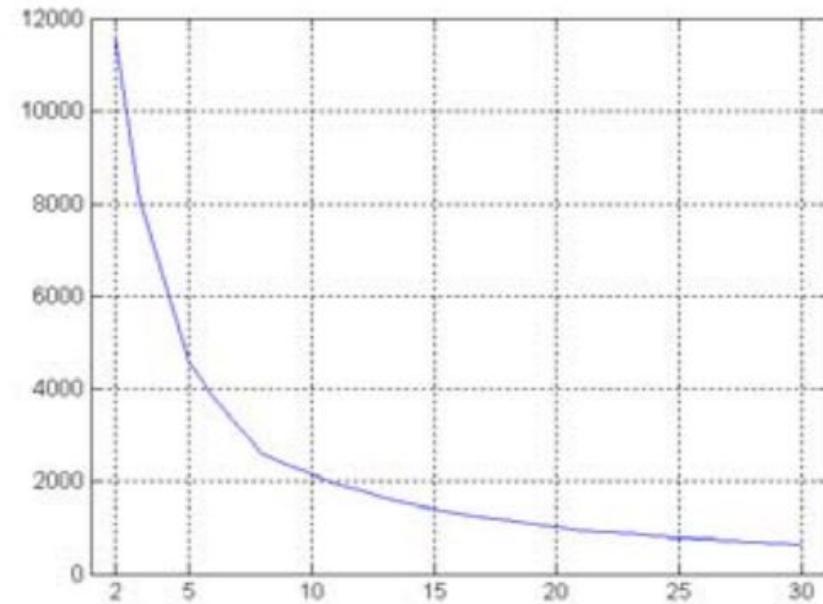
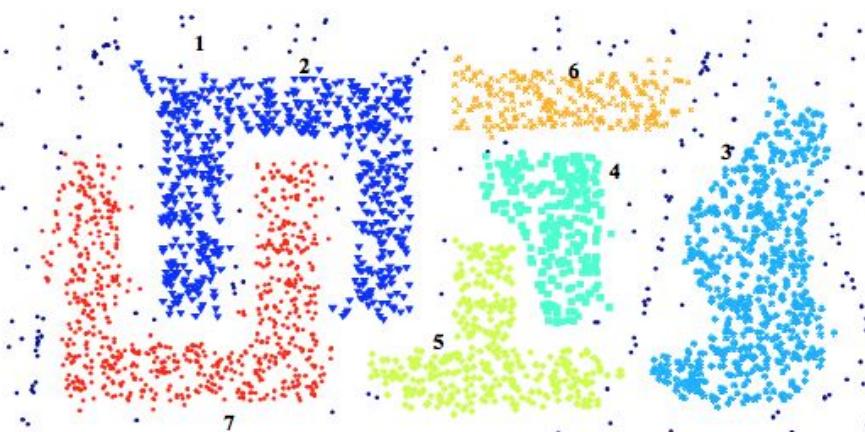
Internal measures: SSE

- Clusters in more complicated figures aren't well separated
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information
 - SSE
- SSE is good for comparing two clusterings or two clusters (average SSE).
- Can also be used to estimate the number of clusters



Internal measures: SSE

- SSE curve for a more complicated data set

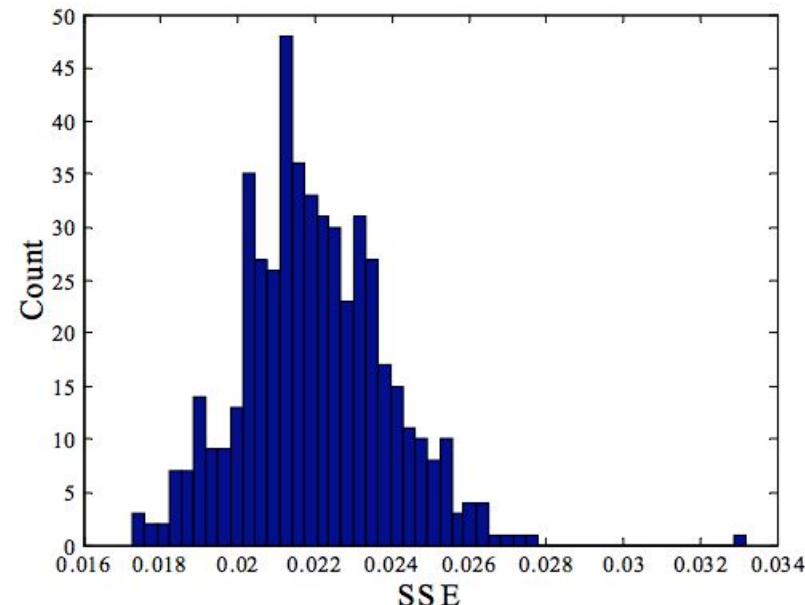
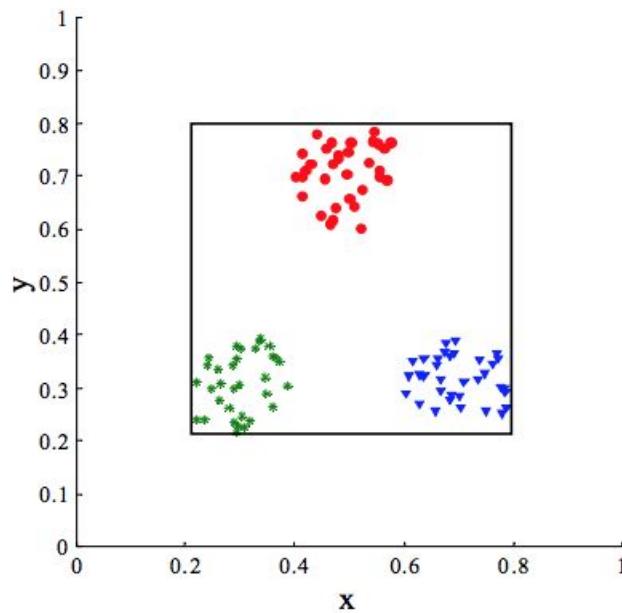


SSE of clusters found using K-means

Internal measures: SSE

● Example

- Compare SSE of 0.005 against three clusters in random data
- Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values



Internal measures: Cohesion and separation

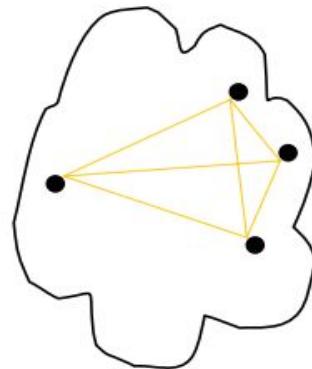
- **Cluster Cohesion:** Measures how closely related are objects in a cluster
 - Example: SSE
- **Cluster Separation:** Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
 - Cohesion is measured by the within cluster sum of squares (SSE)
$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$
 - Separation is measured by the between cluster sum of squares

$$BSS = \sum_i |C_i| (m - m_i)^2$$

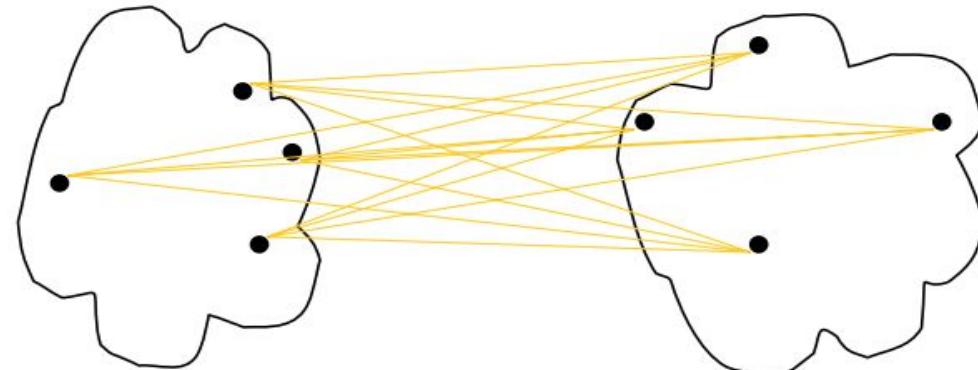
– Where $|C_i|$ is the size of cluster i

Internal measures: Cohesion and separation

- A proximity graph based approach can also be used for cohesion and separation.
 - Cluster cohesion is the sum of the weight of all links within a cluster.
 - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



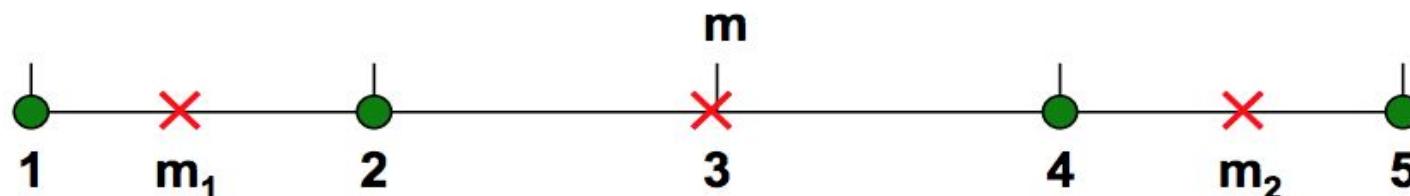
cohesion



separation

Internal measures: Cohesion and separation

- Example: SSE
 - BSS + WSS = constant



K=1 cluster:

$$WSS = (1 - 3)^2 + (2 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$$

$$BSS = 4 \times (3 - 3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

$$WSS = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 = 1$$

$$BSS = 2 \times (3 - 1.5)^2 + 2 \times (4.5 - 3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Individual assignment

Classification, data analysis and visualization in notebook

Next class

Applications of clustering

Resources

- Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics, 2001.



Thank you
Barcelona, 2017