Unsupervised Learning

Nama Pengajar

- Dian Ade Kurnia, M.Kom
- Rusnanda Farhan
- Rusnandi Fikri
- Rika Sahriana

List of Contents

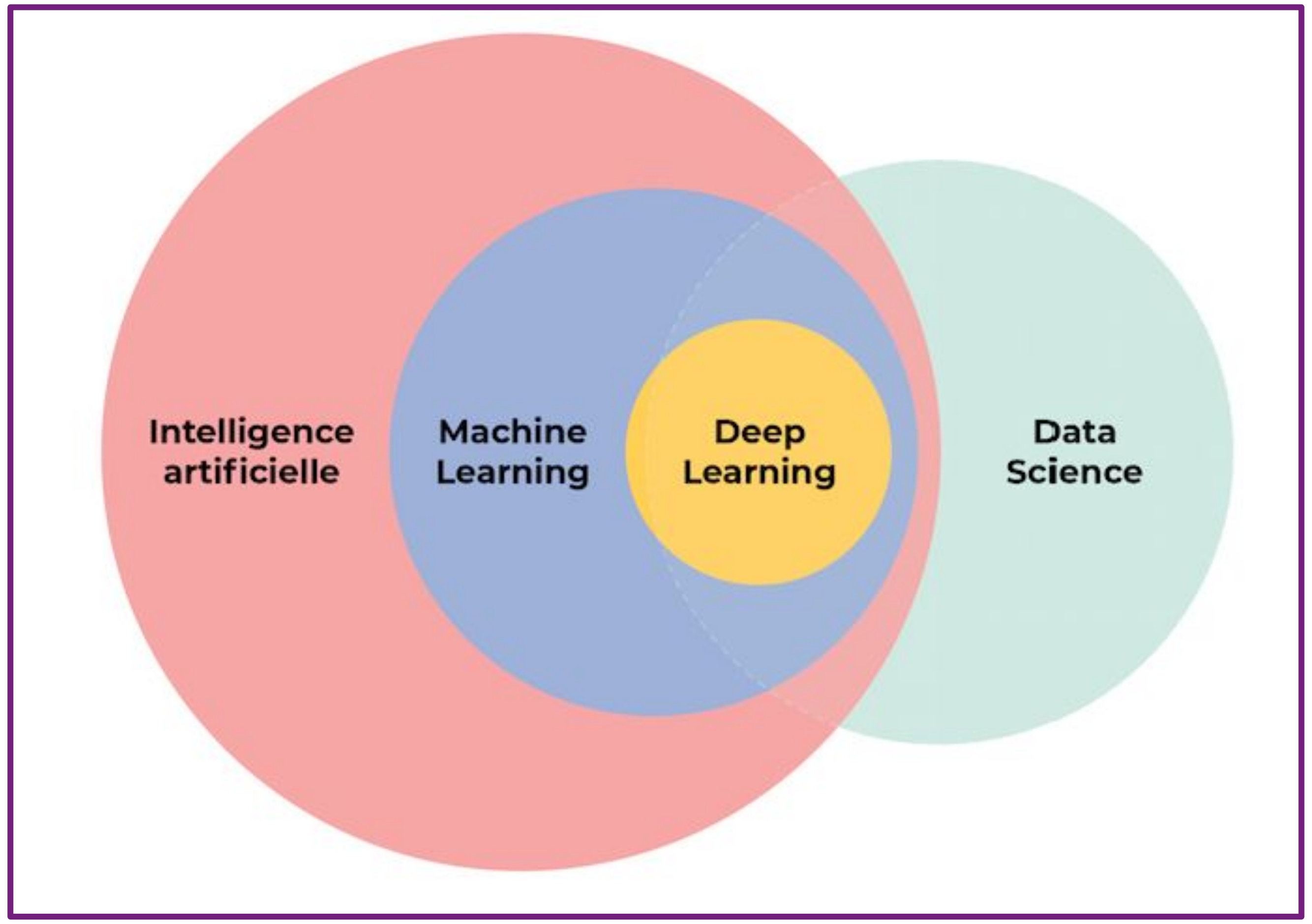
01

Introduction Machine Learning

Introduction Machine Learning

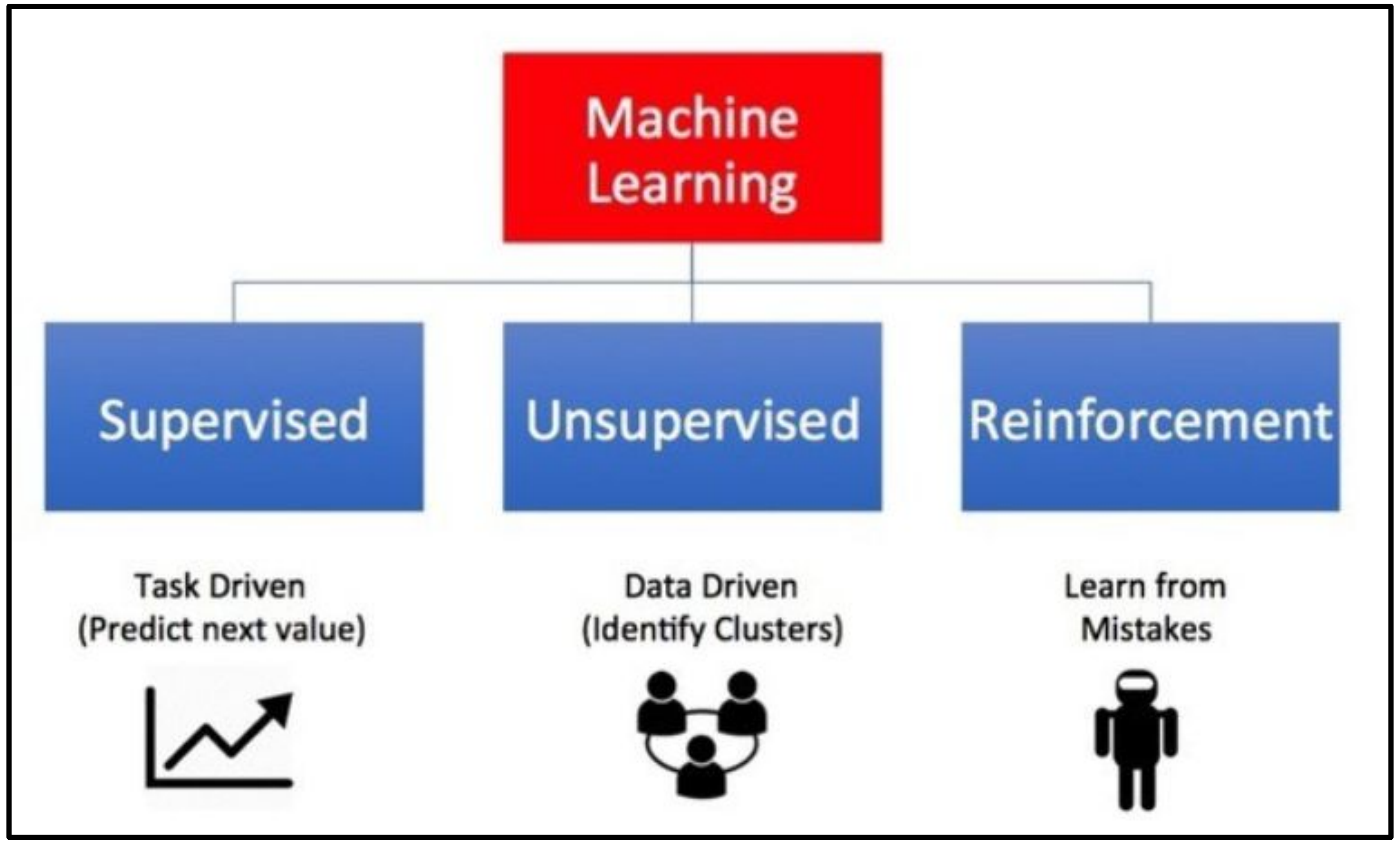
Definition

Machine Learning is part of Artificial Intelligence, and intersects with Data Science. Machine Learning is the study of computer algorithms that can improve automatically through experience and by the use of data (training data). As the result of Machine Learning. it usually builds/produces a model.



Introduction Machine Learning

Kind of Machine Learning



02

Introduction Unsupervised Learning

Introduction Unsupervised Learning

Unsupervised learning is a type of algorithm that learns patterns from untagged data.
- Wikipedia

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

- IBM

Introduction Unsupervised Learning

Unsupervised Learning

- Clustering
- Association Rules
- Dimensionality Reduction
- Autoencoder

03

Introduction Clustering

Introduction Clustering

"Clustering is a process to group data into several clusters so that each data in one cluster has a high (maximum) similarity and data in different clusters has a low (minimum) similarity."

Introduction Clustering

Standard Distance Methods

Minkowski distance or L_p distance, $d_p(X,Y)=\left\{\sum_{i=1}^{n}|x_i-y_i|^p\right\}^{\frac{1}{p}}$

Manhattan distance, $d_1(X,Y)=\sum_{i=1}^{n}|x_i-y_i|$ $(P=1)$

Euclidian distance, $d_2(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$ $(P=2)$

Max distance, $d_\infty(X,Y)=\max_{i=1}^{n}|x_i-y_i|$ $(P=\infty)$

04

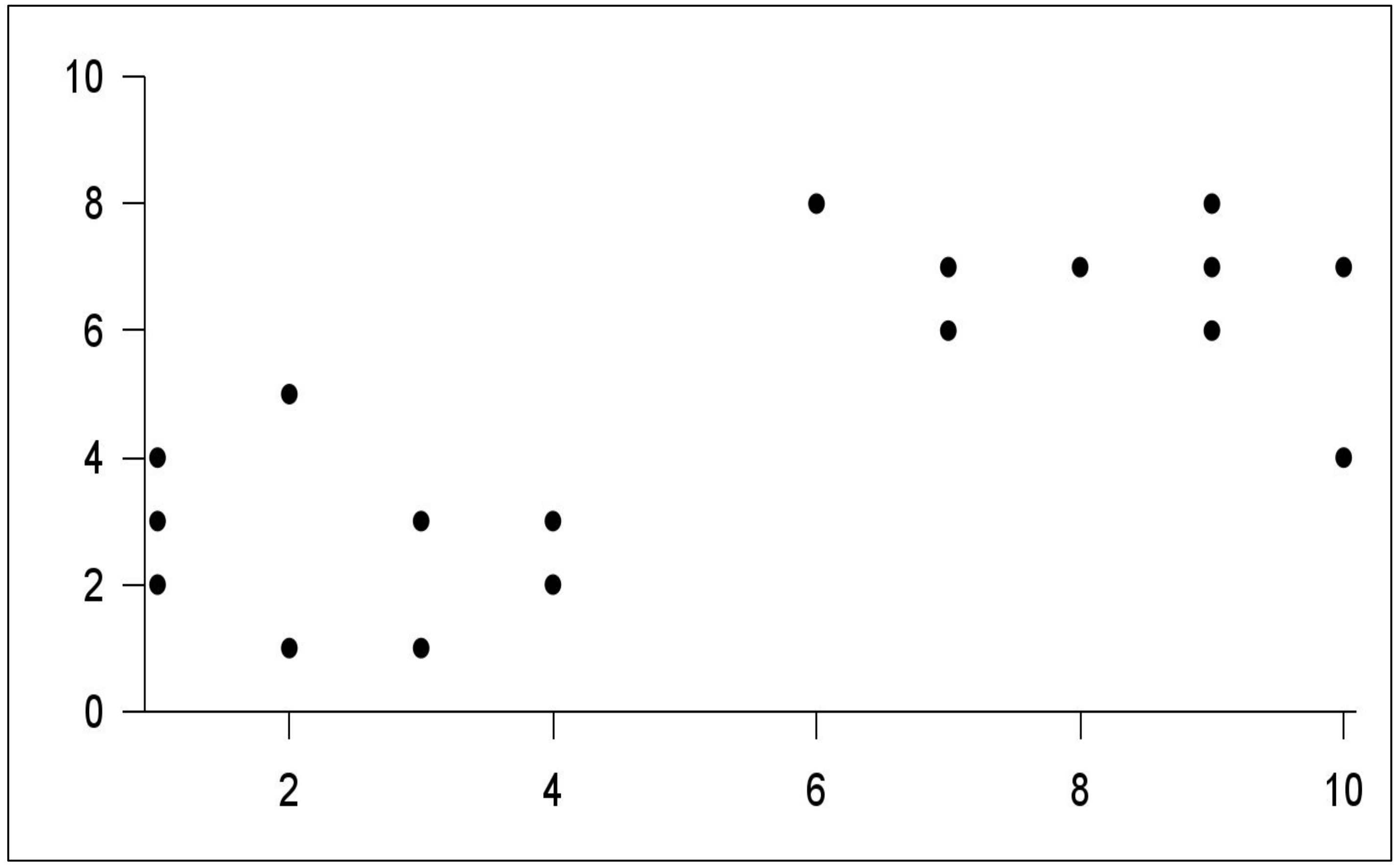
K Means Algorithm

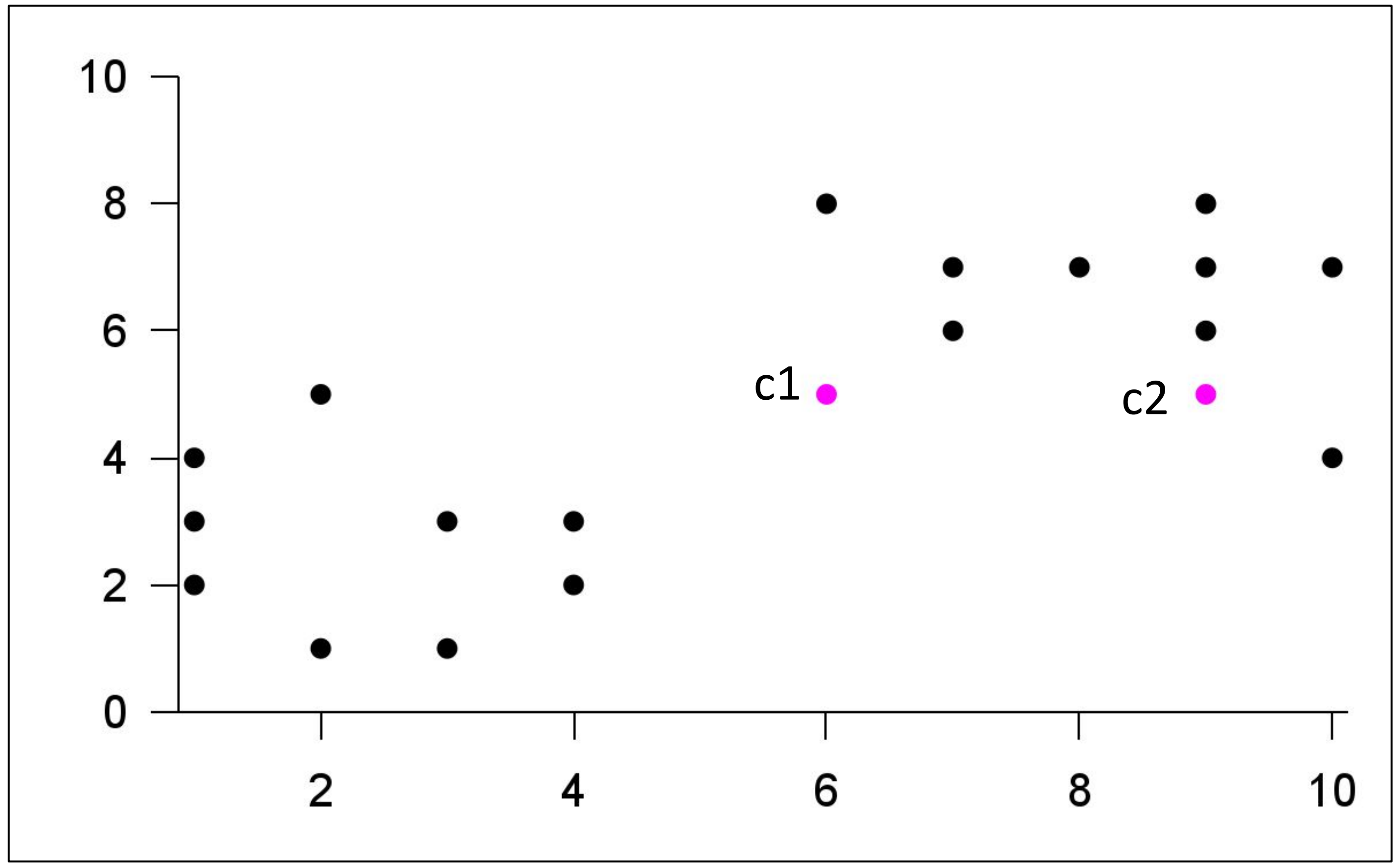
K Means Algorithm

The fundamental idea in K Means is grouping the data into a number of clusters (K>1) which are defined first. By utilizing the concept of distance, each data will be grouped with the nearest cluster center point (centroid).

K Means Algorithm

ID	X1	X2
1	1	5
2	1	6
3	1	4
4	2	5
5	9	7
6	9	8
7	9	6
8	8	7
9	10	7
10	4	4
11	4	6
12	10	4





1. Determine the number of K

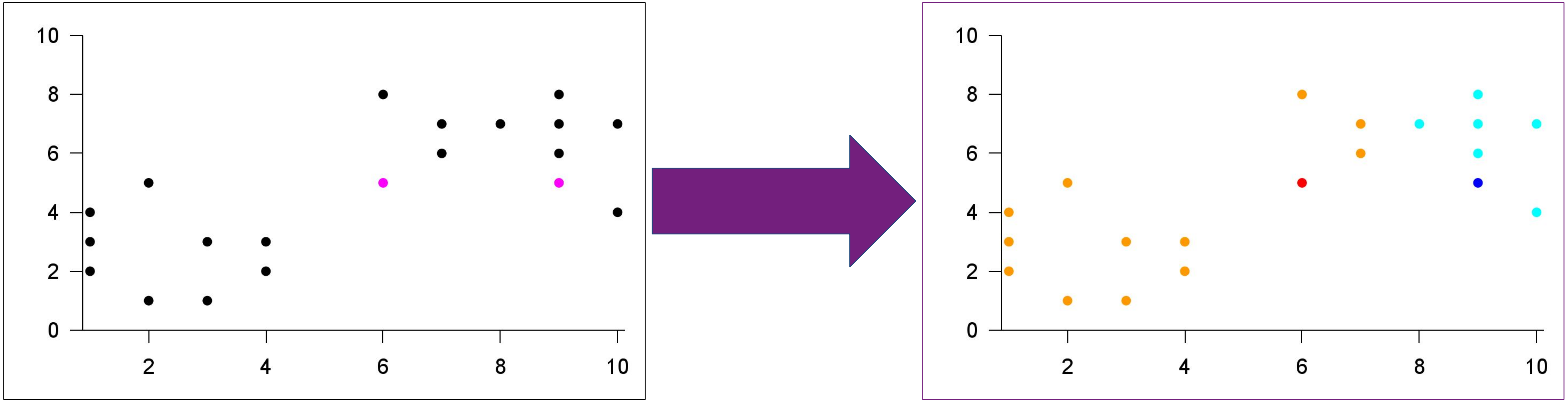
We choose K=2 so we get 2 centroids. The value of centroid determined by random values. We denote the centroids with c1 and c2, as the center of the cluster1 and cluster2.

c1 = (6, 5)
c2 = (9, 5)

K Means Algorithm

2. Calculate the distance to make cluster

Calculate The distance between c1 to each datapoint, and do the same way with c2. If a datapoint closer to c1 than c2, so it will be the member of cluster1, and vice versa.

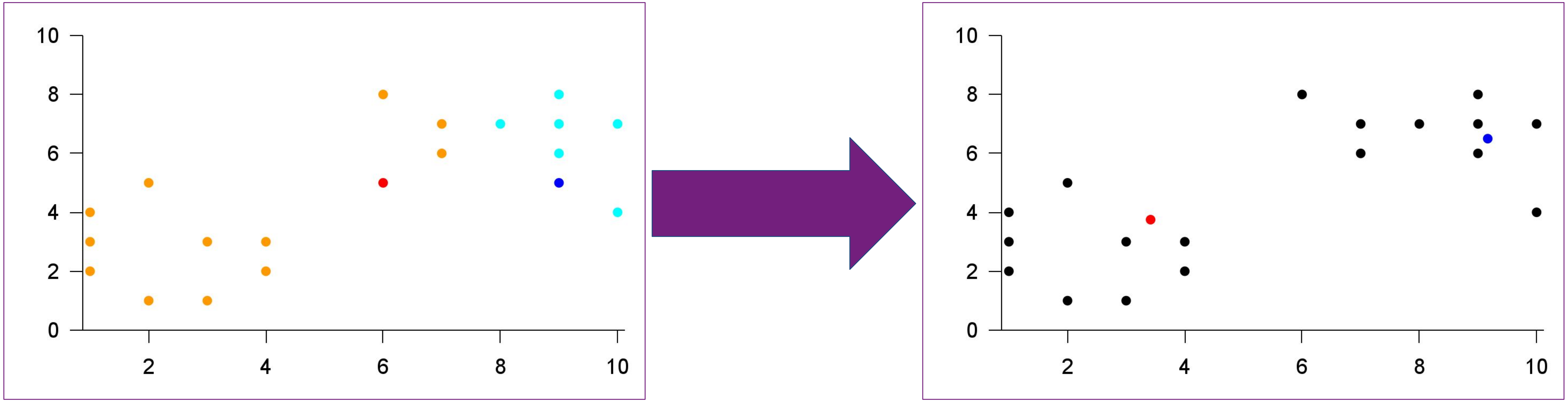


c1 = red, cluster1 = orange
c2 = dark blue, cluster2 = light blue

K Means Algorithm

3. Update the value of centroid

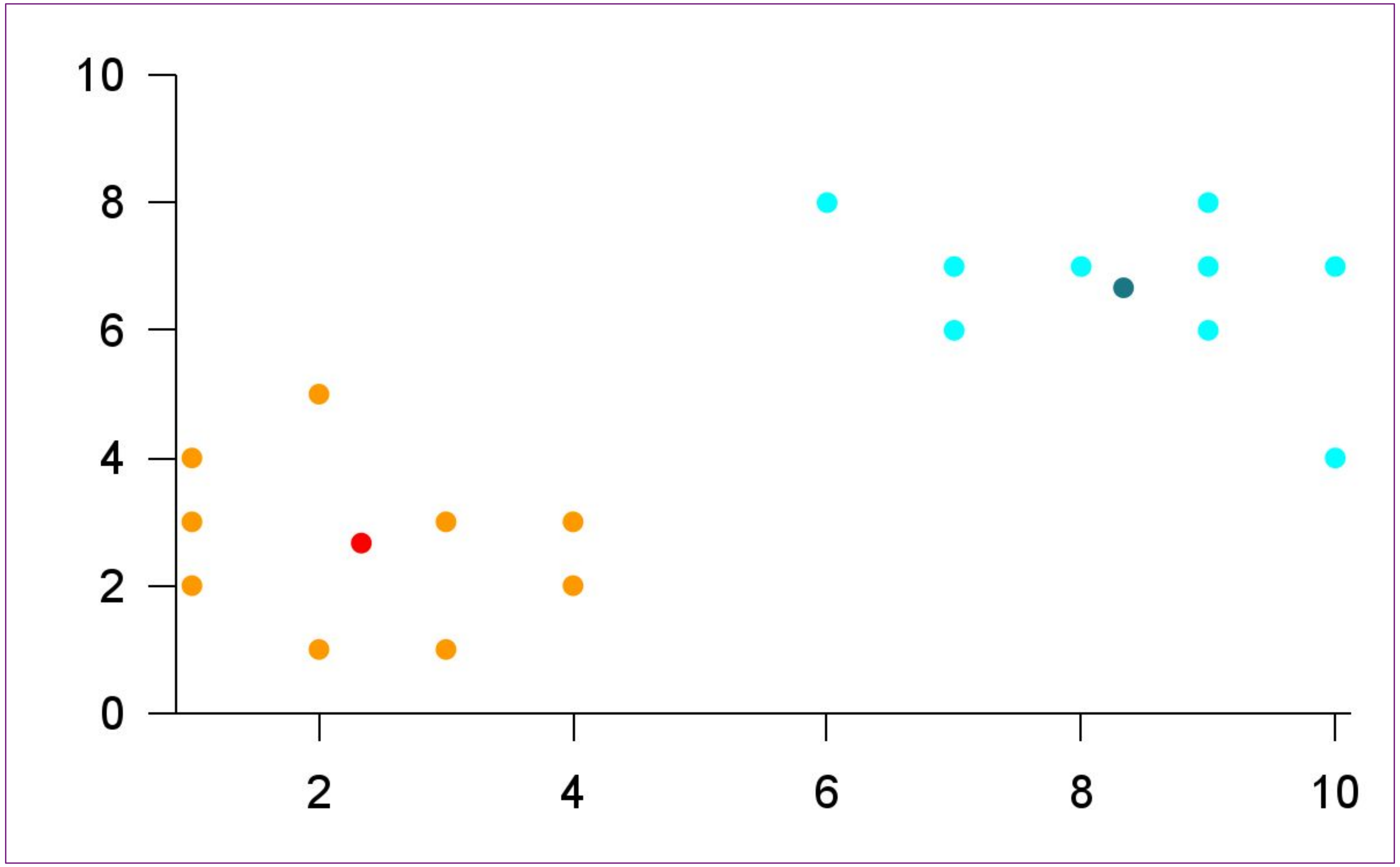
Update the value of c1 and c2 with the mean of cluster1 and mean of cluster2



c1 =(6, 5) -> (3,41666666666667 , 3,75)
c2 = (9, 5) -> (9,16666666666667 , 6,5)

K Means Algorithm

4. Repeat step 2 and 3
Repeat step number 2 to determine member of cluster and step number 3 to update the value of centroids. This loop will stop when the centroid value doesn't change anymore.



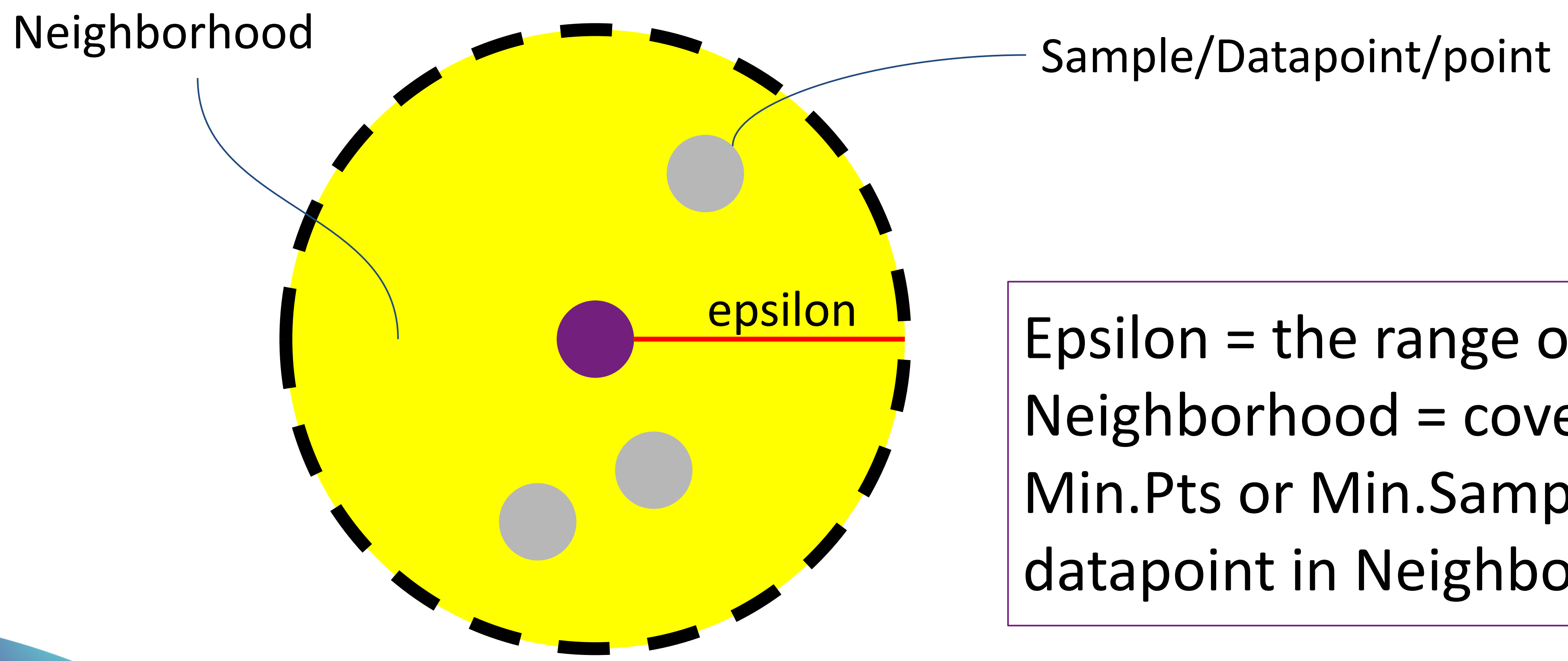
05

DBSCAN Algorithm

DBSCAN Algorithm

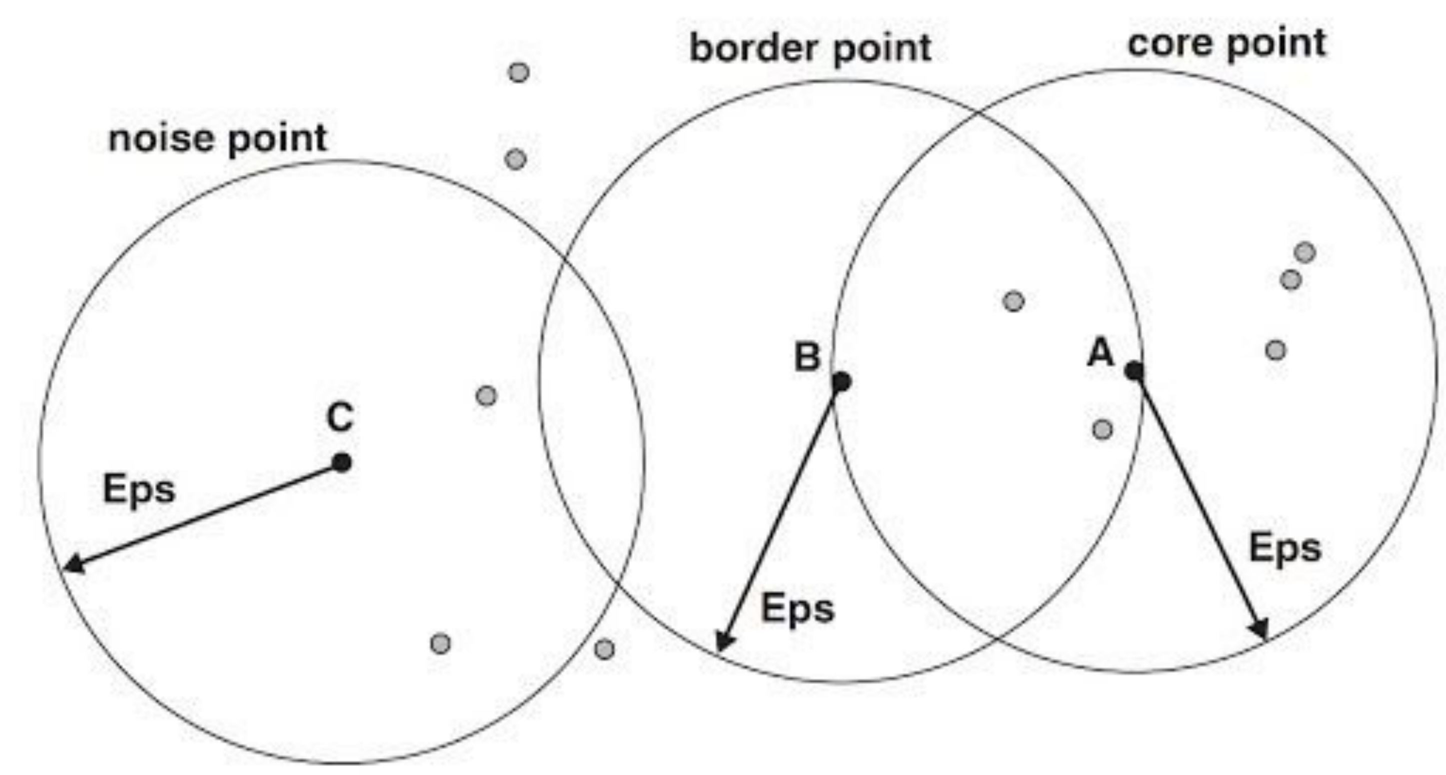
DBSCAN algorithm (Density-based Spatial Clustering of Applications with Noise) is the algorithm that uses density to cluster the data points and it has the ability to identify the noise very well.

DBSCAN Algorithm



Epsilon = the range of datapoint
Neighborhood = coverage area of radius epsilon
Min.Pts or Min.Samples = minimum threshold for another sample or datapoint in Neighborhood

DBSCAN Algorithm



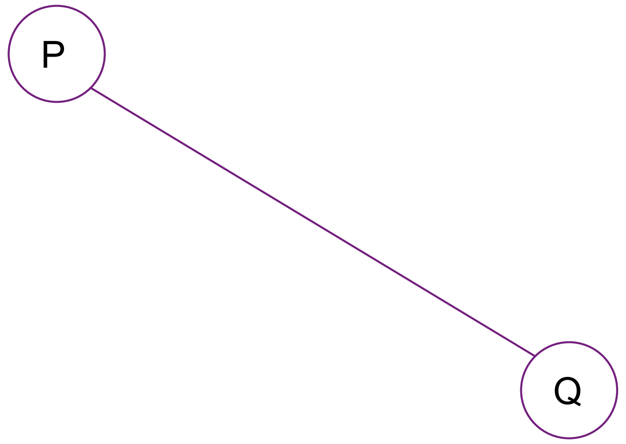
Core Point
Condition : the number of neighbors must be greater than or equal to threshold Min.Samples.

Boundary Point
Condition : the number of neighbors must be less than threshold Min.Samples and this point should be in the neighborhood of a core point.

Noise Point
Condition : the point that not satified the condition in Core Point and Boundary Point. We can say the number of neighbors is less than the threshold Min. Samples and the point not in neighborhood of a core point.

DBSCAN Algorithm

Density Edge



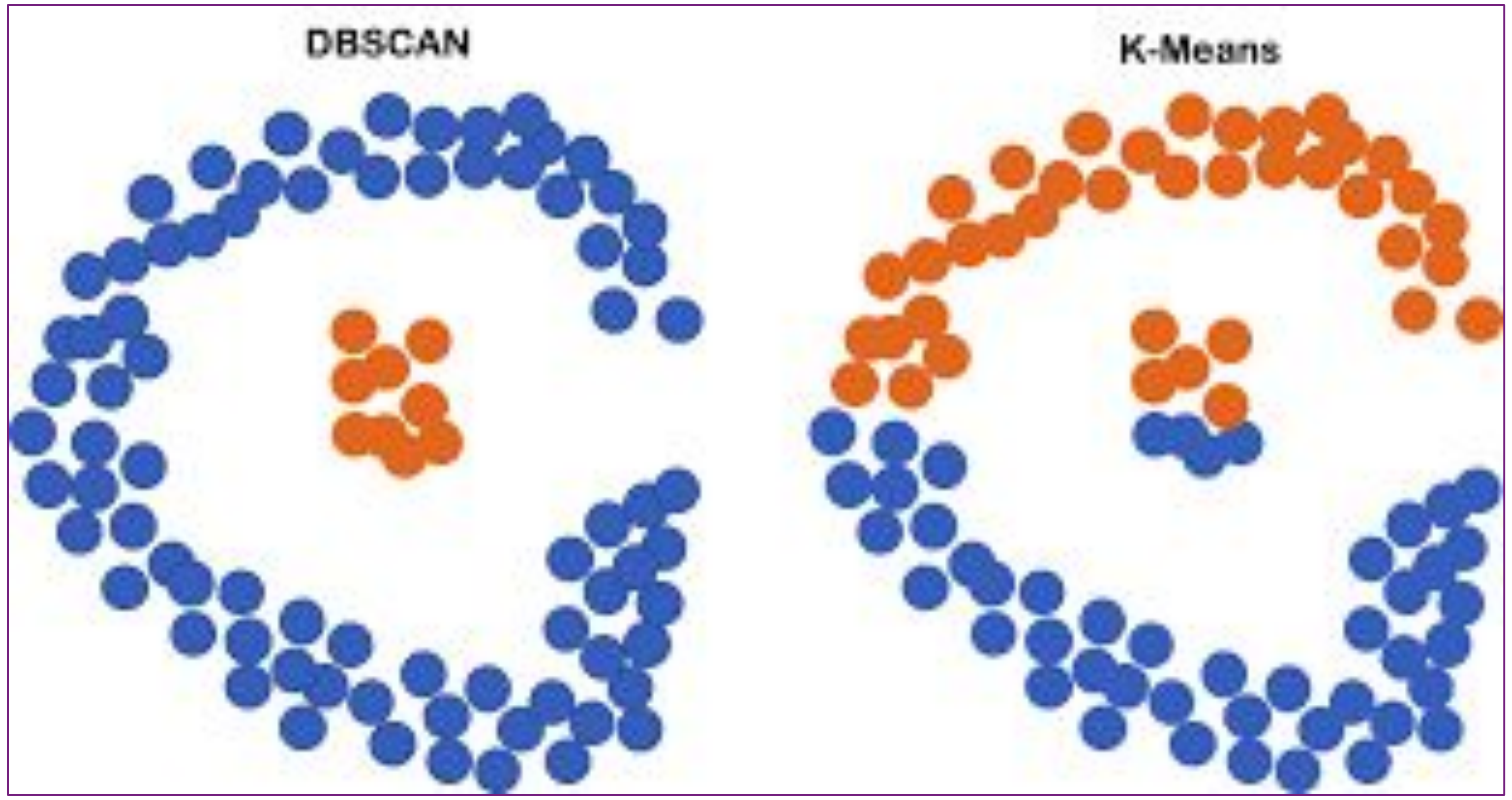
Let P and Q are Core Point and dist(P,Q)<=epsilon. The edge that connecting Q and Q called Density Edge

Density Connected Point

Assume that we have 4 Core Points, denote P, Q, R, S that connected by Density Edge (see the illustration). Q is in neighborhood P but S in not in neighborhood P. For P and S that connected via Density Endge, we can called P and S Density Connection Point

DBSCAN Algorithm

1. Classify the points.
2. Discard noise.
3. Assign cluster to a core point.
4. Color all the density connected points of a core point.
5. Color boundary points according to the nearest core point.



06

Association Rules

Association Rules

transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

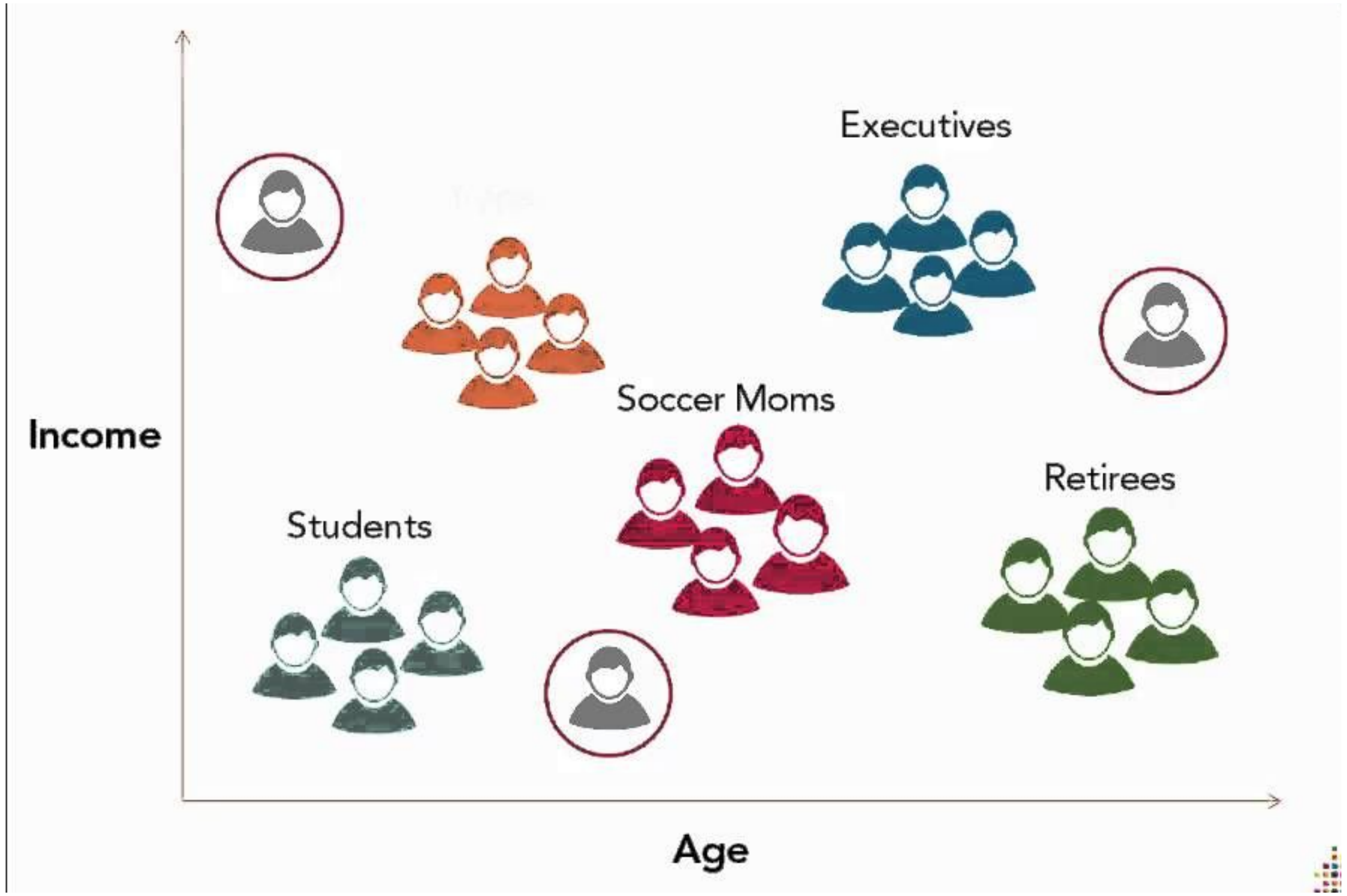
The support *supp(X)* of an itemset X is defined as the proportion of transactions in the data set which contain the itemset. In the example database, the itemset {milk, bread, butter} has a support of 1/5=0.2 since it occurs in 20% of all transactions (1 out of 5 transactions).

The confidence of a rule is defined *conf(X-> Y) = supp(X U Y) / supp(X)*. For example, the rule {milk, bread}->{butter} has a confidence of 0.2/0.4=0.5 in the database, which means that for 50% of the transactions containing milk and bread the rule is correct (50% of the times a customer buys milk and bread, butter is bought as well).

07

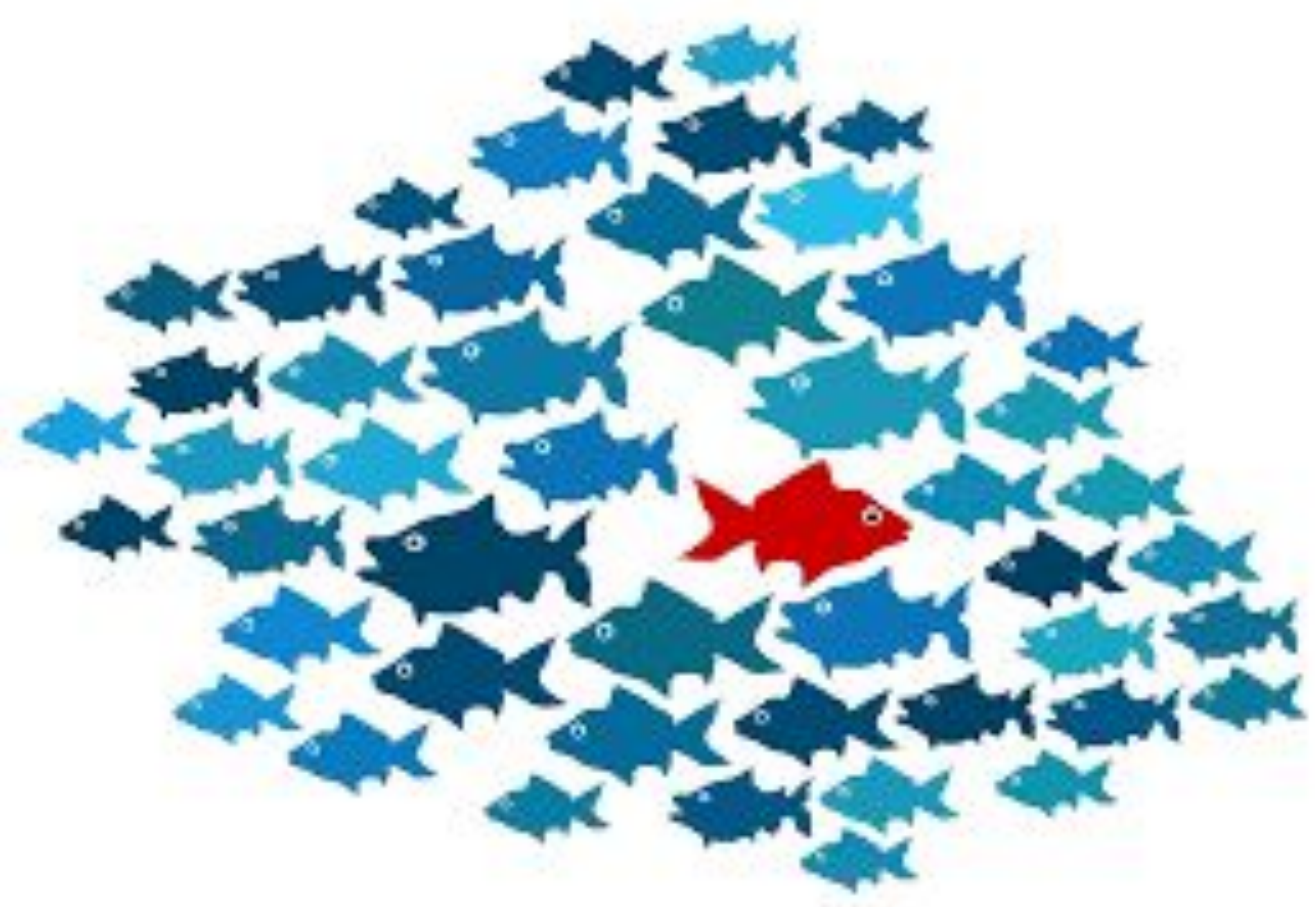
Implementation Unsupervised Learning In Real Life

Customer Segmentation



Recommender System / Building Sales Strategy

Anomaly Detection



Conclusion

- Unsupervised Learning works with data that hasn't label or target
- By using Unsupervised Learning algorithm we can know the similarity of data, or we can see the pattern of our data.
- Another use of UL is that we can detect anomalies in data

References

https://en.wikipedia.org/wiki/Unsupervised_learning

https://www.ibm.com/cloud/learn/unsupervised-learning

https://www.digitalvidya.com/

https://www.kdnuggets.com/