Data Mining & Knowledge Discovery

Clustering

Md. Biplob Hosen

Lecturer, IIT-JU

Email: biplob.hosen@juniv.edu

Supervised vs. Unsupervised Learning

Supervised Machine Learning:

- Supervised learning is a machine learning method in which models are trained using labeled data. In supervised learning, models need to find the mapping function to map the input variable (X) with the output variable (Y).
- Supervised learning needs supervision to train the model, which is similar to as a student learns things in the presence of a teacher. Supervised learning can be used for two types of problems: **Classification and Regression.**
- **Example:** Suppose we have an image of different types of fruits. The task of our supervised learning model is to identify the fruits and classify them accordingly. So to identify the image in supervised learning, we will give the input data as well as output for that, which means we will train the model by the shape, size, color, and taste of each fruit. Once the training is completed, we will test the model by giving the new set of fruit. The model will identify the fruit and predict the output using a suitable algorithm.

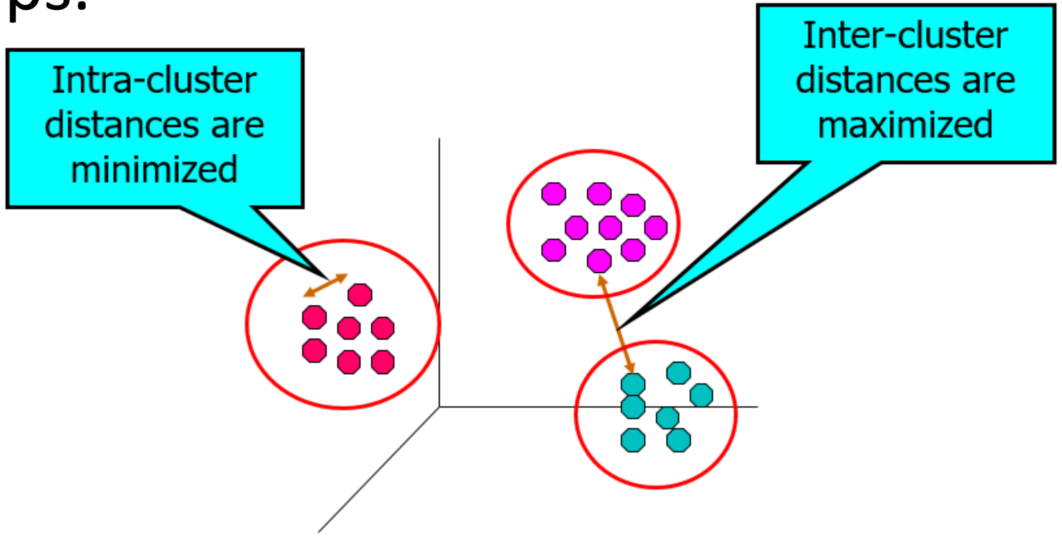
Supervised vs. Unsupervised Learning

Unsupervised Machine Learning:

- Unsupervised learning is another machine learning method in which patterns inferred from the unlabeled input data. The goal of unsupervised learning is to find the structure and patterns from the input data. Unsupervised learning does not need any supervision. Instead, it finds patterns from the data by its own. Unsupervised learning can be used for two types of problems: **Clustering and Association.**
- **Example:** To understand the unsupervised learning, we will use the example given above. So unlike supervised learning, here we will not provide any supervision to the model. We will just provide the input dataset to the model and allow the model to find the patterns from the data. With the help of a suitable algorithm, the model will train itself and divide the fruits into different groups according to the most similar features between them.

What is Cluster Analysis?

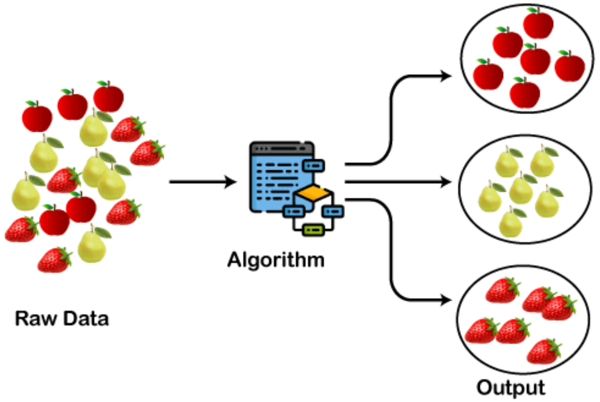
- Given a set of objects, place them in groups such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups.
- **Example:** Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together.


Intra-cluster distances are minimized
Inter-cluster distances are maximized

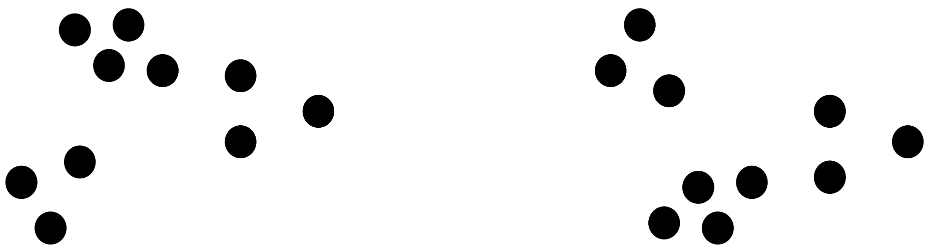
- Such as the t-shirts are grouped in one section, and trousers are at other sections, similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily find out the things. The clustering technique also works in the same way. Other examples of clustering are grouping documents according to the topic.

Clustering Applications

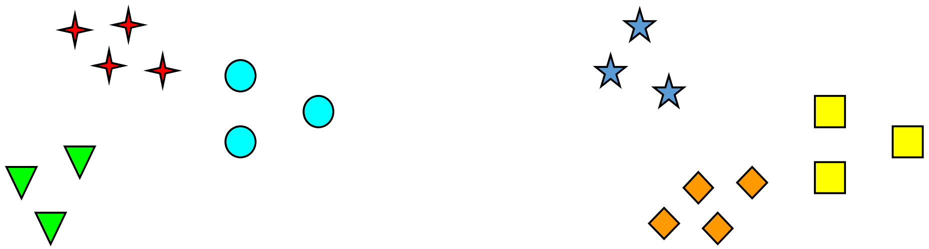
- The clustering technique can be widely used in various tasks. Some most common uses of this technique are:
  1. Market Segmentation
  2. Statistical data analysis
  3. Social network analysis
  4. Image segmentation
  5. Anomaly detection, etc.
- Apart from these general usages, it is used by the **Amazon** in its recommendation system to provide the recommendations as per the past search of products. **Netflix** also uses this technique to recommend the movies and web-series to its users as per the watch history.



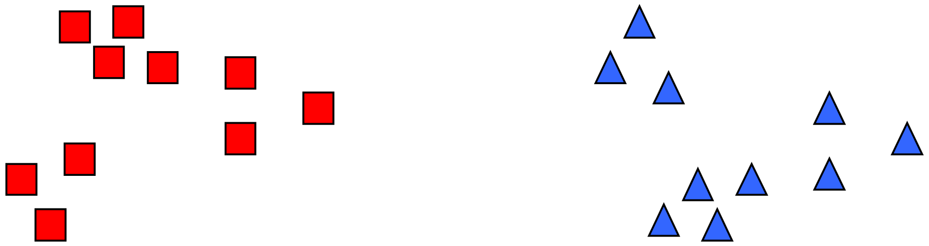
Notion of a Cluster can be Ambiguous



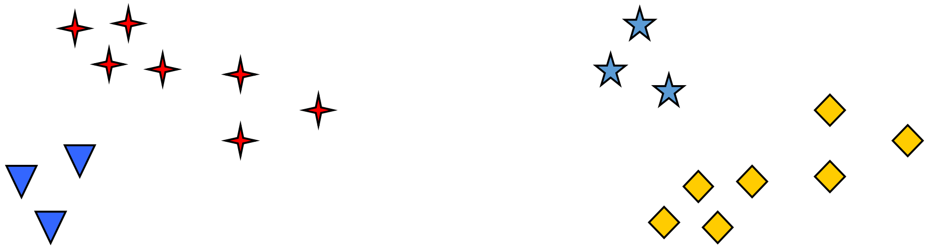
How many clusters?



Six Clusters



Two Clusters

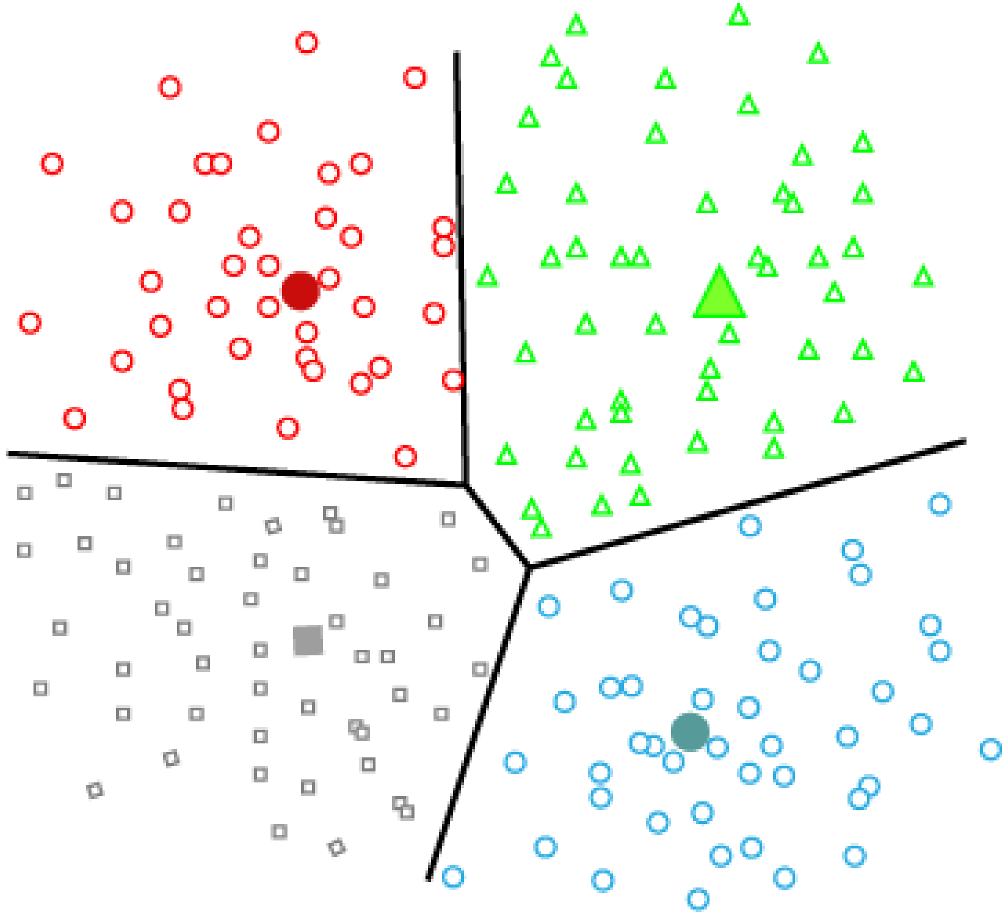


Four Clusters

Types of Clustering Methods – Partitioning Clustering

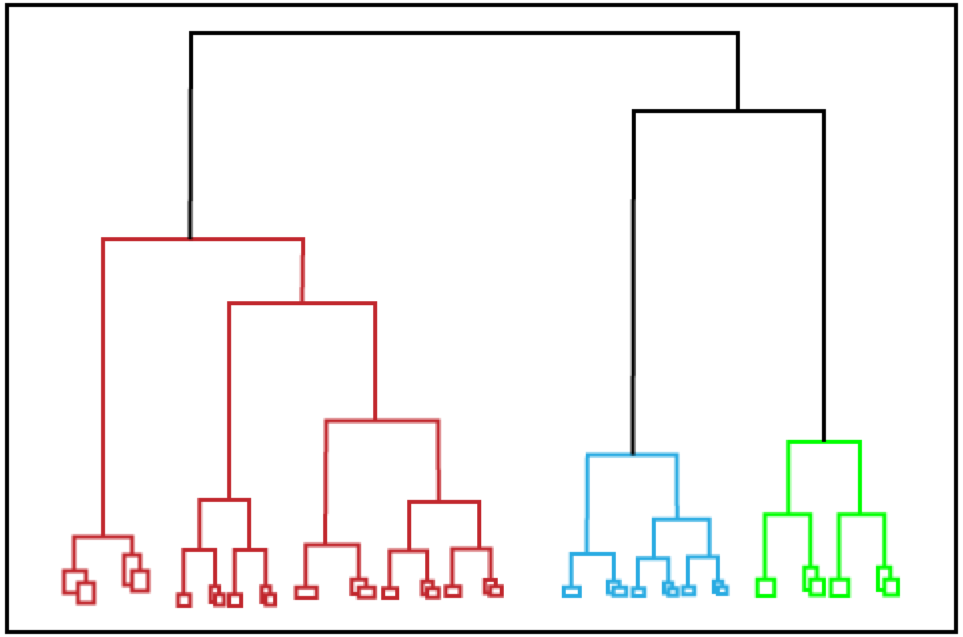
- It is a type of clustering that divides the **centroid-based** data into non-hierarchical group. The most common example of partitioning clustering is the **K-Means Clustering algorithm**.

- In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.



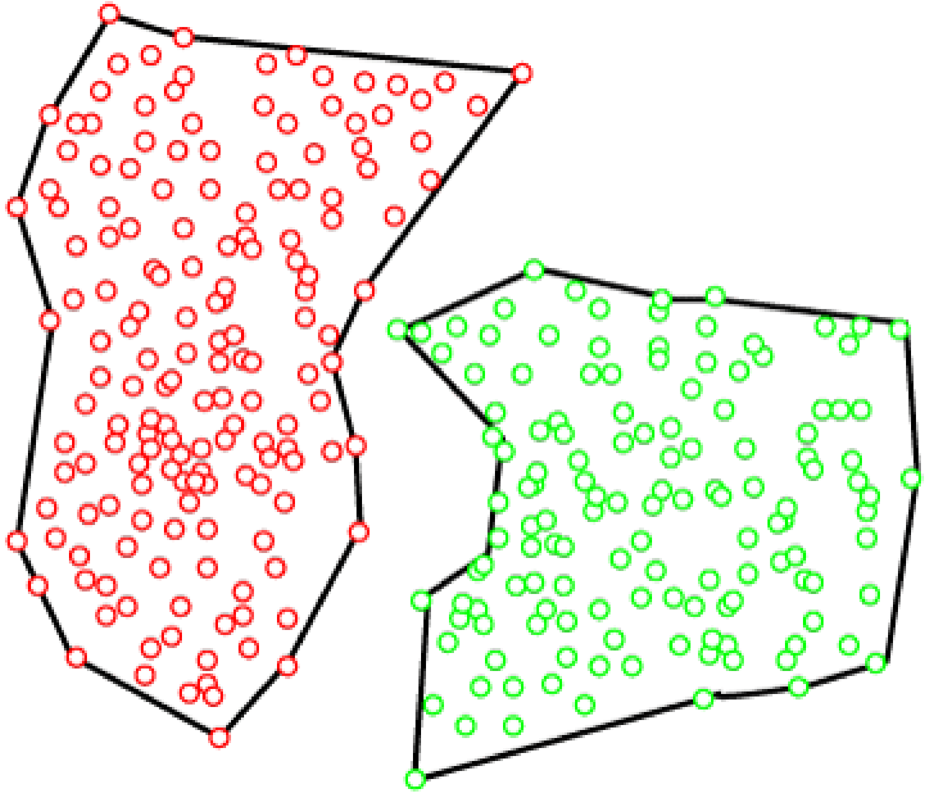
Types of Clustering Methods – Hierarchical Clustering

- Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a dendrogram. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the Agglomerative Hierarchical algorithm.



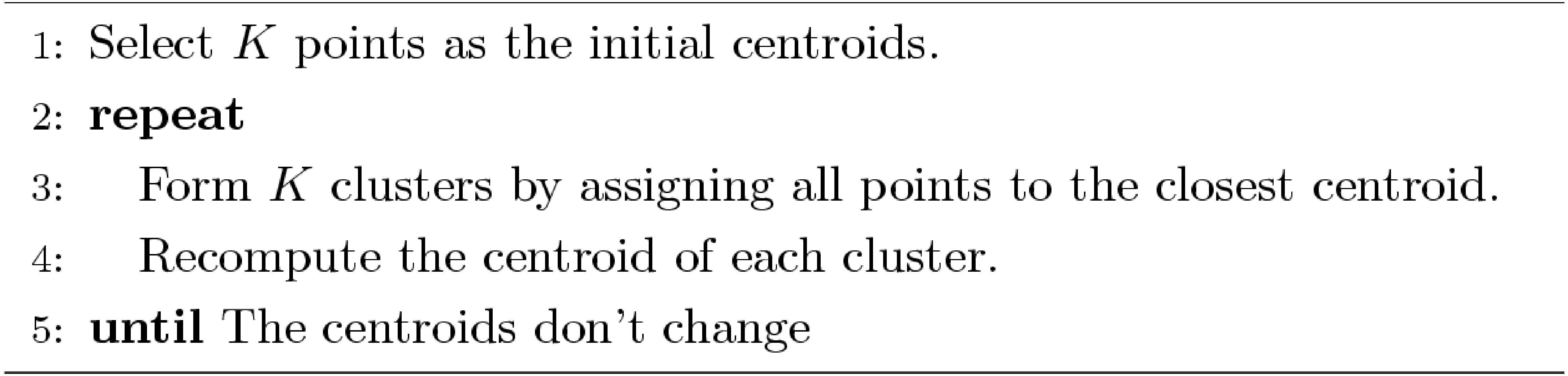
Types of Clustering Methods – Density-Based Clustering

- The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.



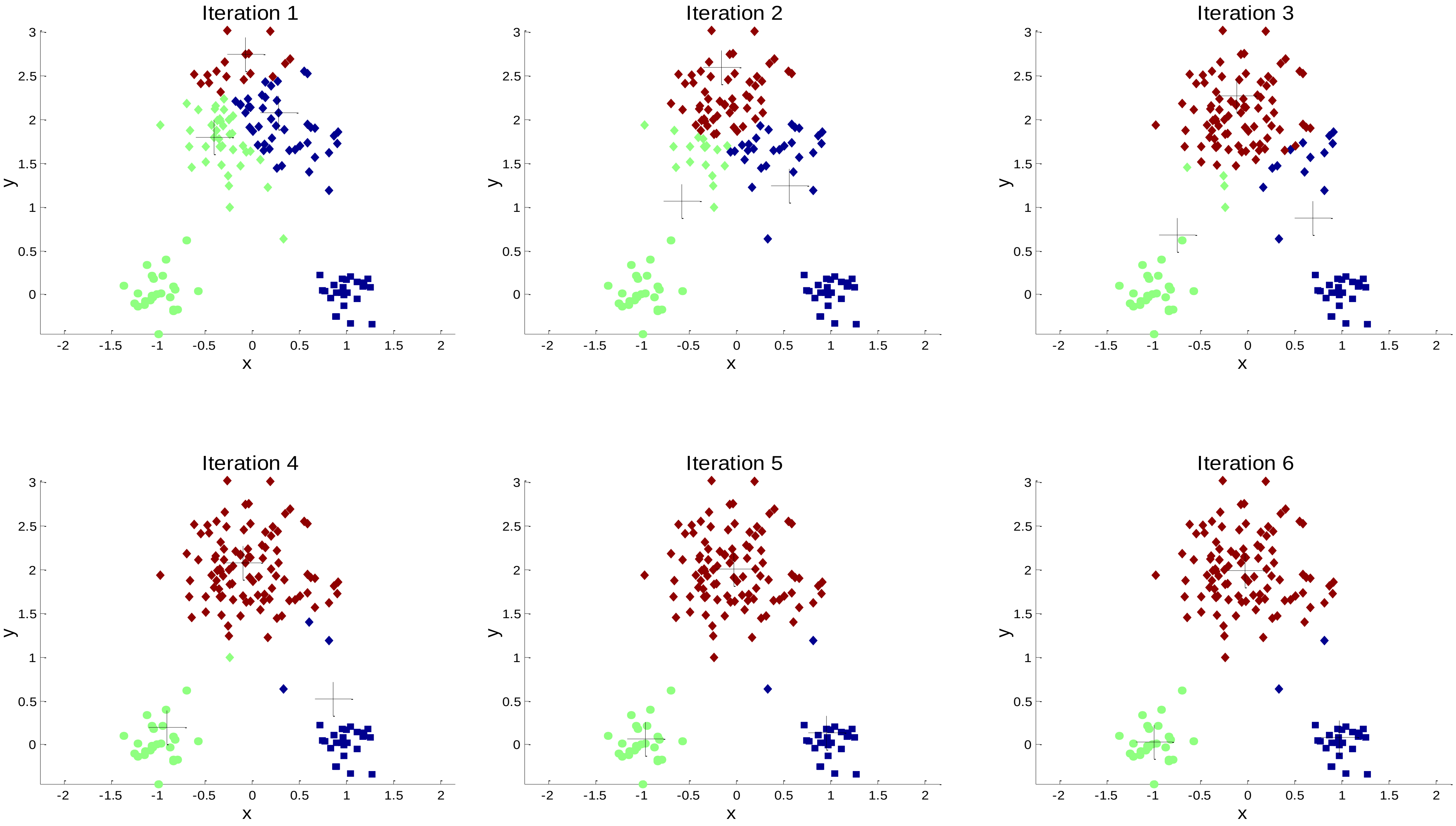
K-means Clustering

- Partitional clustering approach.
- Number of clusters, K, must be specified.
- Each cluster is associated with a centroid (center point).
- Each point is assigned to the cluster with the closest centroid.
- The basic algorithm is very simple.



```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

Example of K-means Clustering



How to choose the value of K?

- The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task.

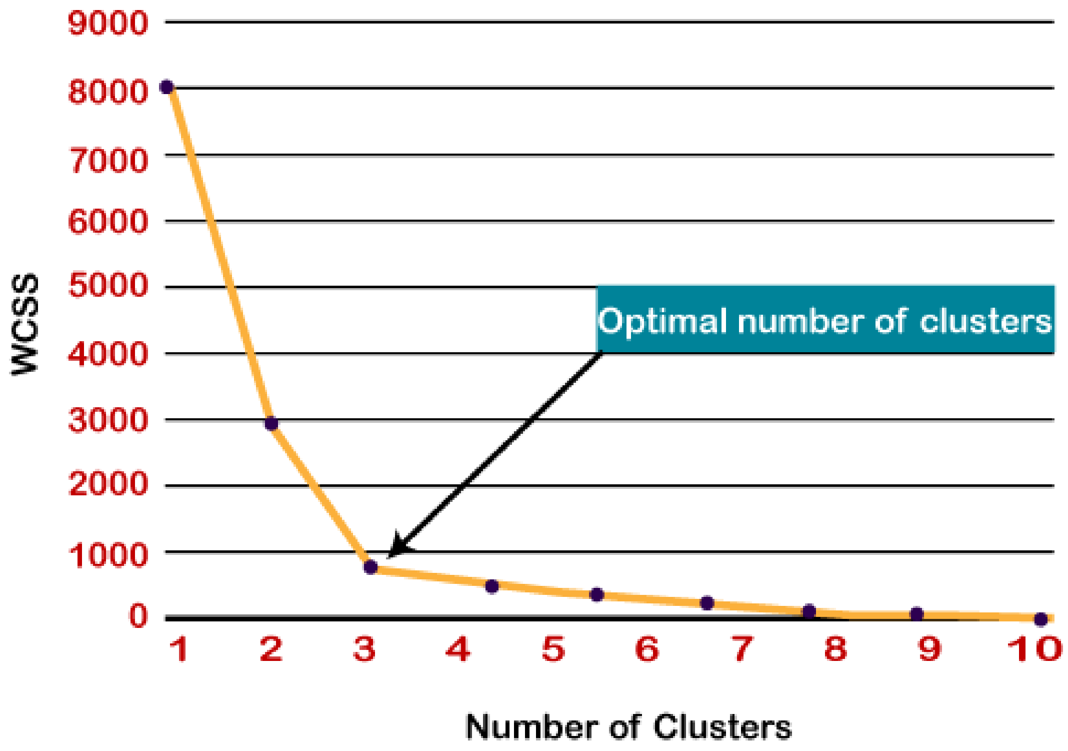
Elbow Method:

- This method uses the concept of **WCSS value**. WCSS stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:
- WCSS= $\sum_{P_i \text{ in Cluster1}}$ distance$(P_i\ C_1)^2$ +$\sum_{P_i \text{ in Cluster2}}$distance$(P_i\ C_2)^2$+$\sum_{P_i \text{ in CLuster3}}$ distance$(P_i\ C_3)^2$
- Here, $\sum_{P_i \text{ in Cluster1}}$ distance$(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.
- To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

How to choose the value of K?

To find the optimal value of clusters, the elbow method follows the steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.
- Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method.



Continue…

- **Iteration-02:**

	A1	A2	A3	A4	A5	A6
C1	1.52	21.13	19.76	6.19	1.52	6.54
C2	21.26	2.24	2.24	14.14	19.10	26.87
Cluster	C1	C2	C2	C1	C1	C1

- **New Center:** C1 ((185+179+182+188)/4, (72+68+72+77)/4)
  = (183.5, 72.25)
- **New Center:** C2 ((170+168)/2, (56+60)/2)
  = (169, 58)

Ans: C1 (A1, A4, A5, A6) and C2 (A2, A3).

Example Problem-01:

- Cluster the following eight points into three clusters: A1(185, 72), A2(170, 56), A3(168, 60), A4(179, 68), A5(182, 72), A6(188, 77). Initial cluster centers are: A1 (C1) and A2 (C2). Use K-means Algorithm to find the two clusters.
- **Iteration-01:**

	A1	A2	A3	A4	A5	A6
C1	0	21.93	20.81	7.21	3	5.83
C2	21.93	0	4.47	15	20	27.66
Cluster	C1	C2	C2	C1	C1	C1

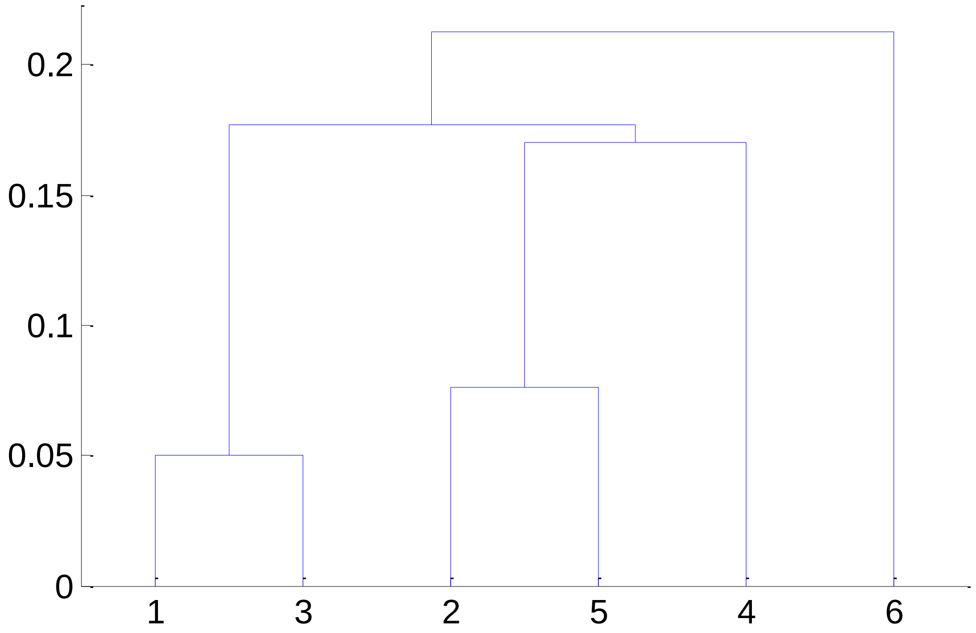
- **New Center:** C1 ((185+179+182+188)/4, (72+68+72+77)/4)
  = (183.5, 72.25)
- **New Center:** C2 ((170+168)/2, (56+60)/2)
  = (169, 58)

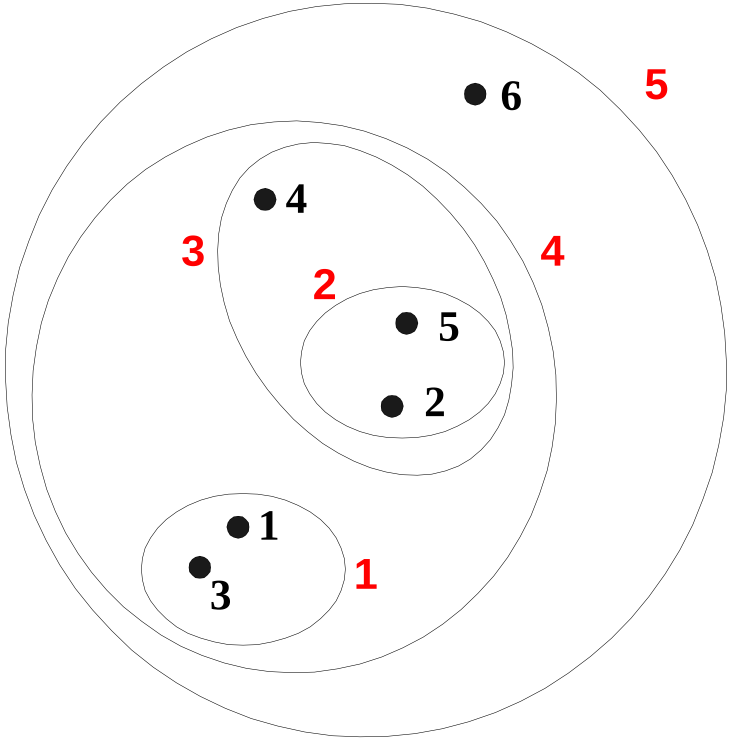
Example Problem-02

- Cluster the following eight points into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9). Initial cluster centers are: A1, A4 and A7. Use K-means Algorithm to find the three cluster centers after the second iteration.
- Do it by yourself.

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree.
- Can be visualized as a dendrogram.
- A tree like diagram that records the sequences of merges or splits.
- Do not have to assume any particular number of clusters.
- Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level.





Hierarchical Clustering

- Two main types of hierarchical clustering:

1. **Agglomerative:**

- Start with the points as individual clusters.
- At each step, merge the closest pair of clusters until only one cluster (or k clusters) left.

2. **Divisive:**

- Start with one, all-inclusive cluster.
- At each step, split a cluster until each cluster contains an individual point (or there are k clusters).

- Traditional hierarchical algorithms use a similarity or distance matrix.
- Merge or split one cluster at a time.

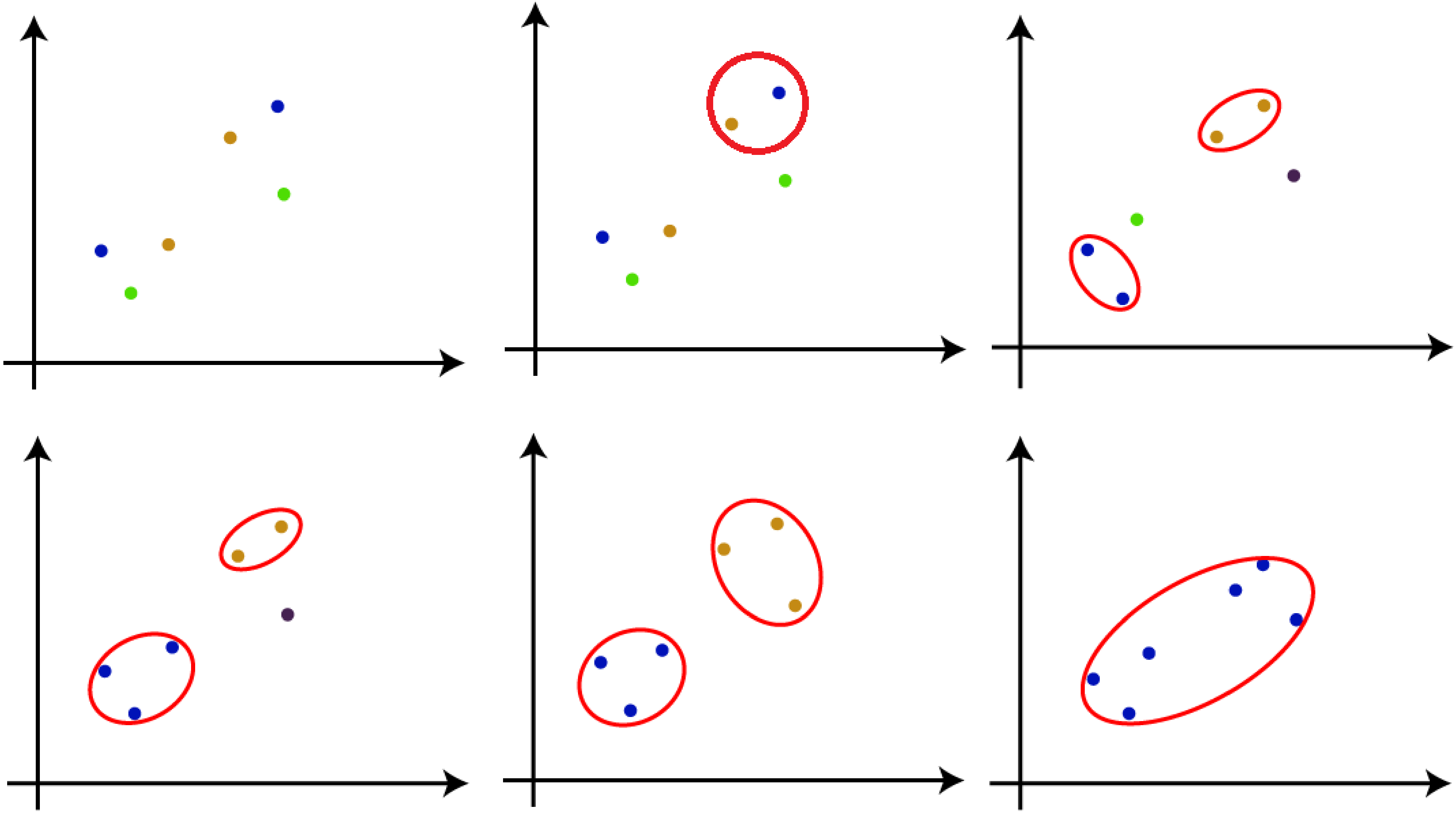
Agglomerative Clustering Algorithm

- **Key Idea:** Successively merge closest clusters
- **Basic algorithm**

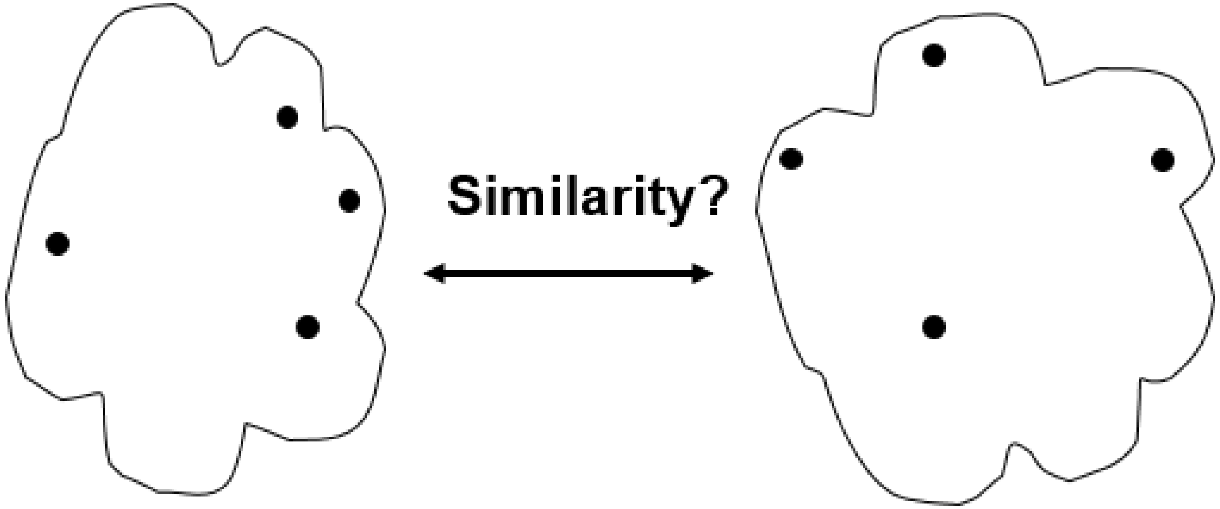
```
1. Compute the proximity matrix.
2. Let each data point be a cluster.
3. Repeat
4.     Merge the two closest clusters
5.     Update the proximity matrix
6. Until only a single cluster remains
```

- Key operation is the computation of the proximity of two clusters.
- Different approaches to defining the distance between clusters distinguish the different algorithms.

Agglomerative Clustering Algorithm



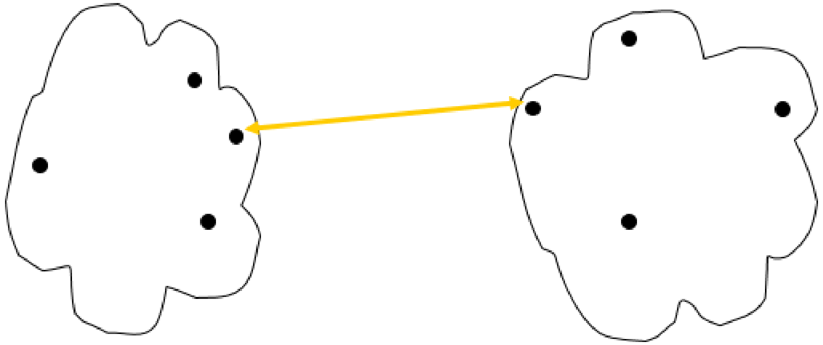
How to Define Inter-Cluster Distance



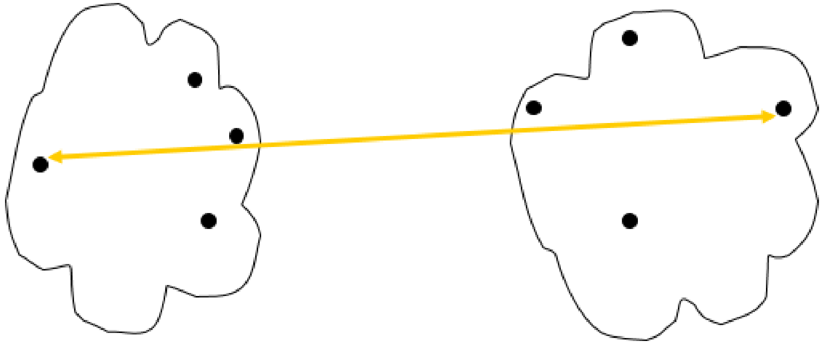
- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Distance

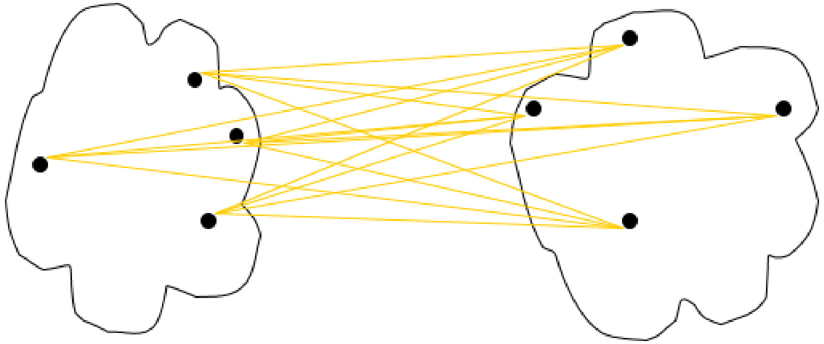
MIN



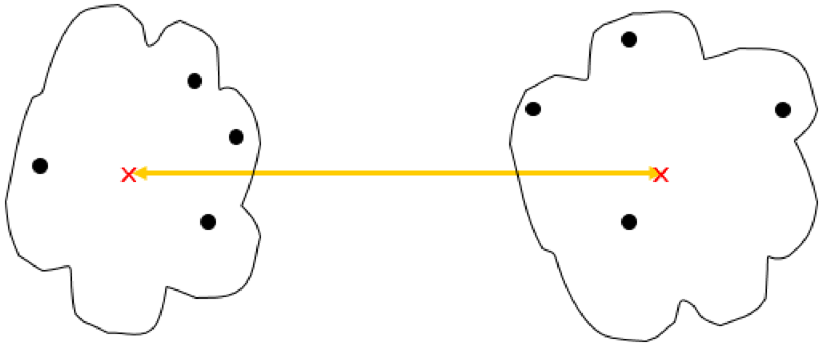
MAX



Group Average

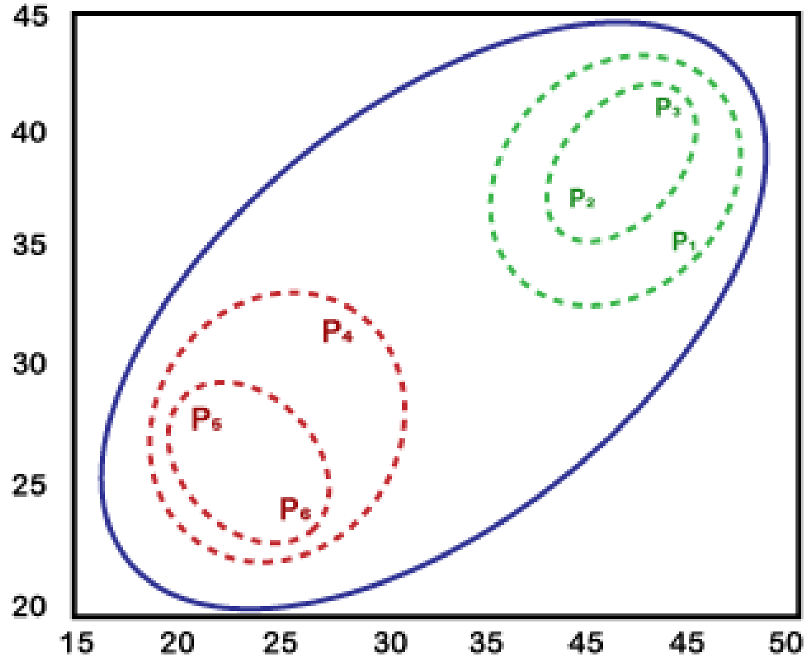


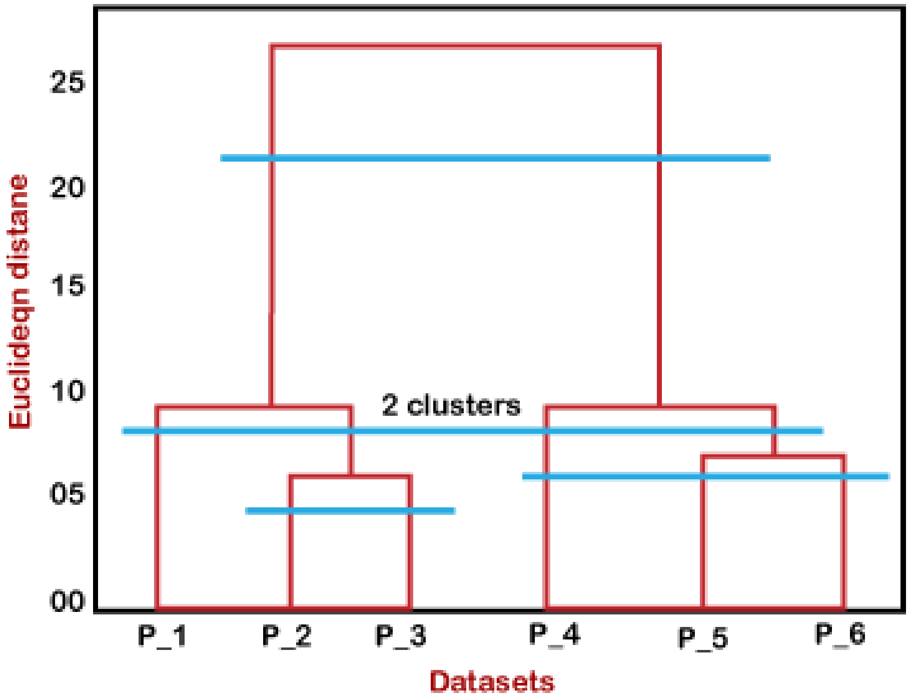
Distance Between Centroids



Agglomerative Clustering - Example

- The dendrogram is a tree-like structure that is mainly used to store each step as a memory that the HC algorithm performs. In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset.
- The working of the dendrogram can be explained using the below diagram:



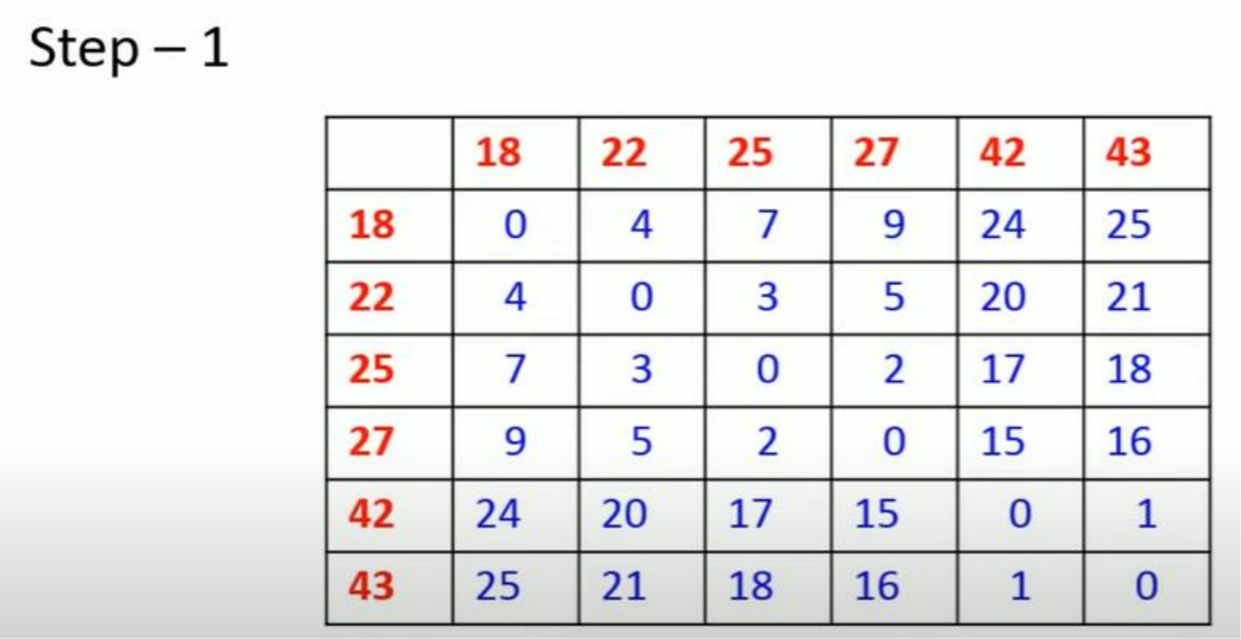


Continue…

- In the above diagram, the left part is showing how clusters are created in agglomerative clustering, and the right part is showing the corresponding dendrogram.
- Firstly, the datapoints P2 and P3 combine together and form a cluster, correspondingly a dendrogram is created, which connects P2 and P3 with a rectangular shape. The hight is decided according to the Euclidean distance between the data points.
- In the next step, P5 and P6 form a cluster, and the corresponding dendrogram is created. It is higher than of previous, as the Euclidean distance between P5 and P6 is a little bit greater than the P2 and P3.
- Again, two new dendrograms are created that combine P1, P2, and P3 in one dendrogram, and P4, P5, and P6, in another dendrogram.
- At last, the final dendrogram is created that combines all the data points together.

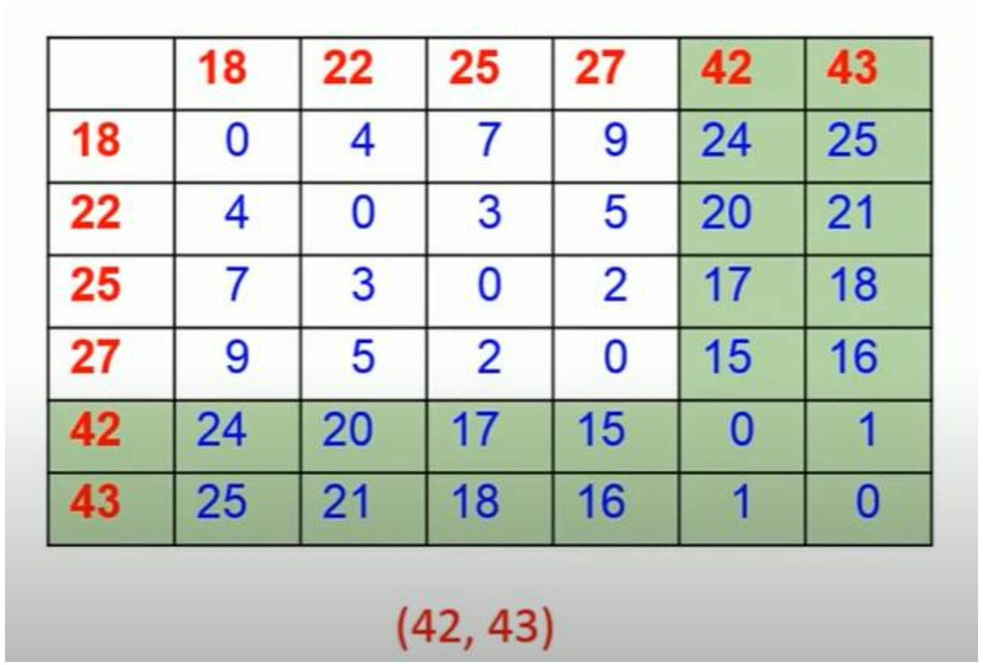
Agglomerative Clustering – Example (2)

- Consider the following set of 6 one dimensional data points 18, 22, 25, 42, 27, 43. Apply the Agglomerative Clustering algorithm to build the hierarchical clustering dendogram. [Use simple distance].



Step – 1

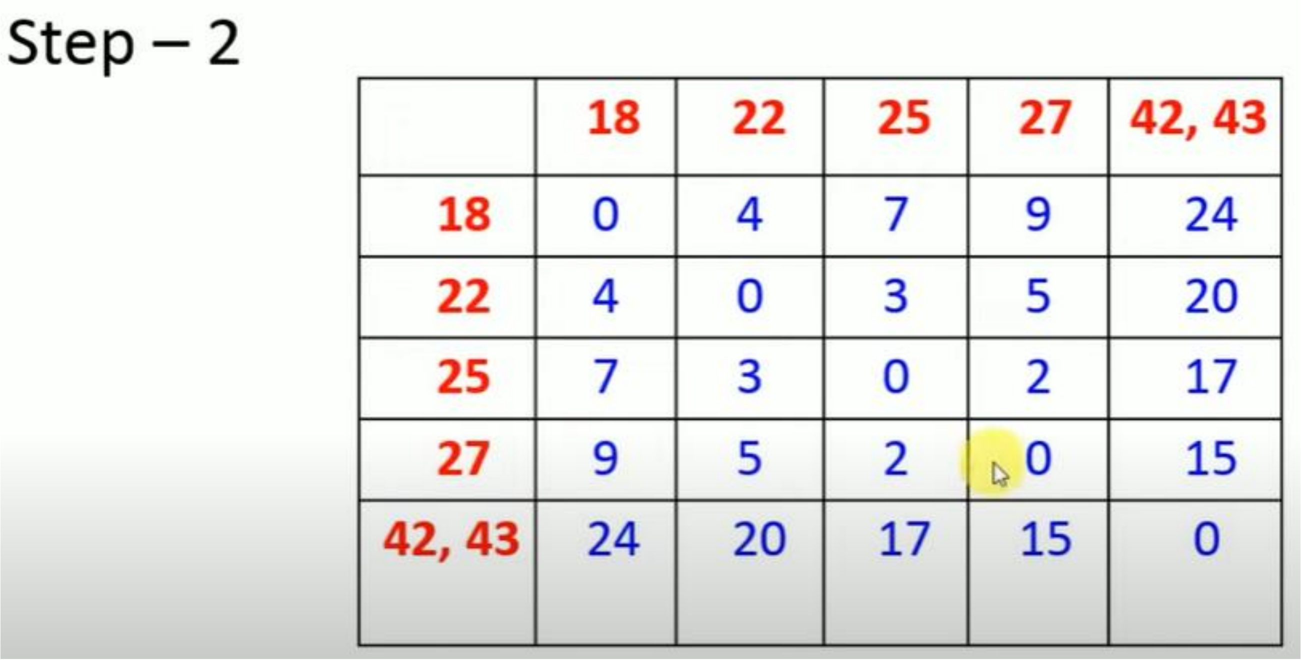
	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0



	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

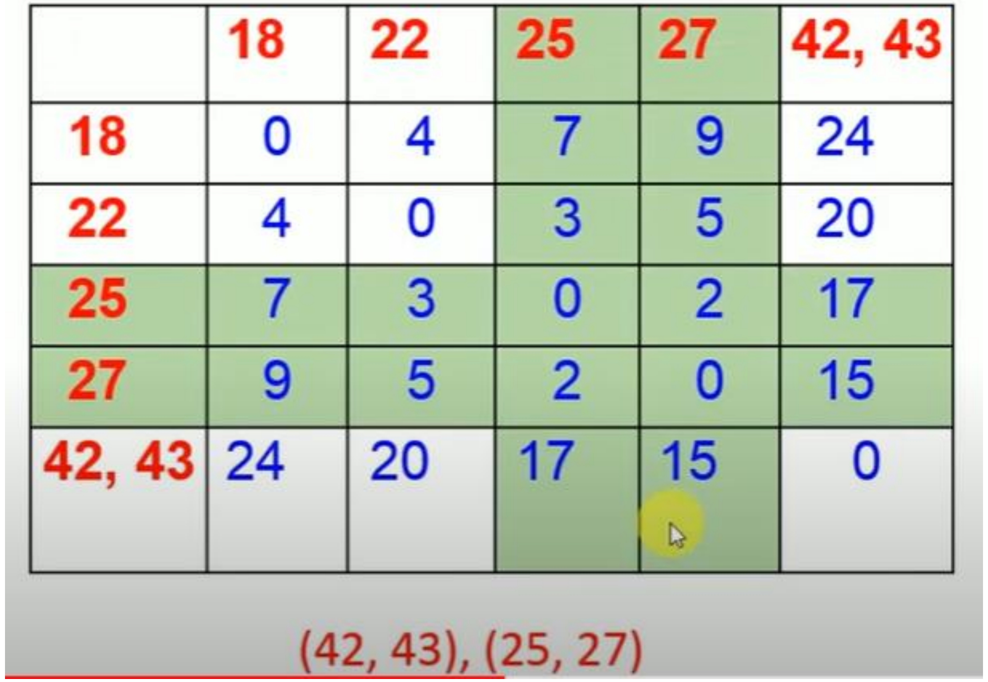
(42, 43)

Continue…



Step – 2

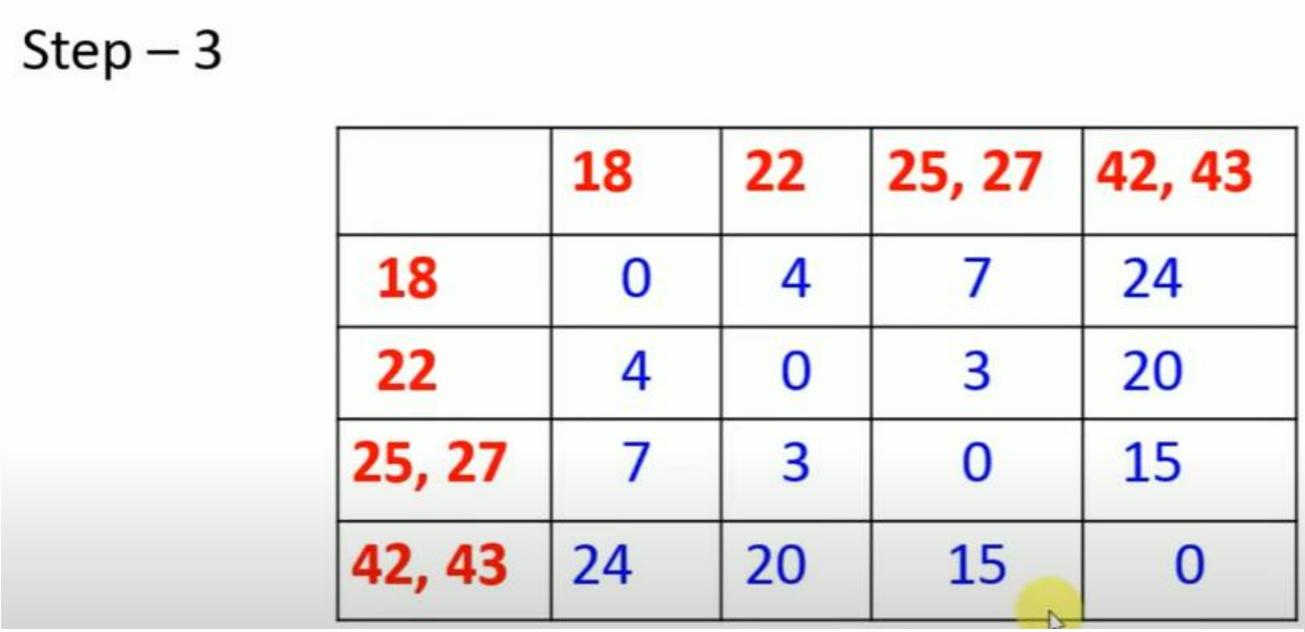
	18	22	25	27	42, 43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42, 43	24	20	17	15	0



	18	22	25	27	42, 43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42, 43	24	20	17	15	0

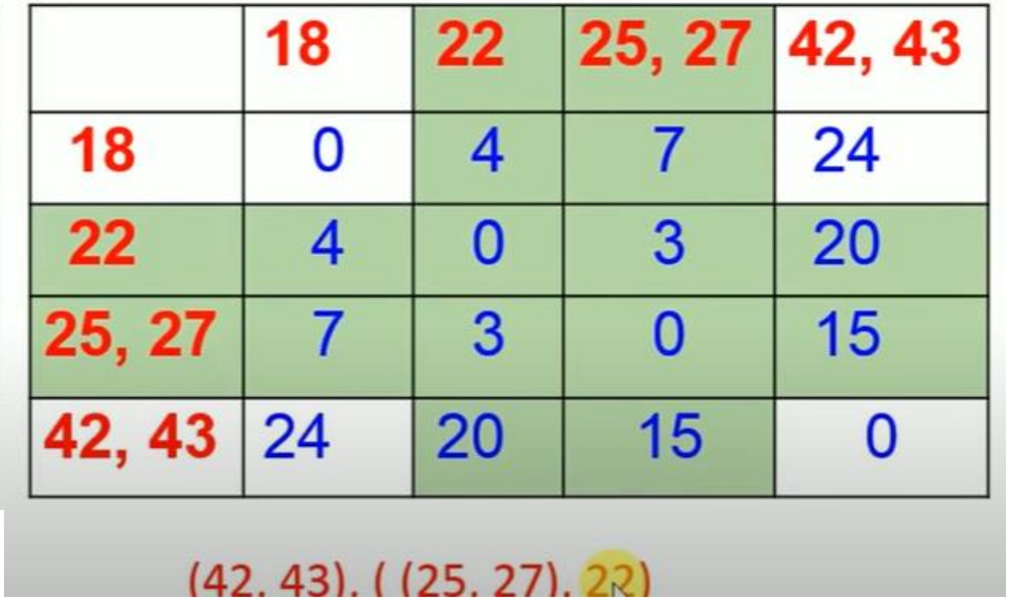
(42, 43), (25, 27)

Continue…



Step – 3

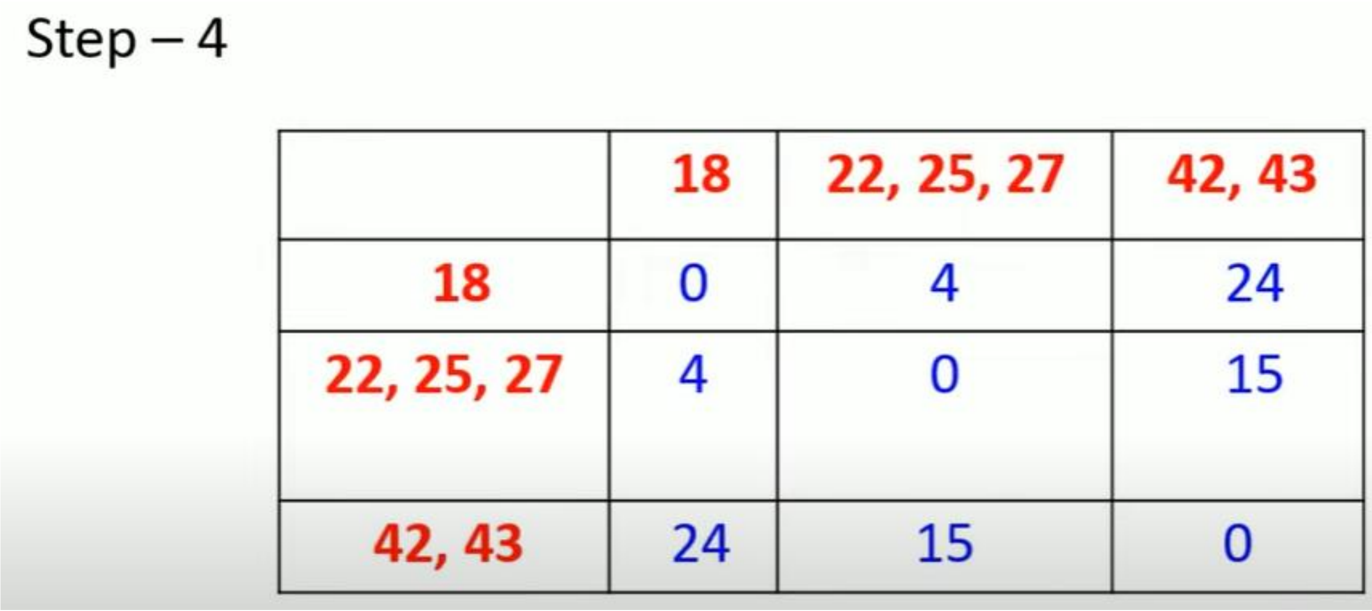
	18	22	25, 27	42, 43
18	0	4	7	24
22	4	0	3	20
25, 27	7	3	0	15
42, 43	24	20	15	0



	18	22	25, 27	42, 43
18	0	4	7	24
22	4	0	3	20
25, 27	7	3	0	15
42, 43	24	20	15	0

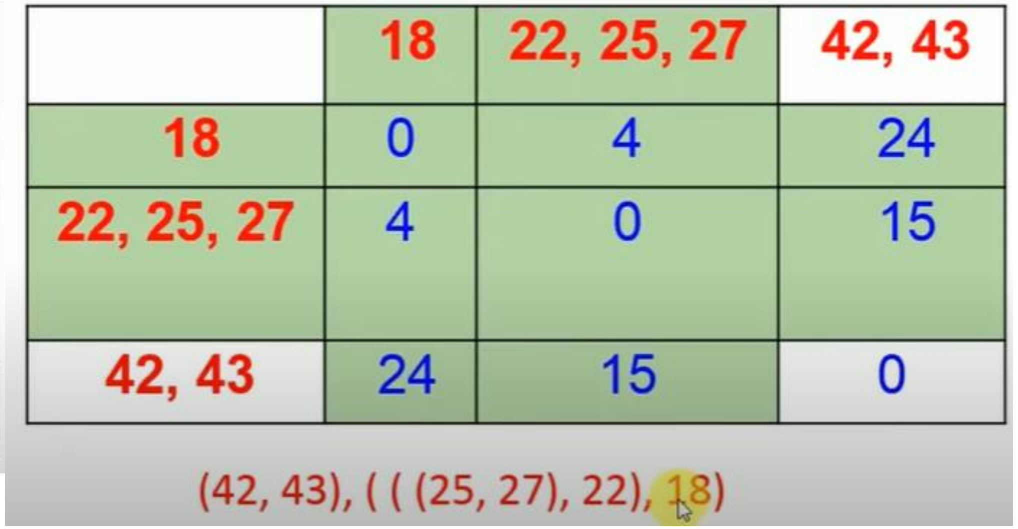
(42, 43), ((25, 27), 22)

Continue…



Step – 4

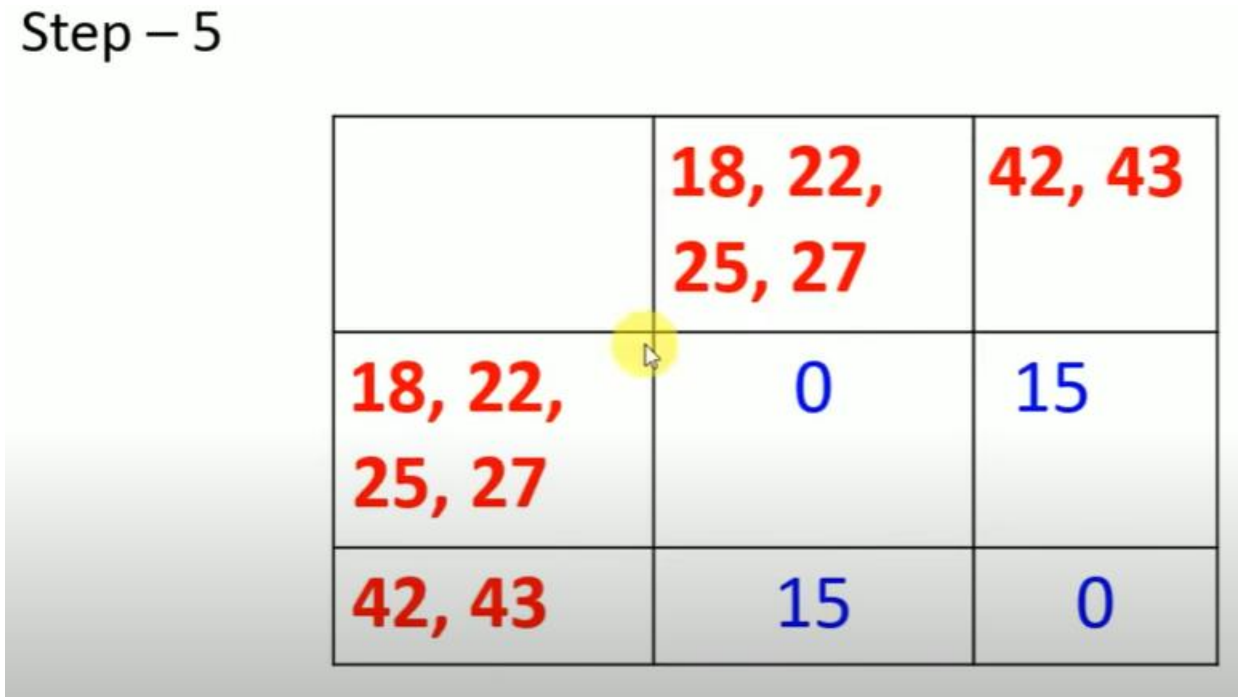
	18	22, 25, 27	42, 43
18	0	4	24
22, 25, 27	4	0	15
42, 43	24	15	0



	18	22, 25, 27	42, 43
18	0	4	24
22, 25, 27	4	0	15
42, 43	24	15	0

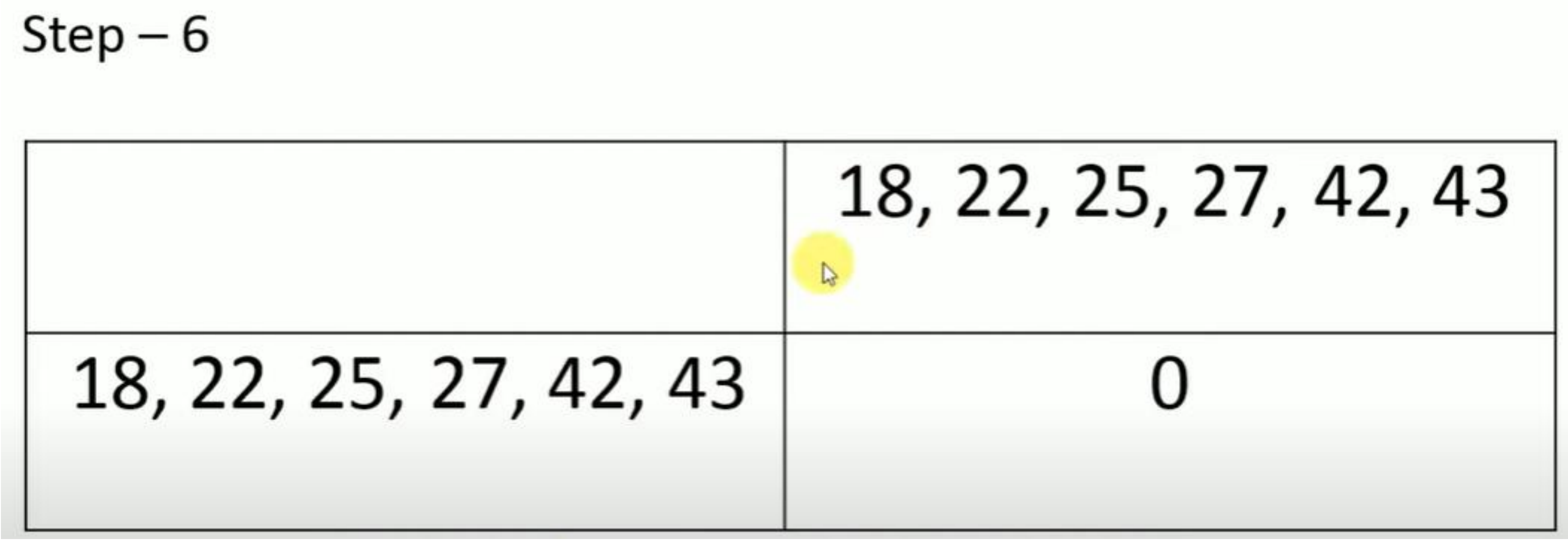
(42, 43), (((25, 27), 22), 18)

Continue…



Step – 5

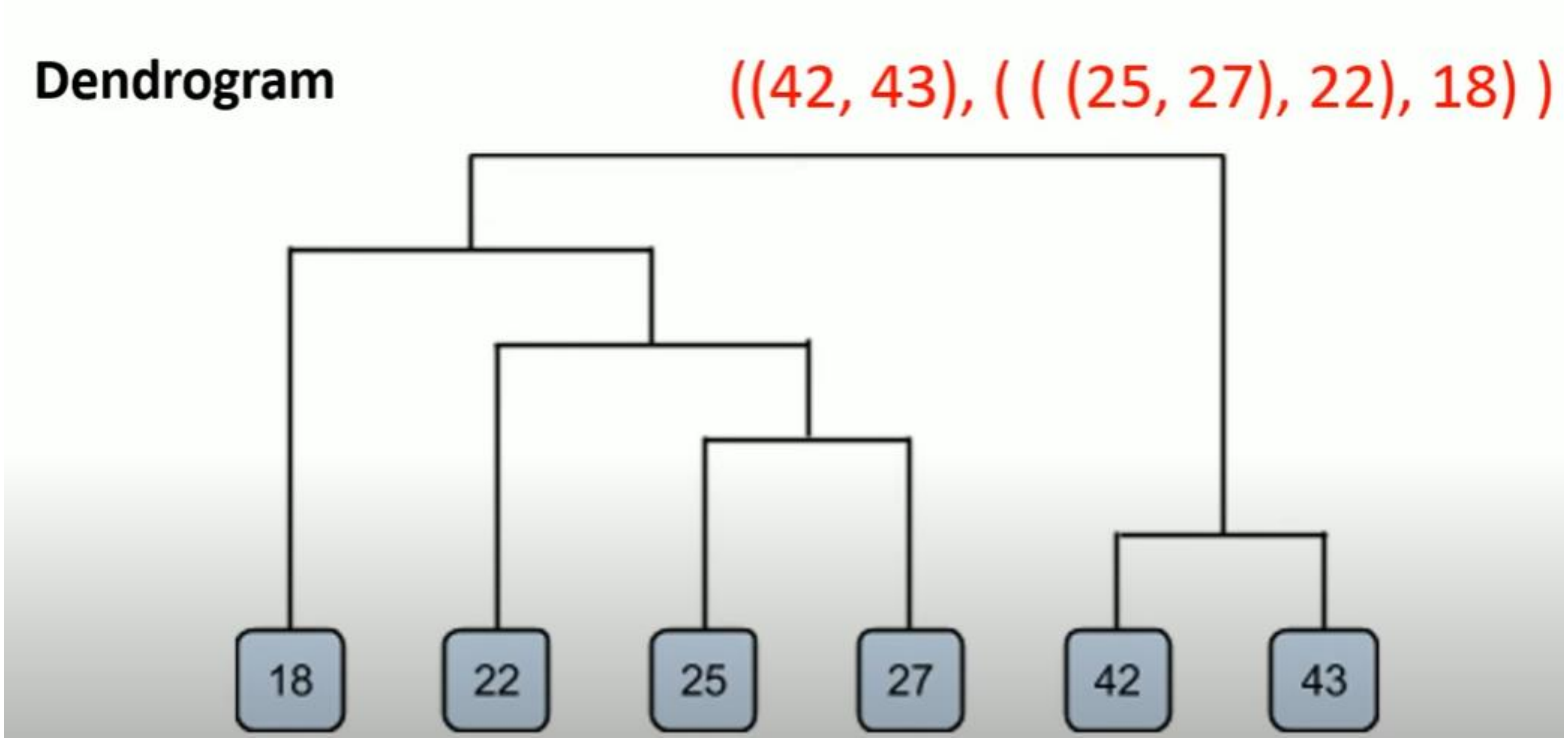
	18, 22, 25, 27	42, 43
18, 22, 25, 27	0	15
42, 43	15	0



Step – 6

	18, 22, 25, 27, 42, 43
18, 22, 25, 27, 42, 43	0

Continue…



Dendrogram

((42, 43), (((25, 27), 22), 18))

Example Problem-03

- Cluster the following eight points into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9). Initial cluster centers are: A1, A4 and A7. Apply the Agglomerative Clustering algorithm to build the hierarchical clustering dendogram. [Use Eucledian Distance & Group Average Inter-Cluster Distance].

- Do it by yourself.

Thank You!