# Data Mining: Cluster Analysis: Basic Concepts and Methods

## What is Cluster Analysis?

- Cluster: A collection of data objects.
  - Similar (or related) to one another within the same group.
  - Dissimilar (or unrelated) to the objects in other groups.
- Cluster Analysis (or clustering, data segmentation, etc.):
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.
- Unsupervised learning: no predefined classes (i.e., *learning by observations* vs. *learning by examples*: supervised).
- Typical applications:
  - As a stand-alone tool to get insight into data distribution.
  - As a preprocessing step for other algorithms.

## Applications of Cluster Analysis

- Data reduction:
  - Summarization: Preprocessing for regression, PCA, classification, and association analysis.
  - Compression: Image processing: vector quantization.
- Hypothesis generation and testing.
- Prediction based on groups.
  - Cluster & find characteristics/patterns for each group.
- Finding K-nearest Neighbors.
  - Localizing search to one or a small number of clusters.
- Outlier detection: Outliers are often viewed as those “far away” from any cluster.

## Clustering: Application Examples

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus, and species.
- Information retrieval: document clustering.
- Land use: Identification of areas of similar land use in an earth observation database.
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- City-planning: Identifying groups of houses according to their house type, value, and geographical location.
- Earth-quake studies: Observed earthquake epicenters should be clustered along continent faults.
- Climate: understanding earth climate, find patterns of atmospheric and ocean.
- Economic Science: market research.

## Basic Steps to Develop a Cluster Task

- Feature Selection:
  - Select info concerning the task of interest.
  - Minimal information redundancy.
- Proximity measure:
  - Similarity of two feature vectors.
- Clustering criterion:
  - Expressed via a cost function or some rules.
- Clustering algorithms:
  - Choice of algorithms.
- Validation of the results:
  - Validation test (also, clustering tendency test).
- Interpretation of the results:
  - Integration with applications.

## Quality: What is Good Clustering?

- A good clustering method will produce high quality clusters:
  - High intra-class similarity: cohesive within clusters.
  - Low inter-class similarity: distinctive between clusters.
- The quality of a clustering method depends on:
  - The similarity measure used by the method.
  - Its implementation, and
  - Its ability to discover some or all the hidden patterns.

## Measure the Quality of Clustering

- Dissimilarity/Similarity metric:
  - Similarity is expressed in terms of a distance function, typically metric: d(i, j).
  - The definitions of distance functions are usually rather different for interval-scaled, Boolean, categorical, ordinal ratio, and vector variables.
  - Weights should be associated with different variables based on applications and data semantics.
- Quality of clustering:
  - There is usually a separate “quality” function that measures the “goodness” of a cluster.
  - It is hard to define “similar enough” or “good enough”. The answer is typically highly subjective.

## Considerations for Cluster Analysis

- Partitioning criteria:
  - Single level.
  - Hierarchical partitioning (often, multi-level hierarchical partitioning is desirable).
- Separation of clusters:
  - Exclusive (e.g., one customer belongs to only one region).
  - Non-exclusive (e.g., one document may belong to more than one class).
- Similarity measure:
  - Distance-based (Euclidian, road network, vector).
  - Connectivity-based (density, contiguity).
- Clustering space:
  - Full space (often when low dimensional).
  - Subspaces (often in high-dimensional clustering).

## Requirements and Challenges

- Scalability:
  - Clustering all the data instead of only on samples.
- Ability to deal with different types of attributes:
  - Numerical, binary, categorical, ordinal, linked, and mixture of these.
- Constraint-based clustering:
  - User may give inputs on constraints.
  - Use domain knowledge to determine input parameters.
- Interpretability and usability.
- Others:
  - Discovery of clusters with arbitrary shape.
  - Ability to deal with noisy data.
  - Incremental clustering and insensitivity to input order.
  - High dimensionality.

## Major Clustering Approaches

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion.
  - E.g., minimizing the sum of square errors.
  - Typical methods: k-means, k-medoids, CLARANS.
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion.
  - Typical methods: Diana, Agnes, BIRCH, CAMELEON.
- Density-based approach:
  - Based on connectivity and density functions.
  - Typical methods: DBSACN, OPTICS, Den Clue.
- Grid-based approach:
  - Based on a multiple-level granularity structure.
  - Typical methods: STING, Wave Cluster, CLIQUE.
- Model-based:
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other.
  - Typical methods: EM, SOM, COBWEB.
- Frequent pattern-based:
  - Based on the analysis of frequent patterns.
  - Typical methods: p-Cluster.
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints.
  - Typical methods: COD (obstacles), constrained clustering.
- Link-based clustering:
  - Objects are often linked together in various ways.
  - Massive links can be used to cluster objects: SimRank, LinkClus.

## Partitioning Algorithms: Basic Concept

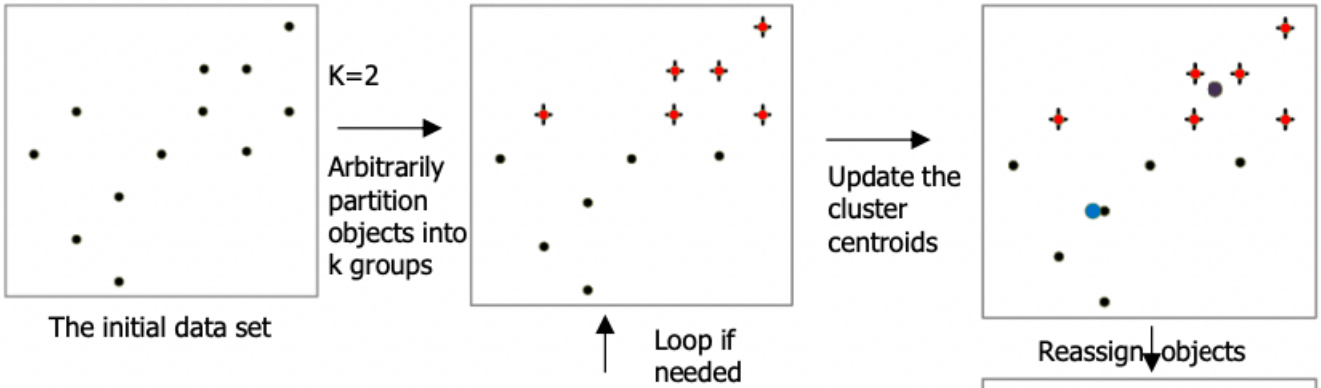
- Partitioning method: Partitioning a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized (where $c_i$ is the centroid of medoid of cluster $C_i$).

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (d(p, c_i))^2$$

- Given k, find a partition of k clusters that optimized the chosen partitioning criterion.
  - Global optimal: exhaustively enumerate all partitions.
  - Heuristic methods: k-means and k-medoids algorithms.
  - K-means (MacQueen'67, Klogd'57/'82): Each cluster is represented by the center of the cluster.
  - K-medoids or PAM (Partition Around Medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster.

## The K-Means Clustering Method.

- Given k, the k-means algorithm is implemented in four steps:
  - Partition objects into k non-empty subsets.
  - Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., mean point of the cluster).
  - Assign each object to the cluster with the nearest seed point.
  - Go back to step 2, stop when the assignment does not change.

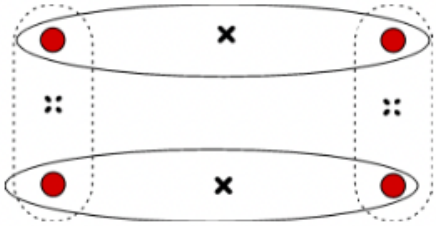



- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

## Comments on the K-means Method

- Strength: Efficient O(t*k*n), where n is number of objects, k is number of clusters, and t is number of iterations. Normally k, t ≪ n.
    - PAM: $O(k(n-k)^2)$
    - CLARA: $O(ks^2 + k(n-k))$
- Comment: Often terminates at a local optimal.
- Weakness:
    - Applicable only to objects in a continuous n-dimensional space.
        - Using the k-modes method for categorical data.
        - In comparison, k-medoids can be applied to a wide range of data.
    - Need to specify $k_i$ the number of clusters, in advance (there are ways to automatically determine the best k (see Hastie et al., 2009)).
    - Sensitive to noisy data and outliers.
    - Not suitable to discover clusters with non-convex shapes.

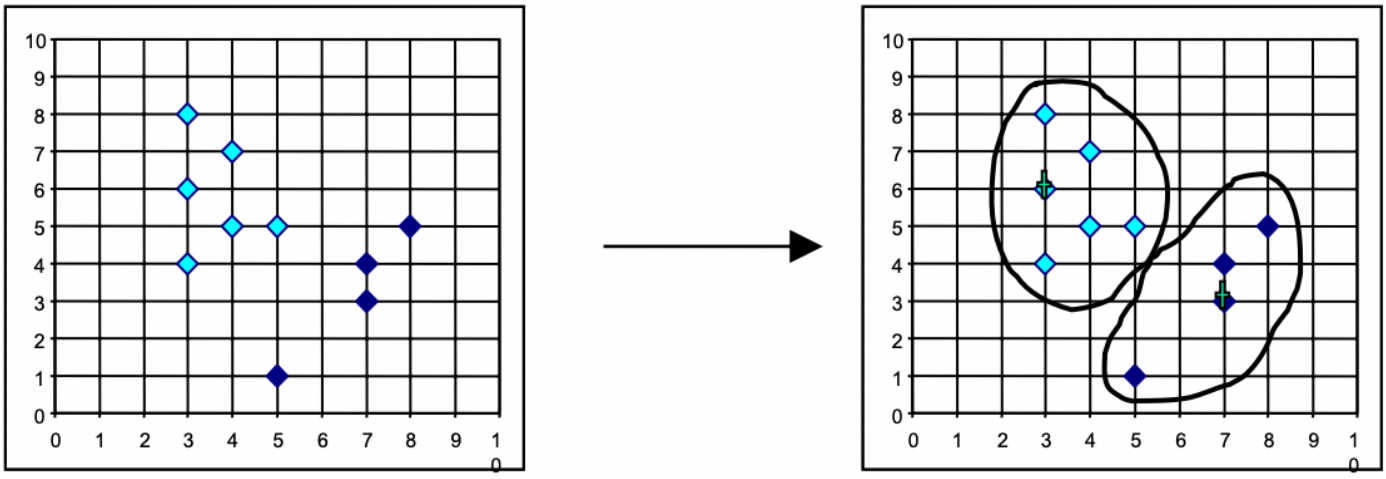
## Variations of the K-Means Method



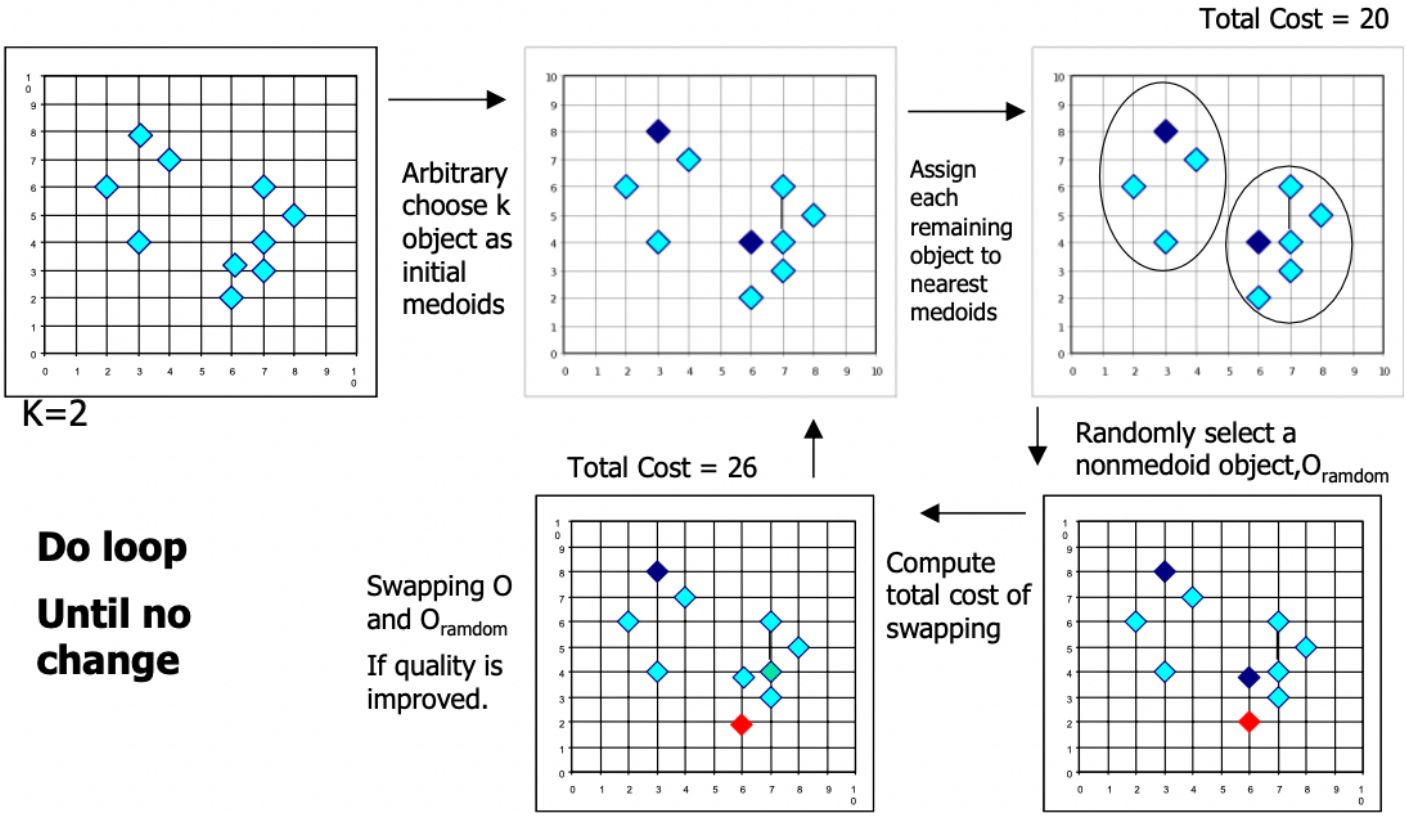
- Most of the variants of the k-means which differ in:
    - Selection of the initial k means.
    - Dissimilarity calculations.
    - Strategies to calculate cluster means.
- Handling categorical data: k-modes.
    - Replacing means of clusters with modes.
    - Using new dissimilarity measures to deal with categorical objects.
    - Using a frequency-based method to update modes of clusters.
    - A mixture of categorical and numerical data: k-prototype method.

## What is the Problem of the K-Means Method

- The k-means algorithm is sensitive to outliers.
    - Since an object with an extremely large value may substantially distort the distribution of the data.
- K-Medoids: Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.

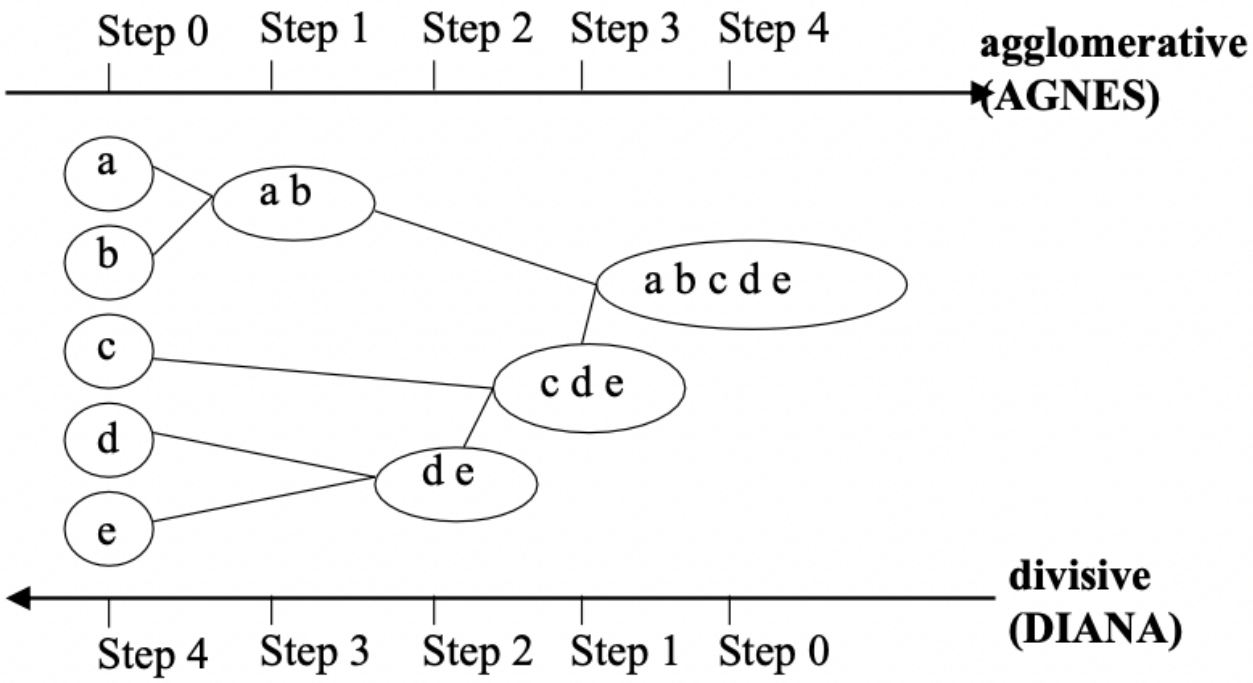


## PAM: A Typical K-Medoids Algorithm



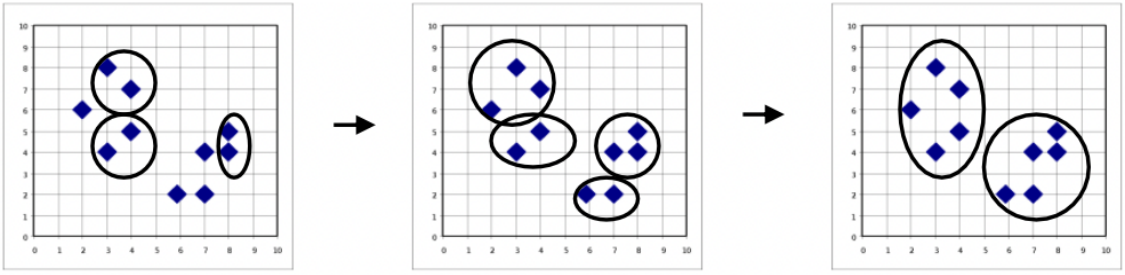

## Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters k as an input but needs a termination condition.

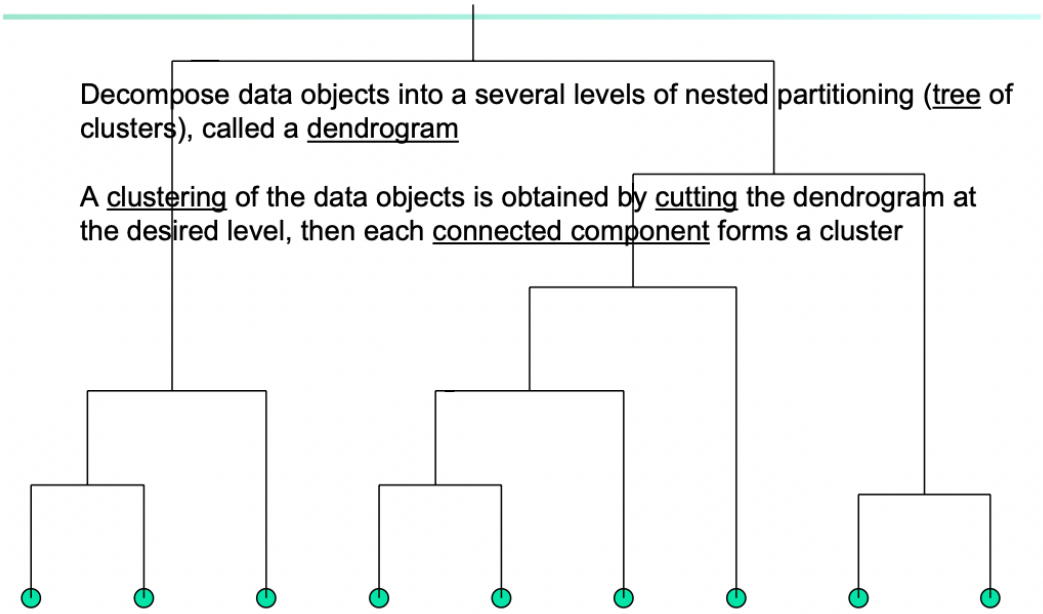


## AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseau (1990).
- Implemented in statistical packages, e.g., Splus.
- Use the single-link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity.
- Go on in a non-descending fashion.
- Eventually all nodes belong to the same cluster.

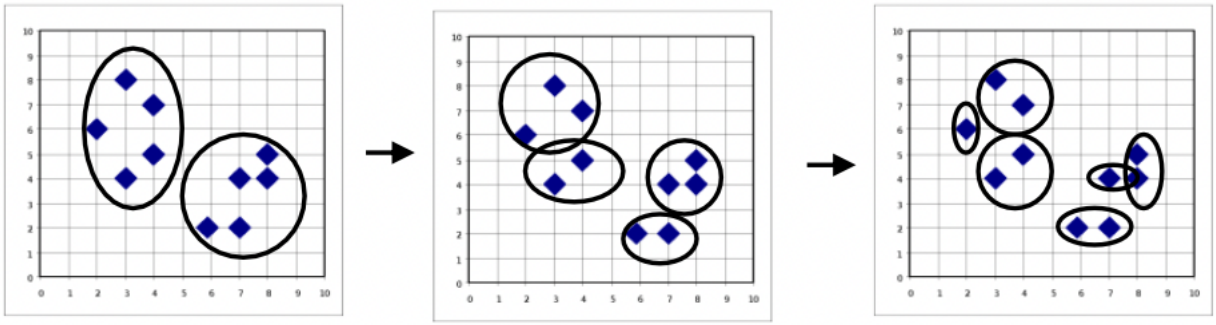


## Dendrogram: Shows how Clusters Are Merged




## DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseau (1990).
- Implemented in statistical analysis package, e.g., Splus.
- Inverse order of AGNES.
- Eventually each node forms a cluster on its own.



## Distance Between Clusters

- Single link: smallest distance between an element in one cluster and an element in the other:
  - dist($K_i$, $K_j$) = min($t_{ip}$, $t_{jq}$)
- Complete link: largest distance between an element in one cluster and an element in the other:
  - dist($K_i$, $K_j$) = max($t_{ip}$, $t_{jq}$)
- Average: avg distance between an element in one cluster and an element in the other:
  - dist($K_i$, $K_j$) = avg($t_{ip}$, $t_{jq}$)
- Centroid: distance between the centroids of 2 clusters:
  - dist($K_i$, $K_j$) = dist($C_i$, $C_j$)
- Medoid: distance between the medoids of two clusters:
  - dist($K_i$, $K_j$) = dist($M_i$, $M_j$)
  - Medoid: a chosen, centrally located object in the cluster.

## Centroid, Radius, and Diameter of a Cluster (for numerical data sets)

- Centroid: the "middle" of a cluster.

$$C_m = \frac{\sum_{i=1}^{N}(t_{ip})}{N}$$

- Radius: Square root of average distance from any point of the cluster to its centroid.

$$R_m = \sqrt{\frac{\sum_{i=1}^{N}(t_{ip} - c_m)^2}{N}}$$

- Diameter: Square root of average mean squared distance between all pairs of points in the cluster.

$$D_m = \sqrt{\frac{\sum_{i=1}^{N}\sum_{i=1}^{N}(t_{ip} - t_{iq})^2}{N(N-1)}}$$

## Density-Based Clustering Methods

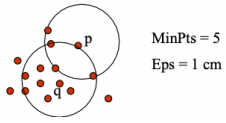
- Clustering based on density (local cluster criterion), such as density-connected points.
- Major features:
  - Discover clusters or arbitrary shape.
  - Handle noise.
  - One scan.
  - Need density parameters as termination condition.
- Several interesting studies:

- 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)

## Density-Based Clustering: Basic Concepts

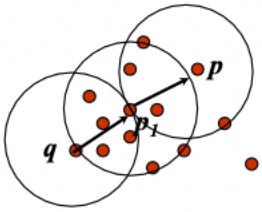
- 2 parameters:
  - Eps: Maximum radius of the neighborhood.
  - MinPts: Minimum number of points in an Eps-neighborhood of that point.
- $N_{Eps}(q)$: {p belongs to D | dist(p, q) ≤ Eps}
- Directed density-reachable: A point p is directly density-reachable from a point q w.r.t Eps, MinPts if:
  - p belongs to $N_{eps}(q)$
  - core point condition:

$$|N_{Eps}(q)| \geq MinPts$$

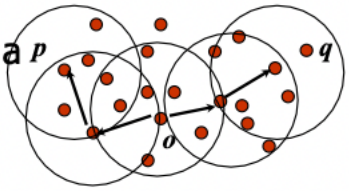



## Density-Reachable and Density-Connected

- Density-reachable:
  - A point *p* is density-reachable from a point *q* w.r.t. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n, p_1 = q, p_n = p$ such that $p_{i+1}$ is directly density-reachable from $p_i$

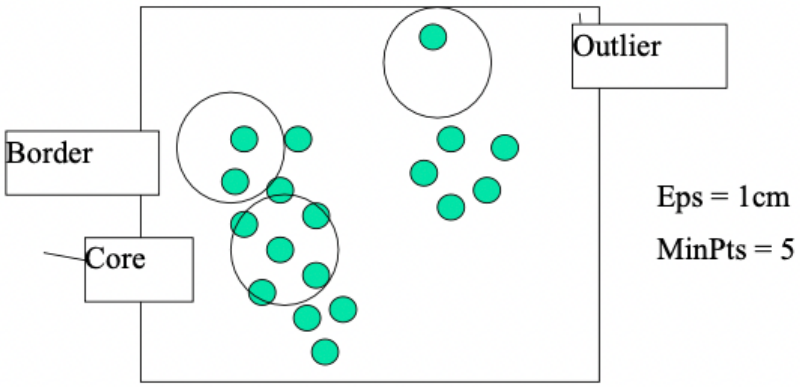



- Density-connected
  - A point *p* is density-connected to a point *q* w.r.t. *Eps*, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* w.r.t. *Eps* and *MinPts*




## DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- 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



## DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

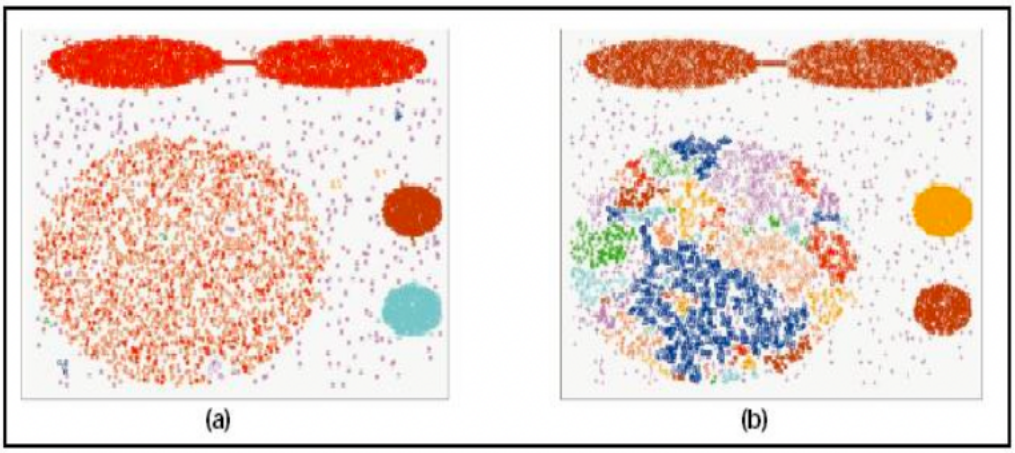
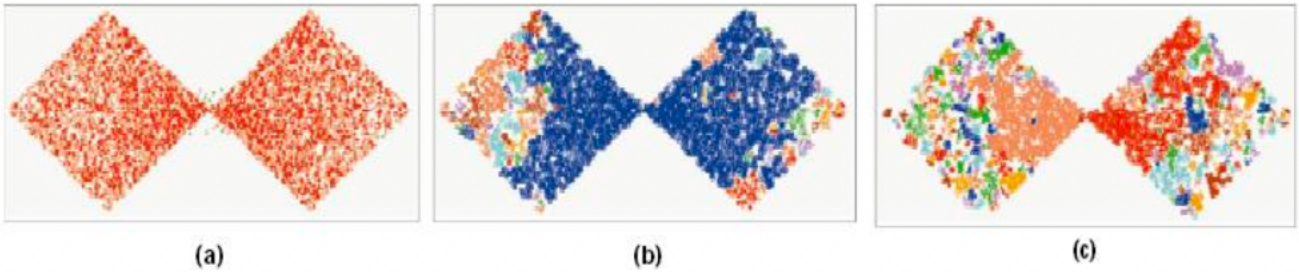



Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.




## OPTICS: A Cluster-Ordering Method (1999)

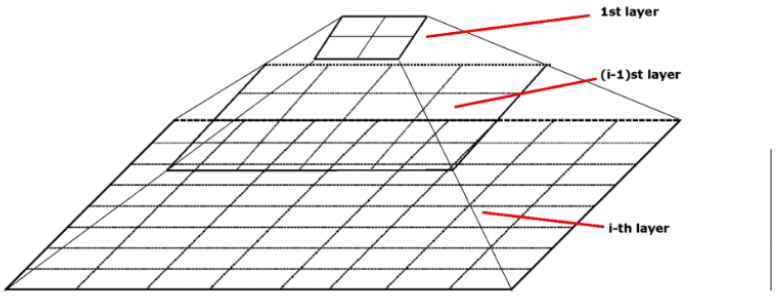
- OPTICS: Ordering Points To Identify the Clustering Structure
  - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
  - Produces a special order of the database wrt its density-based clustering structure
  - This cluster-ordering contains info equiv to the density-based clusterings corresponding to a broad range of parameter settings
  - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
  - Can be represented graphically or using visualization techniques

## Grid-Based Clustering Method

- Using multi-resolution grid data structure
- Several interesting methods
  - STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - CLIQUE: Agrawal, et al. (SIGMOD'98)
    - Both grid-based and subspace clustering
  - WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
    - A multi-resolution clustering approach using wavelet method

## STING: A Statistical Information Grid Approach

- Wang, Yang and Muntz (VLDB'97)
- The spatial area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution




## The STING Clustering Method

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
- Parameters of higher level cells can be easily calculated from parameters of lower level cell
  - *count*, *mean*, *s*, *min*, *max*
  - type of distribution—*normal*, *uniform*, etc.
- Use a top-down approach to answer spatial data queries
- Start from a pre-selected layer—typically with a small number of cells
- For each cell in the current level compute the confidence interval

## STING Algorithm and Its Analysis

- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level
- Repeat this process until the bottom layer is reached
- Advantages:
  - Query-independent, easy to parallelize, incremental update
  - *O(K),* where *K* is the number of grid cells at the lowest level
- Disadvantages:
  - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

## CLIQUE (Clustering In QUEst)

- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98)
- Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space
- CLIQUE can be considered as both density-based and grid-based
  - It partitions each dimension into the same number of equal length interval
  - It partitions an m-dimensional data space into non-overlapping rectangular units
  - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
  - A cluster is a maximal set of connected dense units within a subspace

## CLIQUE: The Major Steps

- Partition the data space and find the number of points that lie inside each cell of the partition.
- Identify the subspaces that contain clusters using the Apriori principle
- Identify clusters
  - Determine dense units in all subspaces of interests
  - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters
  - Determine maximal regions that cover a cluster of connected dense units for each cluster
  - Determination of minimal cover for each cluster

## Determine The Number Of Clusters

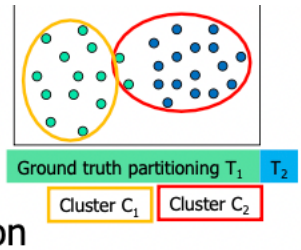
- Empirical method
  - # of clusters: k ≈√n/2 for a dataset of n points, e.g., n = 200, k = 10
- Elbow method
  - Use the turning point in the curve of sum of within cluster variance w.r.t the # of clusters
- Cross validation method
  - Divide a given data set into *m* parts
  - Use *m* – 1 parts to obtain a clustering model
  - Use the remaining part to test the quality of the clustering
    - E.g., For each point in the test set, find the closest centroid, and use the sum of squared distance between all points in the test set and the closest centroids to measure how well the model fits the test set
  - For any k > 0, repeat it *m* times, compare the overall quality measure w.r.t. different *k's*, and find # of clusters that fits the data the best

## Measuring Clustering Quality

- 3 kinds of measures: External, internal and relative
- External: supervised, employ criteria not inherent to the dataset
  - Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure
- Internal: unsupervised, criteria derived from data itself
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are, e.g., Silhouette coefficient
- Relative: directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

## Some Commonly Used External Measures

- Matching-based measures
  - Purity, maximum matching, F-measure
- Entropy-Based Measures
  - Conditional entropy, normalized mutual information (NMI), variation of information
- Pair-wise measures
  - Four possibilities: True positive (TP), FN, FP, TN
  - Jaccard coefficient, Rand statistic, Fowlkes-Mallow measure
- Correlation measures
  - Discretized Huber static, normalized discretized Huber static



## Table of Contents