Advanced Machine learning Mastering Course

Introduced by

George Samuel

Master in computer science
Cairo University

Innovisionray.com

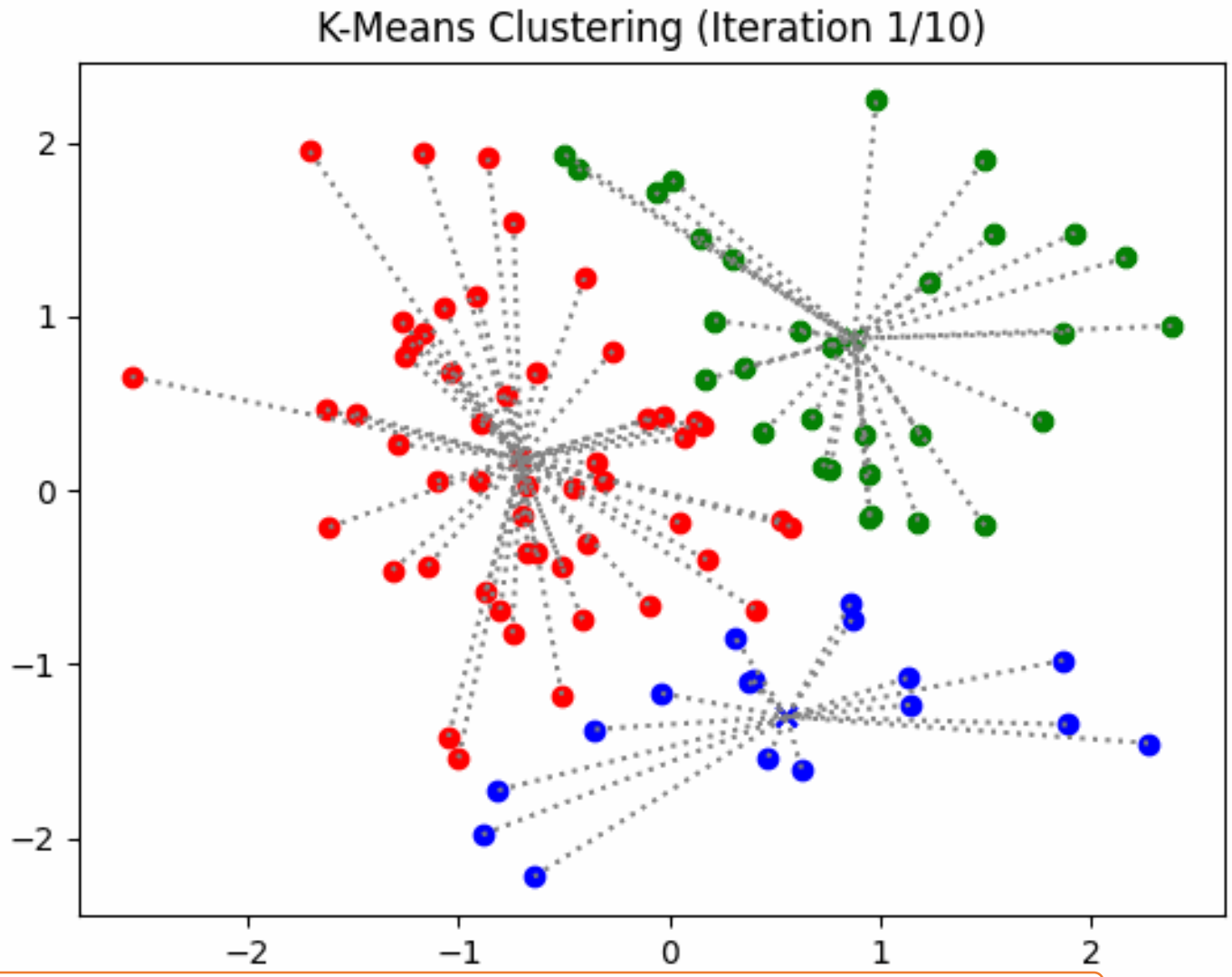
2024

Agenda

1. **Unsupervised Learning**
2. **Introduction** Unsupervised Learning
3. Clustering in Machine Learning
4. K-Means Clustering
5. Hierarchal Clustering 1.
6. DBSCAN Clustering
7. Evaluation of unsupervised learning
8. Association rules
10. Dimensionality reduction

1. Feature selection
2. Principle Component Analysis
3. Revision
4. Final Project
5.
6.
7.
8.
10.

Hierarchal Clustering 1.



HIERARCHICAL CLUSTERING

Hierarchical Clustering (HCA)

Hierarchical Clustering is a type of clustering where we deal with database in hierarchical manner to build a tree-like shape representing the clusters of data points. Hierarchical clustering has mainly two types:

Agglomerative clustering (most common):

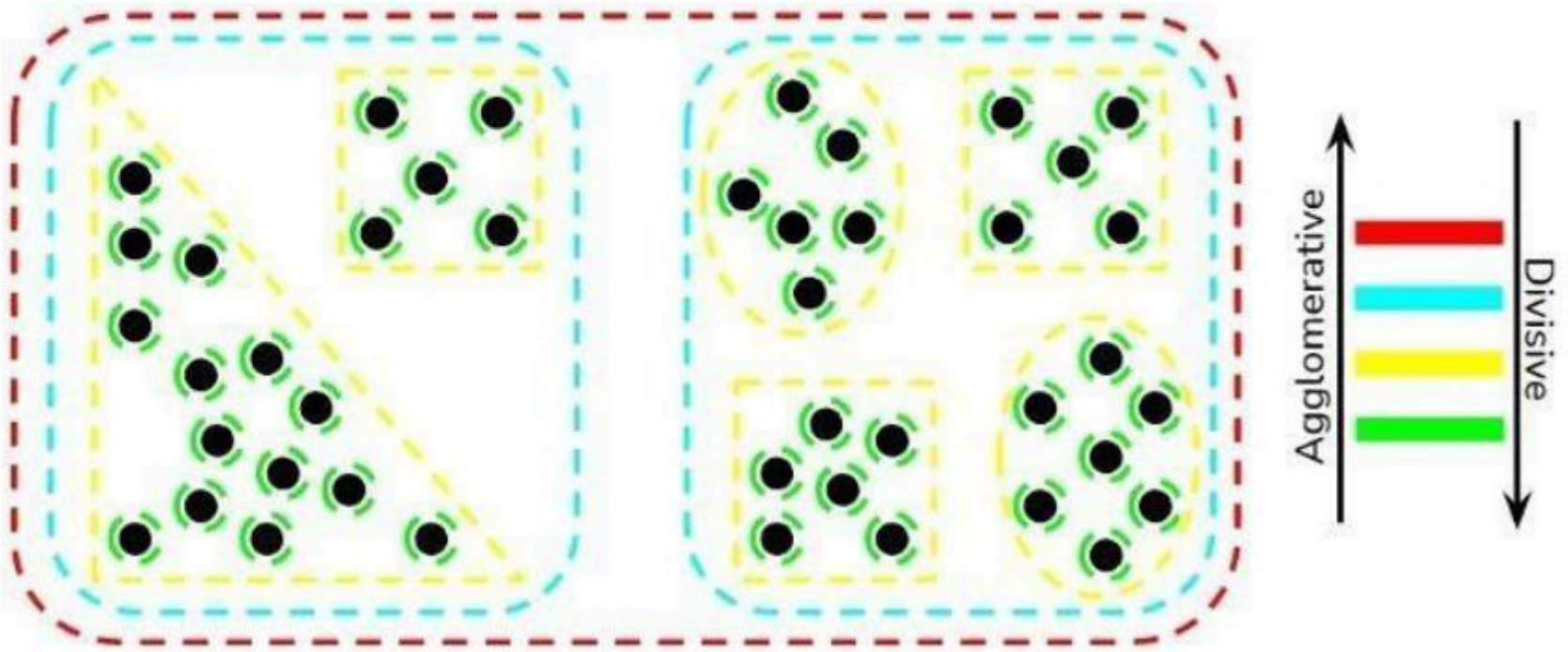
It is a bottom-up approach where we start as if each point is a cluster on its own, then we start linking the closest two clusters together with each other, and iterate in this manner until combining all data points in one cluster.

Divisive clustering:

It is a top-down approach where we start as single cluster having all points inside and start splitting out the farthest point from the center of the clusters each time until reaching that all data points becomes clusters on their own.

Hierarchal Clustering 1.

HIERARCHICAL CLUSTERING



HIERARCHICAL CLUSTERING

Agglomerative Hierarchical Clustering Example

Assume having the following dataset representing the marks of students:

Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

HIERARCHICAL CLUSTERING

Agglomerative Hierarchical Clustering Example

Now we compute what we call a proximity (distance) matrix where it gives us all the available combinations of distances (any distance measurement as Euclidean, Manhattan, Minkowski, ..) between all data points as shown in the following matrix:

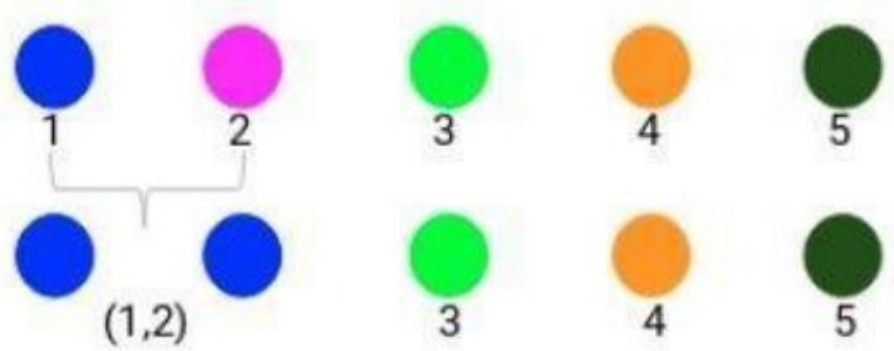
ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

HIERARCHICAL CLUSTERING

Agglomerative Hierarchical Clustering Example

As we have seen from the previous cost matrix, the closest two data points are 1 and 2 with a cost of 3.

ID	1	2	3	4	5
1	0	(3)	18	10	25
2	(3)	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0



HIERARCHICAL CLUSTERING

Agglomerative Hierarchical Clustering Example

Our new dataset now will be ⇒

Student_ID	Marks
(1,2)	10
3	28
4	20
5	35

Note: We took the maximum of the two number in the cluster, however we can take the minimum or the average. We will take about this later in the linkage methods part.

HIERARCHICAL CLUSTERING

Agglomerative Hierarchical Clustering Example Our new distance matrix now will be as follows:

ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0

Which means that the next two closest clusters are 3 and 5 with a cost of 7.

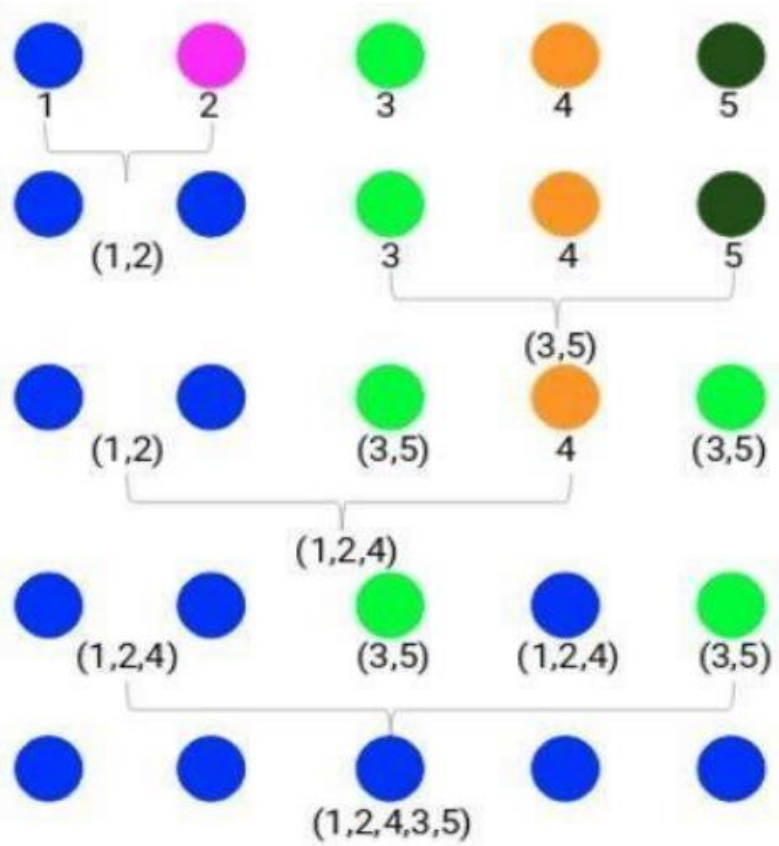
Hierarchal Clustering 1.

HIERARCHICAL CLUSTERING

Agglomerative Hierarchical Clustering Example We keep iterating like this until we have all data points as one cluster as shown:

But here comes the question? What should we do next? How can we decide which number of Clusters is best suited for this data set?

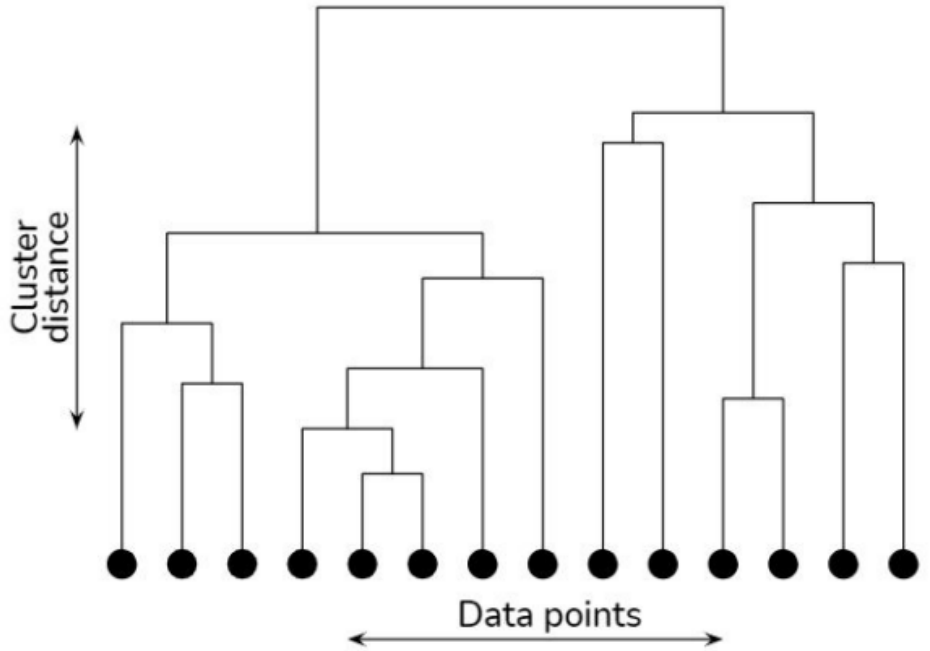
Here comes the idea of dendrogram.



HIERARCHICAL CLUSTERING

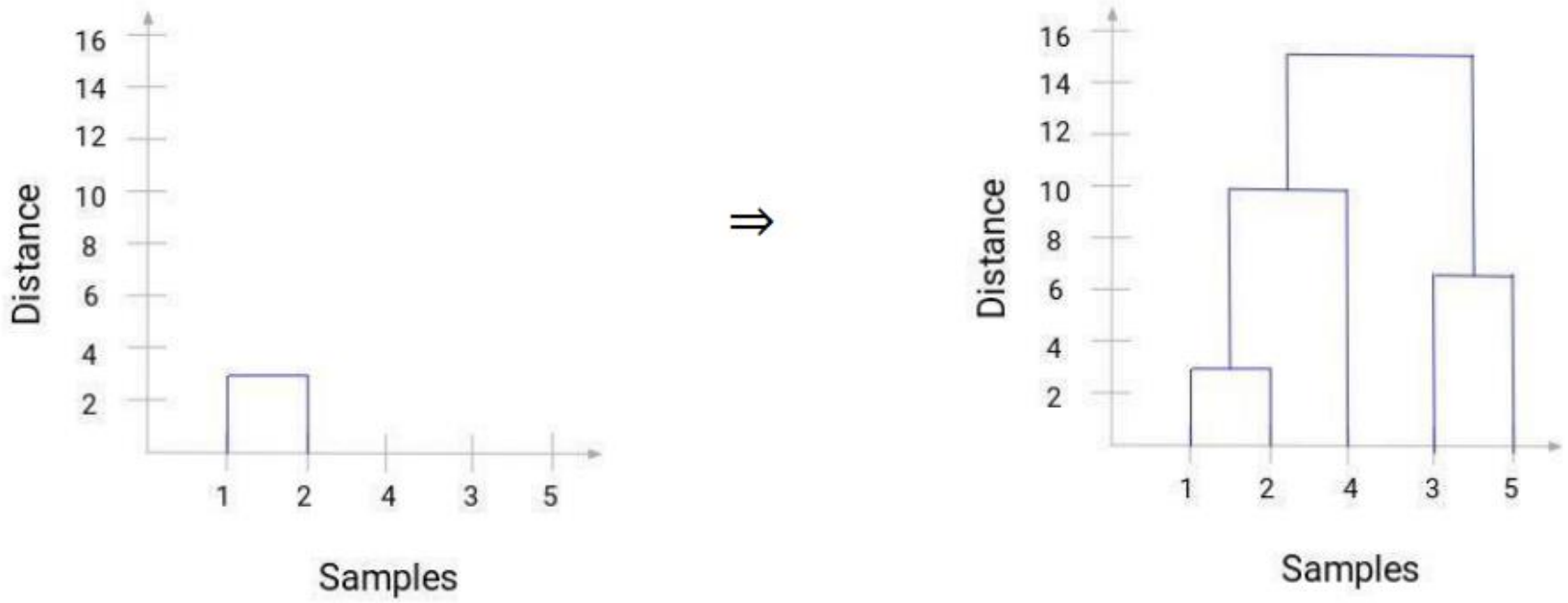
Dendrogram

The whole concept of hierarchical clustering lies in the construction and analysis of the dendrogram which is a tree-like structure that explains the relationship between the data samples and each other based on their distances between their clusters.



HIERARCHICAL CLUSTERING

Dendrogram for our Example

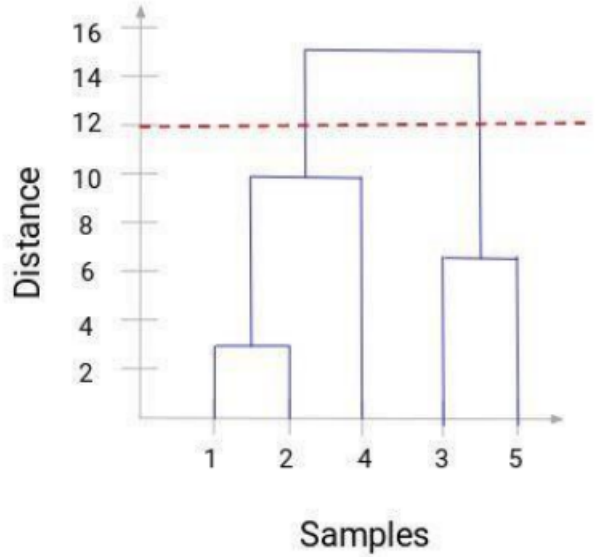


More the distance of the vertical lines in the dendrogram, more the distance between those clusters.

HIERARCHICAL CLUSTERING

Dendrogram for our Example

Now how would we decide the number of clusters? We try to cut our dendrogram in a way that it cuts the tallest vertical line for each number of clusters. This step is done visually by using dendrogram.

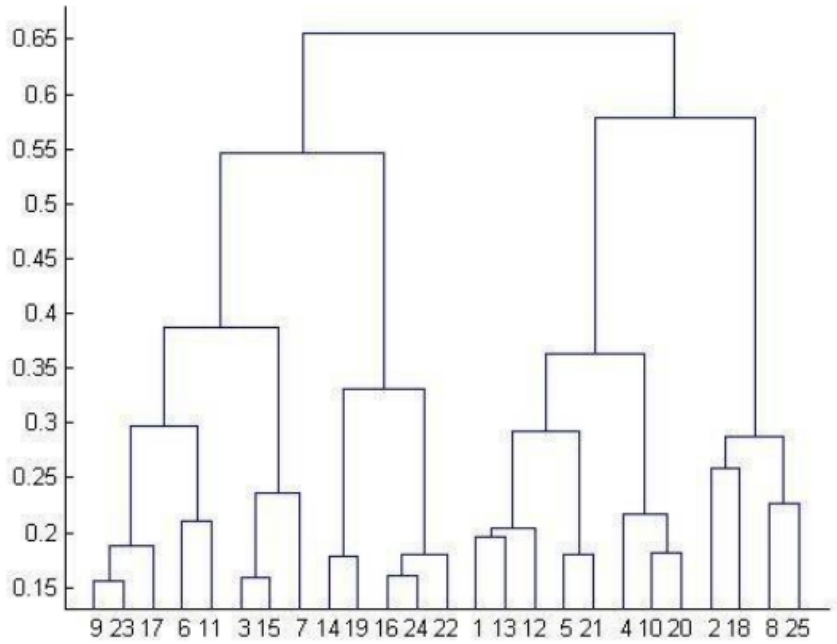


The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold. Hence 2 is the optimum number of clusters.

HIERARCHICAL CLUSTERING

Dendrogram

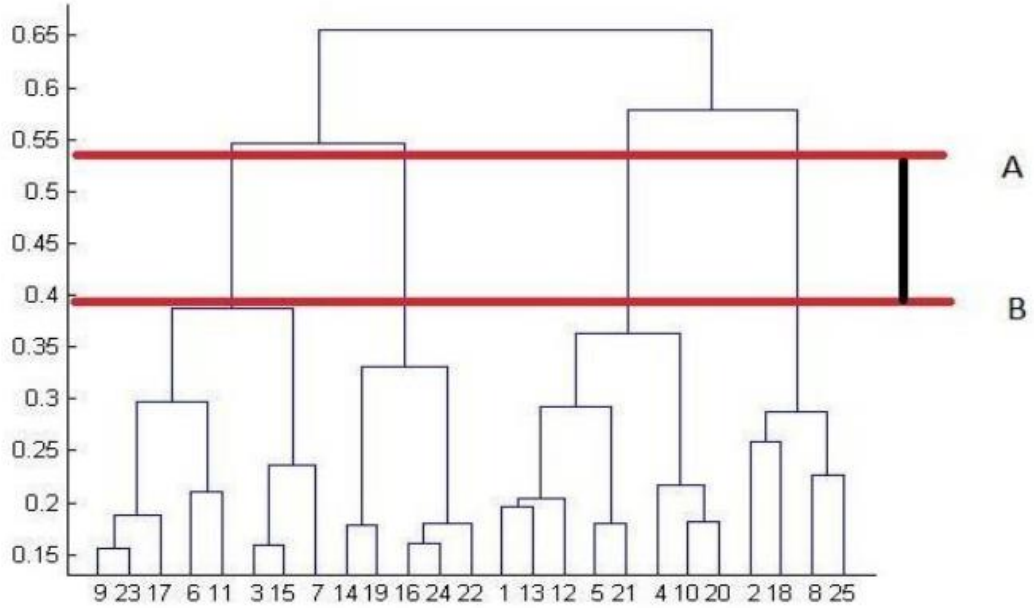
What would be the number of clusters if the dendrogram is as follows?



HIERARCHICAL CLUSTERING

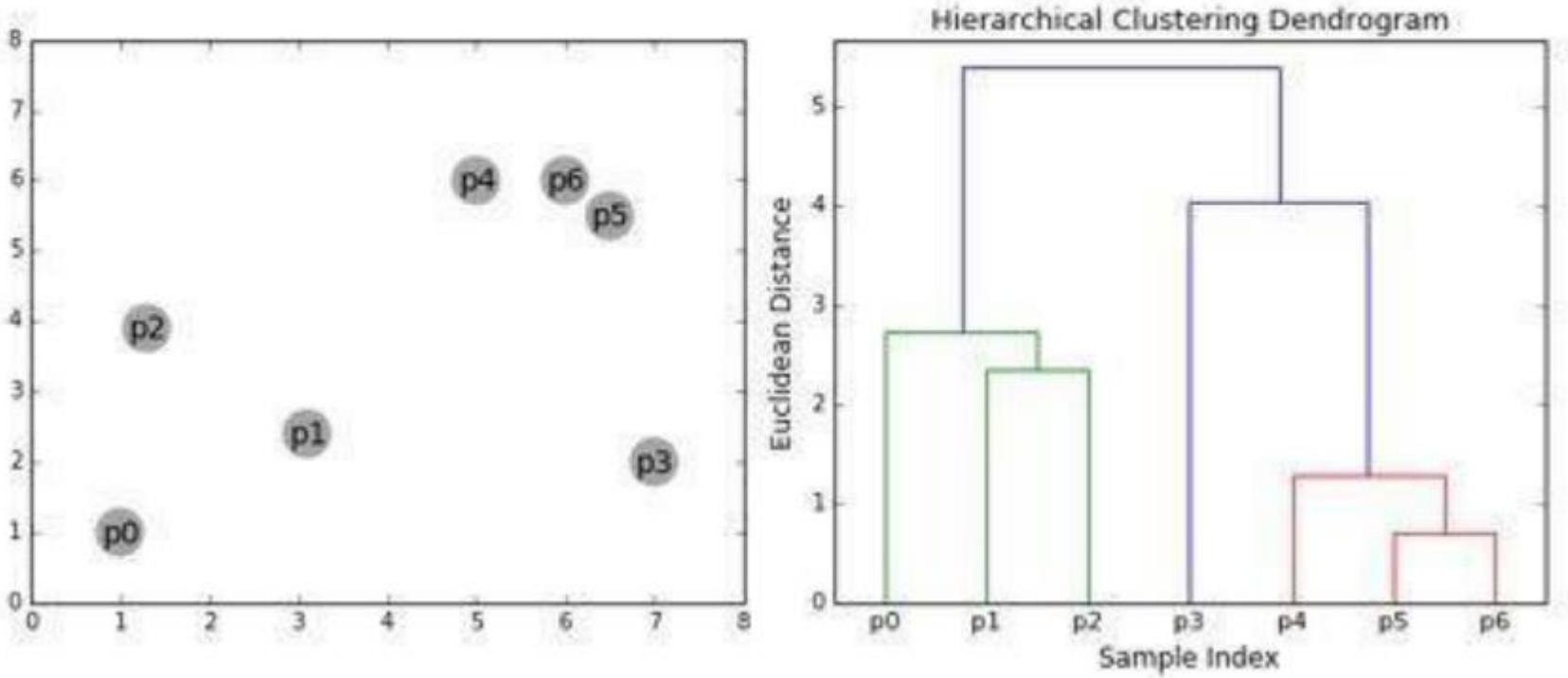
Dendrogram

The best line is the horizontal line that cuts the tallest vertical line in a way that maximizes vertical cut distance along all number of clusters candidates. Hence you can cut in any place between A and B resulting in 4 clusters.



HIERARCHICAL CLUSTERING

Agglomerative Hierarchical Clustering with Dendrogram

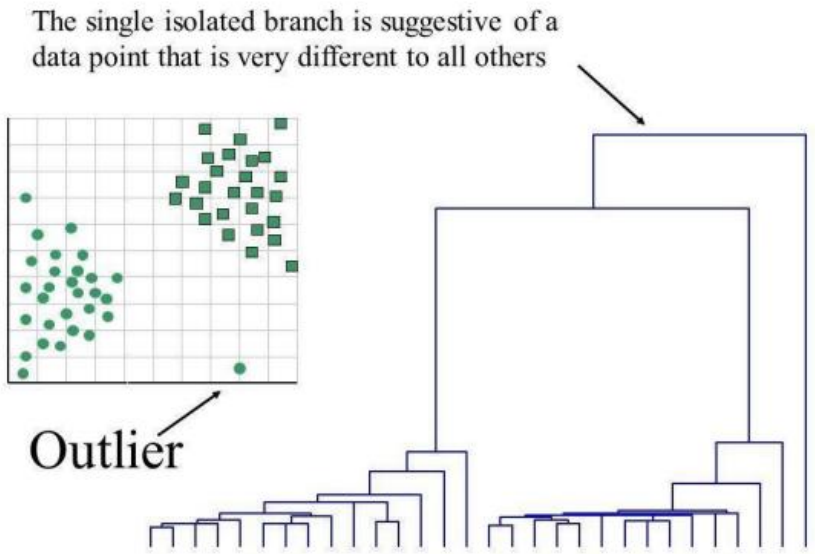


HIERARCHICAL CLUSTERING

Can Hierarchical clustering handles outliers? Sensitive or not to outliers.

Hierarchical clustering is generally sensitive to outliers where it can give a decision of cutting a dendrogram at a certain height which can be misleading.

One potential use of a dendrogram is to detect outliers

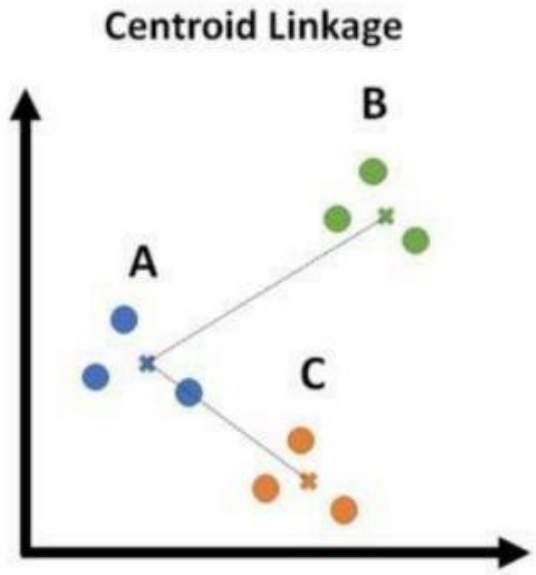


HIERARCHICAL CLUSTERING

Linkage methods

Linkage method defines the way of how we link two clusters together. We have talked about linkage method inherently while dealing with the example taken.

1) **Centroid linkage** is the distance between the centroids of 2 clusters before merging.



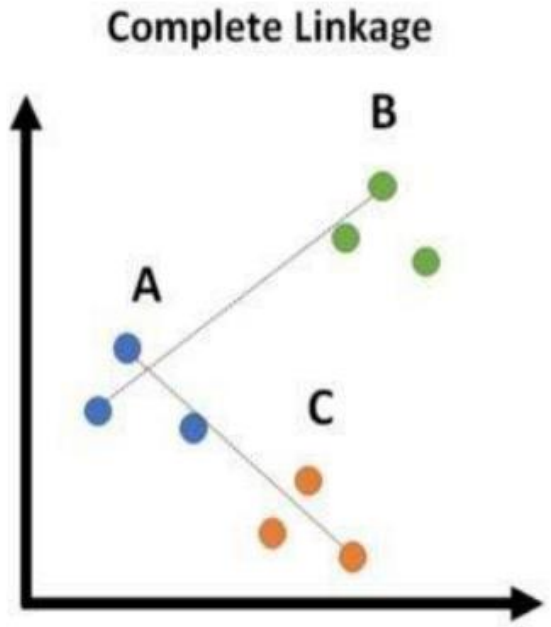
Link cluster A with cluster C because:

Distance from **centroid** of cluster A to centroid of cluster C Is **shorter** than the distance from **centroid** of cluster A to centroid of cluster C

HIERARCHICAL CLUSTERING

Linkage methods

2) Complete linkage is the longest distance among any two points in the 2 clusters. Use it if there is noise between clusters. This method reveals outliers better visually.



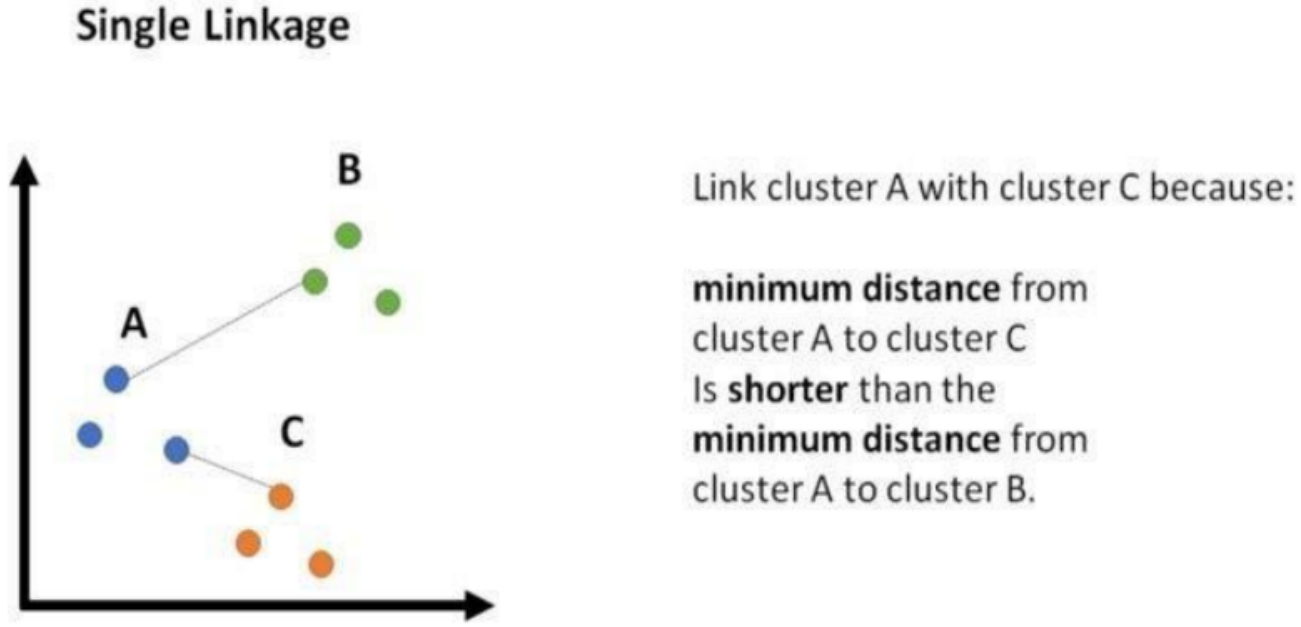
Link cluster A with cluster C because:

maximum distance from cluster A to cluster C Is **shorter** than the **maximum distance** from cluster A to cluster B.

HIERARCHICAL CLUSTERING

Linkage methods

3) Single linkage is the shortest distance among any two points in the 2 clusters. This can(not always) be good if outliers exist in the data.



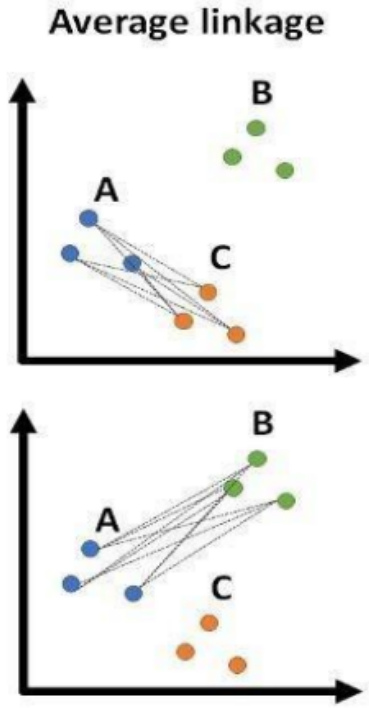
Link cluster A with cluster C because:

minimum distance from cluster A to cluster C
Is **shorter** than the **minimum distance** from cluster A to cluster B.

HIERARCHICAL CLUSTERING

Linkage methods

4) Average linkage is the average distance between each point in a cluster to each point the other cluster.



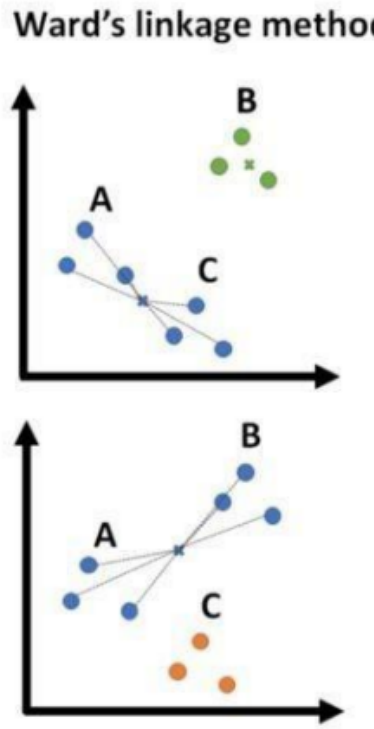
Link cluster A with cluster C because:

mean distance from cluster A to of cluster C Is **shorter** than the **mean distance** from cluster A to cluster B.

HIERARCHICAL CLUSTERING

Linkage methods

5) Ward linkage is the sum of squares of deviation from the mean of the 2 clusters (variance) after fusing them. It handles well the noise between clusters as well.



Link cluster A with cluster C because:

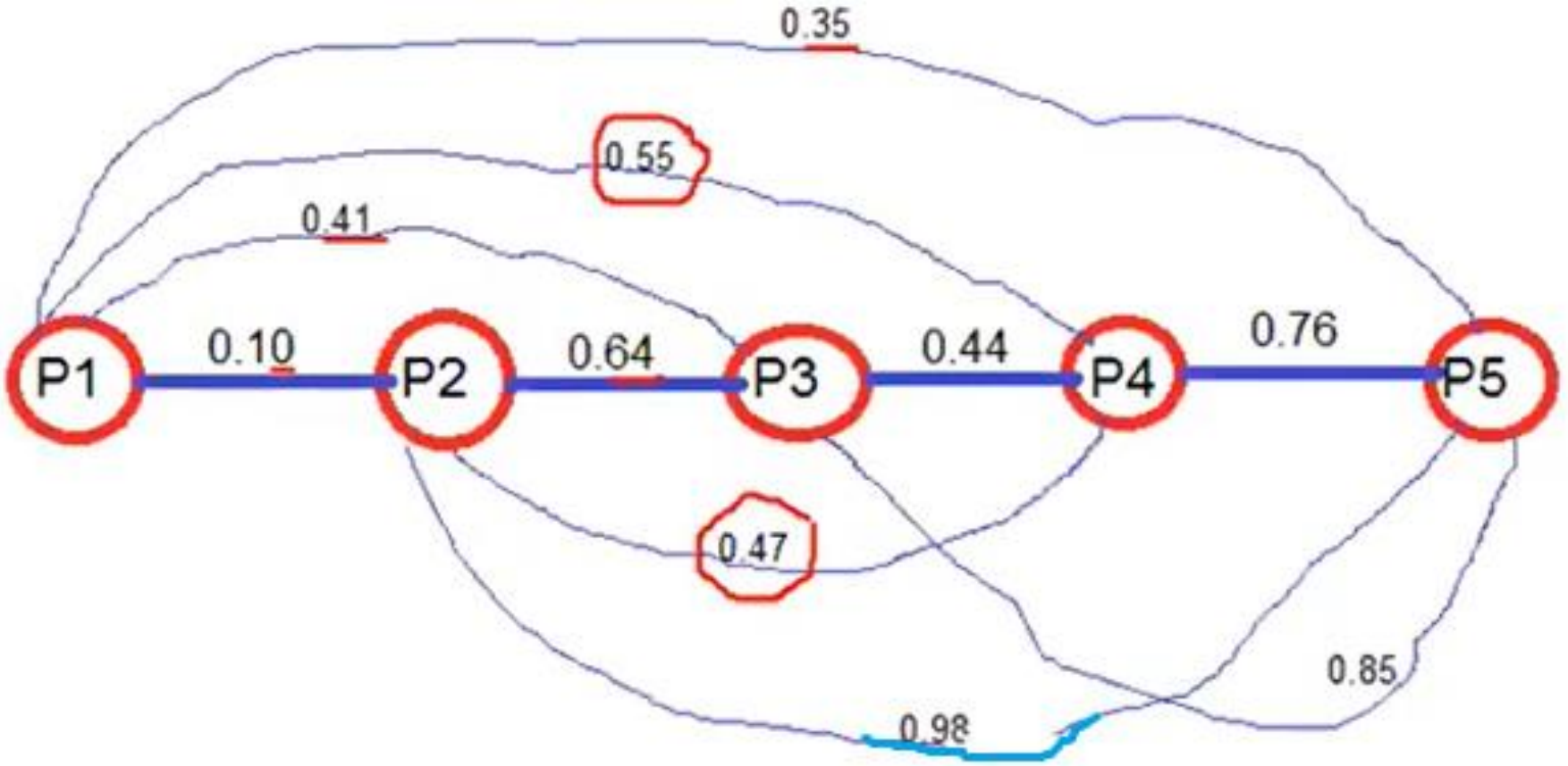
The total **variance** of clusters A and C Is **lower** than the total **variance** of clusters A and B.

Hierarchal Clustering 1.

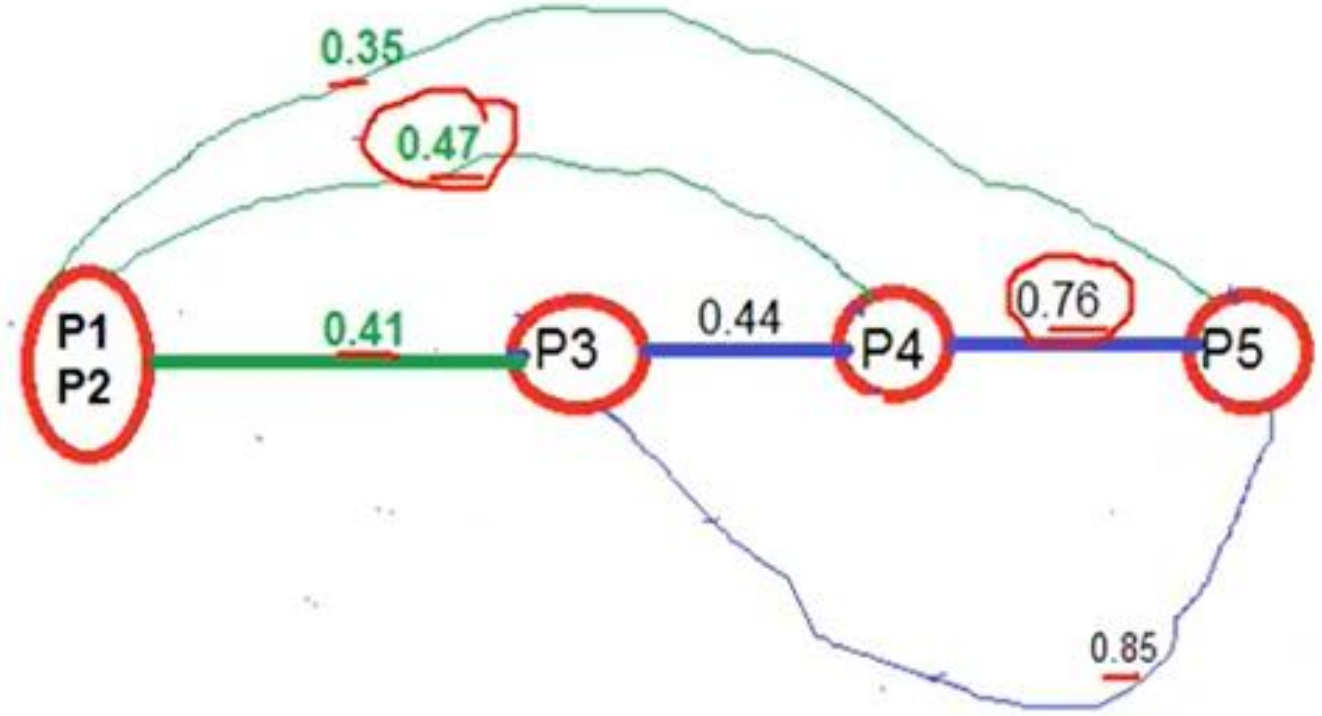
- Use the Euclidean distance matrix to perform MIN agglomerative clustering of five data points p1,p2,p3,p4,p5 (MIN is a cluster proximity measures) Draw the associated Dendogram.

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

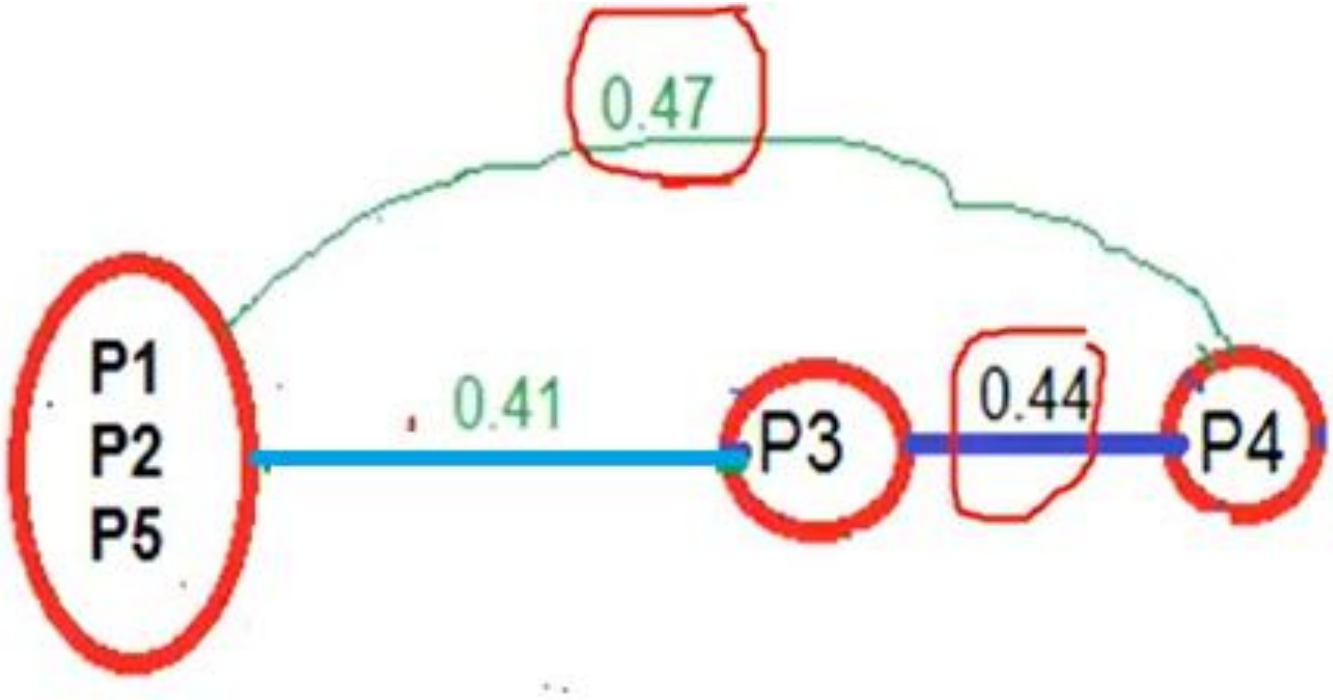
Hierarchal Clustering 1.



Hierarchal Clustering 1.

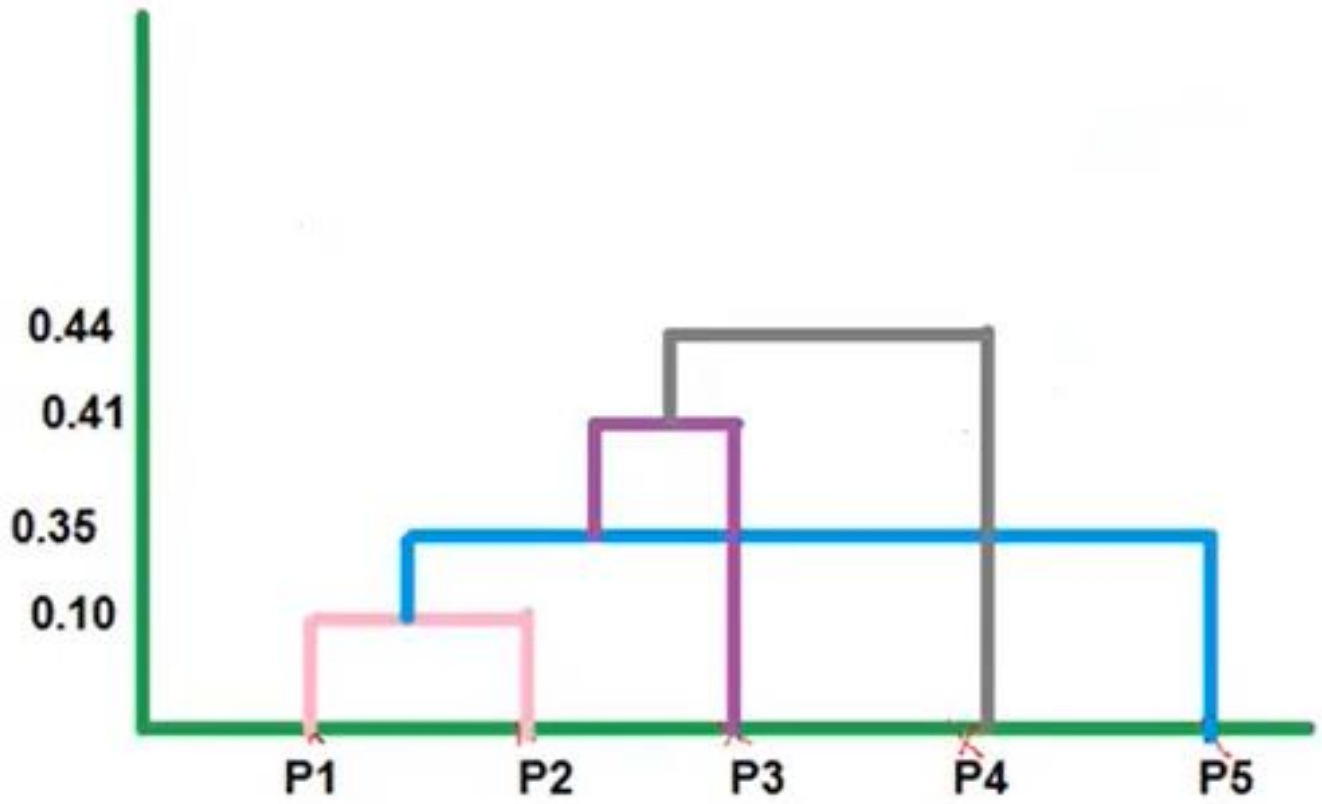


Hierarchal Clustering 1.



Hierarchal Clustering 1.

Hierarchal Clustering 1.



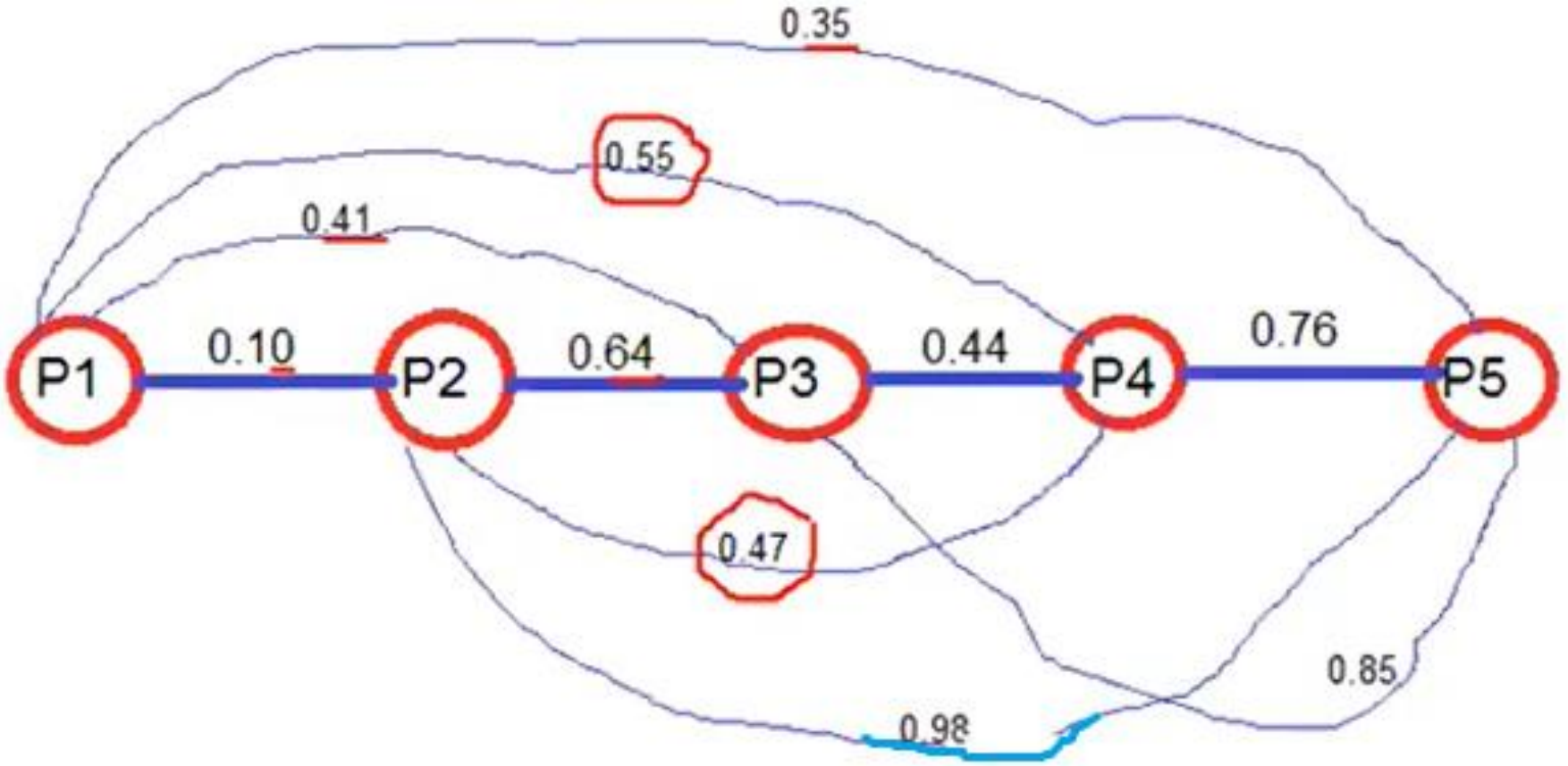
Hierarchal Clustering 1.

- Use the Euclidean distance matrix to perform MAX agglomerative clustering of five data points p1,p2,p3,p4,p5 (MAX is a cluster proximity measures) Draw the associated Dendogram.

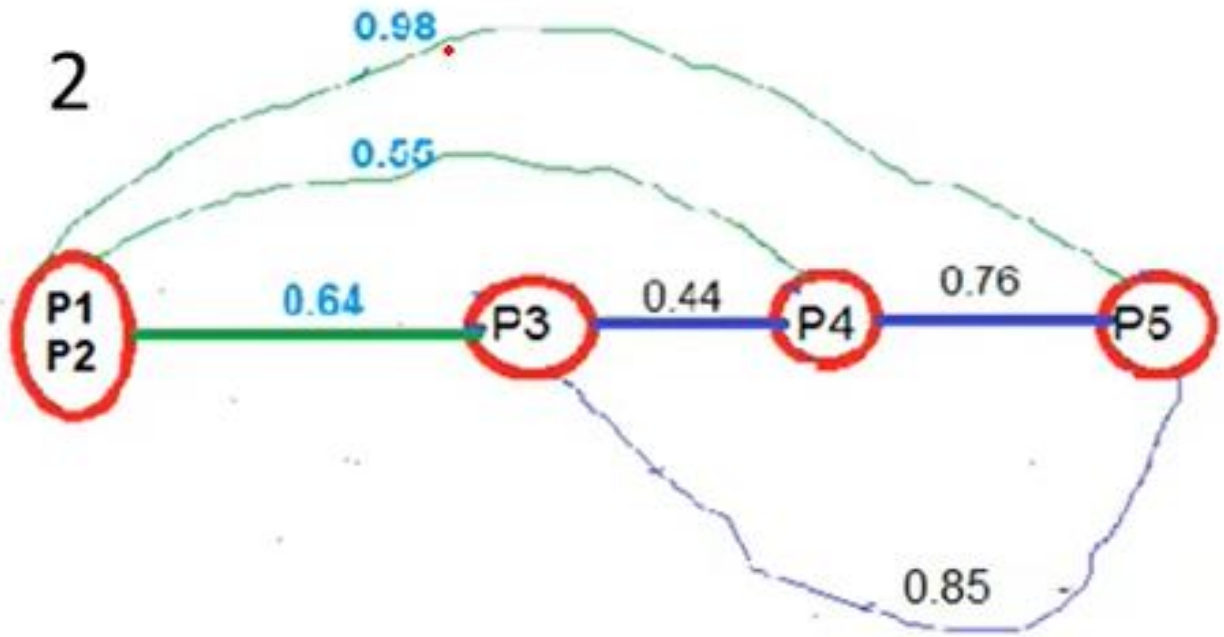
Complete link

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

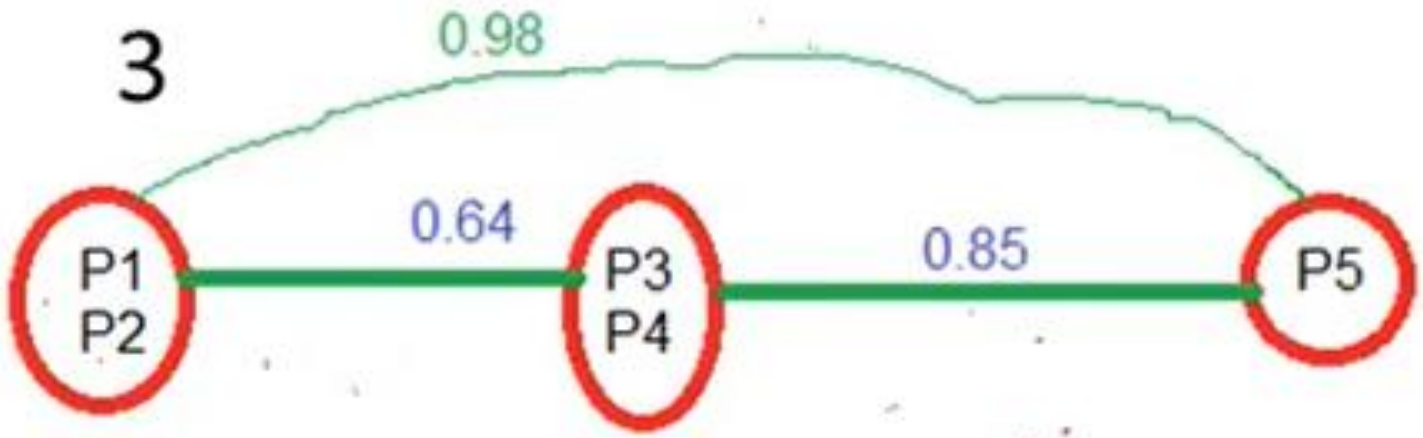
Hierarchal Clustering 1.



Hierarchal Clustering 1.

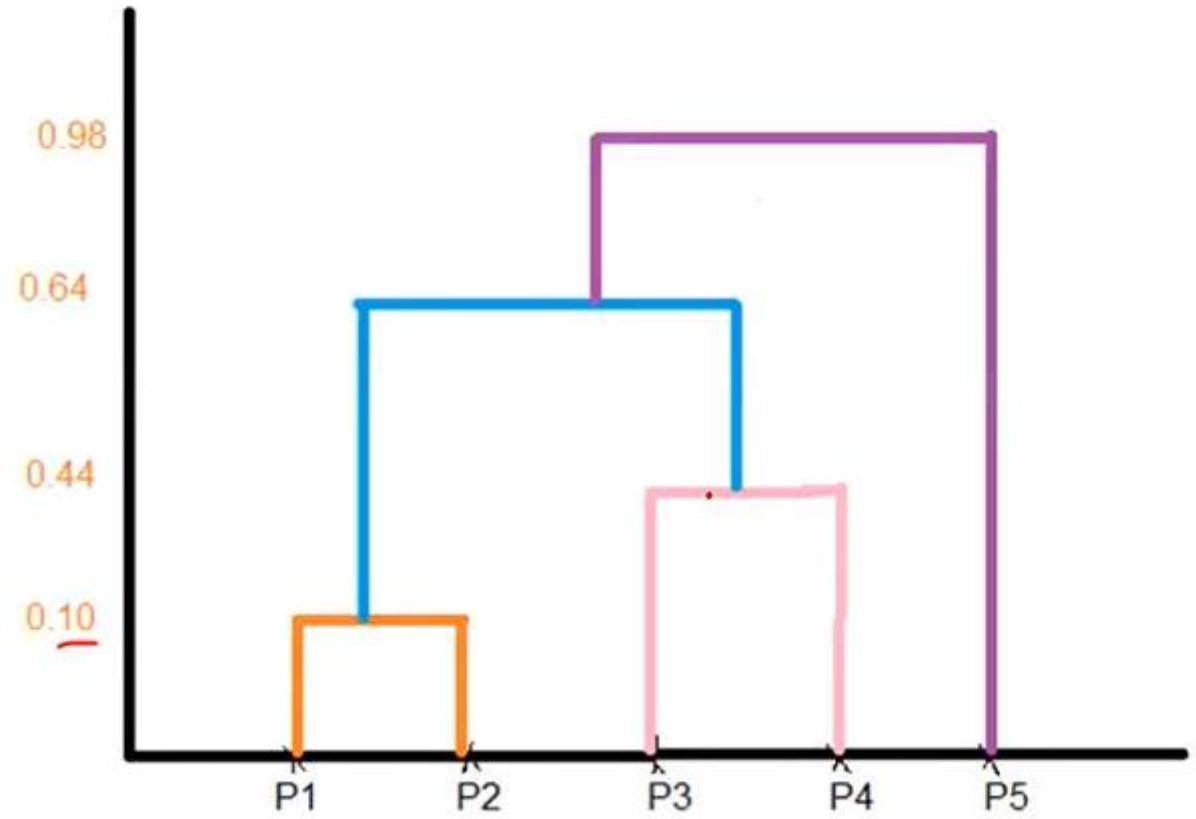


Hierarchal Clustering 1.



Hierarchal Clustering 1.

Hierarchal Clustering 1.



Hierarchal Clustering 1.

Assignment:

Use the Euclidean distance matrix to perform MAX agglomerative clustering of five data points p1,p2,p3,p4,p5 (MAX is a cluster proximity measures) Draw the associated Dendogram.

	P1	P2	P3	P4	P5
P1	0.00	0.55	0.4	0.35	0.2
P2	0.55	0.00	0.7	0.9	0.8
P3	0.4	0.7	0.00	0.85	0.5
P4	0.35	0.9	0.85	0.00	0.6
P5	0.2	0.8	0.5	0.6	0.00

Hierarchal Clustering 1.

Assignment:

Use the Euclidean distance matrix to perform MAX agglomerative clustering of five data points p1,p2,p3,p4,p5 (MAX is a cluster proximity measures) Draw the associated Dendogram.

	P1	P2	P3	P4	P5
P1	0.00	0.55	0.4	0.35	0.2
P2	0.55	0.00	0.7	0.9	0.8
P3	0.4	0.7	0.00	0.85	0.5
P4	0.35	0.9	0.85	0.00	0.6
P5	0.2	0.8	0.5	0.6	0.00

HIERARCHICAL CLUSTERING

Advantages of Hierarchical Clustering:

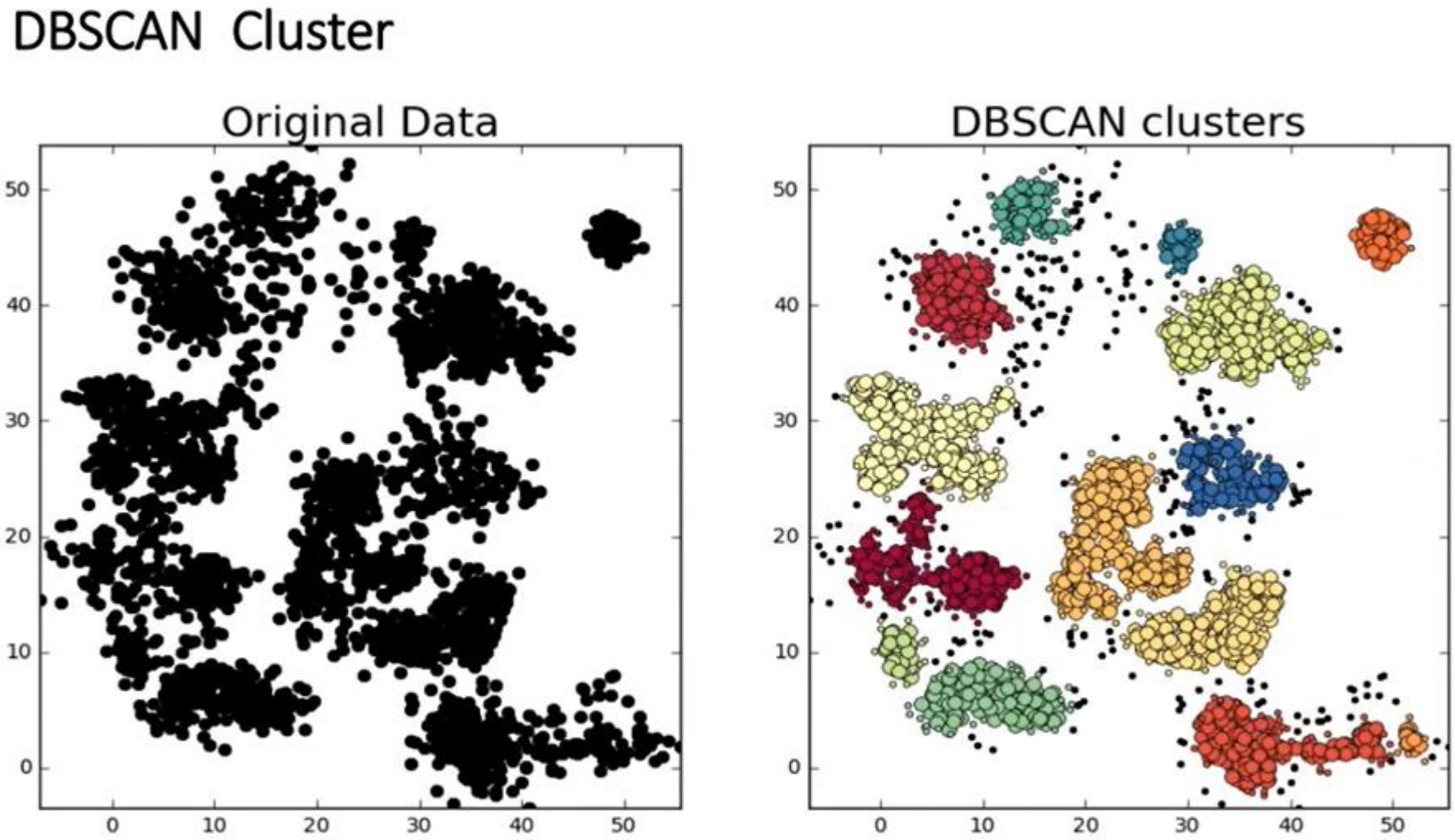
- Don’t have to manually select K prior applying the algorithm.
- Easy to implement.
- Dendrogram is very informative and visually appealing.
- No need to initialize random centroids.

Disadvantages of Hierarchical Clustering:

- Very large datasets can produce difficult dendrogram to be separated.
- Computationally expensive and time consuming more than K-means.
- Sensitivity to outliers.
- Doesn’t work well with missing data.
- Has problems with highly dimensional data because of poor distance measurement quality in the hyper-space.

DBSCAN Cluster

Hierarchal Clustering 1.



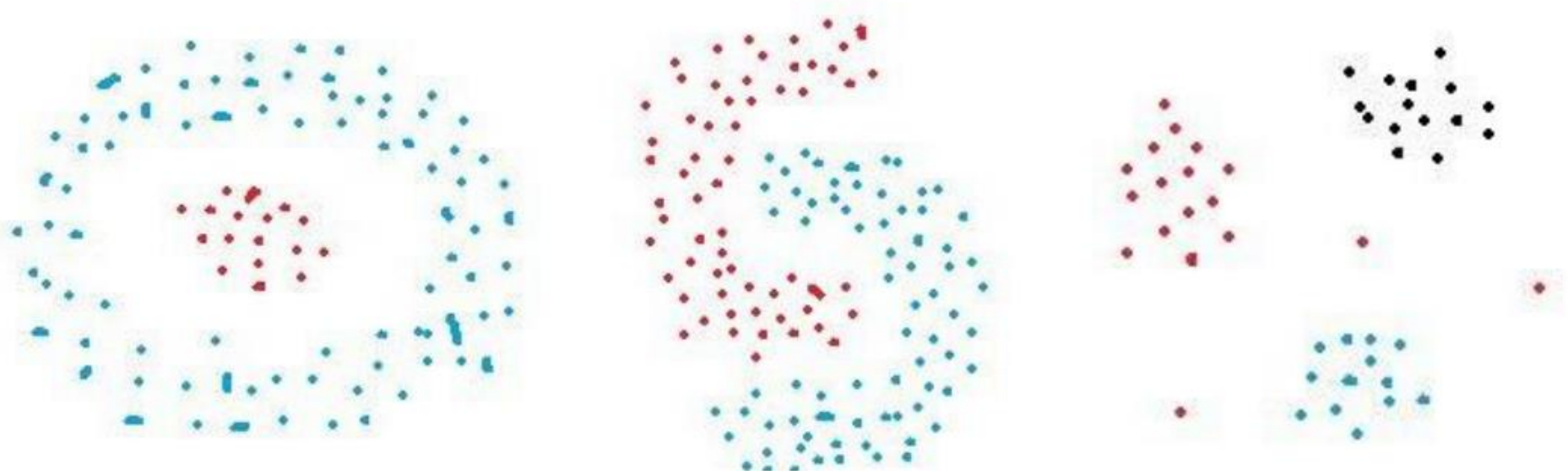
DBSCAN Clustering:

- ➔ It stands for **Density-Based Spatial Clustering of Applications with Noise**.
- ➔ It is an example of a **Density-Based Clustering**.
- ➔ In this algorithm, the areas of high density are separated by the areas of low density.
- ➔ The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.
- ➔ Because of this, the clusters can be found in any arbitrary shape.

Hierarchal Clustering 1.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN):

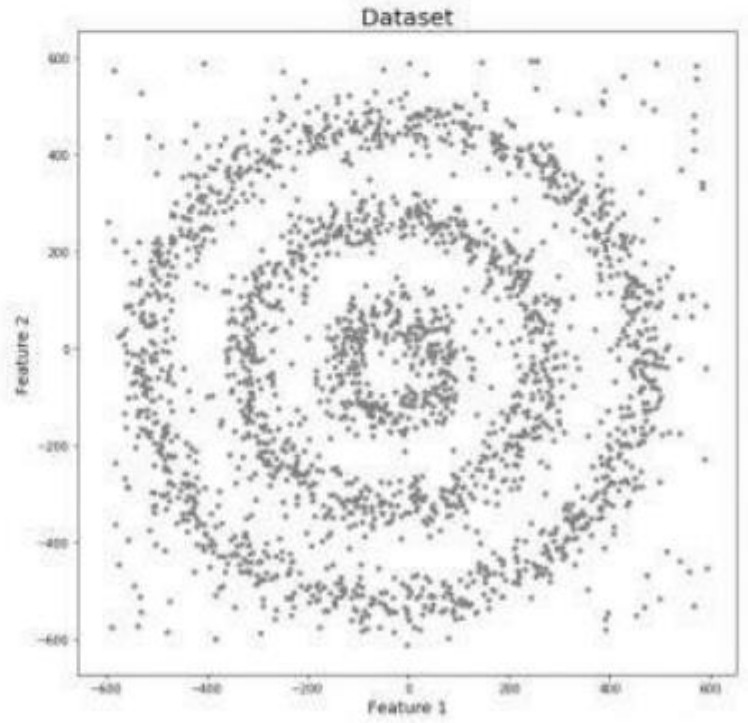
DBSCAN is able to find arbitrary shaped clusters as shown below where the point will belong to a cluster if it is close to many points in this cluster (high density). Moreover it is able to handle outliers easily.



Hierarchal Clustering 1.

Does K-means and Hierarchical clustering fails with arbitrary shaped clusters?

YES, they fail both however Hierarchical clustering is a little bit better than K-means in dealing with arbitrary shaped data but eventually both are weak in comparison with DBSCAN. For example, we have the following data points in a form of concentric (having the same centers) circles as shown in the figure:



Why DBSCAN Clustering?

Problems with **Partitioning** and **Hierarchical** clustering:

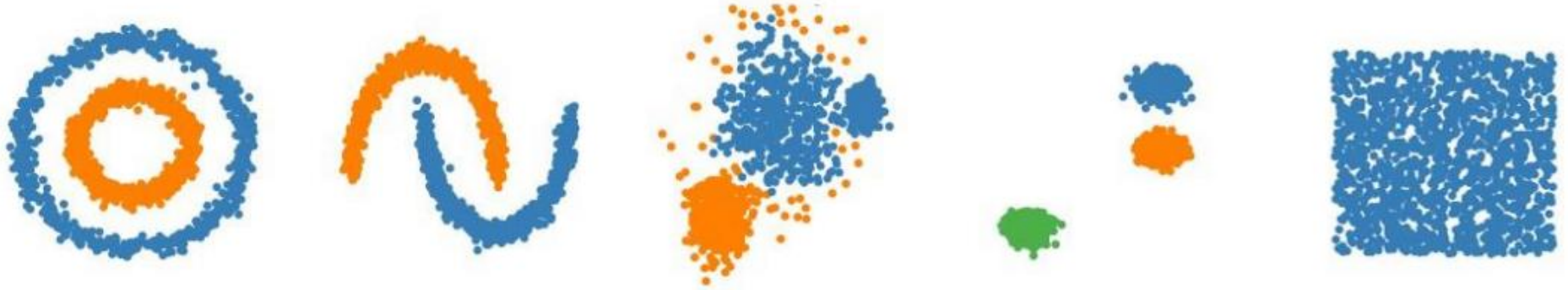
- ➔ In K-Means Clustering, you need to specify the number of clusters to use
- ➔ K-Means Clustering is sensitive towards outlier, so clusters are severely affected by the presence of noise and outliers.
- ➔ Partitioning and Hierarchical clustering work for finding spherical-shaped clusters. They are suitable only for compact and well-separated clusters.

Why DBSCAN Clustering?

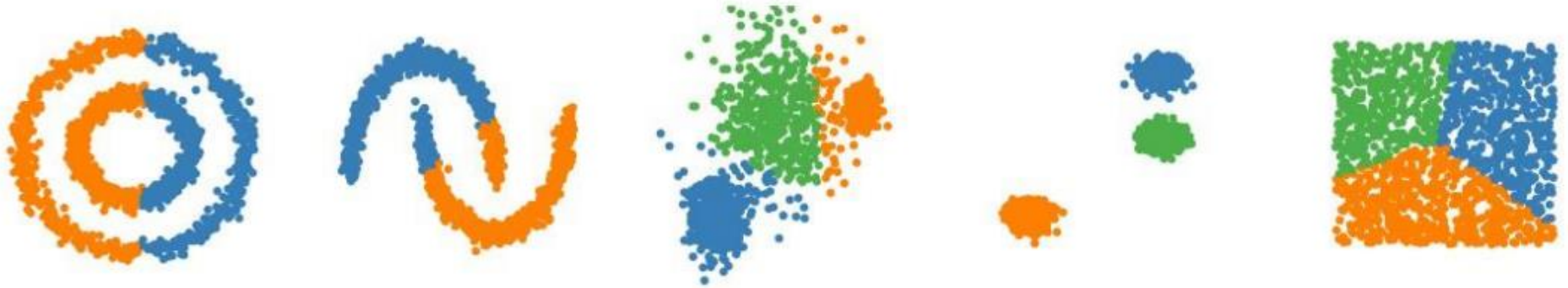
- ➔ These problems are greatly reduced in **DBSCAN** due to the way clusters are formed.
- ➔ Also, in **DBSCAN** we don't have to specify the number of clusters to use.
- ➔ All you need is a function to calculate the distance between values, and some guidance for what amount of distance is considered "close".
- ➔ DBSCAN works well with clusters of arbitrary shapes.
- ➔ It also works well with noisy data.

Why DBSCAN Clustering?

DBSCAN



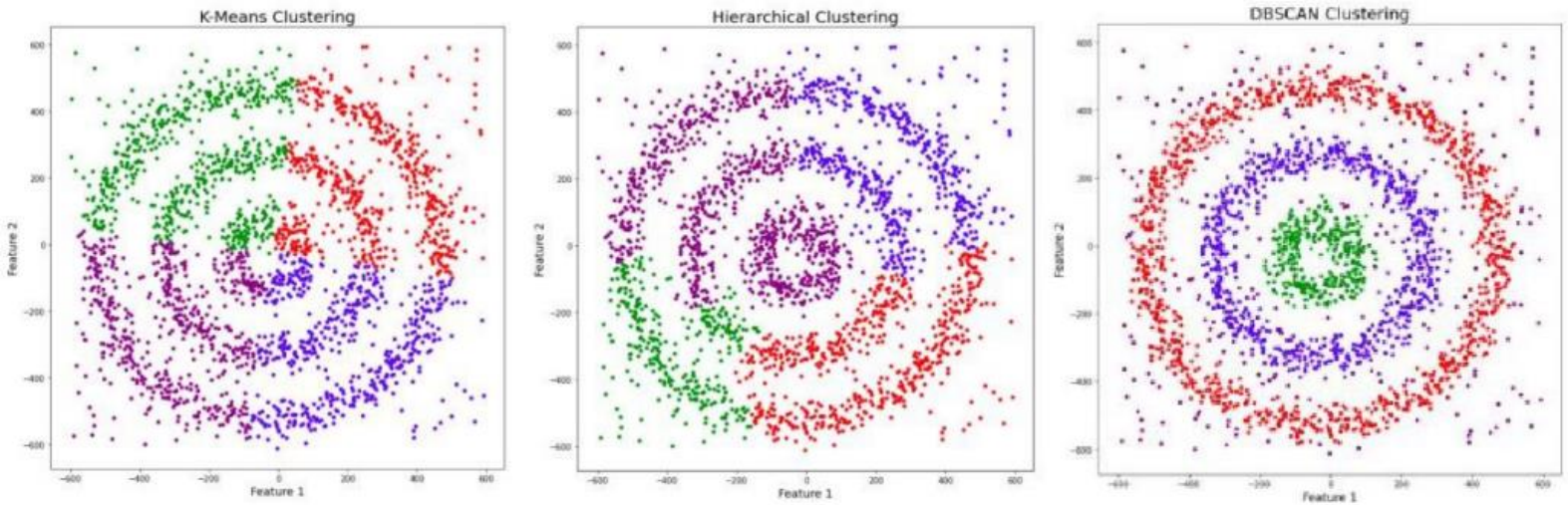
k-means



Hierarchal Clustering 1.

DENSITY-BASED CLUSTERING

What would be the result for each of the three clustering techniques?



We can notice that both K-means and Hierarchical clustering failed to segment the concentric circles to 3 clusters in addition to a noise cluster (Purple) which was easy for the DBSCAN.

DENSITY-BASED CLUSTERING

What DBSCAN needs as pre-defined parameters?

Epsilon(eps) which is the maximum distance that we consider to label a point as a neighbour to this data sample.

Minimum points(minPts) which states the minimum number of points to define a cluster.

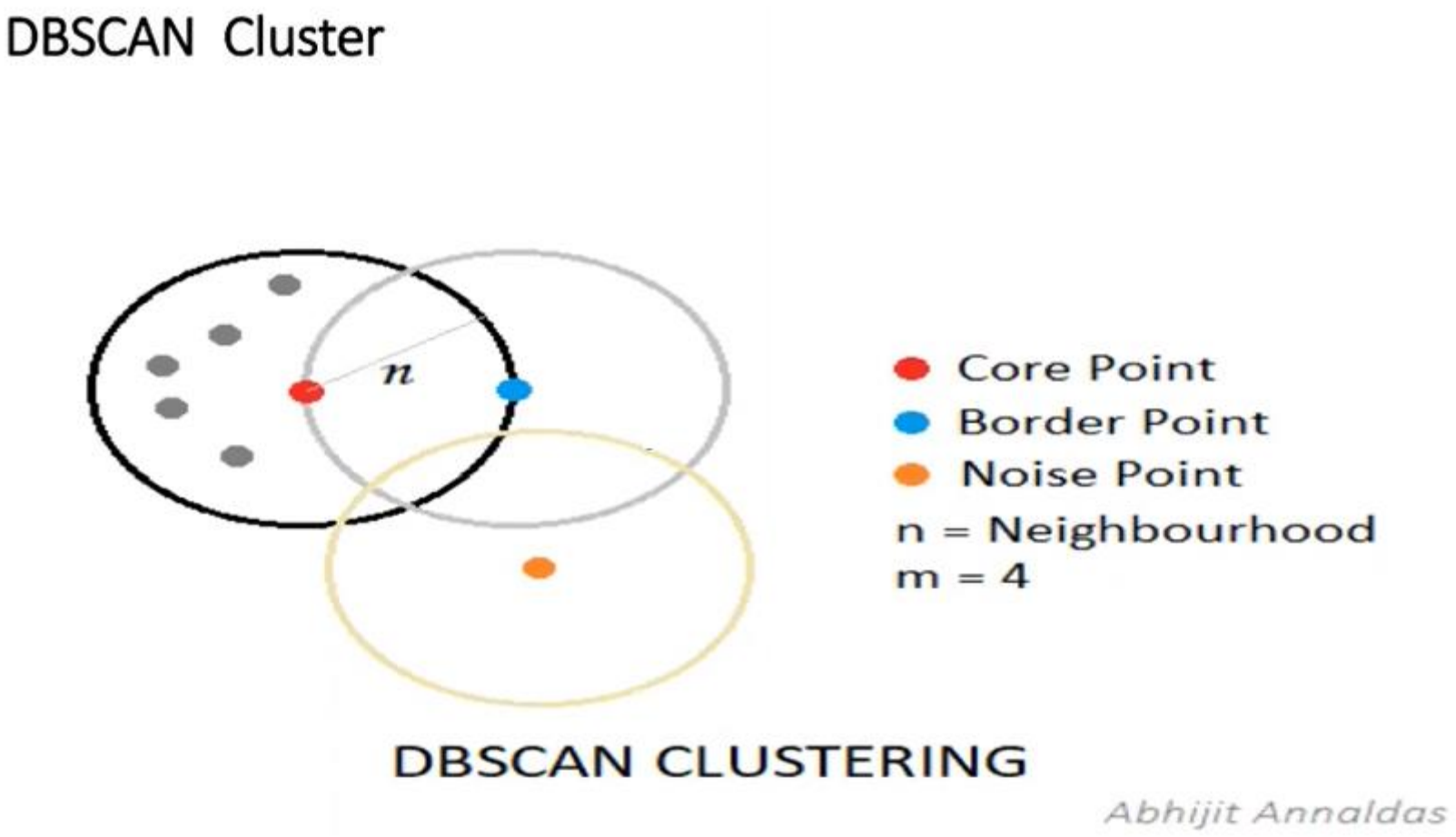
Terminologies:

Core point is a point where there are at least minPts number of points (including the point itself) in the surrounding area with a radius of eps.

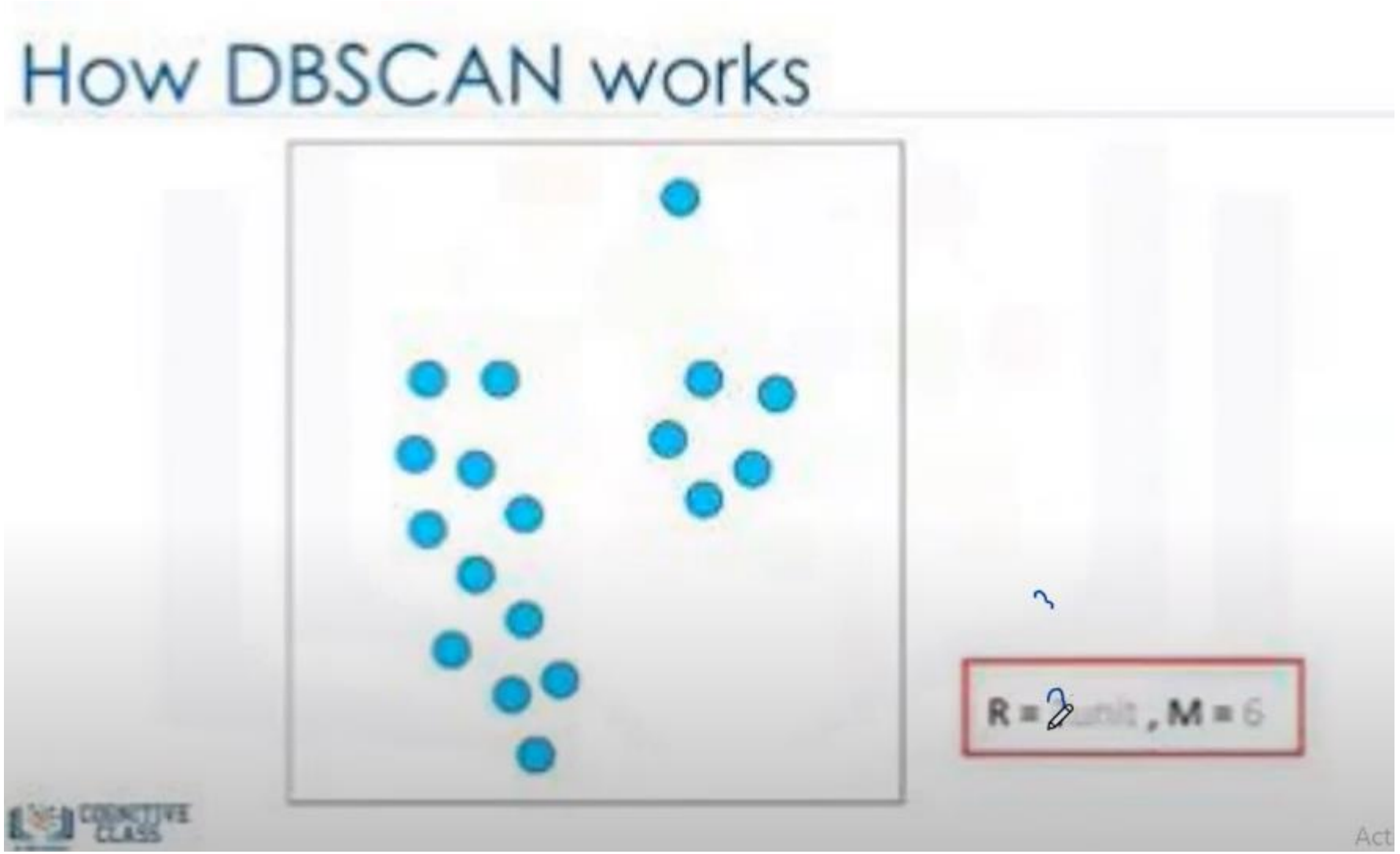
Border point is a point reachable by core point but has less than minPts number of points in the surrounding area with radius of eps.

Outlier/Noise is a point which is not either core or border point.

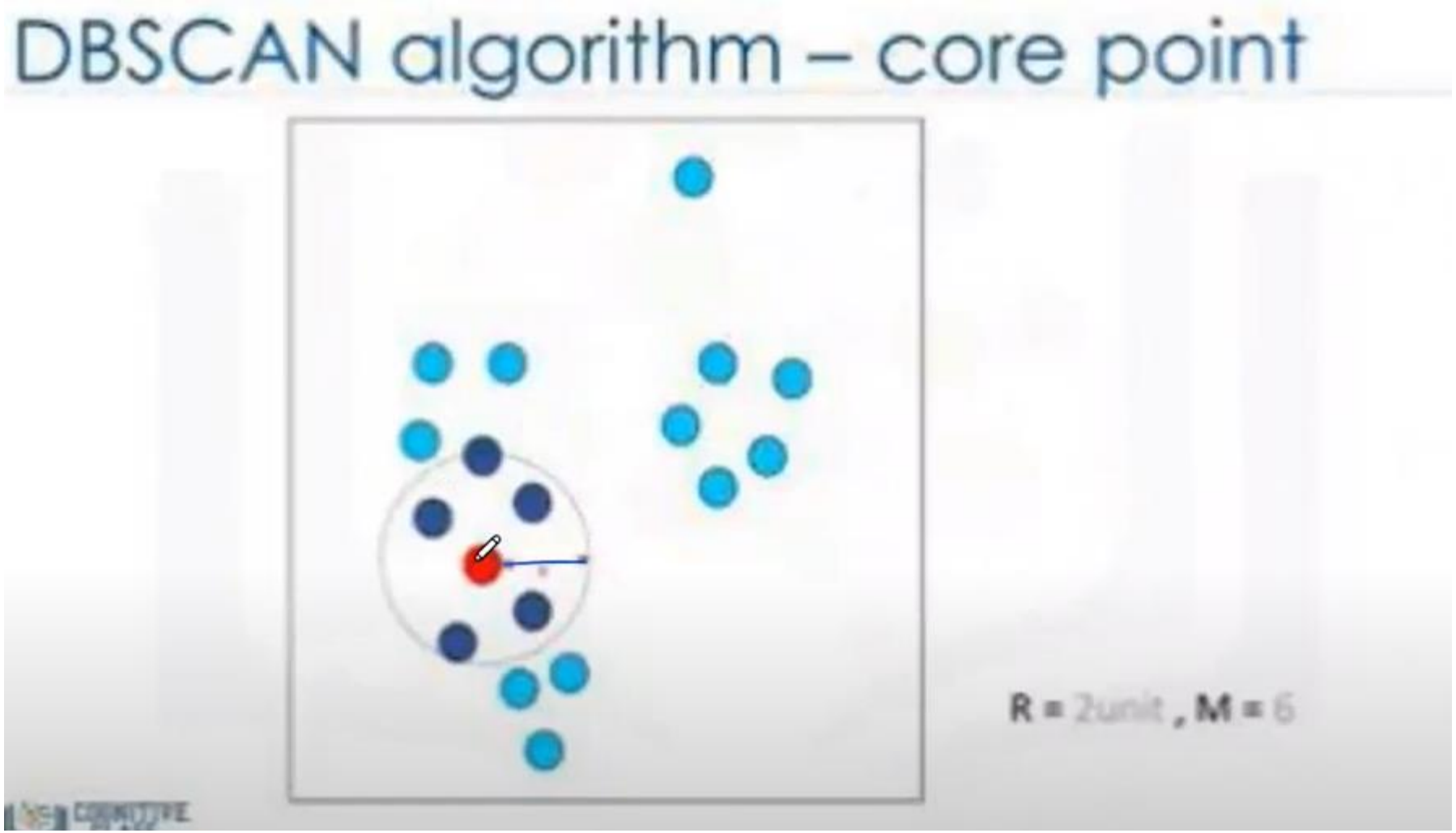
Hierarchal Clustering 1.



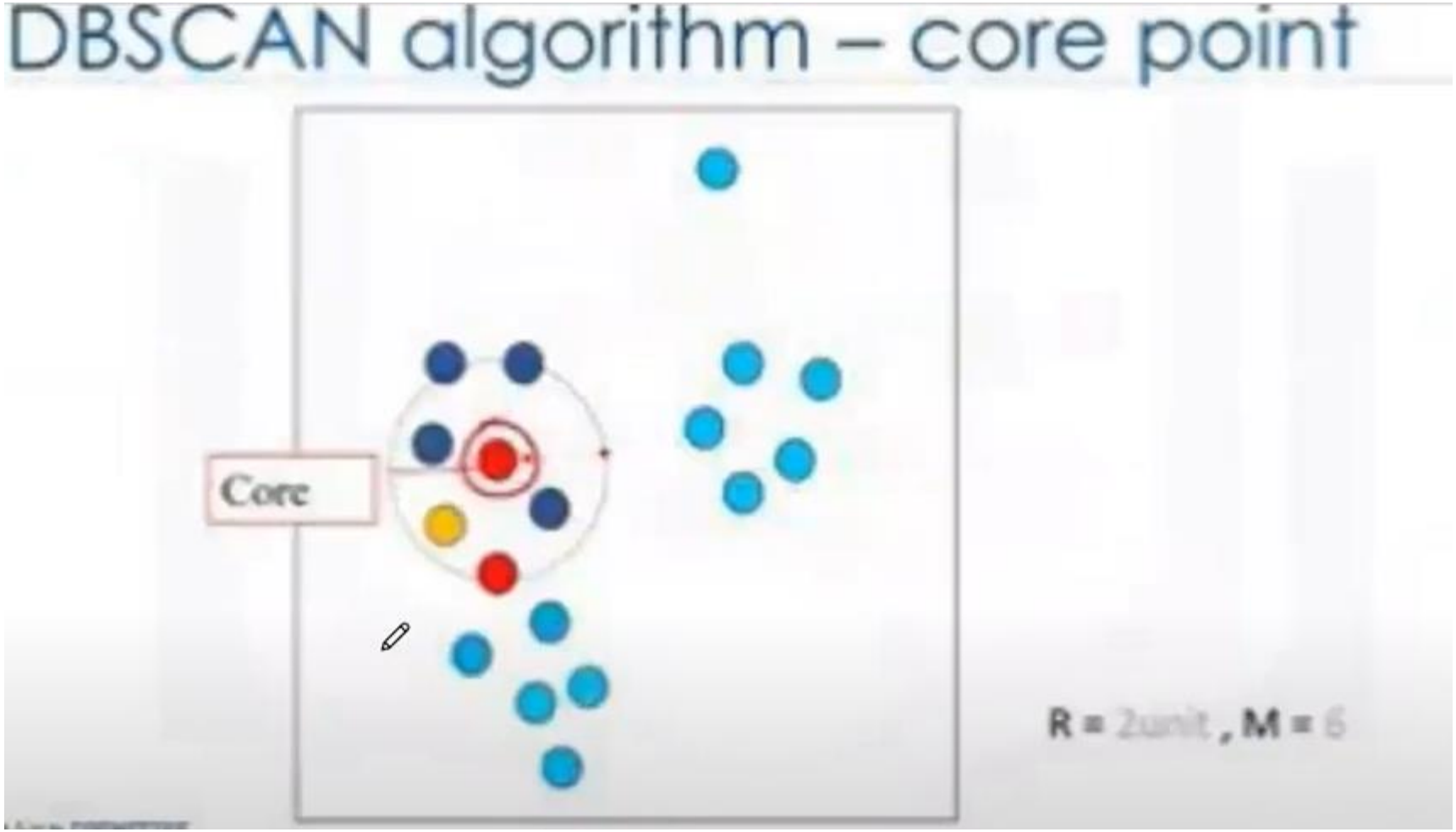
Hierarchal Clustering 1.



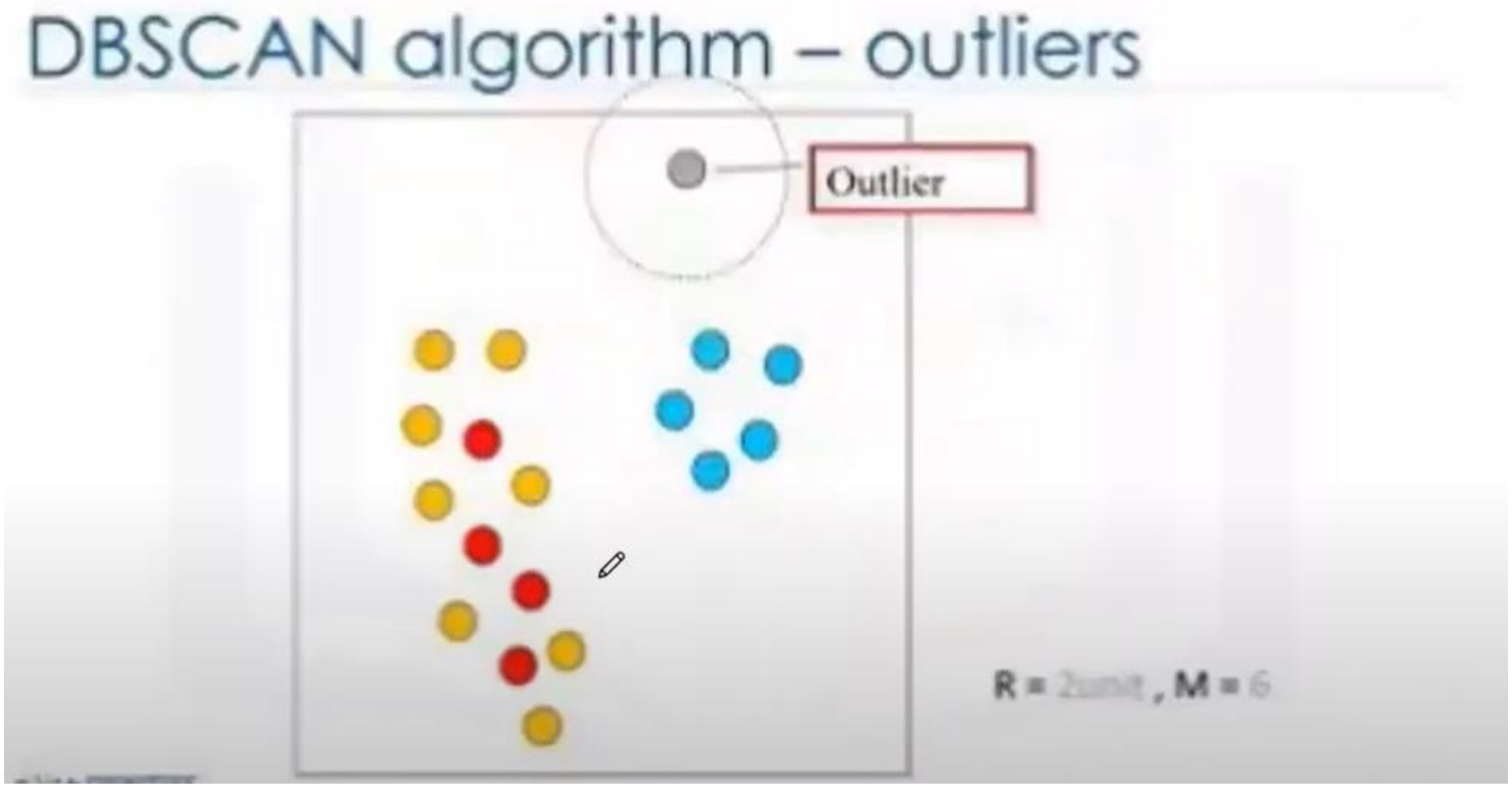
Hierarchal Clustering 1.



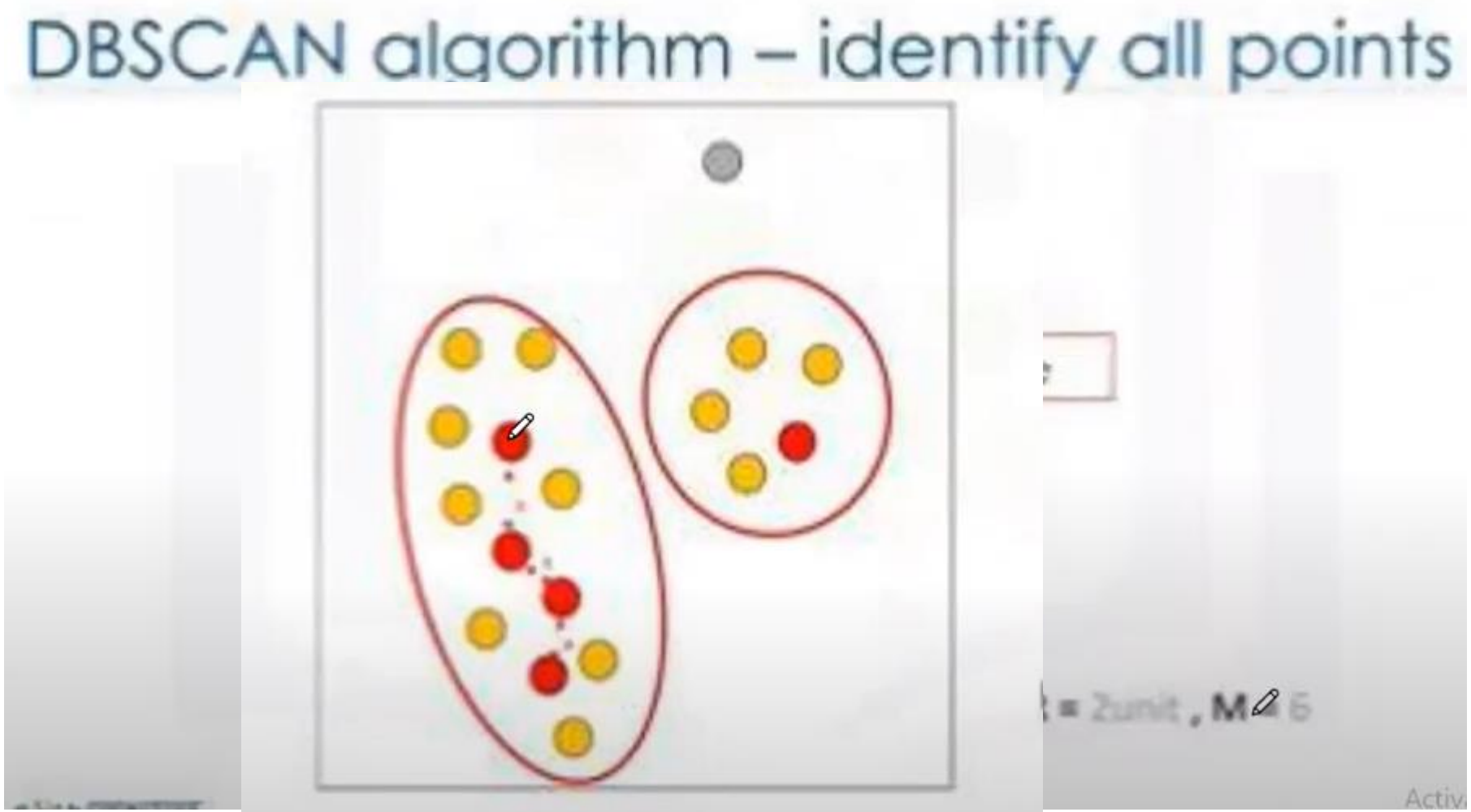
Hierarchal Clustering 1.



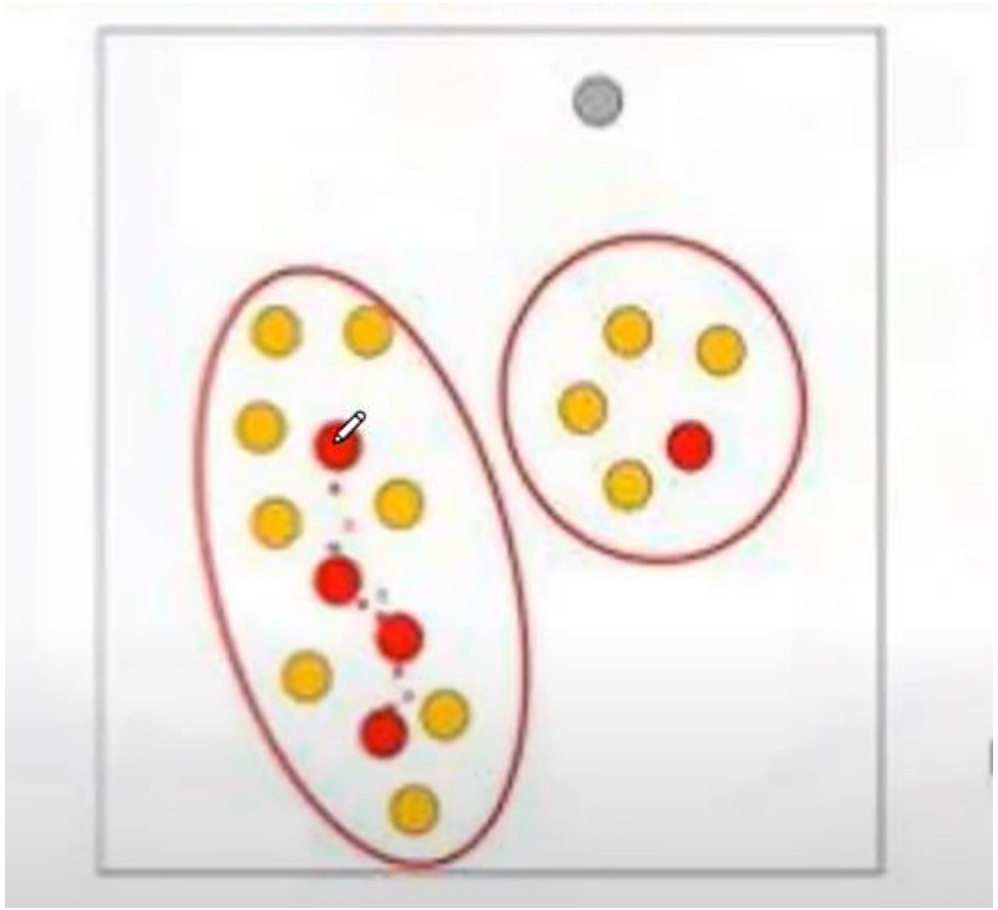
Hierarchal Clustering 1.



Hierarchal Clustering 1.

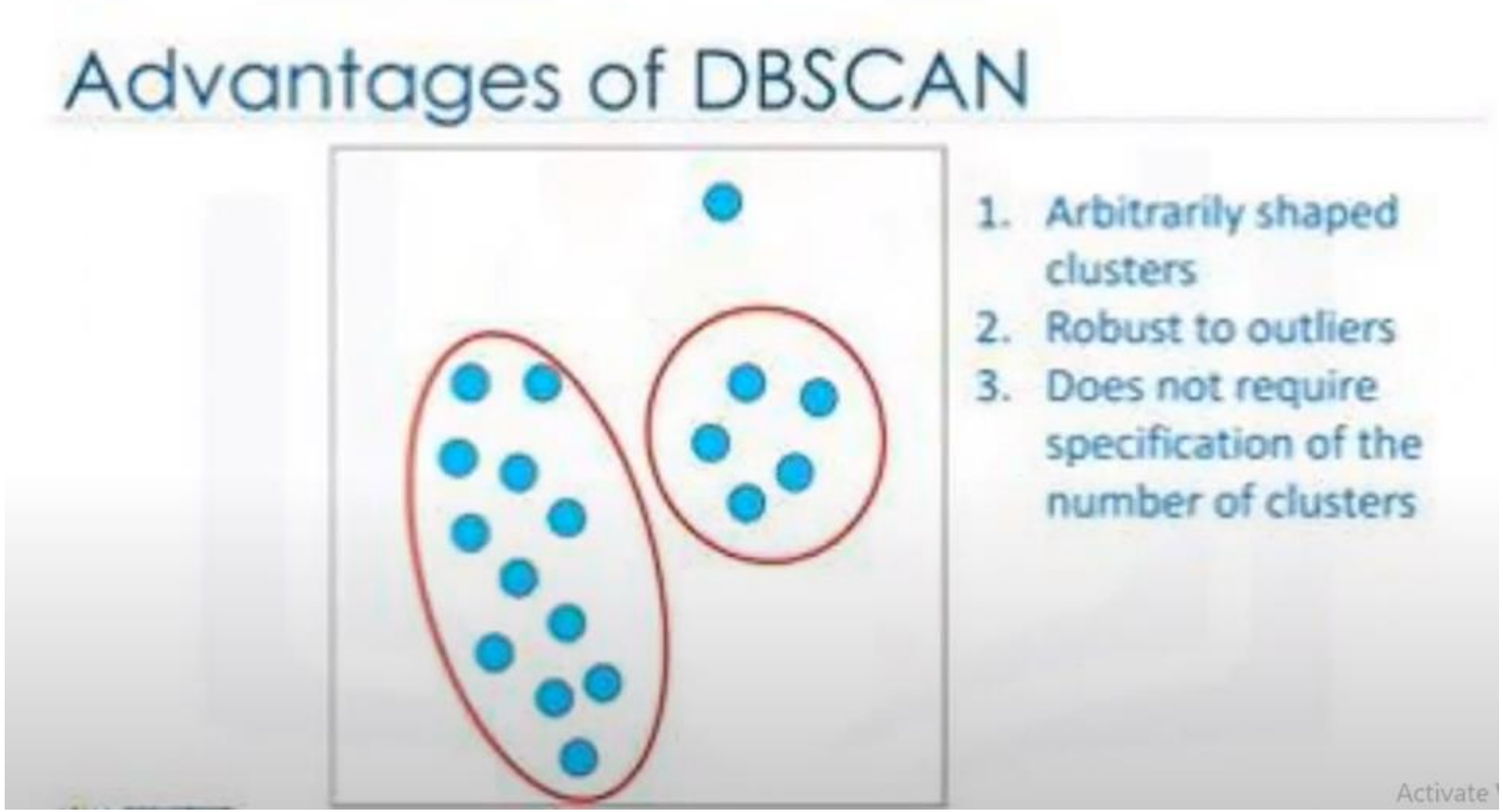


Hierarchal Clustering 1.



Hierarchal Clustering 1.

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/



DENSITY-BASED CLUSTERING

Advantages of DBSCAN:

- Doesn't require prior information of data for determining number of clusters.
- Handles clusters at various shapes and sizes (arbitrary shaped clusters)
- Robust to outliers.

Disadvantages of DBSCAN:

- Sensitive to eps and minPts values. If data is not well understood, estimating such parameters could be problematic.
- Not deterministic as border points that is accessed by several clusters can be either a part of any cluster based on the order the data processed.
- Fails at very high dimensional data due to curse of dimensionality.

Complexity in DBSCAN Clustering:

- ➔ DBSCAN algorithm is somewhat **more difficult to tune** contrasted to K-Means.
- ➔ Parameters like the **epsilon for DBSCAN are less intuitive** to reason about, compared to the number of clusters parameter for K-Means.
- ➔ So, it's more **difficult to choose good initial parameter** values for DBSCAN algorithm.

THANK YOU!