# Cluster Analysis

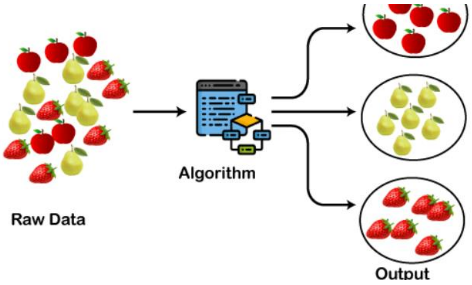
- Clustering in machine learning is a technique for **grouping unlabeled data**.
- It groups data points into clusters **based on similarities**.
- Clusters are groups of data points that are similar to each other and different from data points in other clusters.
- Clustering is an **unsupervised learning method**, meaning that **no labeled da**ta is provided to the algorithm.
- Clustering algorithms **find patterns** in the data and group data points based on those patterns.

- clustering technique is commonly used for **statistical data analysis.**
- Clustering can be used for a variety of tasks, such as:
    - Market segmentation
    - Customer segmentation
    - Fraud detection
    - Image segmentation
    - Natural language processing

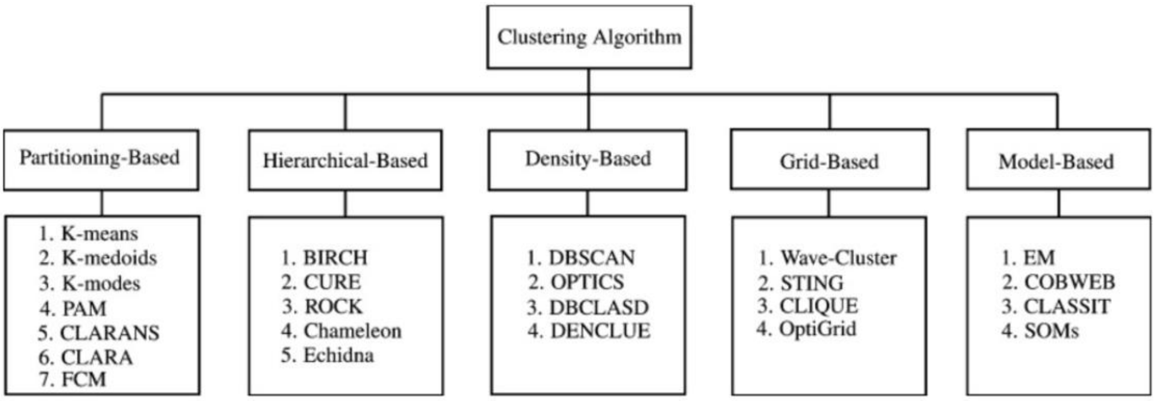
## Example

**In a mall**, similar items are grouped together, such as t-shirts, trousers, and fruits.

This grouping makes it easier for customers to find what they are looking for.

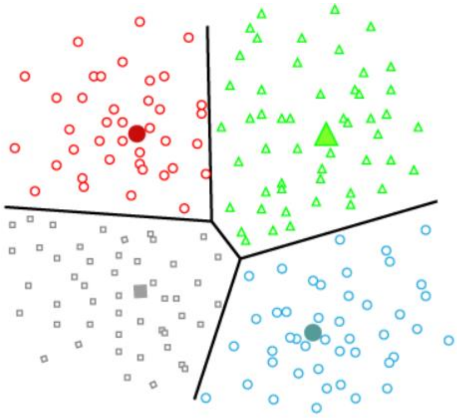


# Types of Clustering Methods



## 1) Partitioning Clustering

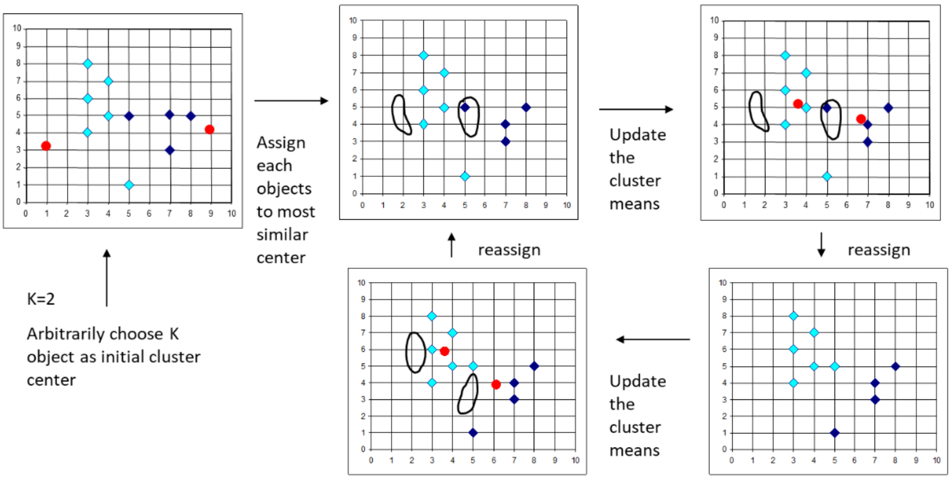
- Partitioning clustering is a type of clustering that **divides data into non-hierarchical groups**.
- It is also known as the **centroid-based method**.
- The most common example of partitioning clustering is the **K-means clustering algorithm**.
- In partitioning clustering, the dataset is divided into a set of k groups, **where k is a pre-defined** number.
- The cluster center is created in such a way that the distance between the data points in one cluster is minimized compared to the distance between data points in other clusters.

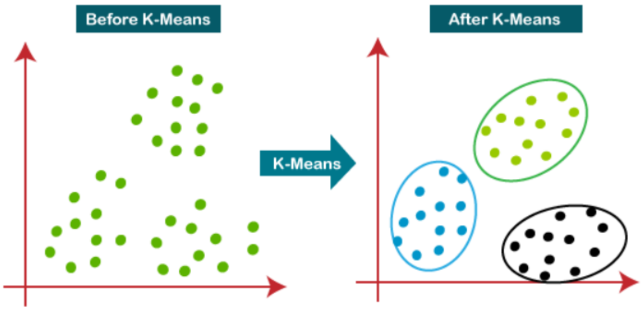


## K-Means Algorithm

- K-means clustering is an unsupervised learning algorithm that groups unlabeled data into k pre-defined clusters.
- The algorithm works by iteratively assigning data points to clusters and recomputing the cluster centroids.
- The goal of the algorithm is to minimize the sum of the squared distances between each data point and its assigned cluster centroid.
- K-means clustering is a simple and effective algorithm that can be used for a variety of tasks.
- The main advantage of k-means clustering is that it is fast and efficient.
- The main disadvantage of k-means clustering is that it requires the user to specify the number of clusters in advance.

- Example


K=2
Arbitrarily choose K object as initial cluster center
Assign each objects to most similar center
Update the cluster means
reassign
reassign
Update the cluster means



K-Means algorithm steps:

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

**Step-4:** Calculate the variance and place a new centroid of each cluster.

**Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

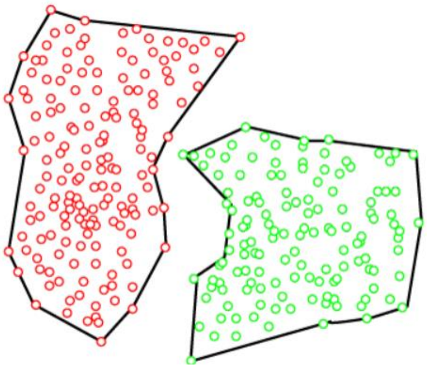
**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

**Step-7**: The model is ready.

- Weakness
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify *k,* the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

## 2) **Density-Based Clustering**

- Density-based clustering **connects high-density** areas into clusters.
- This method can form **arbitrarily shaped clusters**.
- Density-based clustering algorithms can **identify dense areas** in the data space.
- These algorithms can have difficulty clustering data with varying densities and **high dimensions.**
- Examples of density-based clustering algorithms include **DBSCAN** and **OPTICS**.

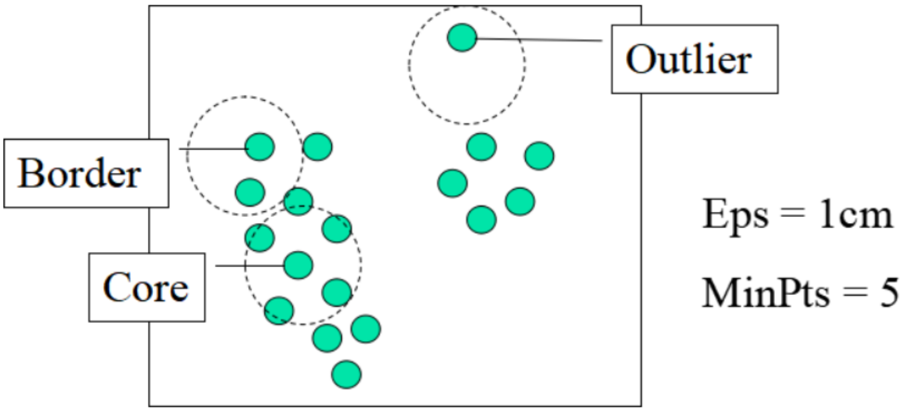


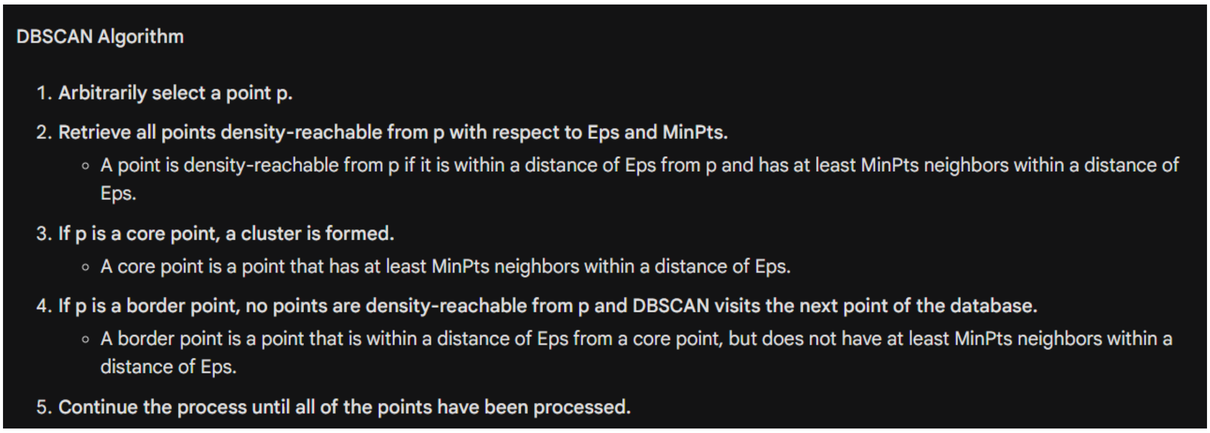
**Major features:**

- It is used to discover clusters of arbitrary shape.
- It is also used to handle noise in the data clusters.
- It is a one scan method.
- It needs density parameters as a termination condition.

## DBSCAN (Density-Based Spatial Clustering Of Applications With Noise )

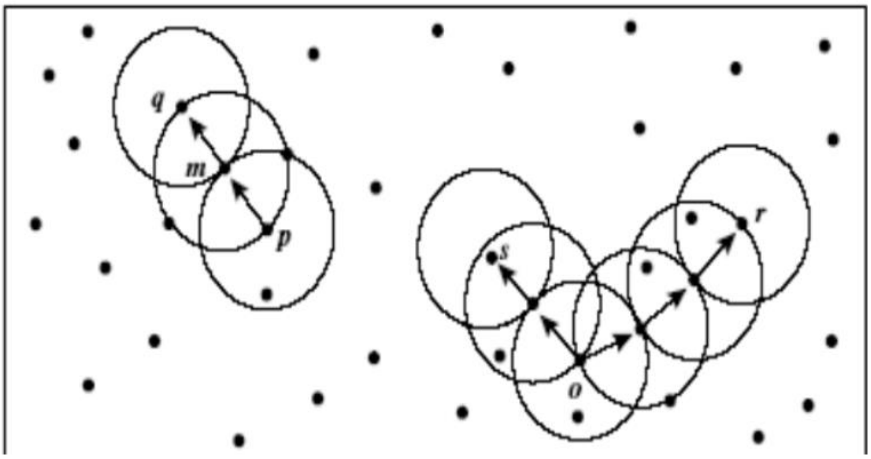
- DBSCAN is a density-based clustering algorithm that identifies clusters as dense regions in the data space.
- The algorithm works by identifying points that have a minimum number of neighbors within a given radius.
- Points that do not have enough neighbors are considered to be noise.
- DBSCAN is able to identify clusters of arbitrary shapes and sizes.
- The algorithm is not sensitive to outliers.

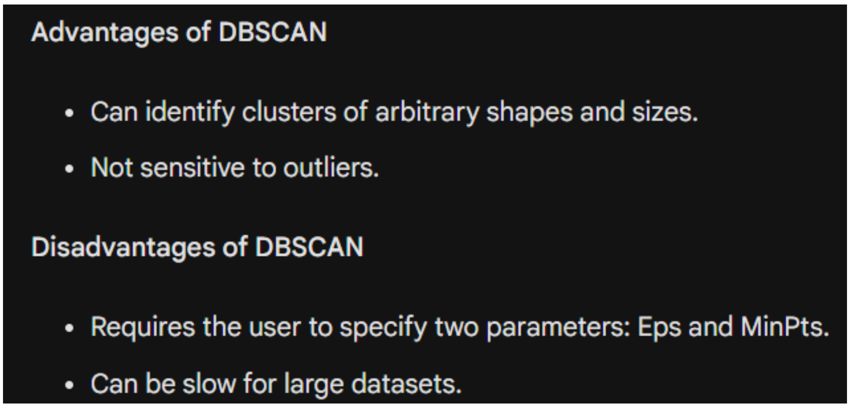




DBSCAN Algorithm

1. Arbitrarily select a point p.
2. Retrieve all points density-reachable from p with respect to Eps and MinPts.
    - A point is density-reachable from p if it is within a distance of Eps from p and has at least MinPts neighbors within a distance of Eps.
3. If p is a core point, a cluster is formed.
    - A core point is a point that has at least MinPts neighbors within a distance of Eps.
4. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
    - A border point is a point that is within a distance of Eps from a core point, but does not have at least MinPts neighbors within a distance of Eps.
5. Continue the process until all of the points have been processed.





Advantages of DBSCAN

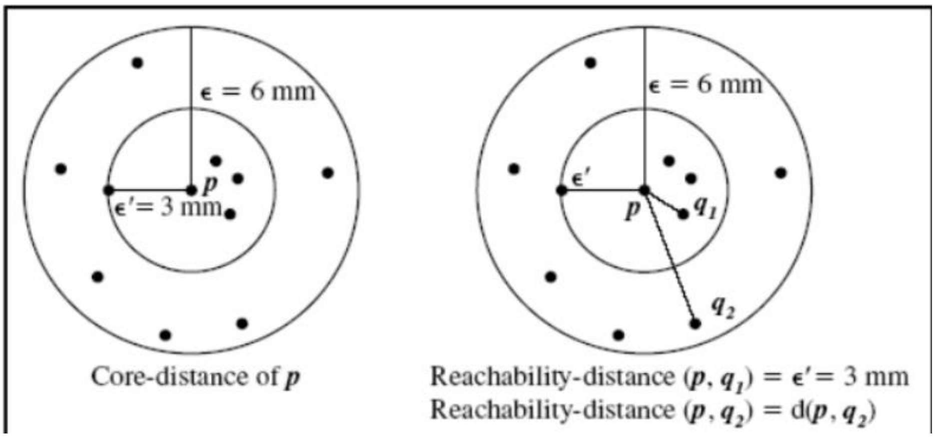
- Can identify clusters of arbitrary shapes and sizes.
- Not sensitive to outliers.

Disadvantages of DBSCAN

- Requires the user to specify two parameters: Eps and MinPts.
- Can be slow for large datasets.

## OPTICS:
## Ordering Points To Identify the Clustering Structure

- OPTICS produces a special order of the database that captures the density-based clustering structure of the data.
- This order can be used to find density-based clusters for a broad range of parameter settings.
- OPTICS is well-suited for both automatic and interactive cluster analysis.
- The results of OPTICS can be visualized graphically.



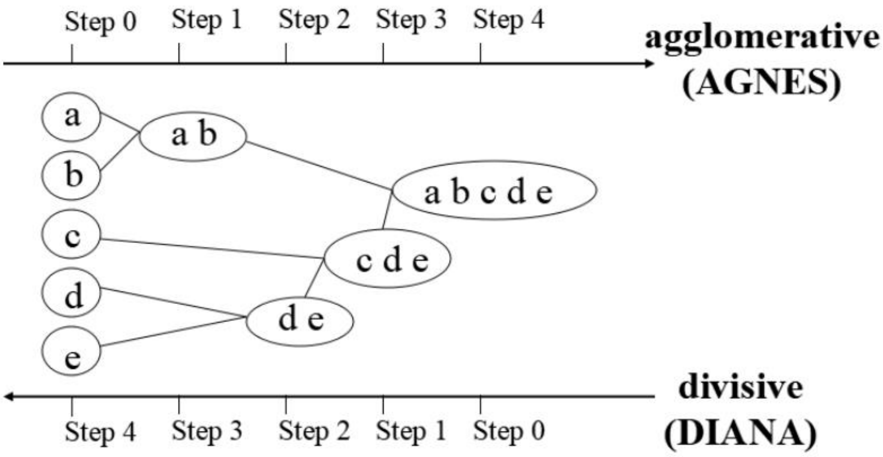
Core-distance of *p*

Reachability-distance $(p, q_1) = \epsilon' = 3$ mm
Reachability-distance $(p, q_2) = d(p, q_2)$

- **Core-distance and reachability-distance:** The figure illustrates the concepts of core-distance and reachability-distance.
- Suppose that e=6 mm and MinPts=5.
  The core distance of p is the distance, e0, between p and the fourth closest data object.
- The reachability-distance of q1 with respect to p is the core-distance of p (i.e., e0 =3 mm) because this is greater than the Euclidean distance from p to q1.
- The reachability distance of q2 with respect to p is the Euclidean distance from p to q2 because this is greater than the core-distance of p.

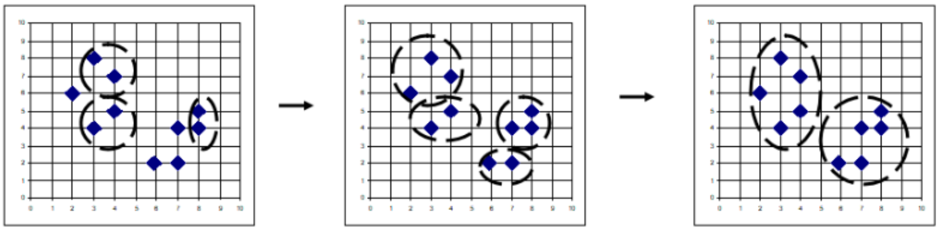
## 3) Hierarchical Clustering

- Hierarchical clustering does **not require pre-specifying** the number of clusters.
- Hierarchical clustering creates a **tree-like structure called a dendrogram**.
- Clusters can be selected by **cutting the dendrogram at the desired level**.
- **AGNES and DIANA** clustering is a common example of this method.
- **ALGO: CURE, Chameleon.**



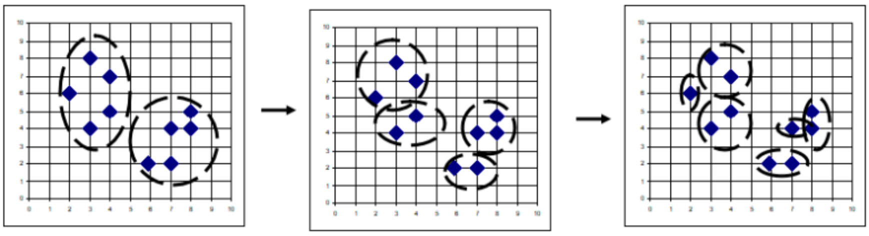
## AGNES (AGGLOMERATIVE NESTING)

- Agglomerative Nesting (AGNES) is a hierarchical clustering algorithm that uses the single-link method.
- The single-link method defines the dissimilarity between two clusters as the smallest dissimilarity between any two points in the two clusters.
- AGNES starts with each data point in its own cluster.
- At each step, the two clusters with the smallest dissimilarity are merged into a single cluster. This process continues until all data points are in a single cluster.
- The resulting dendrogram can be cut at any level to obtain a desired number of clusters. AGNES is a relatively simple and efficient algorithm.
- it can be sensitive to outliers.
- AGNES is a deterministic algorithm.



## Divisive Analysis (DIANA)

- DIANA is the opposite of AGNES.
- DIANA starts with all data points in a single cluster and splits clusters at each step.
- DIANA is less sensitive to outliers than AGNES.
- DIANA has lower computational complexity than AGNES.
- DIANA is a top-down hierarchical clustering algorithm.
- The computational complexity of DIANA is $O(n \log n)$, which is lower than the $O(n^2)$ complexity of AGNES.
- DIANA is a deterministic algorithm.



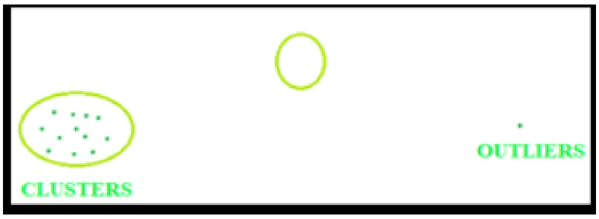
## DISADVANTAGES OF **Hierarchical Clustering**

- Hierarchical clustering can be difficult to choose the merge or split points.
- These decisions can affect the quality of the clusters.

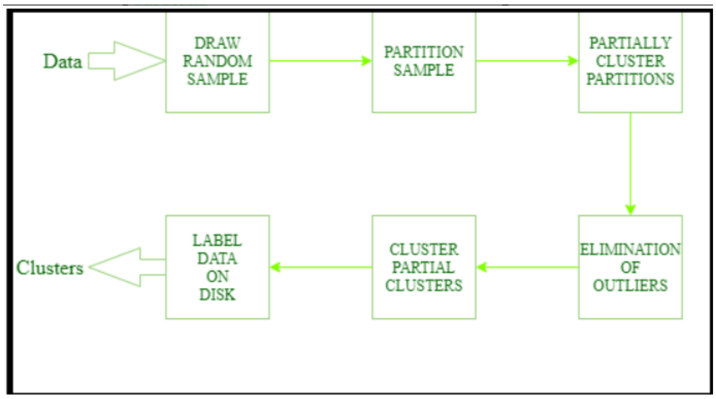
- Hierarchical clustering does not scale well to large datasets.

# CURE Algorithm(Clustering Using Representatives)

- CURE is a hierarchical clustering algorithm that uses a set of representative points to efficiently handle clusters and eliminate outliers.
- CURE is useful for identifying spherical and non-spherical clusters.
- CURE is a middle ground between centroid-based and all-point extremes.
- CURE starts with a single point cluster and merges clusters until the desired number of clusters are formed.
- CURE is useful for discovering groups and identifying interesting distributions in the underlying data.



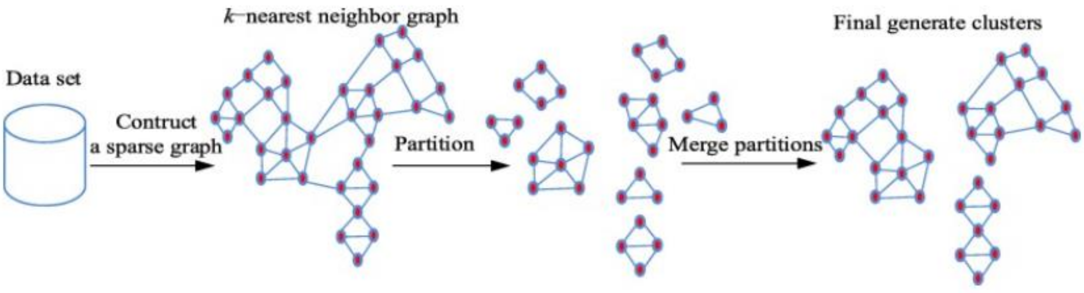
## Six steps in CURE algorithm:



# Chameleon:

## (Hierarchical Clustering Algorithm Using Dynamic Modeling)

- Chameleon is a hierarchical clustering algorithm that uses dynamic modeling.
- Chameleon was derived from ROCK and CURE.
- Chameleon uses a k-nearest-neighbor graph approach to construct a sparse graph. Chameleon uses a graph partitioning algorithm to partition the k-nearest-neighbor graph into subclusters.
- Chameleon uses an agglomerative hierarchical clustering algorithm to merge subclusters. Chameleon takes into account both interconnectivity and closeness of clusters.



Download : Download high-res image (285KB)

Download : Download full-size image

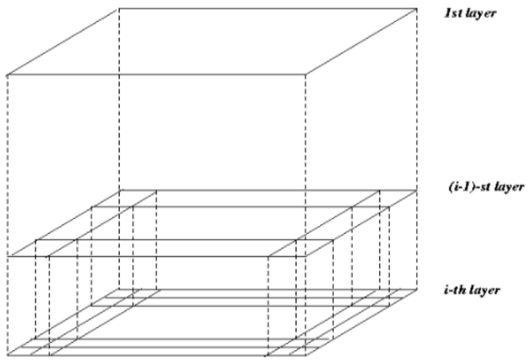
Fig. 1. Chameleon: hierarchical clustering based on *k*-nearest neighbor and dynamic modeling.

## 4) Grid-Based Methods

- It quantizes the object space into a finite number of cells that form a grid structure.
- It is fast and has processing time that is independent of the number of data objects.
- Examples of it include STING, Wave Cluster, and CLIQUE.
- It is well-suited for large datasets.
- It is well-suited for data that is stored in a spatial database.
- It can be used to find clusters of arbitrary shapes.
- It is not sensitive to outliers.

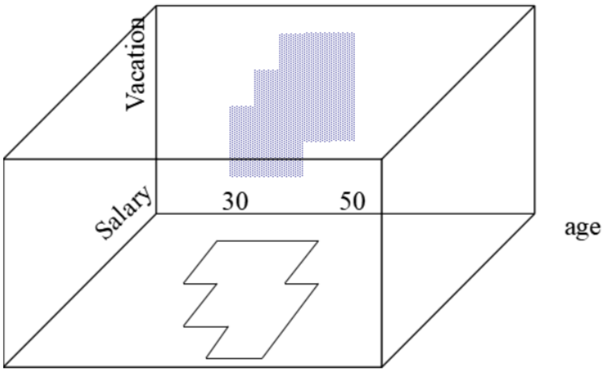
## STING

- STING divides the area into rectangular cells.
- STING stores information about the data in each cell.
- STING is fast because it does not need to look at all the data to answer a query.
- STING is good for large datasets.
- STING is good for data that is stored in a spatial database.
- STING is sensitive to the size of the cells.
- STING cannot handle data that is not rectangular.



# CLIQUE

- CLIQUE starts with single-dimensional subspaces and grows upward to higher-dimensional ones.
- CLIQUE divides each dimension into a grid structure.
- CLIQUE determines whether a cell is dense based on the number of points it contains.
- CLIQUE can be seen as an integration of density-based and grid-based clustering methods.
- CLIQUE identifies the sparse and "crowded" areas in the data space.
- A unit in CLIQUE is considered dense if the fraction of total data points contained in it exceeds an input model parameter.



## WaveCluster

- It was proposed by Sheikholeslami, Chatterjee, and Zhang (VLDB'98).
- It is a multi-resolution clustering approach which applies wavelet transform to the feature space
- A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band.
- It can be both grid-based and density-based method.

› It is an effective removal method for outliers.

› It is of Multi-resolution method.

› It is cost-efficiency.

**Major features:**

› The time complexity of this method is O(N).

› It detects arbitrary shaped clusters at different scales.

› It is not sensitive to noise, not sensitive to input order.

› It only applicable to low dimensional data.

# 5) Model-based clustering

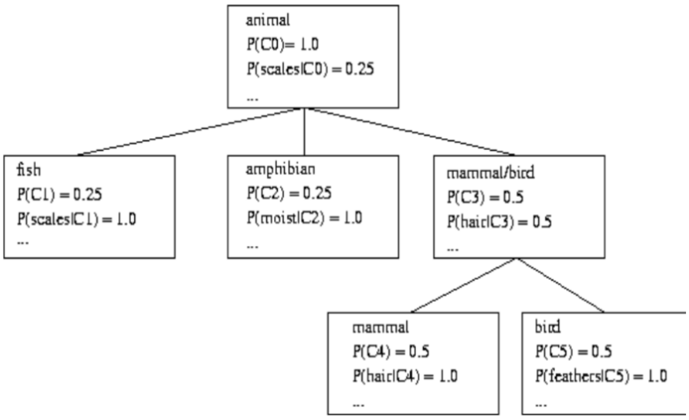
- Model-based clustering optimizes the fit between data and mathematical models.
- Model-based clustering assumes gene expression data comes from a mixture of distributions.
- Each cluster in model-based clustering corresponds to a distribution.
- Model-based clustering estimates the parameters of each distribution.
- K-means clustering is a special case of model-based clustering.
- Model-based clustering provides the probability that each gene belongs in each cluster. Conceptual clustering produces a classification scheme for unlabeled objects.
- Conceptual clustering finds characteristic descriptions for each concept.

- Typical methods
  - Statistical approach
    - EM (Expectation maximization), AutoClass
  - Machine learning approach
    - COBWEB, CLASSIT
  - Neural network approach
    - SOM (Self-Organizing Feature Map)

## COBWEB (Fisher'87)

- COBWEB is a popular a simple method of incremental conceptual learning.
- It creates a hierarchical clustering in the form of a classification tree.
- Each node refers to a concept and contains a probabilistic description of that concept.

### Classification Tree



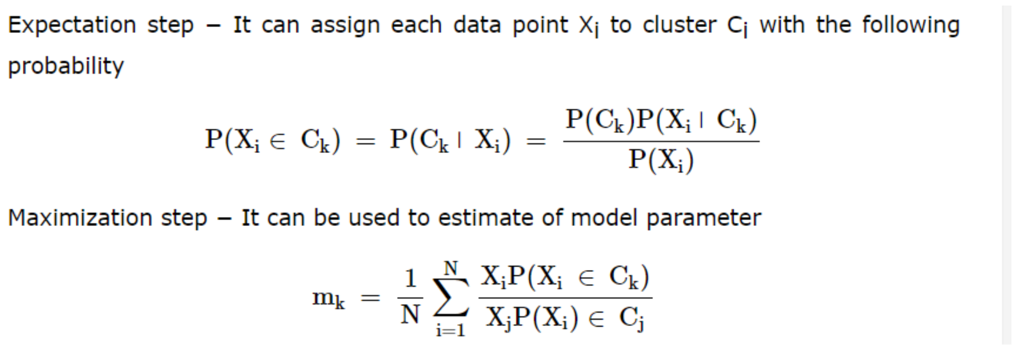
## Limitation

- COBWEB assumes that attributes are independent, but this is often not true.
- COBWEB is not suitable for large databases because it can create skewed trees and expensive probability distributions.

## EM-Algorithm

- Expectation maximization is a popular iterative refinement algorithm.
- It is an extension to k-means clustering.
- It can assign each object to a cluster according to a weight (probability distribution).

- 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

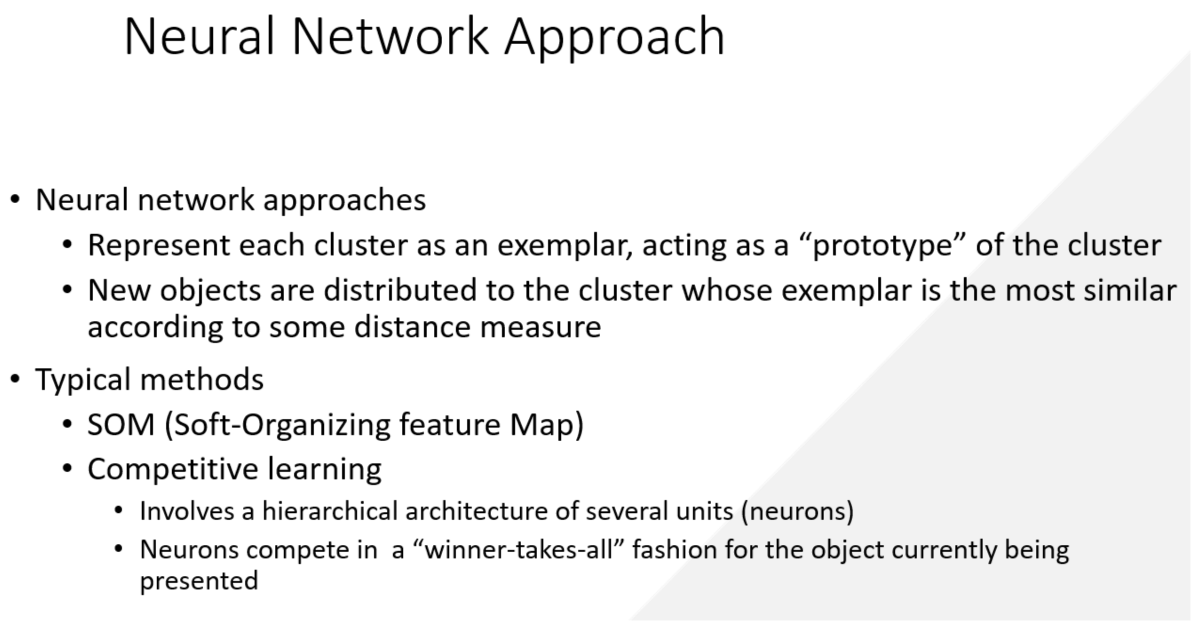


Expectation step – It can assign each data point $X_i$ to cluster $C_i$ with the following probability

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

Maximization step – It can be used to estimate of model parameter

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



# Neural Network Approach

- Neural network approaches
  - Represent each cluster as an exemplar, acting as a “prototype” of the cluster
  - New objects are distributed to the cluster whose exemplar is the most similar according to some distance measure
- Typical methods
  - SOM (Soft-Organizing feature Map)
  - Competitive learning
    - Involves a hierarchical architecture of several units (neurons)
    - Neurons compete in a “winner-takes-all” fashion for the object currently being presented

# Self-Organizing Feature Map (SOM)

- SOMs, also called topological ordered maps, or Kohonen Self-Organizing Feature Map (KSOMs)
- It maps all the points in a high-dimensional source space into a 2 to 3-d target space, s.t., the distance and proximity relationship (i.e., topology) are preserved as much as possible
- Similar to k-means: cluster centers tend to lie in a low-dimensional manifold in the feature space
- Clustering is performed by having several units competing for the current object
  - The unit whose weight vector is closest to the current object wins
  - The winner and its neighbors learn by having their weights adjusted
- SOMs are believed to resemble processing that can occur in the brain
- Useful for visualizing high-dimensional data in 2- or 3-D space

# What Is Outlier Discovery?

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