Master Degree in Artificial Intelligence for Science and Technology

Introduction to Cluster Analysis

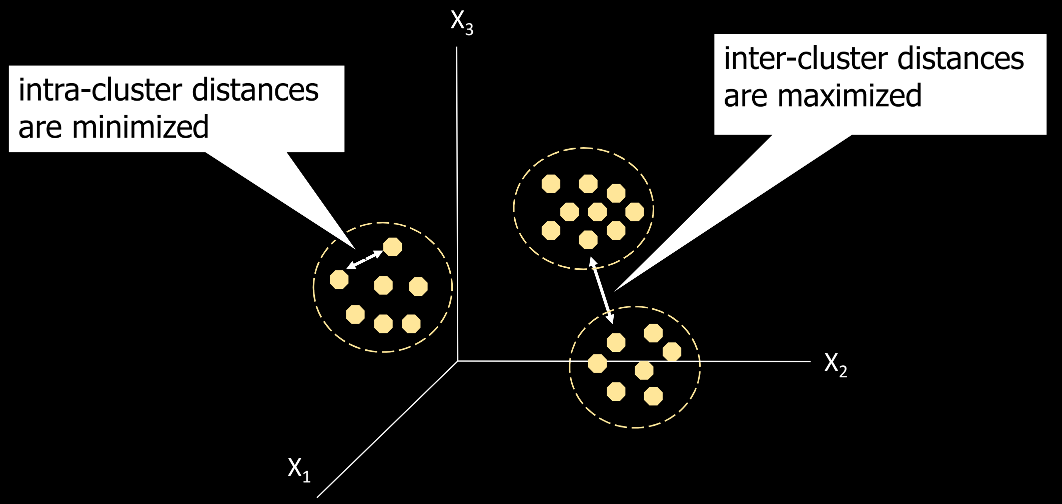
Fabio Stella
Department of Informatics, Systems and Communication
University of Milan-Bicocca
fabio.stella@unimib.it

OUTLOOK

- CLUSTER ANALYSIS
- UNDERSTANDING AND SUMMARIZING
- TYPES OF CLUSTERING
- TYPES OF CLUSTERS
- COMPONENTS OF CLUSTER ANALYSIS

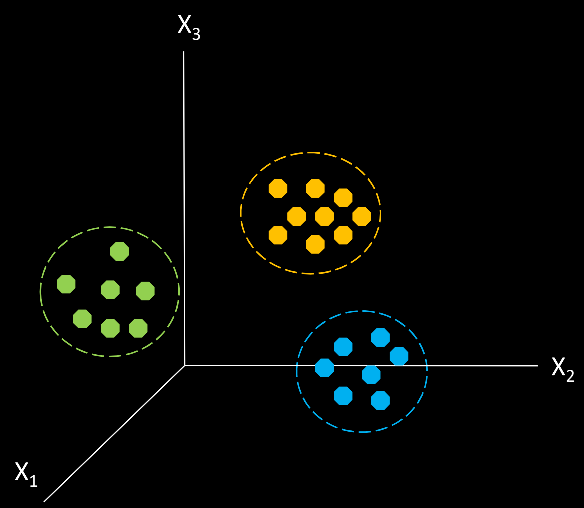
WHAT IS CLUSTER ANALYSIS?

Given a set of objects, place them in groups such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups.



WHAT IS CLUSTER ANALYSIS?

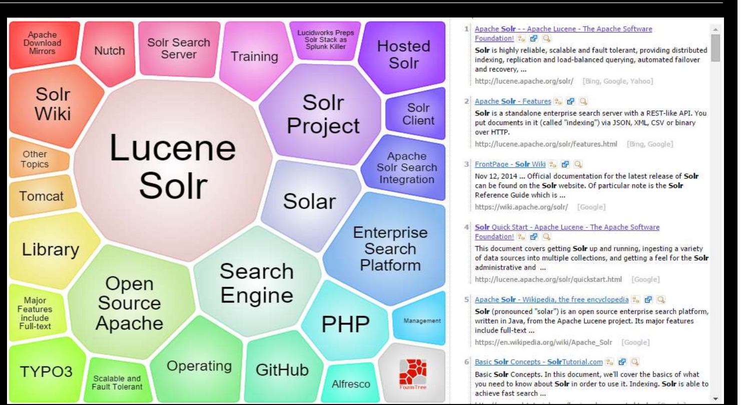
Given a set of objects, place them in groups such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups.

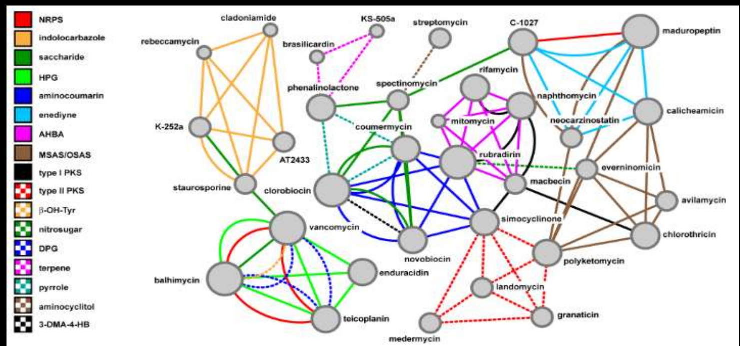


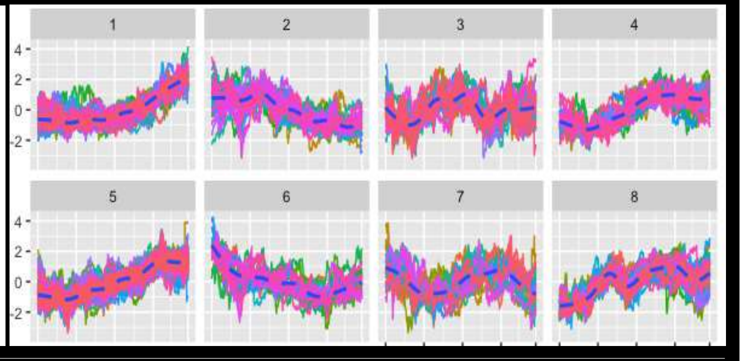
WHICH APPLICATIONS OF CLUSTER ANALYSIS?

- **UNDERSTANDING**

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



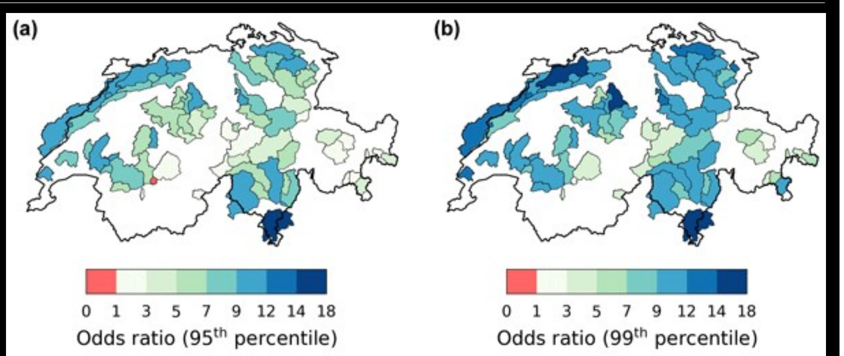


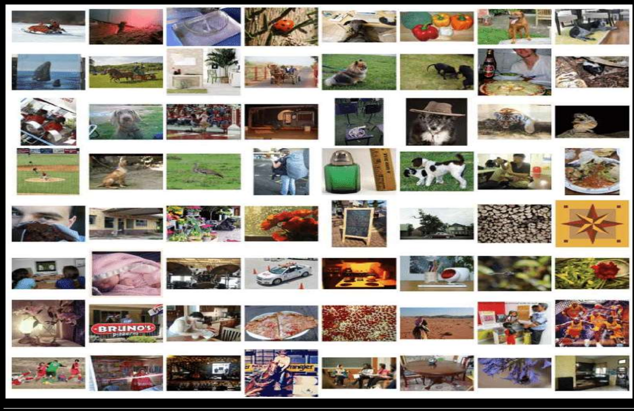


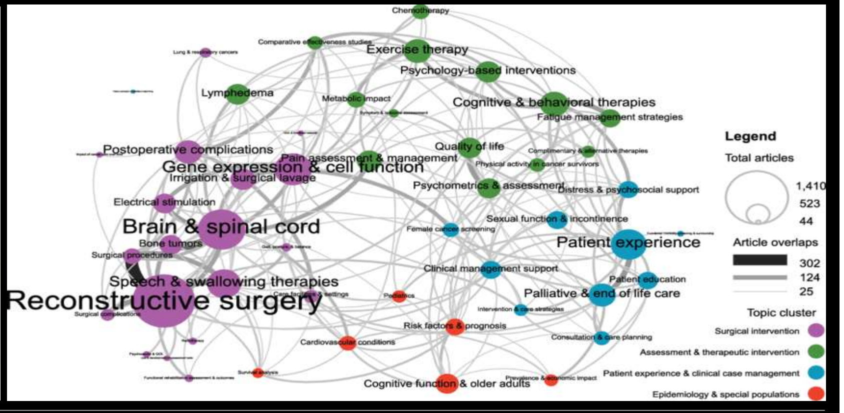
WHICH APPLICATIONS OF CLUSTER ANALYSIS?

- **SUMMARIZATION**

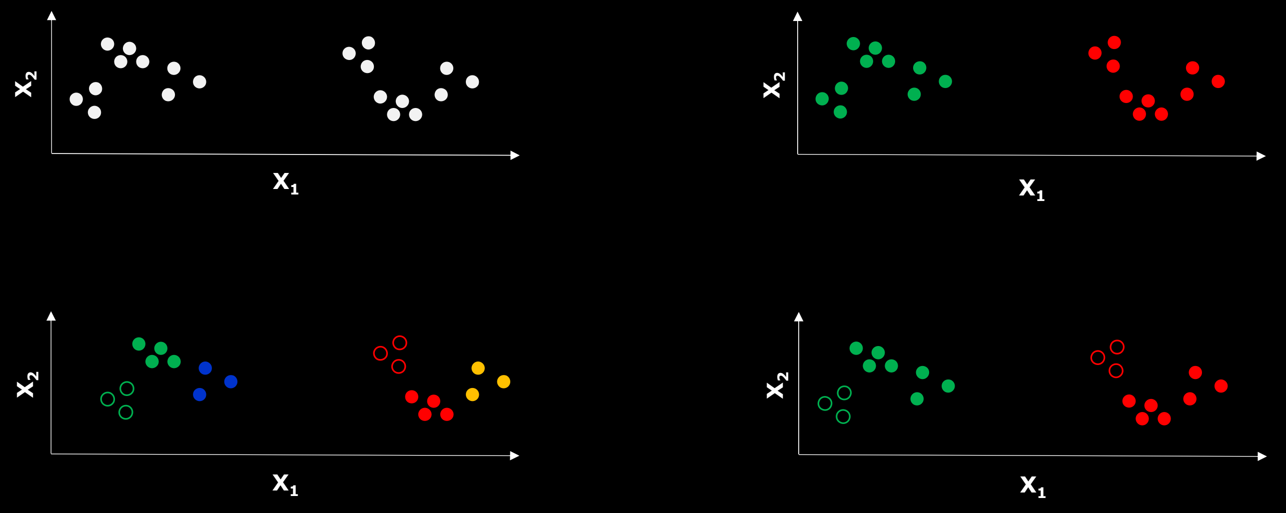
  Reduce the size of large data sets.







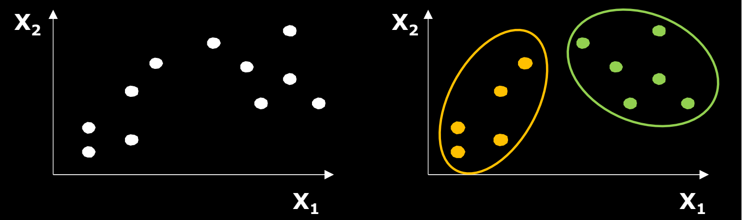
WHAT DO WE MEAN BY CLUSTER?



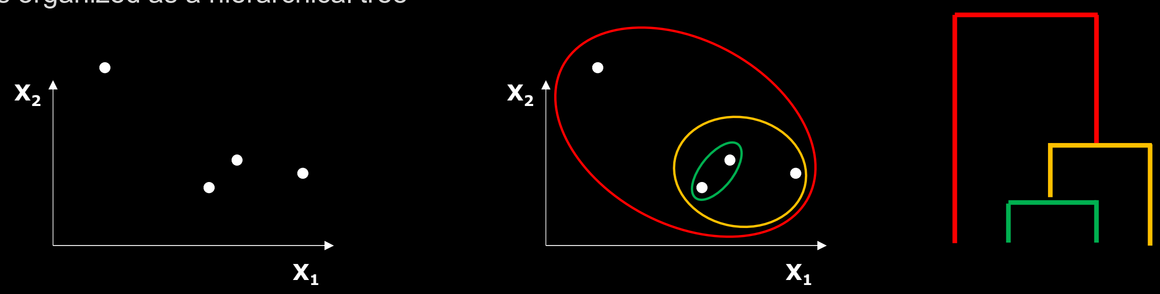
The definition of cluster is imprecise and the best definition depends on the nature of data and the desired results.

TYPES OF CLUSTERING

- A **CLUSTERING** is a set of clusters
- Important distinction between **PARTITIONAL** and **HIERARCHICAL** sets of clusters
  - **PARTITIONAL CLUSTERING:** a division of data objects into non-overlapping subsets
  - **HIERARCHICAL CLUSTERING:** a set of nested clusters organized as a hierarchical tree



PARTITIONAL CLUSTERING



HIERARCHICAL CLUSTERING

OTHER DISTINCTIONS BETWEEN SETS OF CLUSTERS

- **EXCLUSIVE** versus **NON-EXCLUSIVE**
  - In non-exclusive clusterings, points may belong to multiple clusters
  - Can belong to multiple classes or could be ‘**BORDER**’ points
- **FUZZY CLUSTERING** (ONE TYPE OF NON-EXCLUSIVE)
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- **PARTIAL** versus **COMPLETE**
  - In some cases, we only want to cluster some of the data
  - Data can contain **OUTLIERS** or **ANOMALOUS OBSERVATIONS**

TYPES OF CLUSTERS

- Well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function

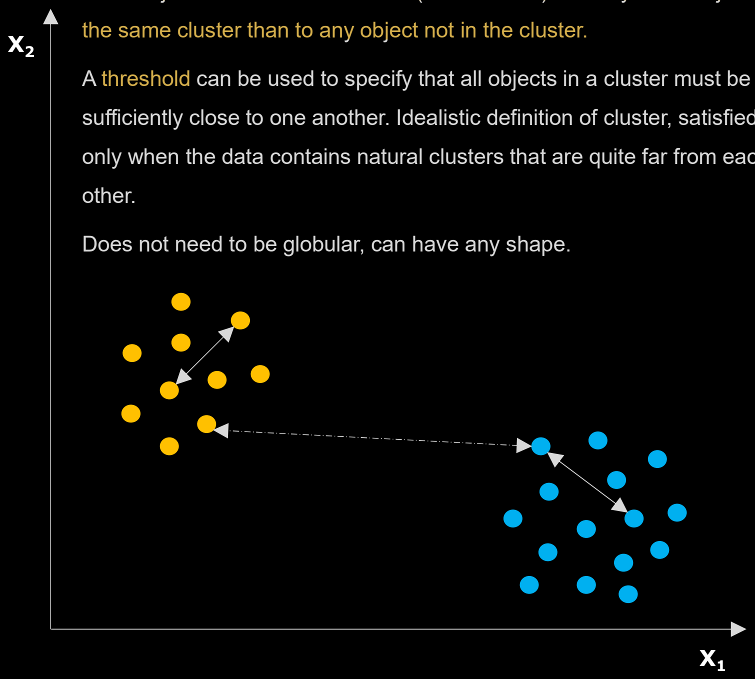
TYPES OF CLUSTERS

- Well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function

Each object in the cluster is closer (more similar) to every other object in the same cluster than to any object not in the cluster.

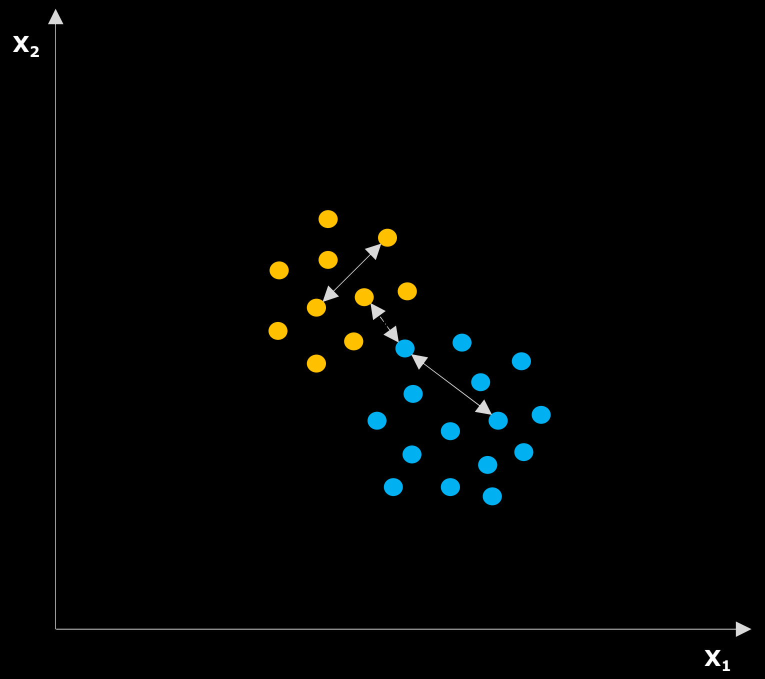
A threshold can be used to specify that all objects in a cluster must be sufficiently close to one another. Idealistic definition of cluster, satisfied only when the data contains natural clusters that are quite far from each other.

Does not need to be globular, can have any shape.



TYPES OF CLUSTERS

- Not well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function



TYPES OF CLUSTERS

- Well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function

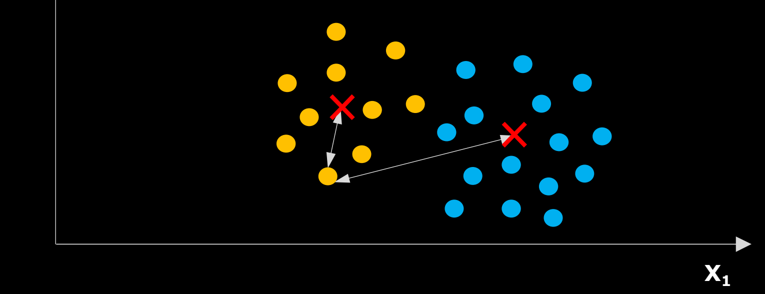
Each object is closer (more similar) to the **PROTOTYPE** that defines the cluster than to the prototype of any other cluster.

For continuous attributes, the prototype of a cluster is often the **CENTROID**, i.e. the average of all the objects in the cluster.

When the centroid is not representative (categorical attributes) then the **MEDOID** (most representative object of the cluster) is used.

Tends to be globular.

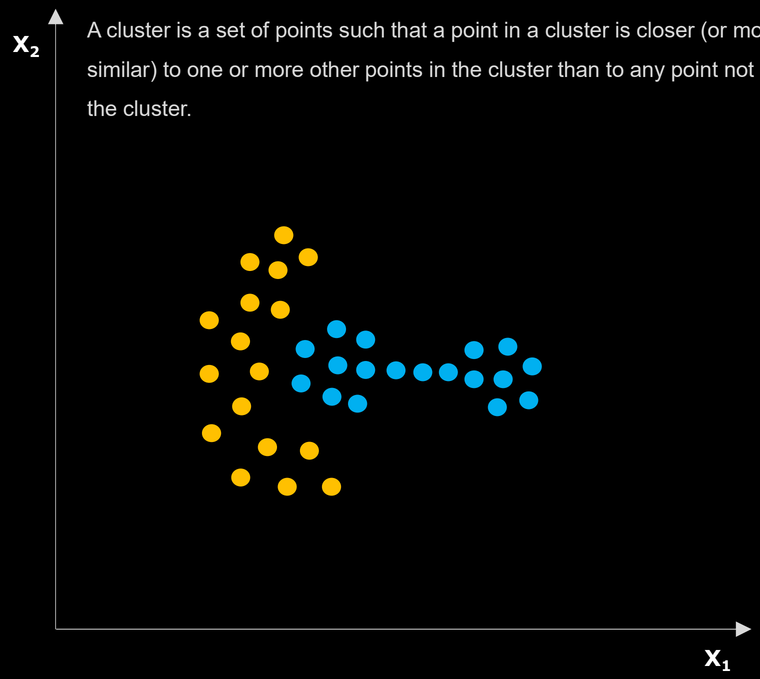
✕ **PROTOTYPE**



TYPES OF CLUSTERS

- Well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function

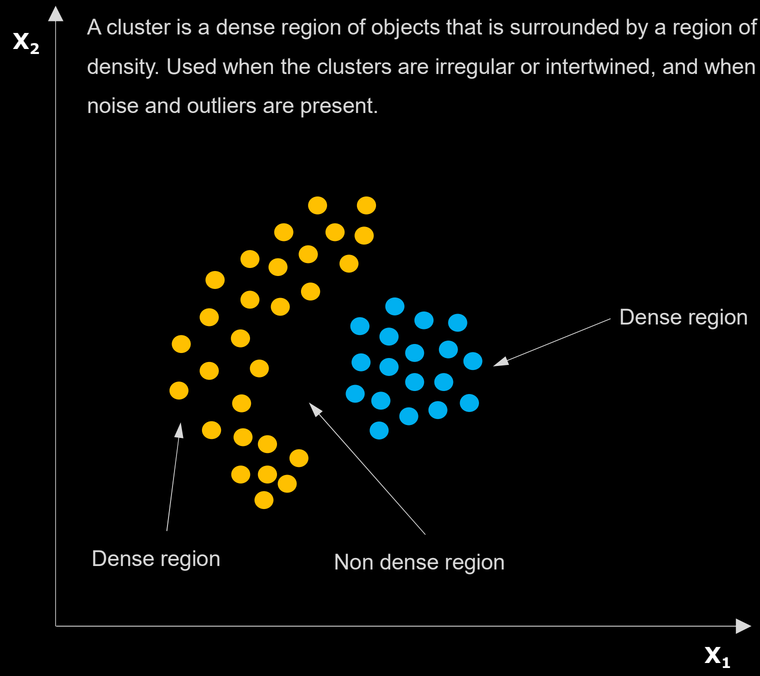
A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



TYPES OF CLUSTERS

- Well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function

A cluster is a dense region of objects that is surrounded by a region of low density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.



TYPES OF CLUSTERS

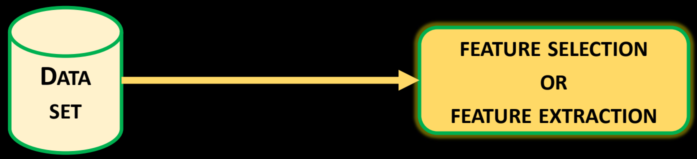
- Well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function

- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function. (NP Hard).
- Can have global or local objectives.
  - Hierarchical clustering algorithms typically have local objectives
  - Partitional algorithms typically have global objectives
- A variation of the global objective function approach is to fit the data to a parameterized model.
  - Parameters for the model are determined from the data.
  - Mixture models assume that the data is a ‘mixture' of a number of statistical distributions.

CHARACTERISTICS OF THE INPUT DATA ARE IMPORTANT

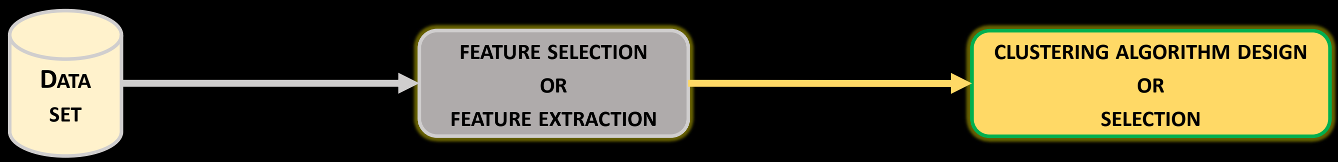
- Type of PROXIMITY or DENSITY MEASURE
  - Central to clustering
  - Depends on data and application
- Data characteristics that affect proximity and/or density are
  - DIMENSIONALITY
    - Sparseness
  - ATTRIBUTE TYPE
  - SPECIAL RELATIONSHIPS IN THE DATA
    - For example, autocorrelation
  - DISTRIBUTION OF THE DATA
- NOISE AND OUTLIERS
  - Often interfere with the operation of the clustering algorithm
- Clusters of DIFFERING SIZES, DENSITIES, and SHAPES

COMPONENTS OF CLUSTER ANALYSIS



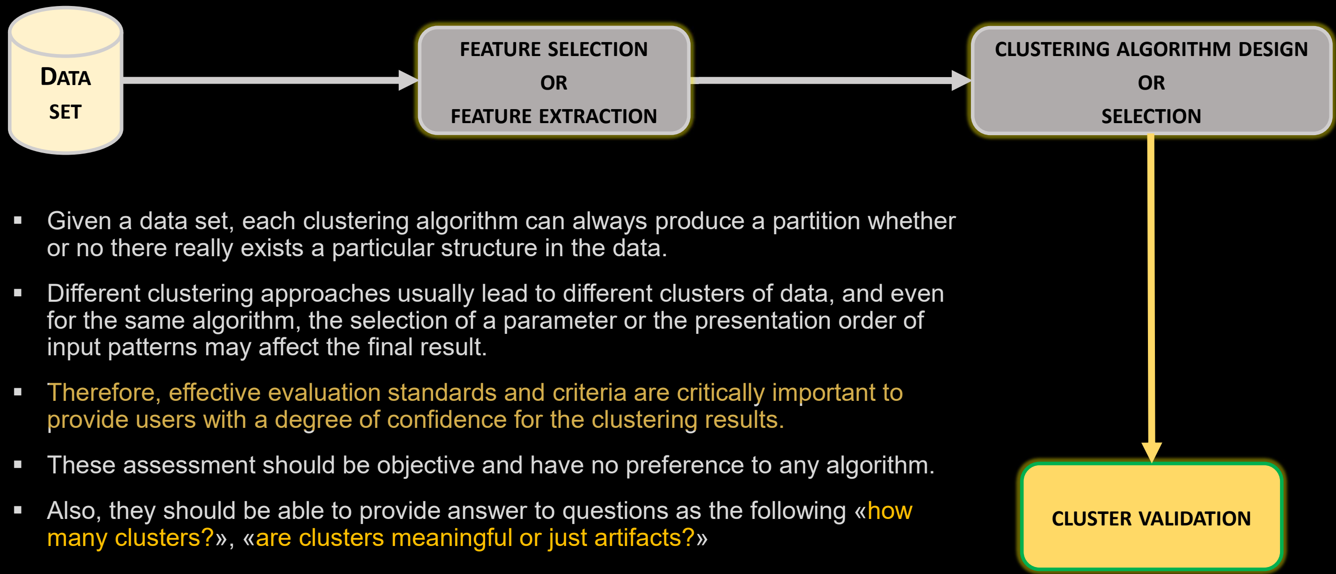
- **FEATURE SELECTION** assures the retention of the meaning of the original attributes.
- **FEATURE EXTRACTION** is capable of producing features that could be of better use in uncovering the data structure. However, it may generate features that are difficult to interpret.
- Ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, and easy to obtain and interpret.

COMPONENTS OF CLUSTER ANALYSIS



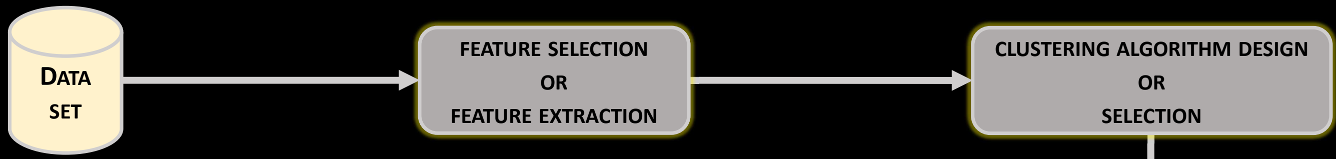
- Determining the **PROXIMITY** measure and constructing a **CRITERION FUNCTION**.
- Once a **PROXIMITY MEASURE** is determined, clustering can be formulated as an **OPTIMIZATION PROBLEM** with a specific **OBJECTIVE FUNCTION**.

COMPONENTS OF CLUSTER ANALYSIS



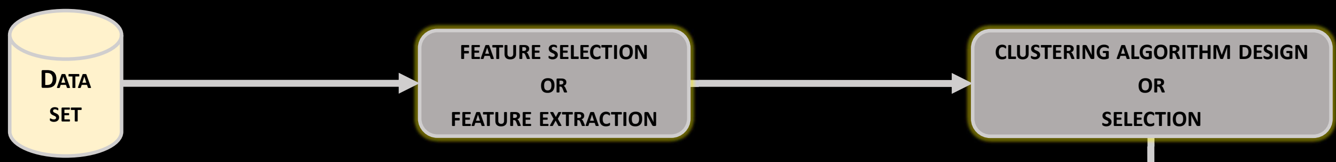
- Given a data set, each clustering algorithm can always produce a partition whether or no there really exists a particular structure in the data.
- Different clustering approaches usually lead to different clusters of data, and even for the same algorithm, the selection of a parameter or the presentation order of input patterns may affect the final result.
- Therefore, effective evaluation standards and criteria are critically important to provide users with a degree of confidence for the clustering results.
- These assessment should be objective and have no preference to any algorithm.
- Also, they should be able to provide answer to questions as the following «how many clusters?», «are clusters meaningful or just artifacts?»

COMPONENTS OF CLUSTER ANALYSIS

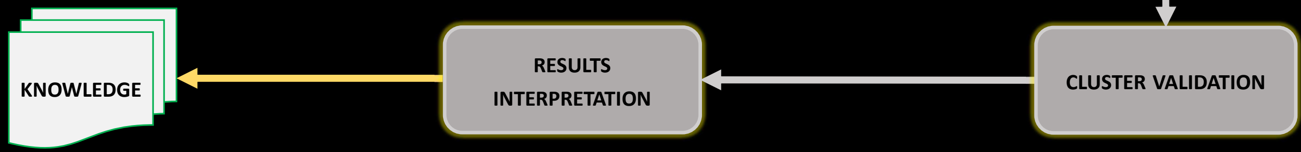


- The ultimate goal of clustering is to provide users with meaningful insights from the original data so that they can develop a clear understanding of the data and therefore effectively solve the problem encountered.
- We can say that «cluster analysis is a device for suggesting hypotheses».
- A set of clusters is not itself a finished result but only a possible outline.

COMPONENTS OF CLUSTER ANALYSIS



- Experts in the relevant fields are encouraged to interpret the data partition, integrating other experimental evidence and domain information without restricting their observations and analysis to any specific clustering results.
- Further analysis and experiments may be required.



RECAP

- CLUSTER ANALYSIS
- UNDERSTANDING AND SUMMARIZING
- TYPES OF CLUSTERING
- TYPES OF CLUSTERS
- COMPONENTS OF CLUSTER ANALYSIS