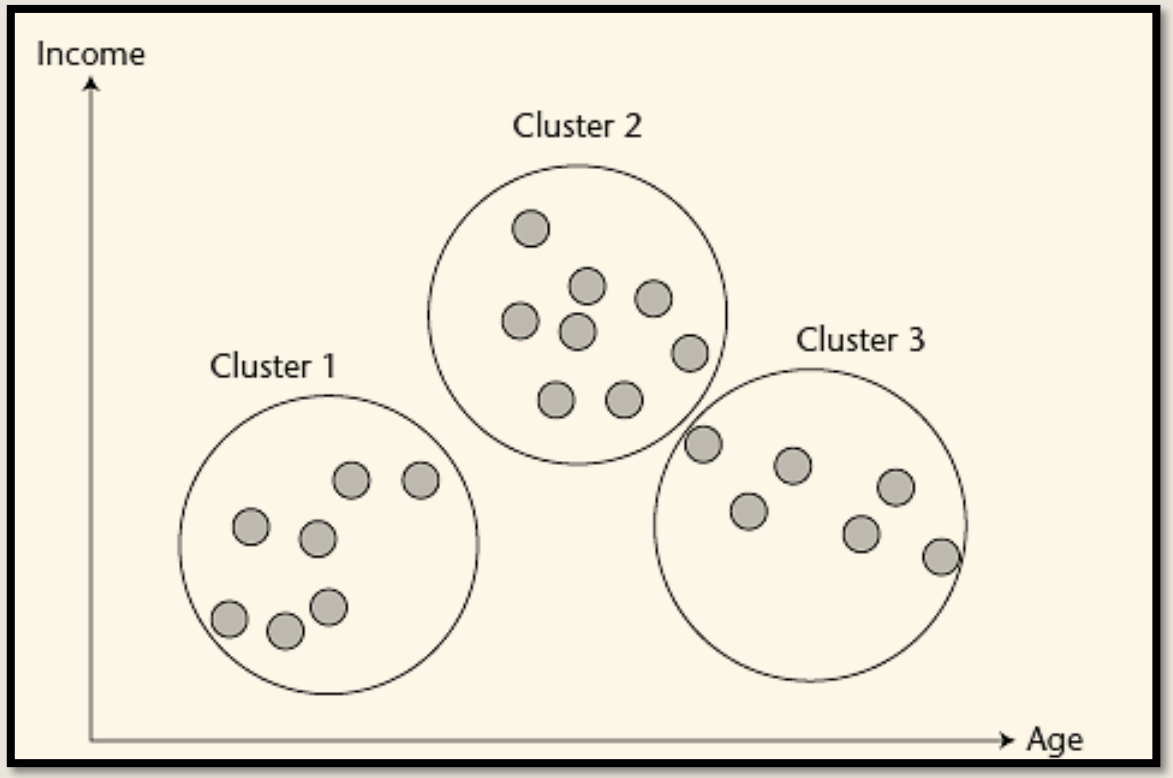
CLUSTER ANALYSIS IN DATA MINING

Subject Incharge: Priya Sachdeva, Assistant Professor (CSE)

Clustering

- Clustering is a task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics.
- Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters.
- Clusters are the grouping of data points such that the distance between the data points within the clusters is minimal.

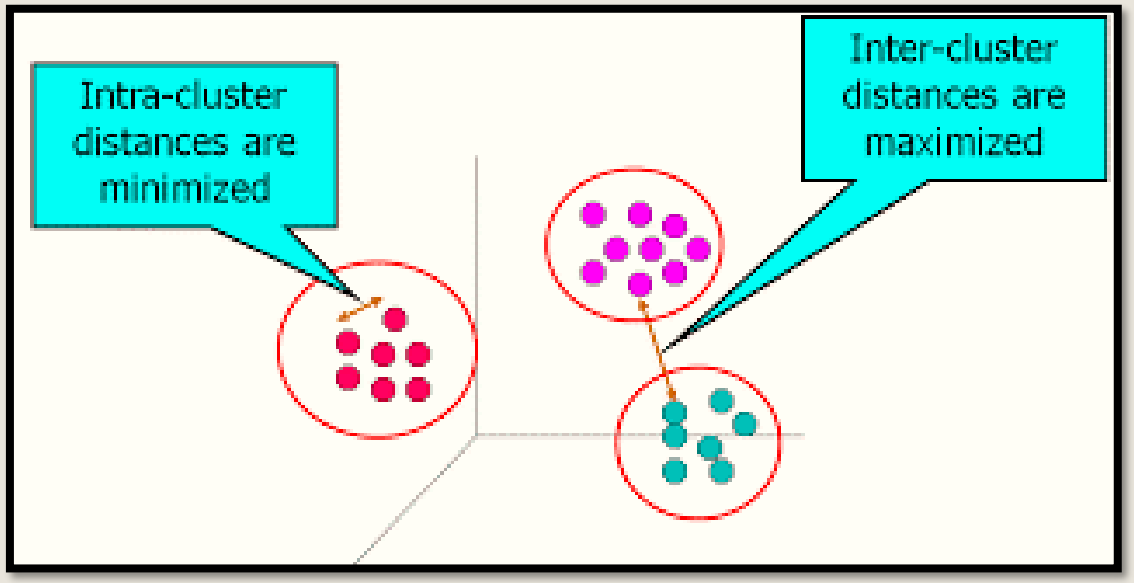


•The distance between any of the two objects in the cluster is less than the distance between any object in the cluster and any object that is not located inside it. In other words, the clusters are regions where the density of similar data points is high.

•A good clustering algorithm aims to obtain clusters whose:

✓The intra-cluster similarities are high: cohesive within clusters. i.e. the data present inside the cluster is similar to one another.

✓The inter-cluster similarity is low : distinctive between clusters. i.e. each cluster holds data that is not similar to other data.



- The quality of a clustering method depends on the similarity measure used by the method, its implementation, and ability to discover some or all of the hidden patterns.
- Generally, the clusters are seen in a spherical shape, but not necessary.

Requirements & Challenges of Clustering

- **Scalability:** We need highly scalable clustering algorithms to deal with large databases. Scalability in clustering implies that as we boost the amount of data objects, the time to perform clustering should approximately scale to the complexity order of the algorithm.
- **Ability to deal with different kinds of attributes:** Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape:** The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality:** The clustering algorithm be able to handle data in all dimensions.
- **Ability to deal with noisy data:** Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability:** The clustering results should be interpretable, comprehensible, and usable.

Applications of Clustering

- Clustering analysis is widely used in data analysis, market research, pattern recognition, and image processing.
- It assists marketers to find different groups in their client base and based on the purchasing patterns. They can characterize their customer groups.
- It helps in allocating documents on the internet for data discovery.
- Clustering is also used in tracking applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to analyze the characteristics of each cluster.
- In terms of biology, it can be used to determine plant and animal taxonomies, categorization of genes with the same functionalities and gain insight into structure inherent to populations.
- It helps in the identification of areas of similar land that are used in an earth observation database and the identification of house groups in a city according to house type, value, and geographical location.

Approaches of Clustering

Partitioning approach

- Construct various partitions and then evaluate them by some criterion.
- Typical methods: k-means, k-medoids, CLARA, CLARANS

Hierarchical approach

- Create a hierarchical decomposition of the set of data (objects) using some criterion.
- Typical methods: Diana, Agnes, CURE, BIRCH, CAMELEON

Density based approach

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

	X	Y	Dissimilarity from C1	Dissimilarity from C2
0	8	7	6	2
1	3	7	3	7
2	4	9	4	8
3	9	6	6	2
4	8	5	-	-
5	5	8	4	6
6	7	3	5	3
7	8	4	5	1
8	7	5	3	1
9	4	5	-	-

	X	Y	Dissimilarity from C1	Dissimilarity from C2
0	8	7	6	3
1	3	7	3	8
2	4	9	4	9
3	9	6	6	3
4	8	5	4	1
5	5	8	4	7
6	7	3	5	2
7	8	4	-	-
8	7	5	3	2
9	4	5	-	-

Step 2: Calculating cost. The dissimilarity of each non-medoid point with the medoids is calculated and tabulated.

Each point is assigned to the cluster of that medoid whose dissimilarity is less.

The points 1, 2, 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2.

Cost = (3+4+4) + (3+1+1+2+2) = 20

Step 3: Randomly select one non-medoid point (8, 4) and recalculate the cost.

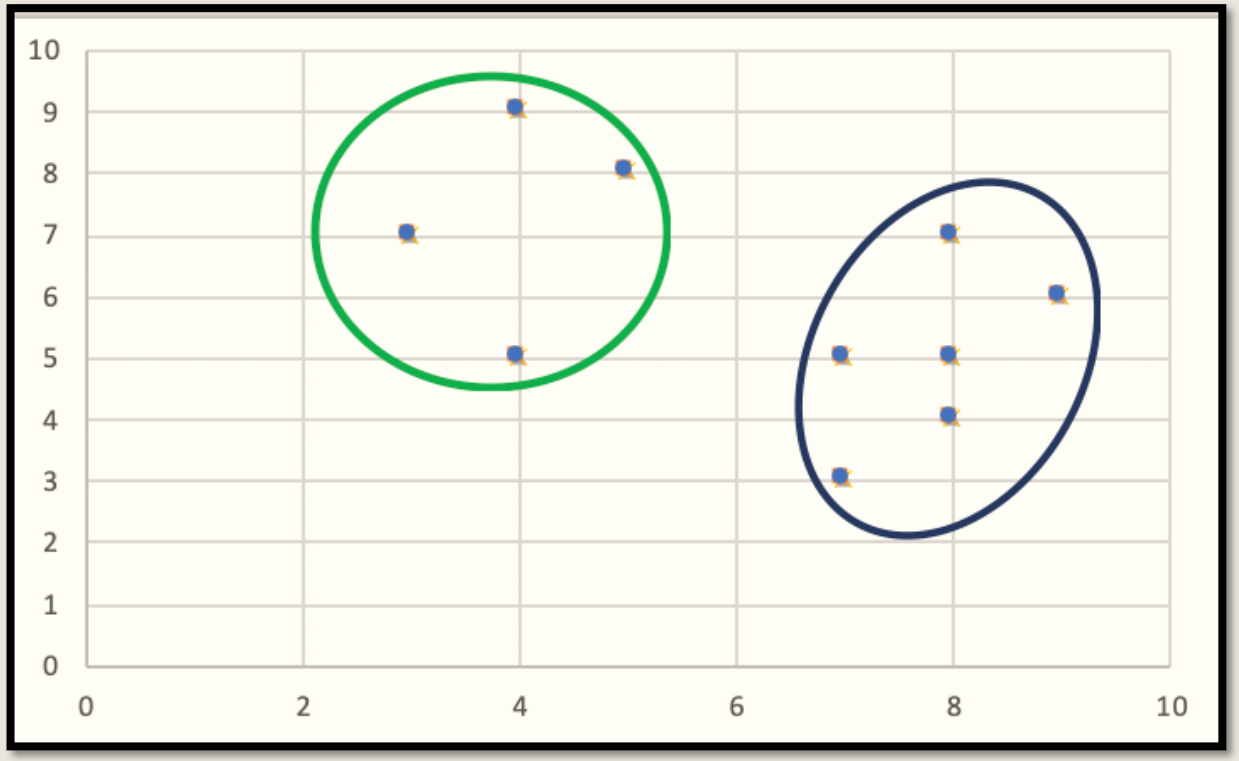
The dissimilarity of each non-medoid point with the medoids C1(4, 5) and C2(8, 4) is calculated and tabulated.

Each point is assigned to that cluster whose dissimilarity is less. So, the points 1, 2, 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2.

New cost = (3 + 4 + 4) + (2 + 2 + 1 + 3 + 3) = 22

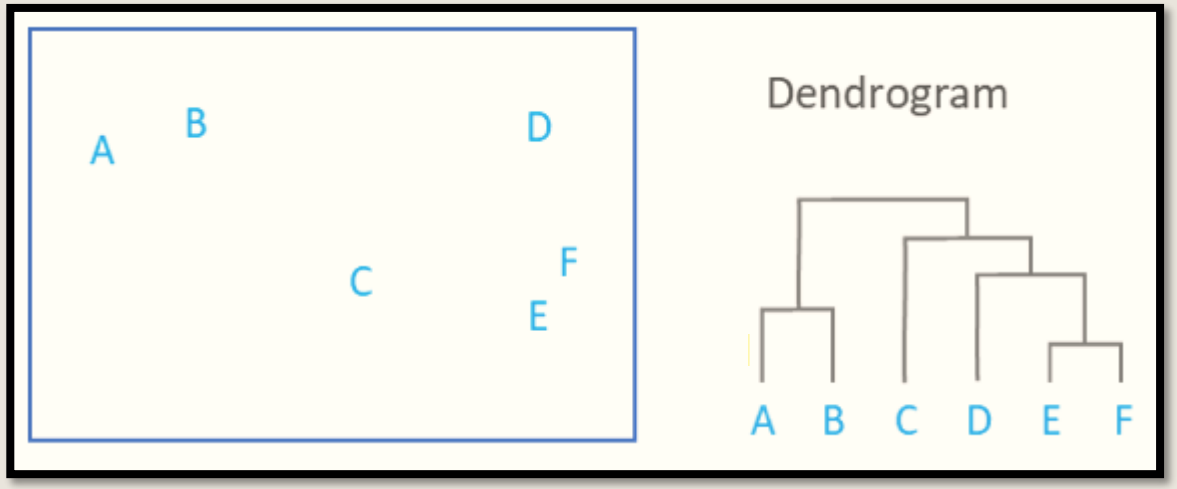
Swap Cost = New Cost – Previous Cost = 22 – 20 and 2 >0

As the swap cost is not less than zero, we undo the swap.

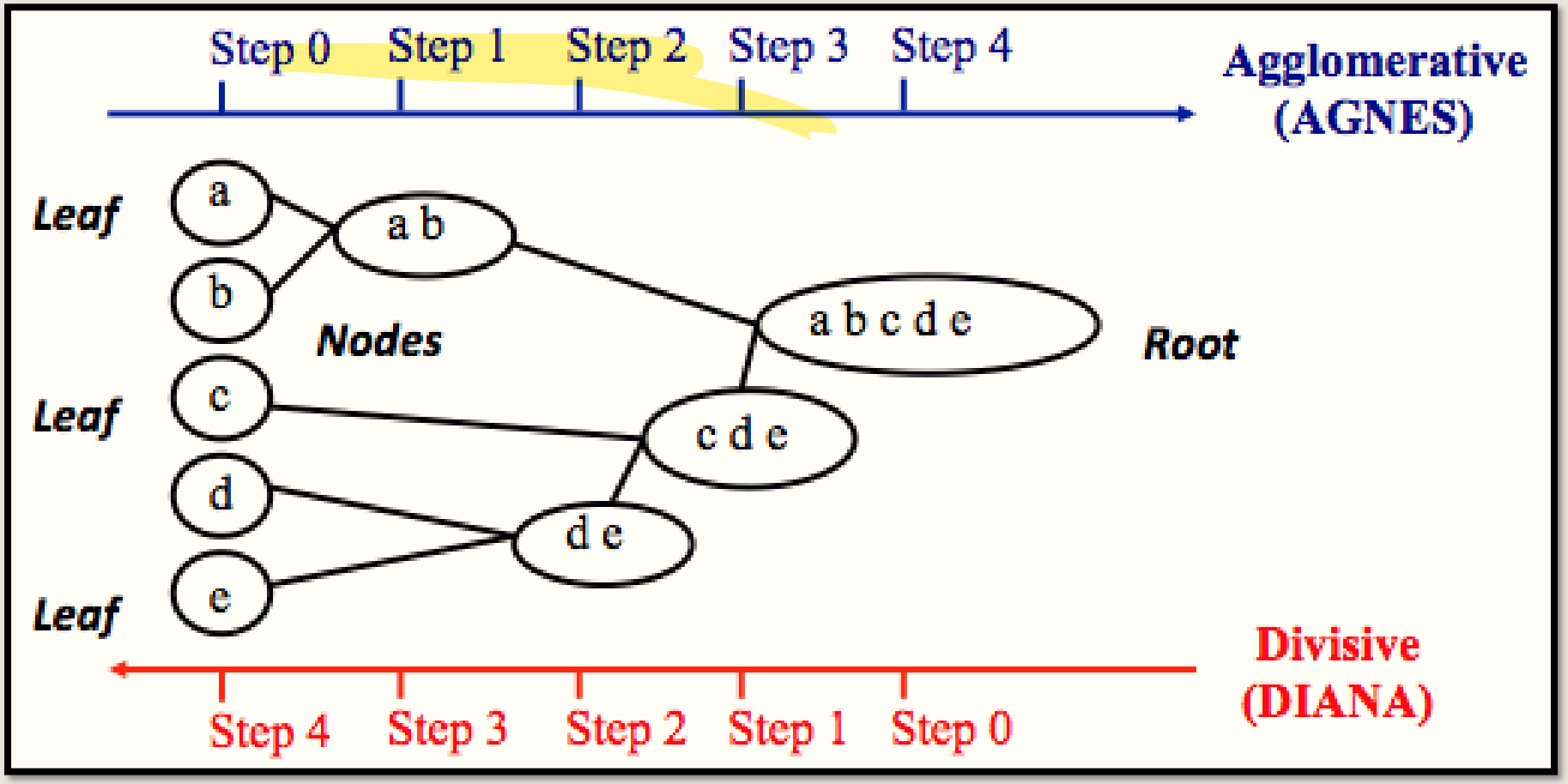


Hierarchical Clustering Methods

- A Hierarchical clustering method works via grouping data into a tree of clusters.
- **Dendrogram** (It is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top-down view).



- There are two types of approaches for the creation of hierarchical decomposition:
  1. Divisive Approach
  2. Agglomerative Approach


Step 0
Step 1
Step 2
Step 3
Step 4
Agglomerative
(AGNES)
Leaf
a
b
a b
Nodes
a b c d e
Root
Leaf
c
c d e
d
d e
Leaf
e
Divisive
(DIANA)
Step 4
Step 3
Step 2
Step 1
Step 0

Divisive Approach

It is also called top-down approach. At the beginning of this method, we consider that all the data points belong to one large cluster and try to divide the data into smaller groups based on a termination logic or, a point beyond which there will be no further division of data points. This termination logic can be based on GINI coefficient. One cannot undo after the group is split or merged, and that is why this method is not so flexible.

Agglomerative Approach

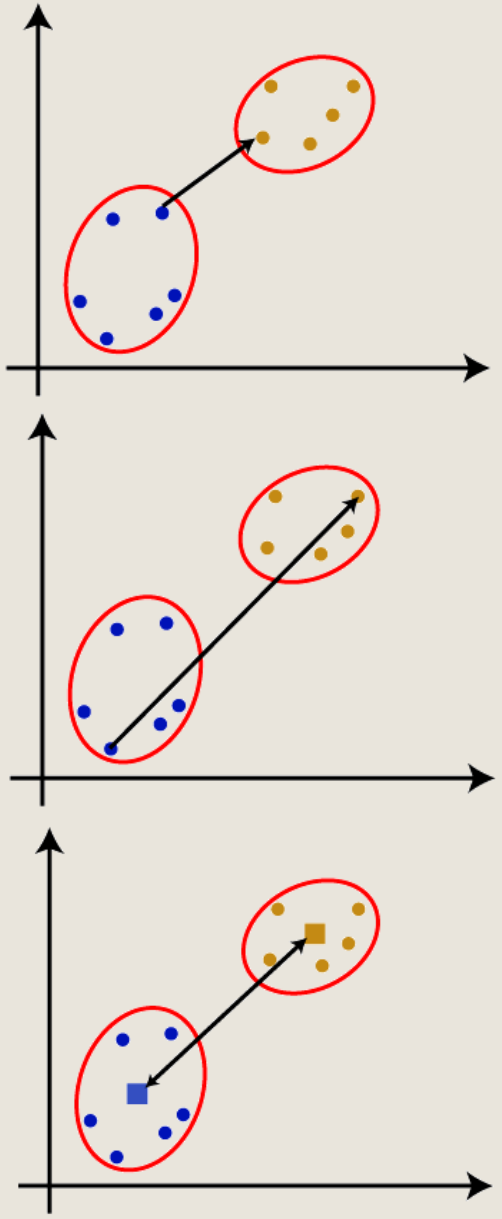
It is also called bottom-up approach. All the groups are separated in the beginning. Then it keeps on merging until all the groups are merged, or condition of termination is met.

1. Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix).
2. Consider every data point as an individual cluster.
3. Merge the clusters which are highly similar or close to each other.
4. Recalculate the proximity matrix for each cluster.
5. Repeat Step 3 and 4 until only a single cluster remains.

Measure for the distance between two clusters

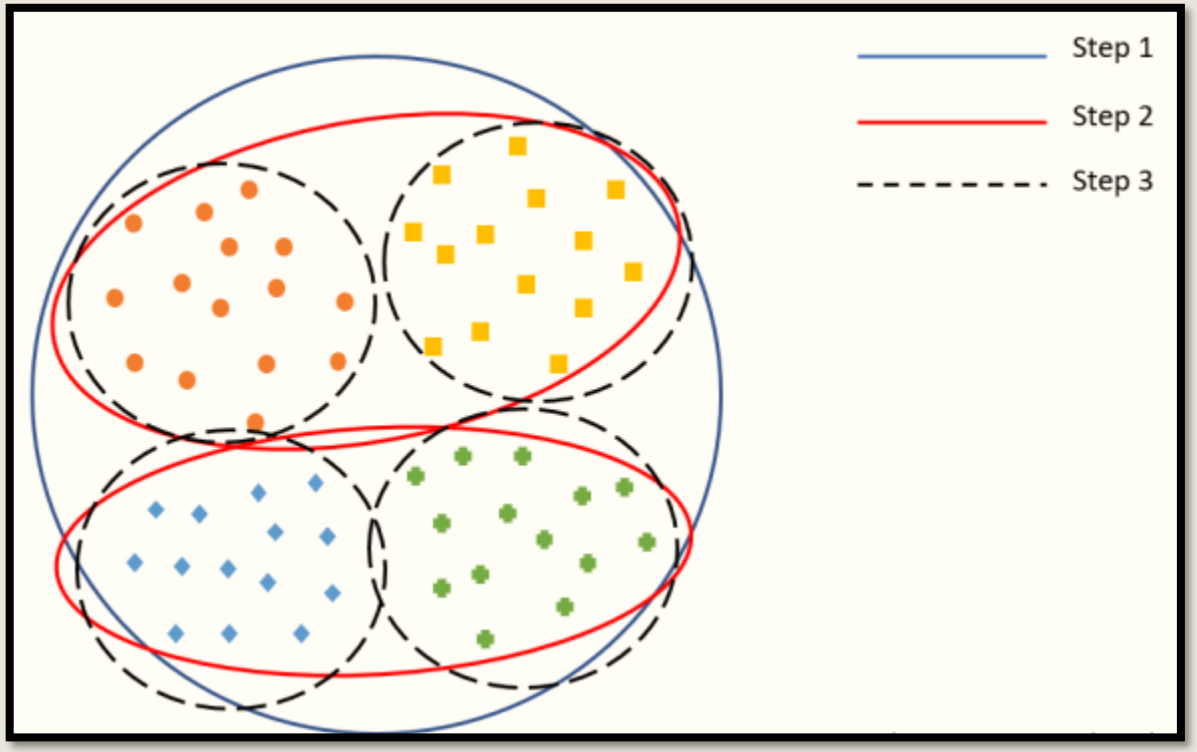
The closest distance between the two clusters is crucial for the hierarchical clustering. There are various ways to calculate the distance between two clusters, and these ways decide the rule for clustering. These measures are called Linkage methods. Some of the popular linkage methods are given below:

1. **Single Linkage:** It is the shortest distance between the closest points of the clusters.
2. **Complete Linkage:** It is the farthest distance between the two points of two different clusters. It forms tighter clusters than single-linkage.
3. **Average Linkage:** In this, distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters.
4. **Centroid Linkage:** In this, the distance between the centroid of the clusters is calculated.



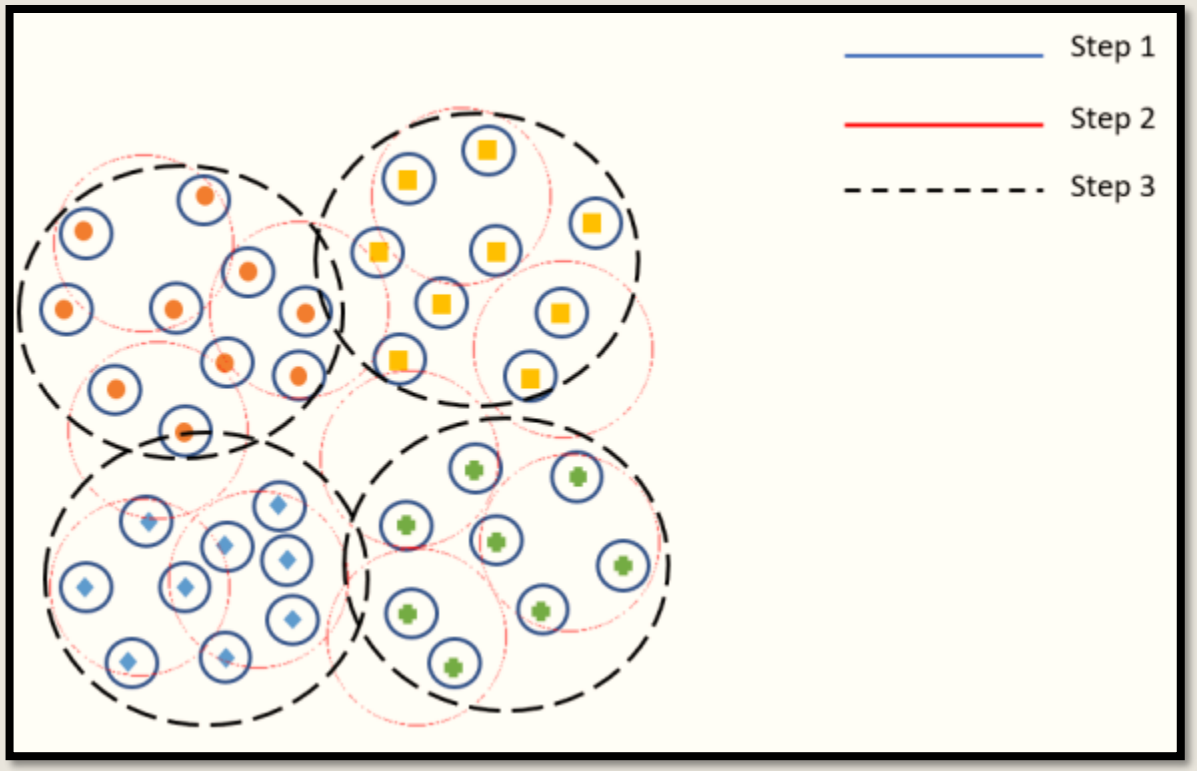
DIANA Hierarchical Clustering

- DIANA is **DI**visie **ANA**lysis clustering algorithm.
- It is the top-down approach form of hierarchical clustering where all data points are initially assigned a single cluster.
- Further, the clusters are split into two least similar clusters. This is done recursively until clusters groups are formed which are distinct to each other.



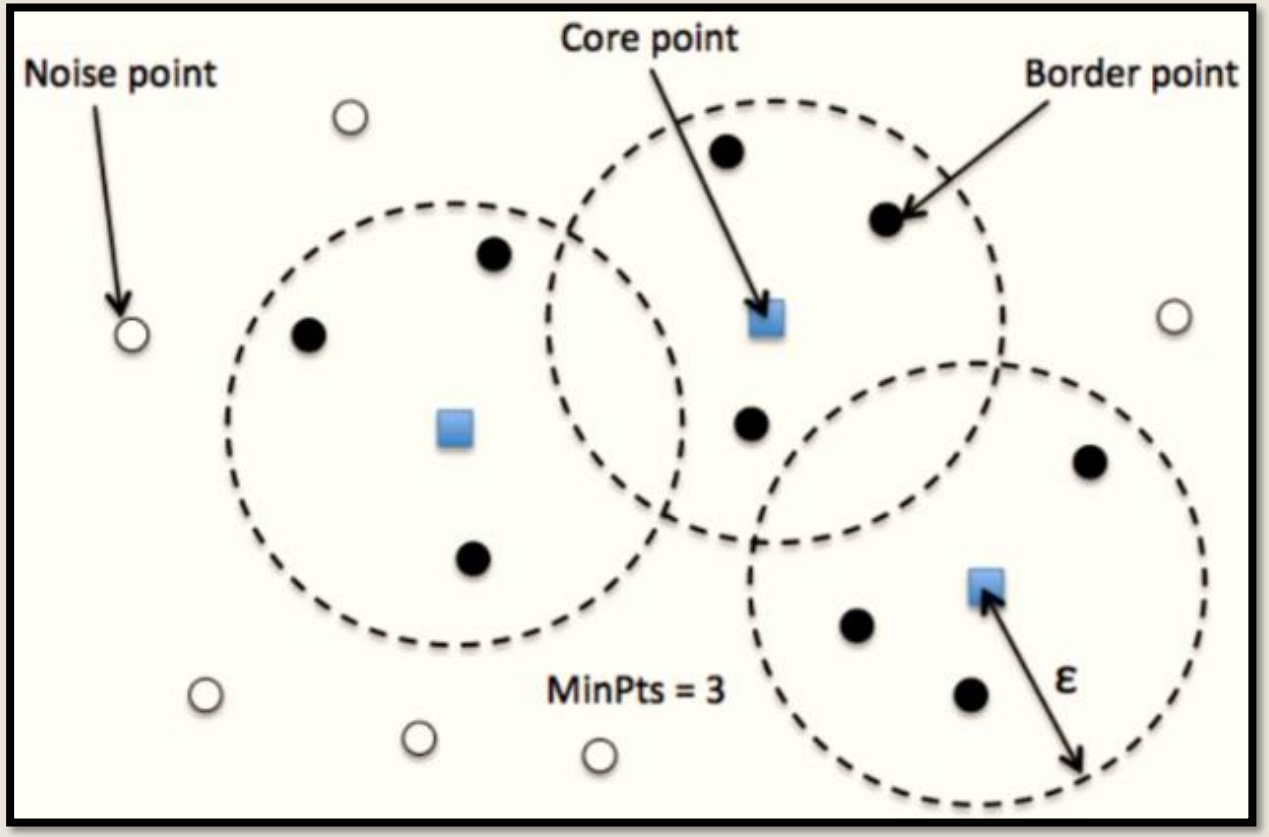
AGNES Hierarchical Clustering

- **AG**glomerative **NE**Sting hierarchical clustering algorithm is exactly opposite of the DIANA.
- It is an inside-out or bottoms-up approach.
- Here every data point is assigned as a cluster initially if there are n data points n clusters will be formed initially.
- In the next iteration, similar clusters are merged (again based on the density and distances), this continuous until similar points are clustered together and are distinct for other clusters.



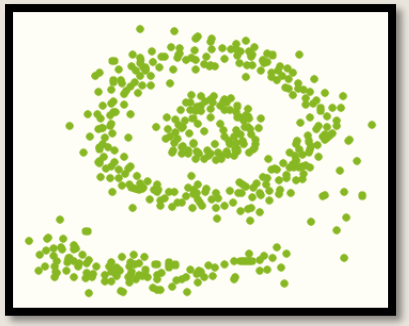
Density based Clustering

- Density-Based Clustering refers unsupervised learning methodologies used in machine learning algorithms.
- The data points in the region separated by two clusters of low point density are considered as noise.
- The surroundings with a radius ε of a given object are known as the ε neighborhood of the object.
- If the ε (epsilon) neighborhood of the object comprises at least a minimum number, MinPts of objects, then it is called a core object.



DBSCAN

- It stands for Density-Based Spatial Clustering of Applications with Noise.
- It is a clustering method utilized for separating high-density clusters from low-density clusters.
- It divides the data points into many groups so that points lying in the same group will have the same properties. It can not cluster data sets with large differences in their densities.
- It is very robust in detection of outliers in data set.
- Partitioning methods (K-means, PAM clustering) and hierarchical clustering work for finding spherical-shaped clusters or convex clusters. In other words, they are suitable only for compact and well-separated clusters. Moreover, they are also severely affected by the presence of noise and outliers in the data. Real life data may contain irregularities, like Clusters can be of arbitrary shape or data may contain noise.



The DBSCAN algorithm uses two parameters:

- **minPts:** The minimum number of points (a threshold) clustered together for a region to be considered dense.
- **eps (ε):** A distance measure that will be used to locate the points in the neighborhood of any point is called epsilon. It means that if the distance between two points is lower or equal to this value, these points are considered neighbors.

These parameters can be understood if two concepts called Density Reachability and Density Connectivity are explored.

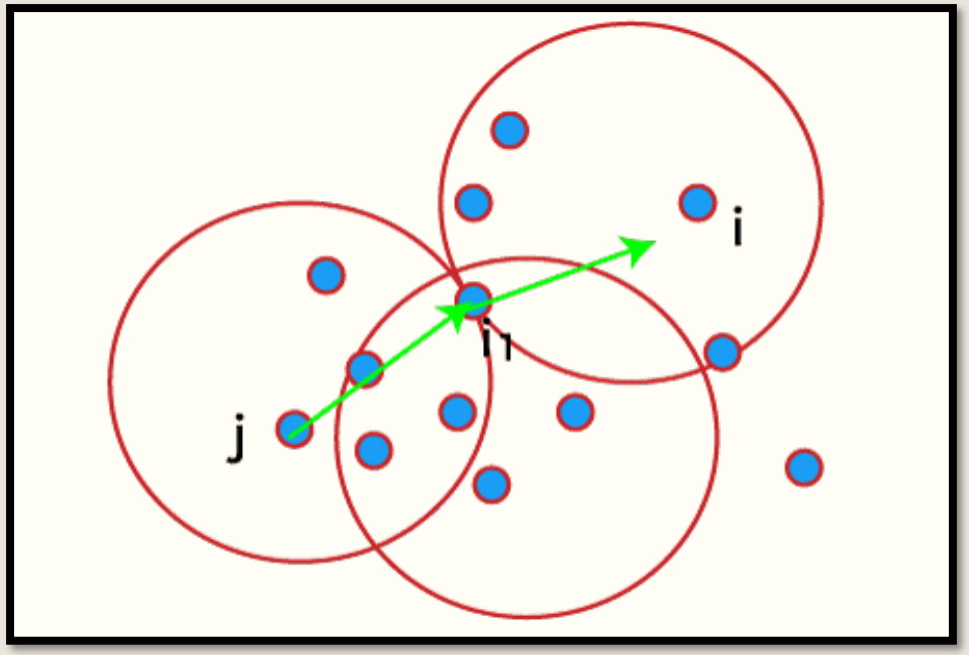
✓**Reachability** in terms of density establishes a point to be reachable from another if it lies within a particular distance (eps) from it.

✓**Connectivity** involves a transitivity based chaining-approach to determine whether points are located in a particular cluster. For example, p and q points could be connected if p->r->s->t->q.

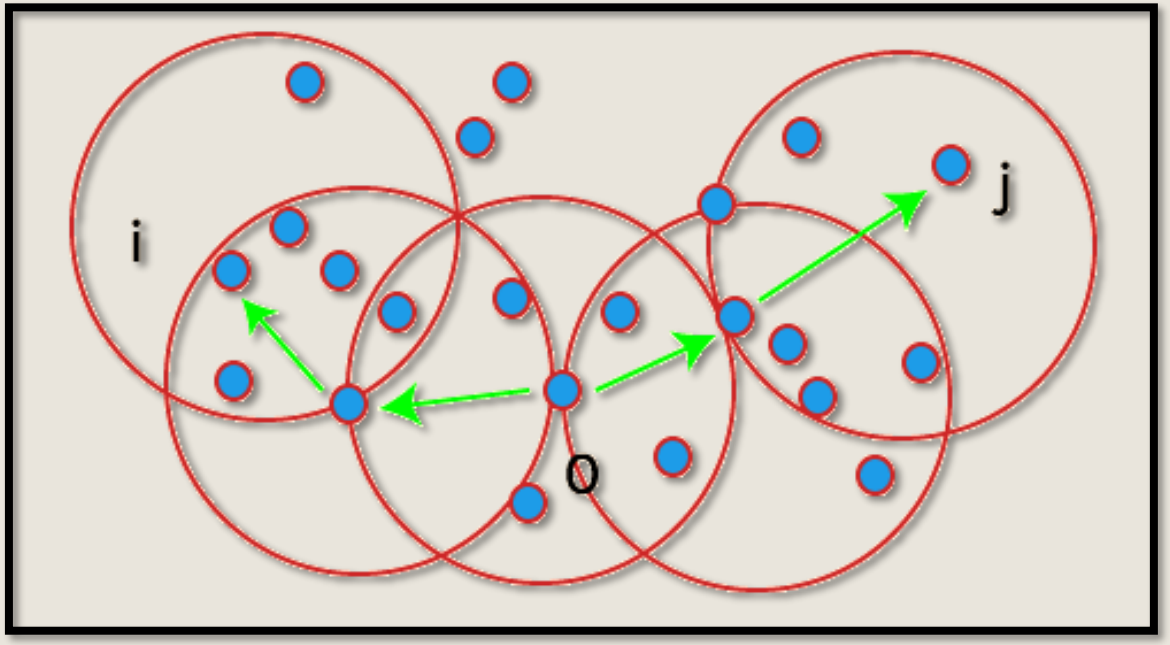
There are three types of points after the DBSCAN clustering is complete:

- **Core point**— This is a point that has at least *m* points within distance *n* from itself.
- **Border point** — This is a point that has at least one Core point at a distance *n*.
- **Noise point (outlier)** — This is a point that is neither a Core nor a Border. And it has less than *m* points within distance *n* from itself.

Density reachable



Density connected



Algorithm for DBSCAN Clustering

Input: N objects to be clusters and global parameters ε and MinPts.
Output: Clusters of objects
1. Arbitrary select a point P.
2. Retrieve all point density reachable from P wrt ε and MinPts.
3. If P is a core point a cluster is formed.
4. If P is a border point, then there is no point that is density-reachable and DBSCAN moves to the next point.
5. This process is continued until all the points are processed.

OPTICS

- OPTICS stands for Ordering Points To Identify Cluster Structure. OPTICS works like an extension of DBSCAN.
- The only difference is that it does not assign cluster memberships but stores the order in which the points are processed. So for each object stores: Core distance and Reachability distance.
  - ✓**Core Distance:** The minimum value of ε which is present in the ε-neighborhood of a P is a core distance. Of course, it's needed to hold the minimum MinPts objects.
  - ✓**Reachability Distance:** Reachability distance between p and q is defined as the least radius value that formulates p density reachable from q.

Advantage

- It does not require density parameters.
- The clustering order is useful to extract the basic clustering information.

Disadvantage

- It only produces a cluster ordering.
- It can't handle high dimensional data.

Algorithm for OPTICS Clustering

1. Randomly selects a point P.
2. Selects all point's density reachable from P w.r.t eps, MinPts.
3. Assign core distance & reachability distance = NULL
4. If P is not a core point, Move next point.
5. If P is a core point, for each object q, in the ε — neighborhood of P, update reachability distance from P.


Thank
you