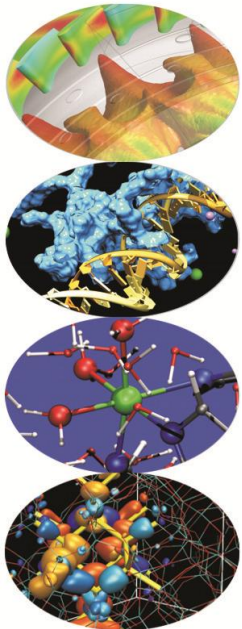




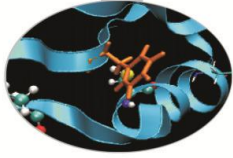
# Unsupervised techniques

Giorgio Pedrazzi, *CINECA-SCAI*  
*School of Scientific Data Analytics and*  
*Visualisation*

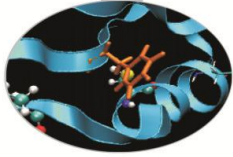
*Bologna, 21/06/2016*



# Unsupervised Learning



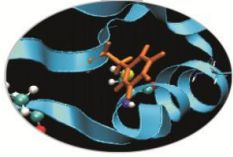
- Cluster Analysis
  - Basic concept and introduction
  - Major clustering approaches
  - Distance measures
- Association rules and sequential patterns
  - Association rules
  - Sequential patterns



# Clustering

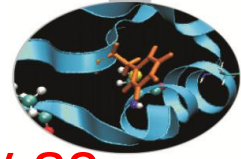
- **Cluster:** A collection/group of data objects/points
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- **Cluster analysis**
  - find *similarities* between data according to characteristics underlying the data and grouping similar data objects into clusters
- **Clustering Analysis: Unsupervised learning**
  - no predefined classes for a training data set
  - Two general tasks: **identify the “natural” clustering number** and **properly grouping objects into “sensible” clusters**
- **Typical applications**
  - as a **stand-alone tool** to gain an insight into data distribution
  - as a **preprocessing step** of other algorithms in intelligent systems

# Typical applications



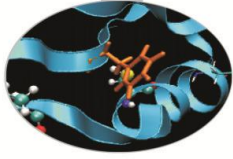
- Scientific applications
  - **Gene expression data:** Discover genes with similar functions in DNA microarray data.
  - ...
- Business applications
  - **Customer segmentation:** Discover distinct groups in customer bases (insurance, bank, retailers) to develop targeted marketing programs.
  - ...
- Internet applications
  - **Social network analysis:** in the study of social networks, clustering may be used to recognize communities within large groups of people.
  - **Search result grouping:** in the process of intelligent grouping of the files and websites, clustering may be used to create a more relevant set of search results.
  - ...

# Data Matrix



The problem **must be formulated in a mathematical way as a matrix of data** containing information on  $N$  objects (cases or observations ; rows of the matrix) specified by the values assigned to  $V$  variables (columns of the matrix)

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

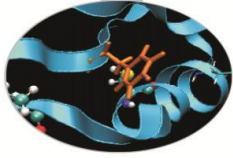


# Cluster Analysis steps

- Pre processing
- Select a clustering algorithm
- Select a distance or a similarity measure (\*)
- Determine the number of clusters (\*)
- Validate the analysis

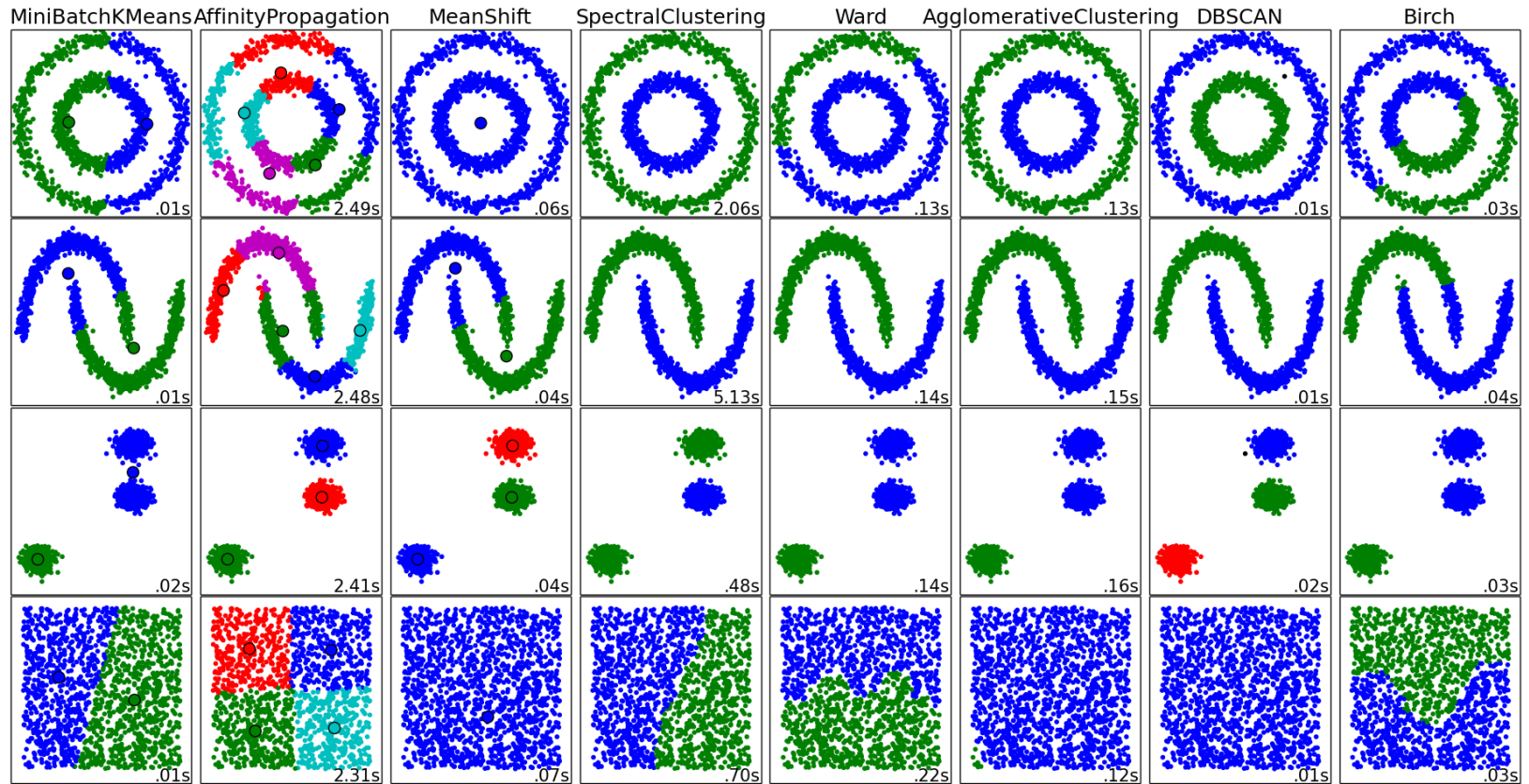
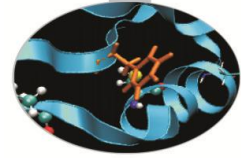
(\*) if needed by the method used

# Classification of methods



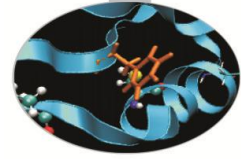
- **Distance-based**
  - partitioning algorithms
  - hierarchical algorithms
- Density based (DBSCAN)
- Model based
- Spectral clustering
- Combination of methods

# Comparison of algorithms





# Distance measure



## Minkowski distance ( $L_p$ Norm)

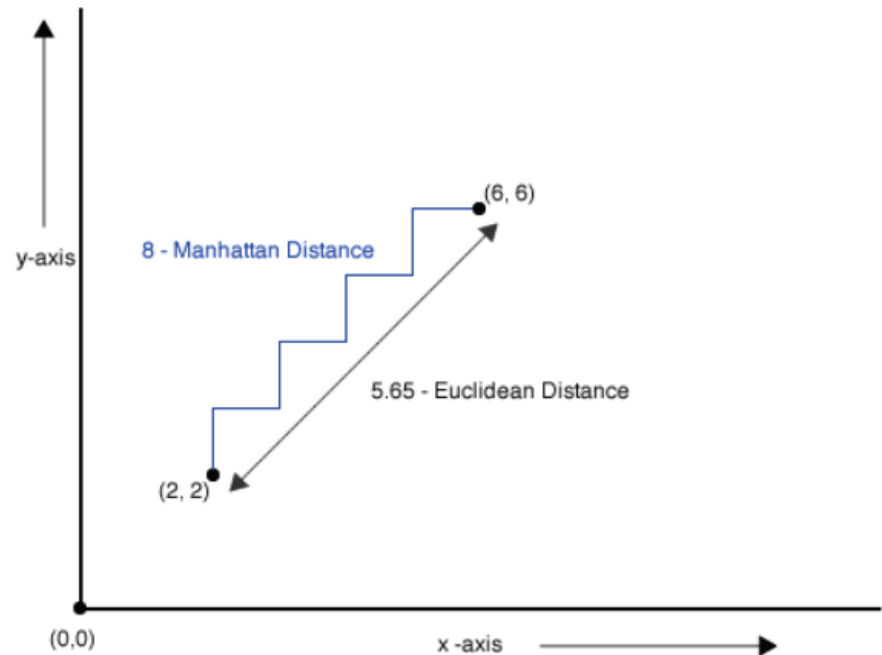
$$d(i, k) = \left[ \sum_{j=1}^d |x_{ij} - x_{kj}|^p \right]^{1/p}$$

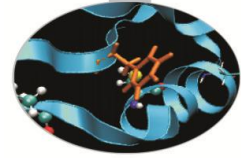
## Euclidean distance ( $L_2$ Norm)

$$d(i, k) = \left[ \sum_{j=1}^d (x_{ij} - x_{kj})^2 \right]^{1/2}$$

## Manhattan distance (city block distance)

$$d(i, k) = \sum_{j=1}^d |x_{ij} - x_{kj}|$$





# Distance Measures

- Cosine Measure (Similarity vs. Distance)

For  $\mathbf{x} = (x_1 \ x_2 \ \dots \ x_n)$  and  $\mathbf{y} = (y_1 \ y_2 \ \dots \ y_n)$

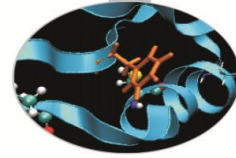
$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \bullet \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{x_1 y_1 + \dots + x_n y_n}{\sqrt{x_1^2 + \dots + x_n^2} \sqrt{y_1^2 + \dots + y_n^2}}$$

$$d(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \bullet \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

$$0 \leq d(\mathbf{x}, \mathbf{y}) \leq 2$$

- Nonmetric vector objects: keywords in documents, gene features in micro-arrays, ...
- Applications: information retrieval, biologic taxonomy, ...

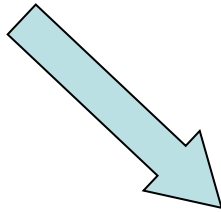
# Similarity measures



Correspondent 1's

$$x_k: \begin{matrix} 0 & 1 & 1 & 0 & 1 \end{matrix}$$

$$x_j: \begin{matrix} 1 & 1 & 0 & 1 & 1 \end{matrix}$$



	1	0
1	$a_{11}$	$a_{10}$
0	$a_{01}$	$a_{00}$



	1	0
1	2	2
0	1	0

Jaccard:

$$d(i,k) = (a_{11}) / (a_{11} + a_{10} + a_{01})$$

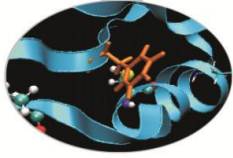
Condorcet:

$$d(i,k) = a_{11} / [a_{11} + 0.5(a_{10} + a_{01})]$$

Dice bis:

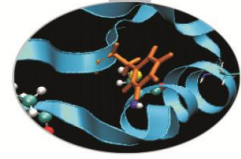
$$d(i,k) = a_{11} / [a_{11} + 0.25(a_{10} + a_{01})]$$

# Partitioning Approach

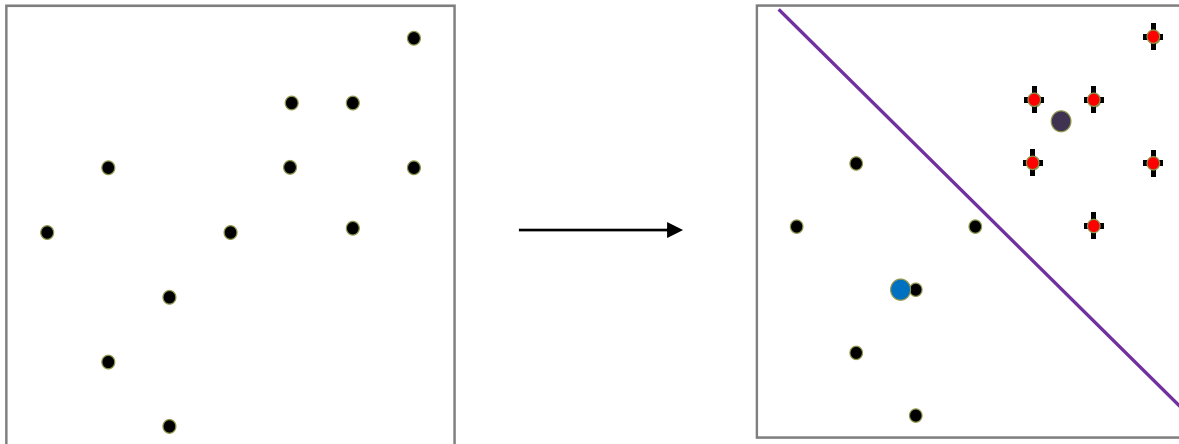


- Partitioning method: Discovering the groupings in the data by optimizing a specific objective function and iteratively improving the quality of partitions
- K-partitioning method: Partitioning a dataset  $D$  of  $n$  objects into a set of  $K$  clusters so that an objective function is optimized (e.g., the sum of squared distances is minimized, where  $c_k$  is the centroid or medoid of cluster  $C_k$ )
  - A typical objective function: Sum of Squared Errors (SSE)
- Problem definition: Given  $K$ , find a partition of  $K$  clusters that optimizes the chosen partitioning criterion
  - Global optimal: Needs to exhaustively enumerate all partitions
  - Heuristic methods (i.e., greedy algorithms): K-Means, K-Medians, K-Medoids, etc.

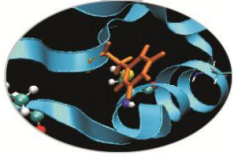
# Partitioning Approach



- Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square distance cost
- Typical methods: K-Means, K-Medoids, K-Medians, .....

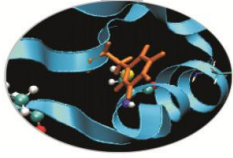


# Density based approach

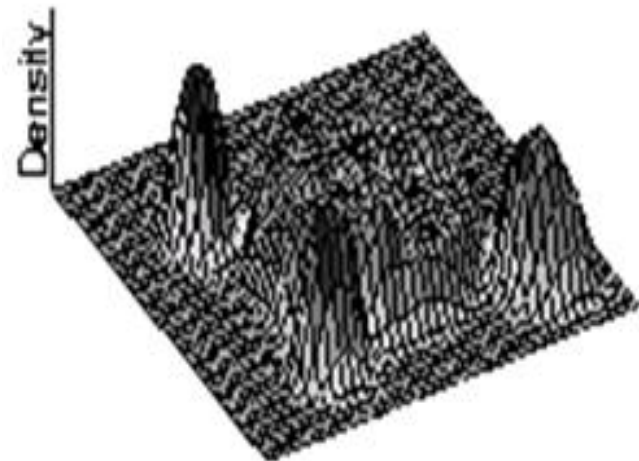
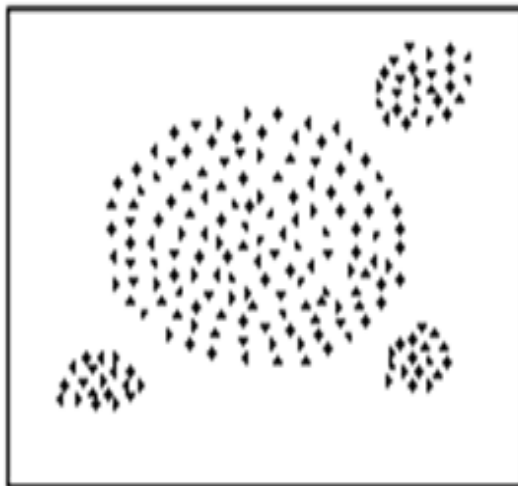


Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature

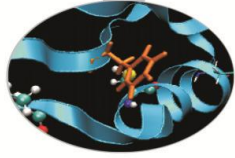
# Density based Approach



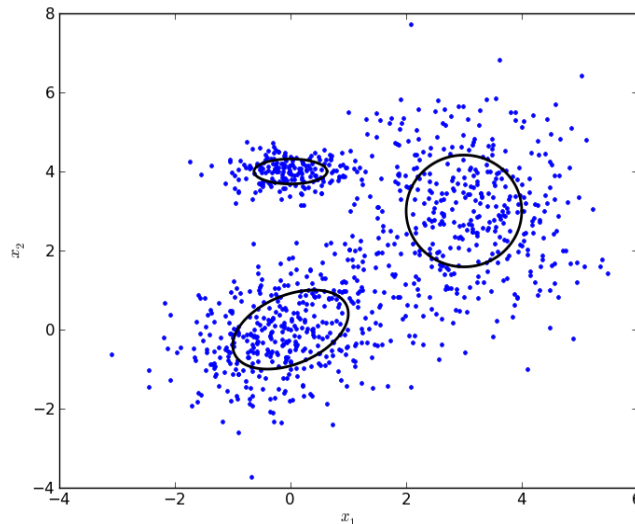
- Based on connectivity and density functions
- Typical methods: DBSCAN, OPTICS, DenClue, .....



# Model-based Approach

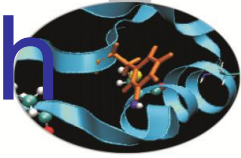


- Probabilistic model-based clustering
  - In model-based clustering, it is assumed that the data are generated by a mixture of underlying probability distributions in which each component represents a different group or cluster.
  - Cluster: Data points (or objects) that most likely belong to the same distribution
  - Clusters are created so that they will have a maximum likelihood fit to the model by a mixture of  $K$  component distributions (i.e.,  $K$  clusters)



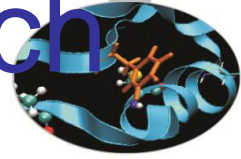


# Spectral Clustering Approach

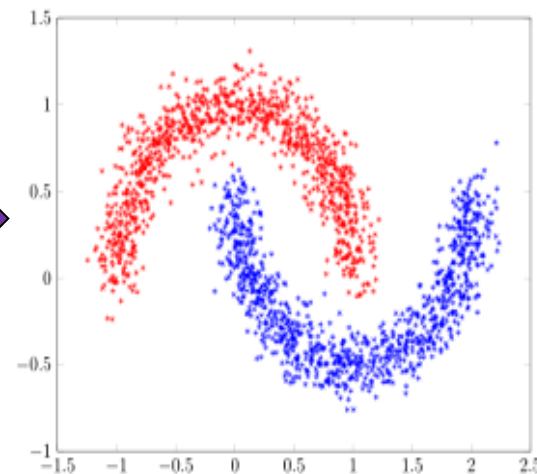
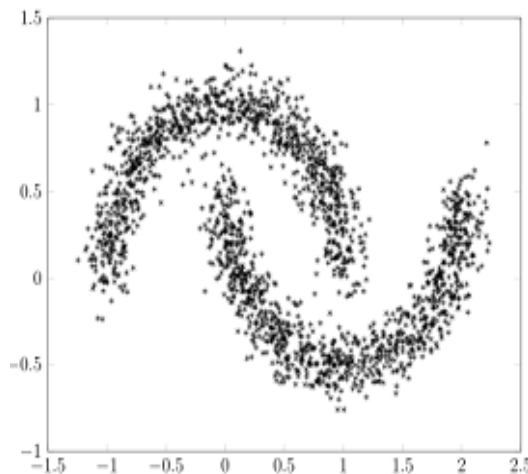
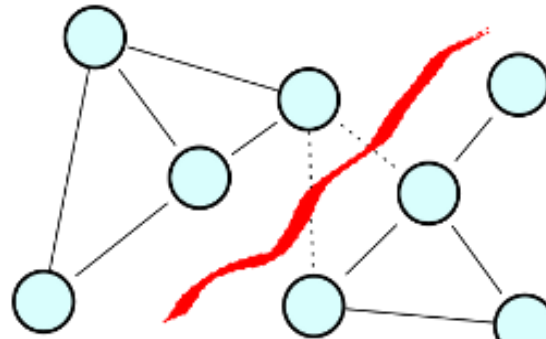


- In multivariate statistics, spectral clustering techniques make use of eigenvalue decomposition (spectrum) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset.
- In application to image segmentation, spectral clustering is known as segmentation-based object categorization.

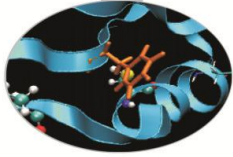
# Spectral Clustering Approach



- Convert data set into weighted graph (vertex, edge), then cut the graph into sub-graphs corresponding to clusters via spectral analysis
- Typical methods: Normalised-Cuts .....



# Combination of methods

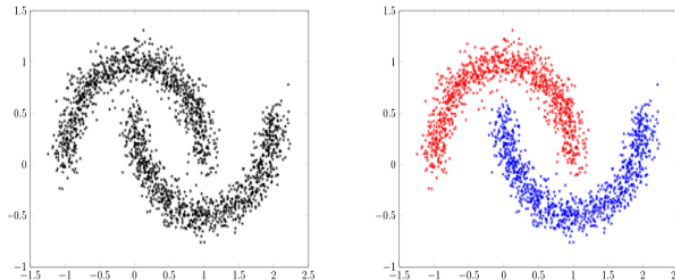


Using different methods can be useful for overcome the drawbacks of a single methods.

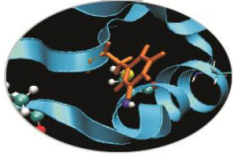
For example it is possible to generate a large number of clusers with K-means and then cluster them together using a hierarchical method.

It is important using the “single-link” method, in which the distance between two clusters is defined by the distance between the two closest data points we can find, one from each cluster.

This method has been applied to find cluster in non-convex set.



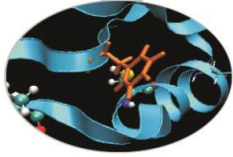
# Clustering validation



Clustering validation is a technique to find a set of clusters that best fits natural partitions (number of clusters  $K$ ). Generally speaking, there are two types of validation techniques, which are based on internal criteria and external criteria.

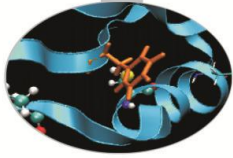
- Internal validation: based on the information intrinsic to the data alone
- External validation: based on previous knowledge about data

# Summary



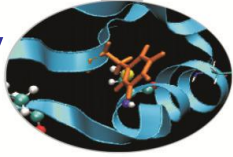
- **Clustering analysis** groups objects based on their (dis)similarity and has a broad range of applications.
- Measure of **distance** (or **similarity**) plays a critical role in clustering analysis and distance-based learning.
- Clustering algorithms can be categorized into partitioning, hierarchical, density-based, model-based, spectral clustering as well as combination approaches.

# Market Basket Analysis



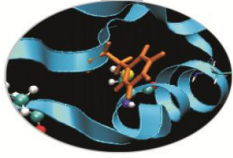
- Basket data consist of collection of transaction date and items bought in a transaction
- Retail organizations interested in generating qualified decisions and strategy based on analysis of transaction data
  - what to put on sale, how to place merchandise on shelves for maximizing profit, customer segmentation based on buying pattern
- **Examples.**
  - Rule form: LHS  $\rightarrow$  RHS [confidence, support].
  - diapers  $\rightarrow$  beers [60%, 0.5%]
  - “90% of transactions that purchase bread and butter also purchase milk”
  - bread and butter  $\Rightarrow$  milk [90%, 1%]
- [https://www.youtube.com/watch?v=N5WurXNec7E&feature=player\\_embedded](https://www.youtube.com/watch?v=N5WurXNec7E&feature=player_embedded)

# Association Rules discovery



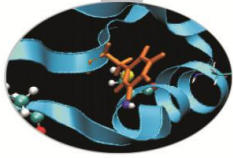
- The problem of mining association rules:  
Generate all association rules that have support and confidence greater than the user-specified *minsup* and *minconf*
- **Minimum support  $s$  (*minsup*)** --- the union of items in the LHS and RHS of the rule is present in a minimum of  $s\%$  of transactions in the database
- **Minimum confidence  $c$  (*minconf*)** --- at least  $c\%$  of transactions in the database that satisfy the LHS of the rule also satisfy the RHS of the rule

# Association rule discovery problem



- Two sub-problems in discovering all association rules:
  - Find all sets of items (itemsets) that have transaction support above minimum support → Itemsets with minimum support are called *large itemsets*, and all others small itemsets.
  - Generate from each large itemset, rules that use items from the large itemset.
    - Given a large itemset  $Y$ , and  $X$  is a subset of  $Y$
    - Take the support of  $Y$  and divide it by the support of  $X$
    - If the ratio is at least *minconf*, then  $X \Rightarrow (Y - X)$  is satisfied with confidence factor  $c$

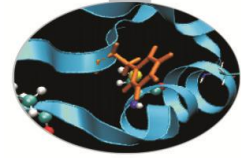




# Discovering Large Itemsets

- Algorithm for discovering large itemsets make multiple passes over the data
  - In the first pass: count the support of individual items and determine which of them are large.
  - In each subsequent pass:
    - start with a set of itemsets found to be large in the previous pass.
    - This set is used for generating new potentially large itemsets, called *candidate* itemsets
    - counts the actual support for these candidate itemsets during the pass over the data.
  - This process continues until no new large itemsets are found.

# Generate rules from large itemsets

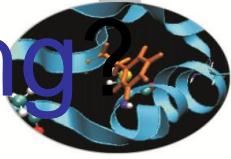


$Y = \{\text{Bread, Butter, Milk}\}, \quad X = \{\text{Bread, Butter}\}$

$conf = \text{support}(Y) / \text{support}(X) = \{\text{Bread, Butter, Milk}\} / \{\text{Bread, Butter}\}$

if  $conf \geq minconf$  then the rule  $\{\text{Bread, Butter}\} \Rightarrow \text{Milk}$  holds

# What Is Sequential Pattern Mining



- Given a set of sequences and support threshold, find the complete set of *frequent* subsequences

A sequence : < (ef) (ab) (df) c b >

A sequence database

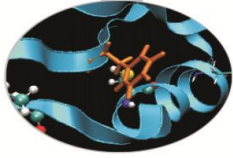
SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

An element may contain a set of items. Items within an element are unordered and we list them alphabetically.

<a(bc)dc> is a subsequence of <a(abc)(ac)d(cf)>

Given support threshold  $min\_sup = 2$ , <(ab)c> is a sequential pattern

# Applications of sequential pattern mining



- Customer shopping sequences:
  - First buy computer, then CD-ROM, and then digital camera, within 3 months.
- Medical treatments, natural disasters (e.g., earthquakes), science & eng. processes, stocks and markets, etc.
- Telephone calling patterns, Weblog click streams
- DNA sequences and gene structures