

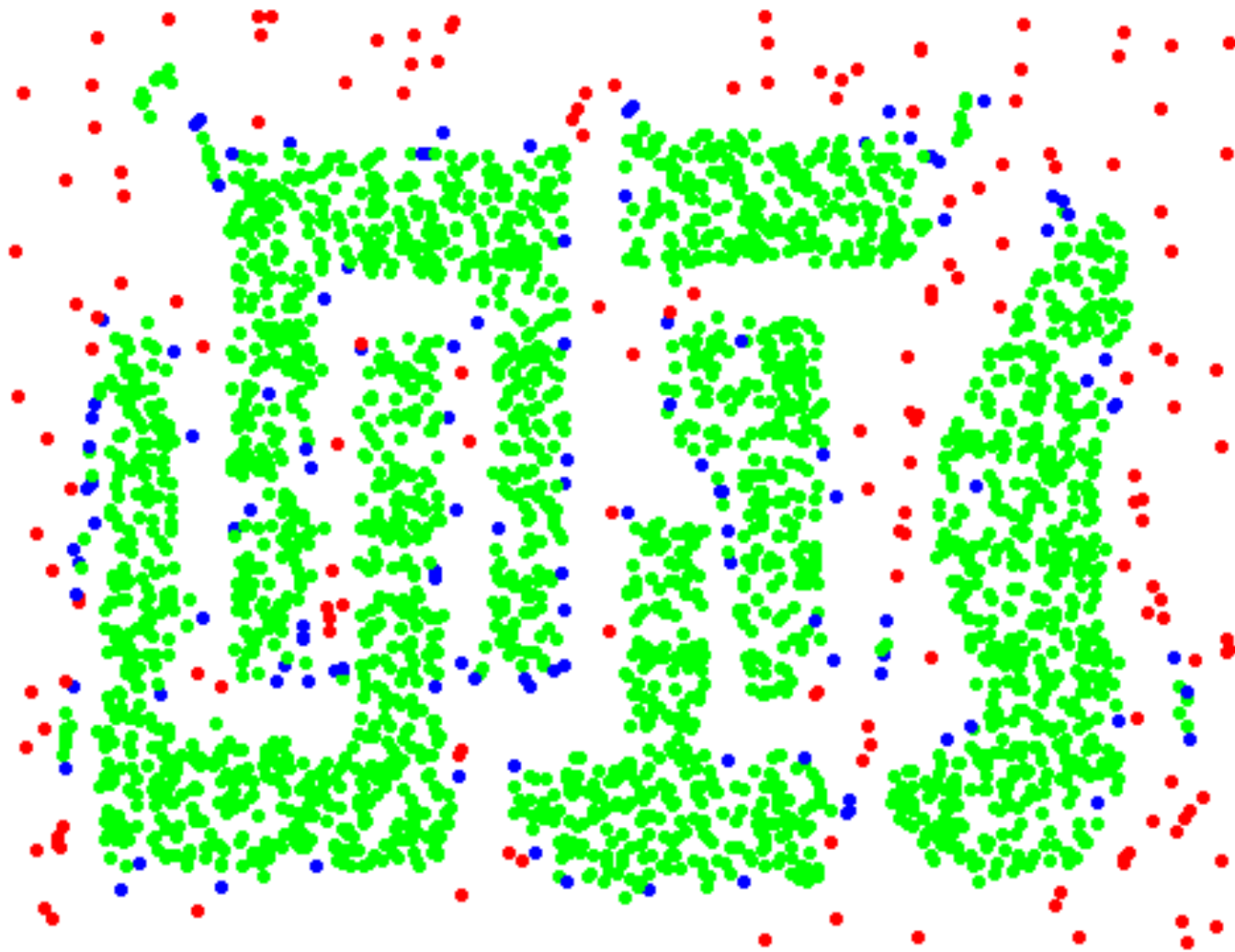


Clustering: Other Methods

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

Density-Based Clustering

THE
LIFE
OF
THE
LIFE



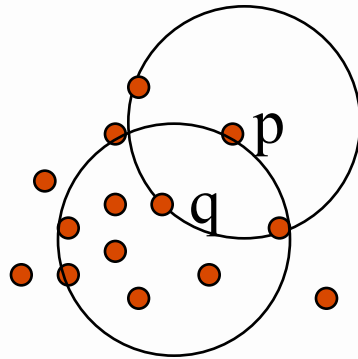
- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
 - Discover clusters of arbitrary shape
 - Handle noise
 - One scan
 - Need density parameters as termination condition
- Several interesting studies:
 - DBSCAN: Ester, et al. (KDD'96)
 - OPTICS: Ankerst, et al (SIGMOD'99).
 - DENCLUE: Hinneburg & D. Keim (KDD'98)
 - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

- The neighborhood within a radius ϵ of a given object is called the ϵ -neighborhood of the object
- If the ϵ -neighborhood of an object contains at least MinPts objects, then the object is a core object
- An object p is directly density-reachable from object q if p is within the ϵ -neighborhood of q and q is a core object
- An object p is density-reachable from object q if there is a chain of object p_1, \dots, p_n where $p_1 = p$ and $p_n = q$ such that $p_i + 1$ is directly density reachable from p_i
- An object p is density-connected to q with respect to ϵ and MinPts if there is an object o such that both p and q are density reachable from o

- Density = number of points within a specified radius (Eps)
- A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
- A noise point is any point that is not a core point or a border point
- A density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability

Density-Reachable & Density-Connected

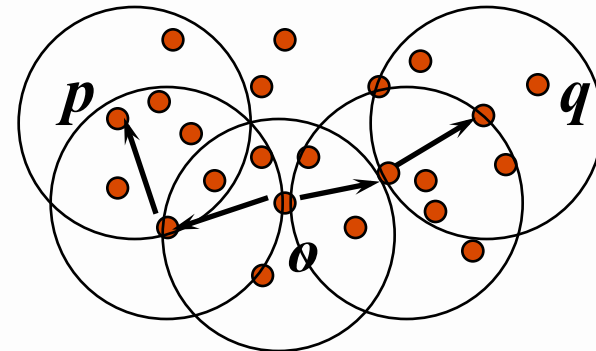
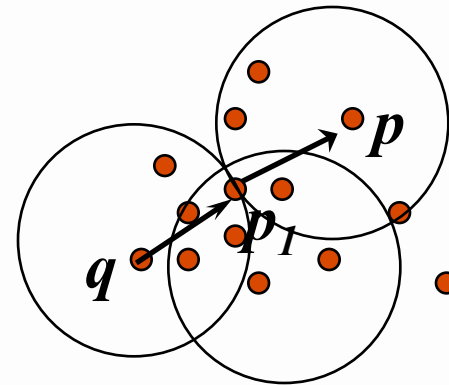
- Directly density-reachable



MinPts = 5

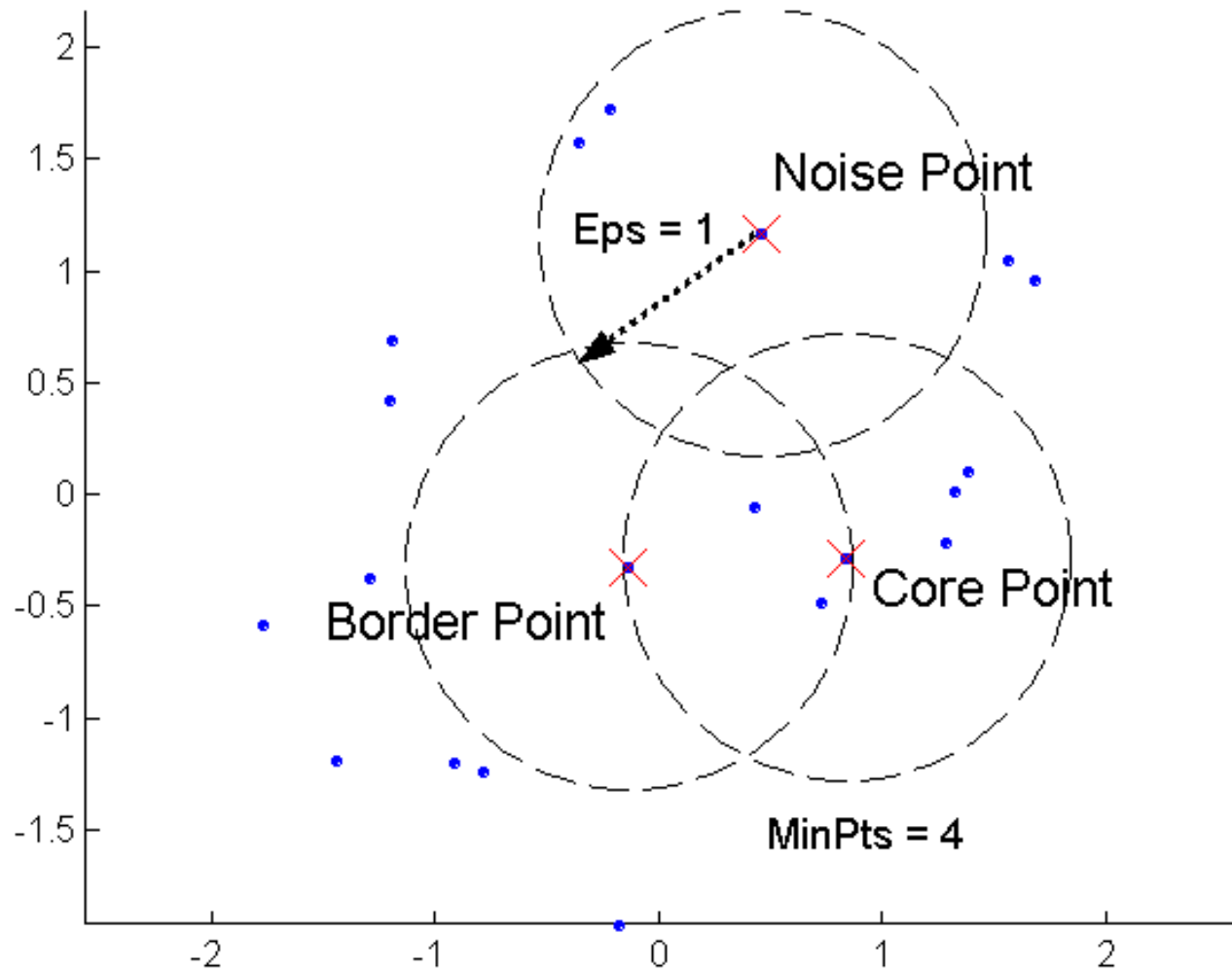
Eps = 1 cm

- Density-reachable



- Density-connected

DBSCAN: Core, Border, and Noise Points



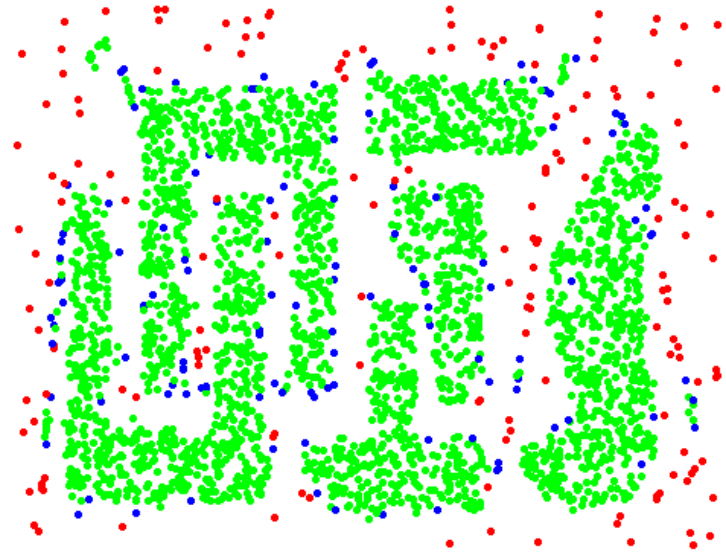
Density Based Spatial Clustering

- Relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise
- The Algorithm
 - Arbitrary select a point p
 - Retrieve all points density-reachable from p given Eps and $MinPts$.
 - If p is a core point, a cluster is formed.
 - If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database
 - Continue the process until all of the points have been processed

DBSCAN: Core, Border and Noise Points



Original Points

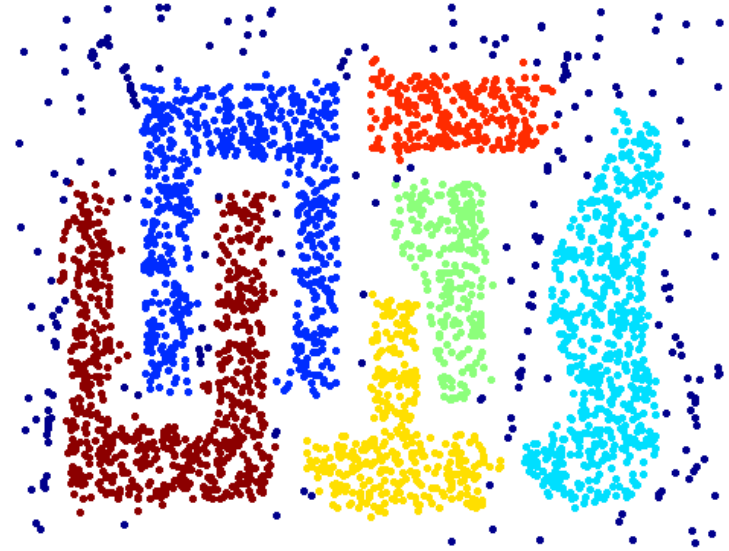


Point types: core,
border and noise

Eps = 10, MinPts = 4

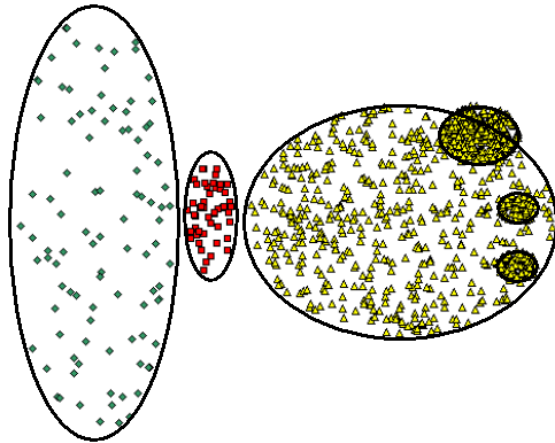


Original Points



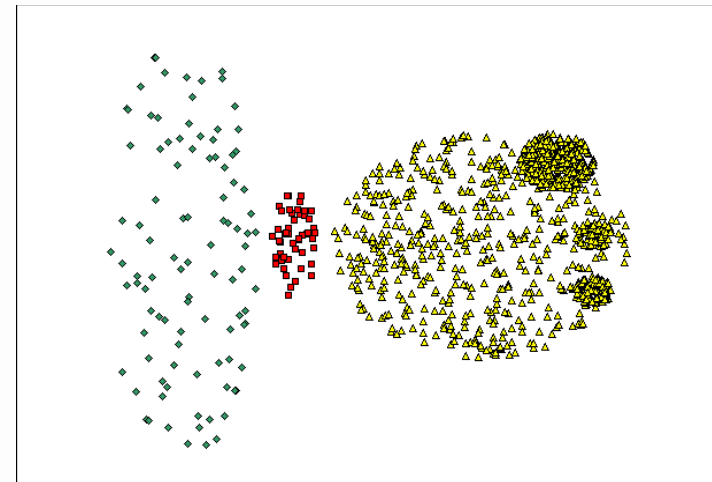
Clusters

- Varying densities

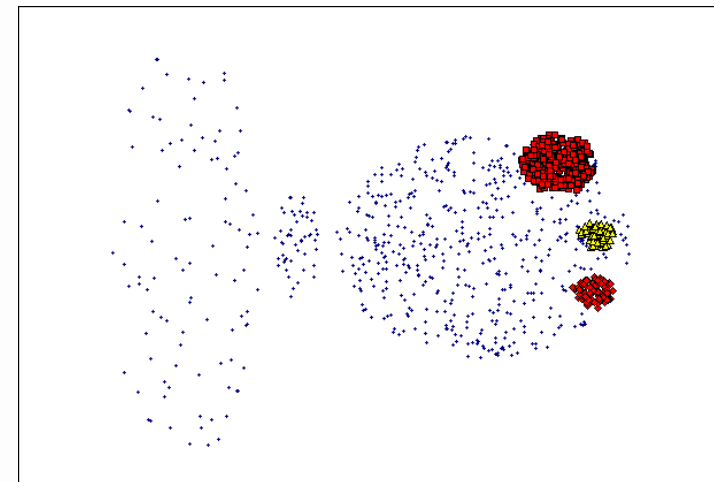


ata

Original Points

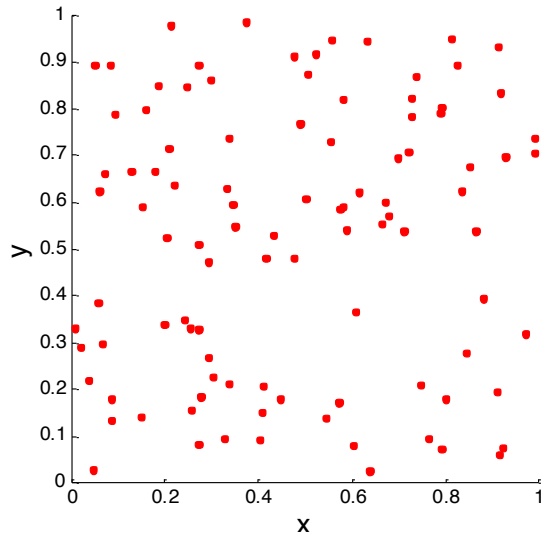


$(\text{MinPts}=4, \text{Eps}=9.75)$.

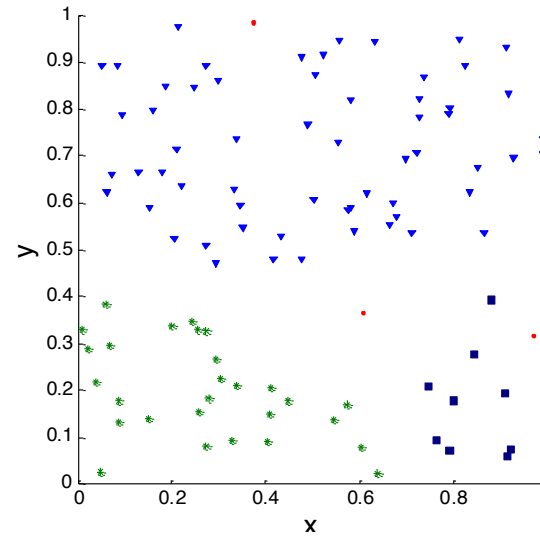


$(\text{MinPts}=4, \text{Eps}=9.92)$

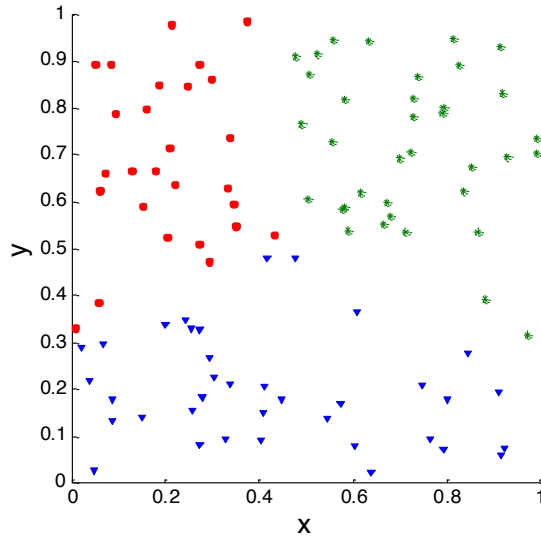
Random
Points



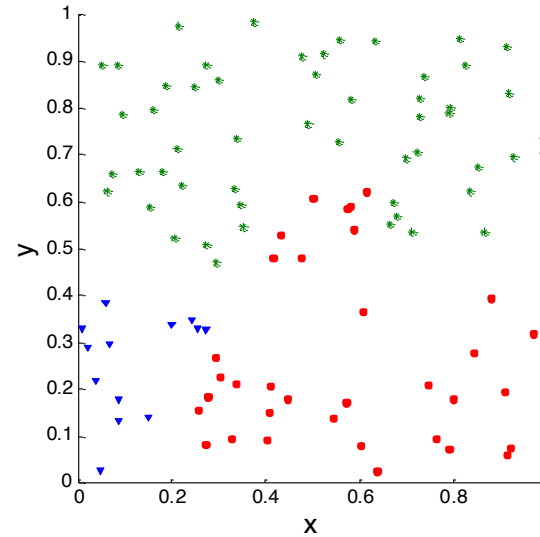
DBSCAN



K-means



Complete
Link



Model-Based Clustering

- Based on the assumption that data are generated by a mixture of underlying probability distribution
- Attempt to optimize the fit between the data and some mathematical model
- Typical methods
 - Statistical approach
EM (Expectation maximization), AutoClass
 - Machine learning approach
COBWEB, CLASSIT
 - Neural network approach
SOM (Self-Organizing Feature Map)

Expectation Maximization (EM) Clustering

- Extension to k-means
 - Assign each object to a cluster according to a weight (prob. distribution)
 - New means are computed based on weighted measures
- General idea
 - Starts with an initial estimate of the parameter vector
 - Iteratively rescores the patterns against the mixture density produced by the parameter vector
 - The rescored patterns are used to update the parameter updates
 - Patterns belonging to the same cluster, if they are placed by their scores in a particular component
- Algorithm converges fast but may not be in global optima

The EM (Expectation Maximization) Algorithm

- Initially, randomly assign k cluster centers
- Iteratively refine the clusters based on two steps
- Expectation step: assign each data point X_i to cluster C_i with the following probability

$$P(X_i \in C_k) = p(C_k|X_i) = \frac{p(C_k)p(X_i|C_k)}{p(X_i)},$$

where $p(x_i|C_k) = N(m_k, \Sigma_k(x_i))$ follows the normal distribution.

- This step calculates the probability of cluster membership of x_i for each C_k
- Maximization step: Estimation of model parameters

$$m_k = \frac{1}{N} \sum_{i=1}^N \frac{X_i P(X_i \in C_k)}{\sum_j P(X_i \in C_j)}.$$

Outlier Discovery

- What are outliers?
 - The set of objects are considerably dissimilar from the remainder of the data
- Problem: Define and find outliers in large data sets
- Applications
 - Credit card fraud detection
 - Telecom fraud detection
 - Customer segmentation
 - Medical analysis

Outlier Discovery: Statistical Approaches

- Assume a model underlying distribution that generates data set (e.g. normal distribution)
- Use discordancy tests depending on
 - Data distribution
 - Distribution parameter (e.g., mean, variance)
 - Number of expected outliers
- Drawbacks
 - most tests are for single attribute
 - In many cases, data distribution may not be known

Outlier Discovery: Distance-Based Approach

- Introduced to counter the main limitations imposed by statistical methods
- We need multi-dimensional analysis without knowing data distribution
- Distance-based outlier: A $DB(p, D)$ -outlier is an object O in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than D from O
- Algorithms for mining distance-based outliers
 - Index-based algorithm
 - Nested-loop algorithm
 - Cell-based algorithm

Summing Up

- For supervised classification we have a variety of measures to evaluate how good our model is
- 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

- 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 (SSE)
 - Relative Index: Used to compare two different clusterings or clusters (Often an external or internal index)
- 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.

- “The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

Algorithms for Clustering Data, Jain and Dubes

- Cluster analysis groups objects based on their similarity and has wide applications
- Measure of similarity can be computed for various types of data
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- Outlier detection and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches
- There are still lots of research issues on cluster analysis