Clustering (分群)

Prof. Chia-Yu Lin
National Central University

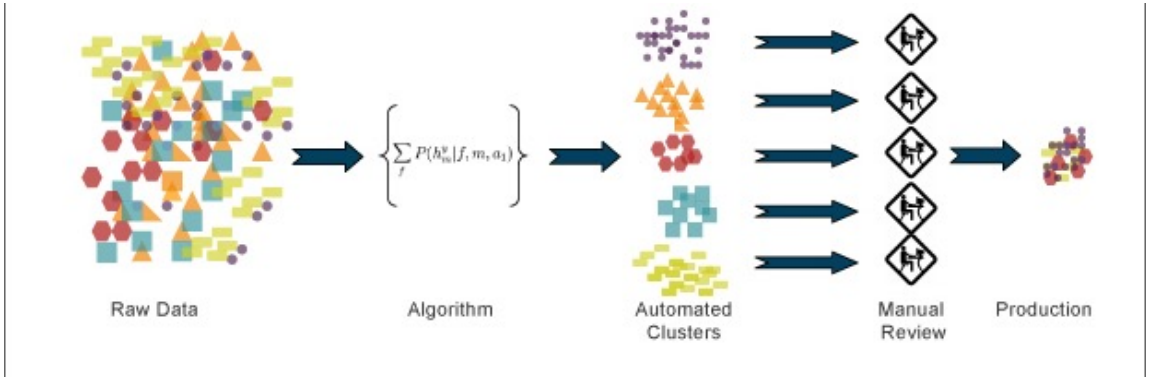
2022 Fall

Outline

- K-means
- K-medioids
- Hierarchical Clustering
- Density Based Clustering (DBSCAN)

Unsupervised Learning

- Unsupervised Learning is the second type of machine learning, in which unlabeled data are used to train the algorithm, which means it used against data that has no historical labels.
- The purpose is to explore the data and find some structure within.
  - Clustering
  - Anomaly Detection
  - Association Rule
  - Autoencoder



K-means Algorithm

- Groups data items into k clusters, where k is user defined.
- Each cluster is defined by a centroid point.
- All points in a cluster are closer (with respect to some distance measure) to their centroid as compared to the centroids of neighboring clusters.

Steps of K-means

- The Goal of K-means attempts to determine k partitions that minimize the square-error function

$$E = \sum_{i-1}^{k} \sum_{p \in C_i} (p - m_i)^2$$

E is the sum of absolute error
C_j is cluster
p is the node in C_j
m_i is the mean of C_j

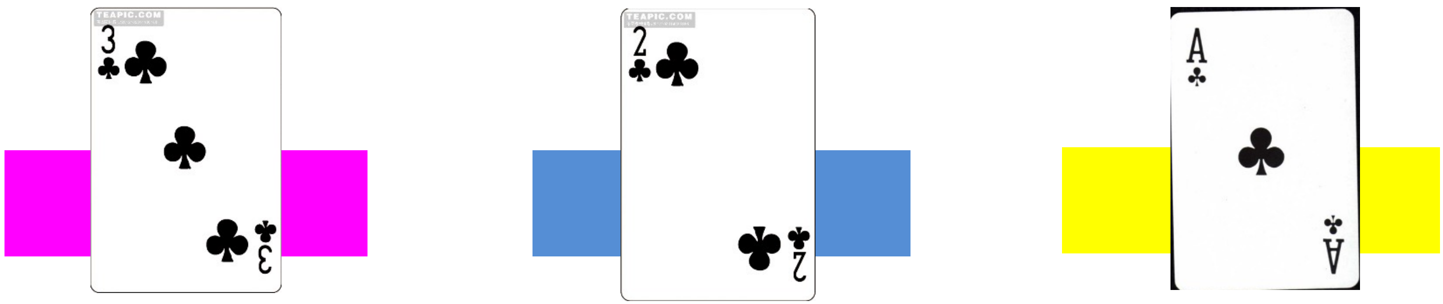
- Step1: Given n objects, initialize k cluster centers.
- Step2: Assign each object to its closest cluster center.
- Step3: Update the center for each cluster.
- Step4: Repeat 2 and 3 until no change in each cluster center.

K-means Demo

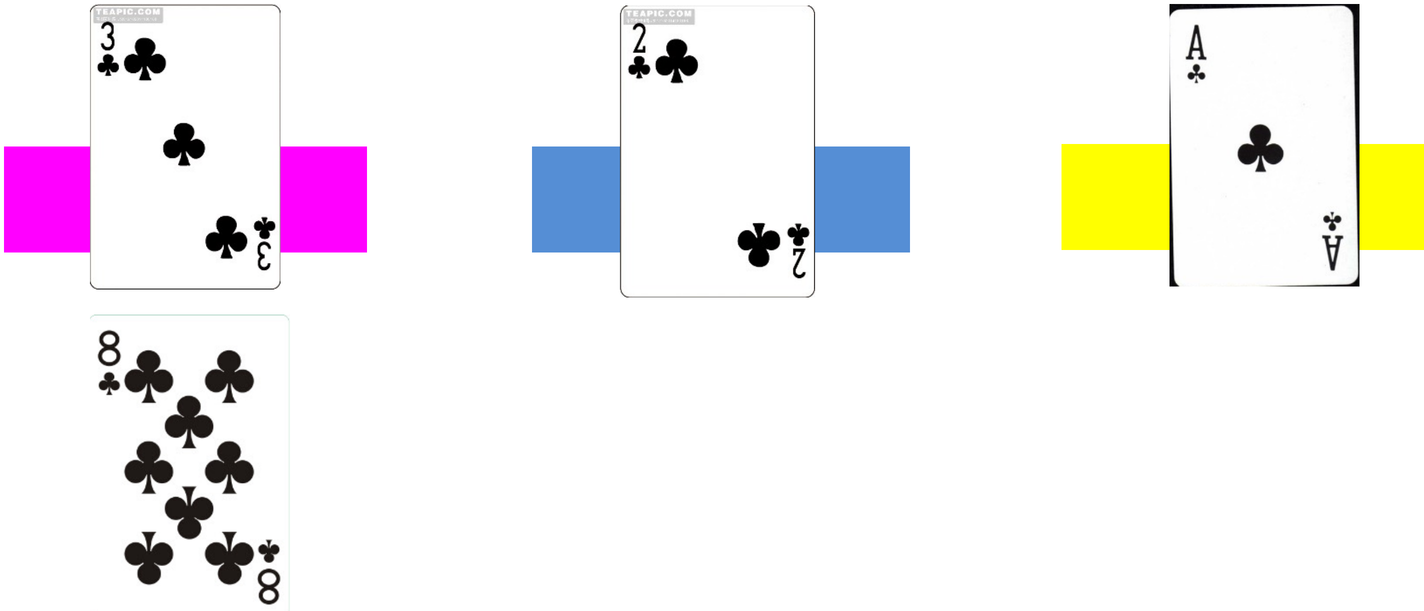
- K=3
- Group Pink　　Group Blue　　Group Yellow

Step1: Give k initial centers

- Random draw three cards as initial centers
- Initial center: 3, 2 ,1



Step2: Assign Each Card to Its Closest Center(1/2)



The node is “8”

Find the closest centroid:

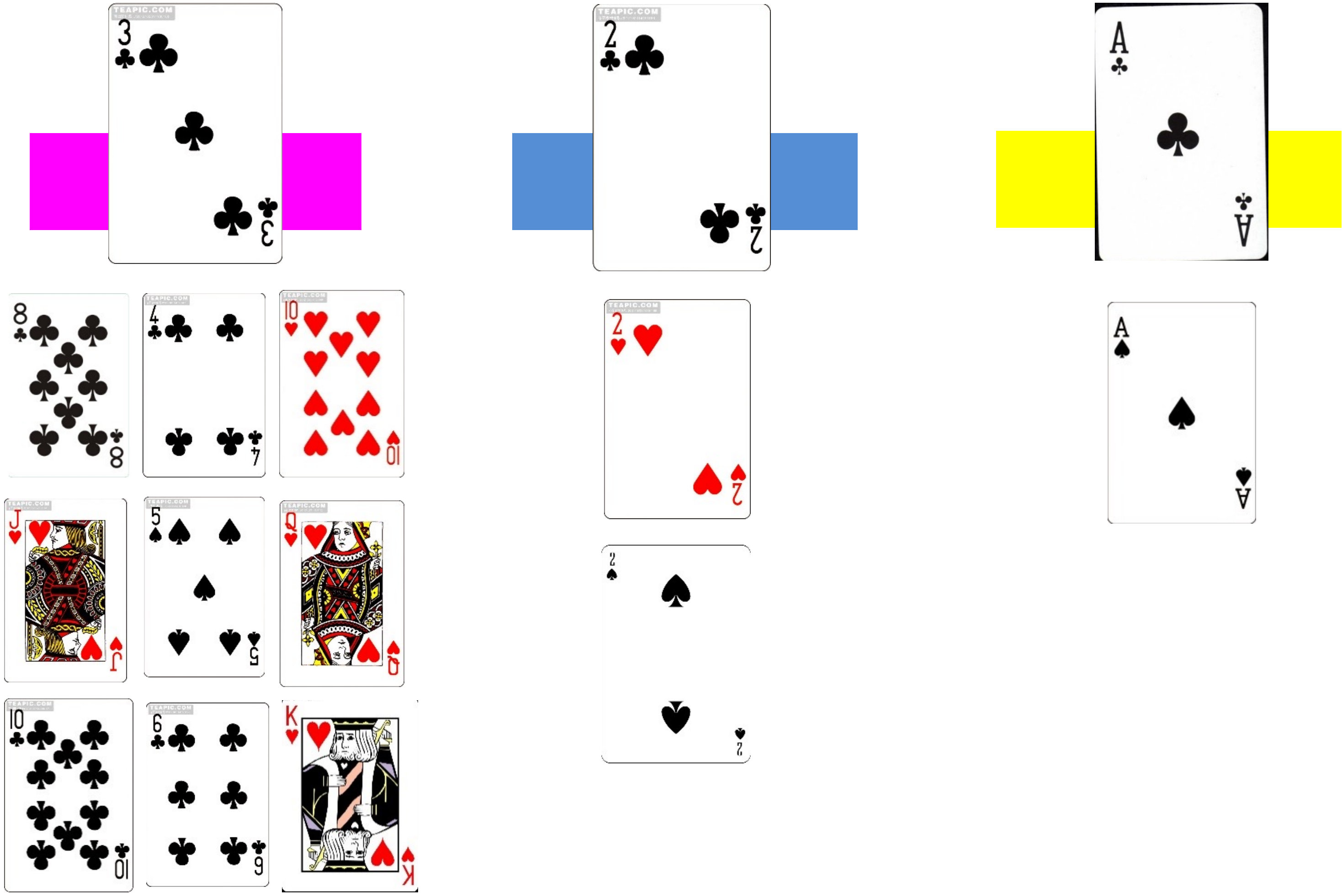
Current centorids:3,2,1

8-3=5 (Closest)

8-2=6

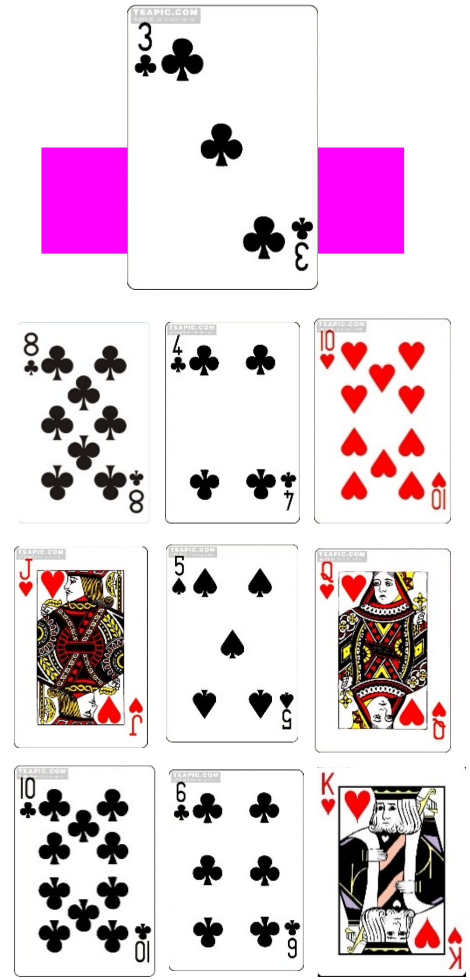
8-1=7

Step2: Assign Each Card to Its Closest Center(2/2)

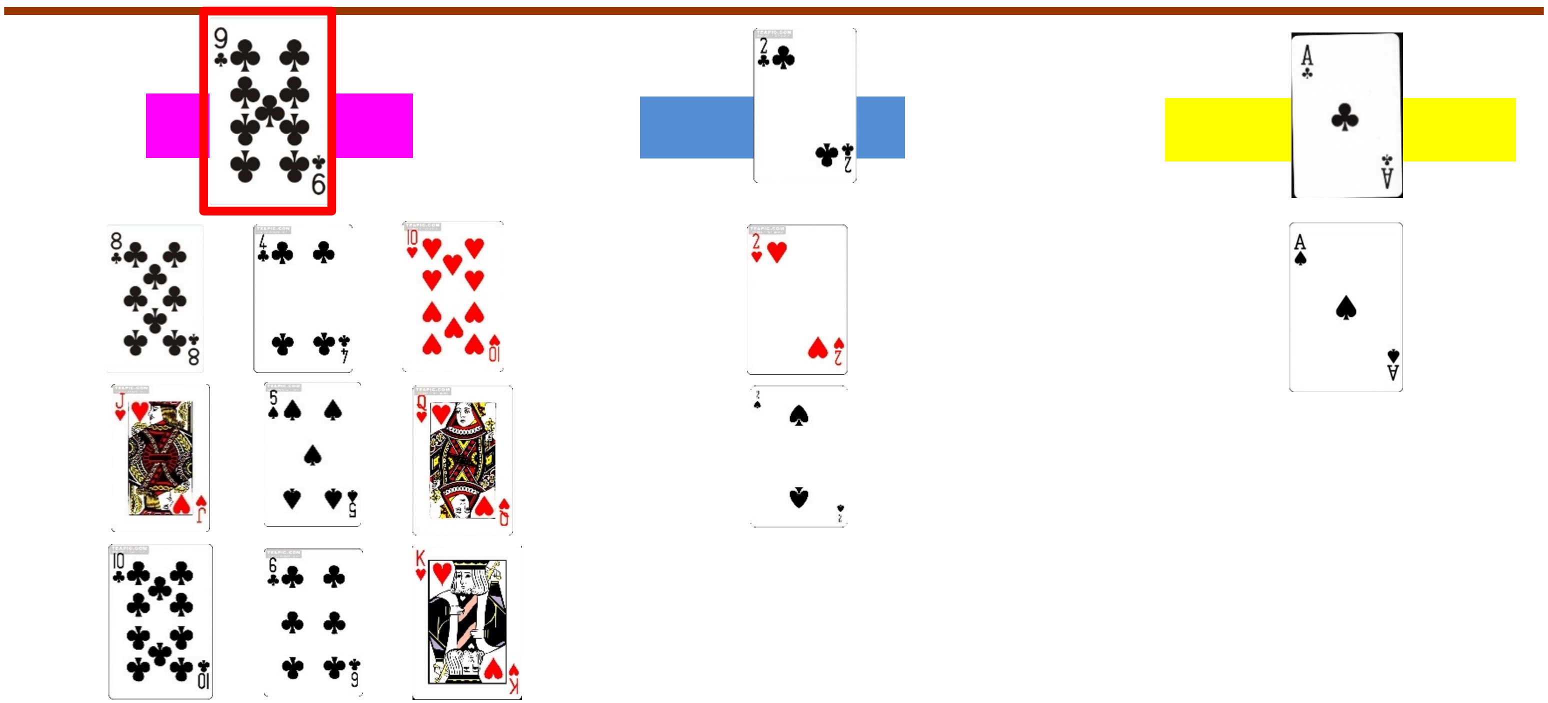


Step3:Update The Center for Each Group (1/2)

- Pink Group:
  - 8,4,10,11,5,12,10,6,13
  - Sum:8+4+10+11+5+12+10+6+13=79
  - # cards=9
  - Mean=79/9 => About 9



Step3:Update The Center for Each Group



Pink Group:

8,4,10,11,5,12,10,6,13
Sum:79
# cards=9
Mean=79/9 => About 9

Blue Group:

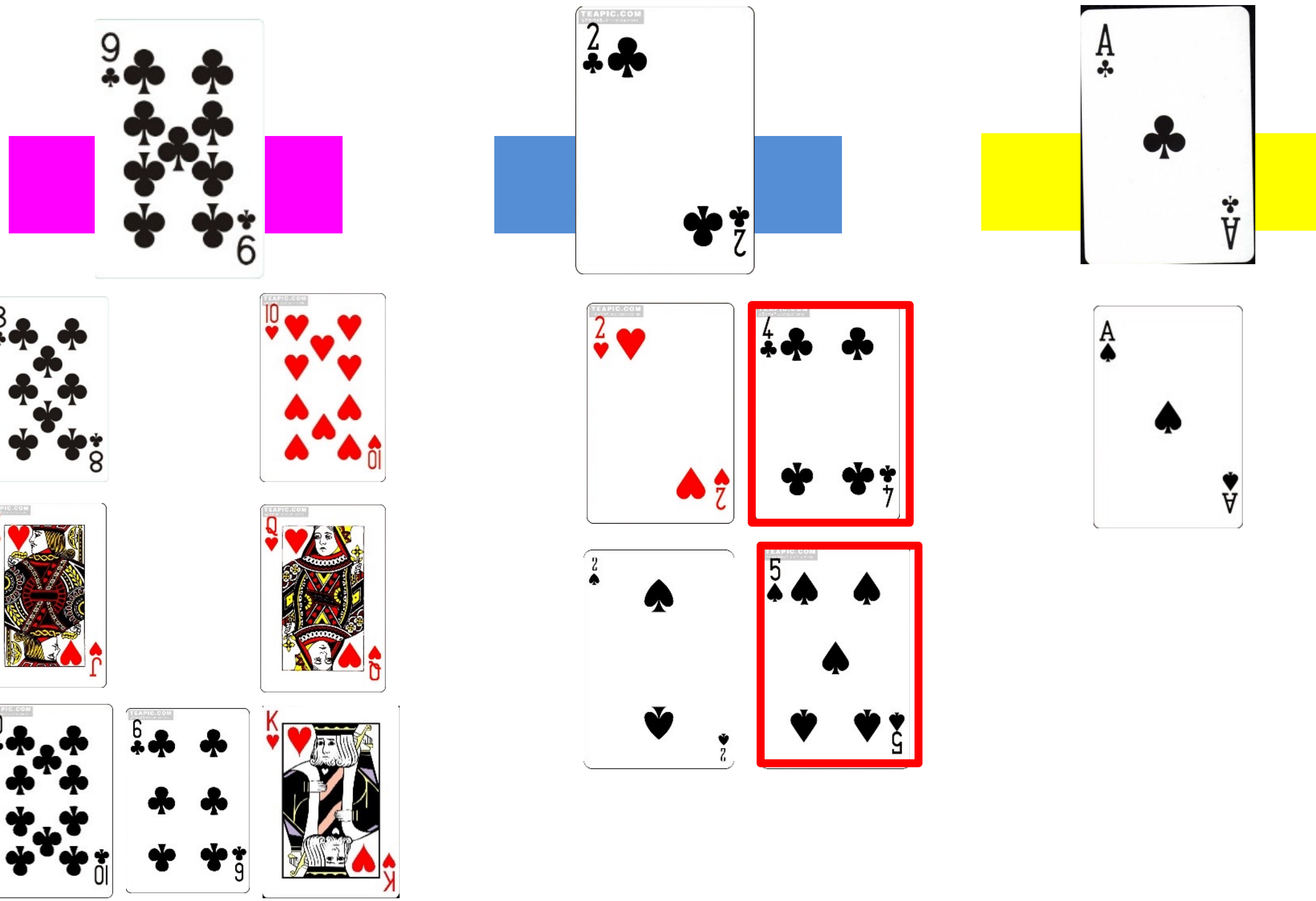
2,2
Sum:4
# cards=2
Mean=4/2 => 2

Yellow Group:

1
Sum:1
# cards=1
Mean=1/1 => 1

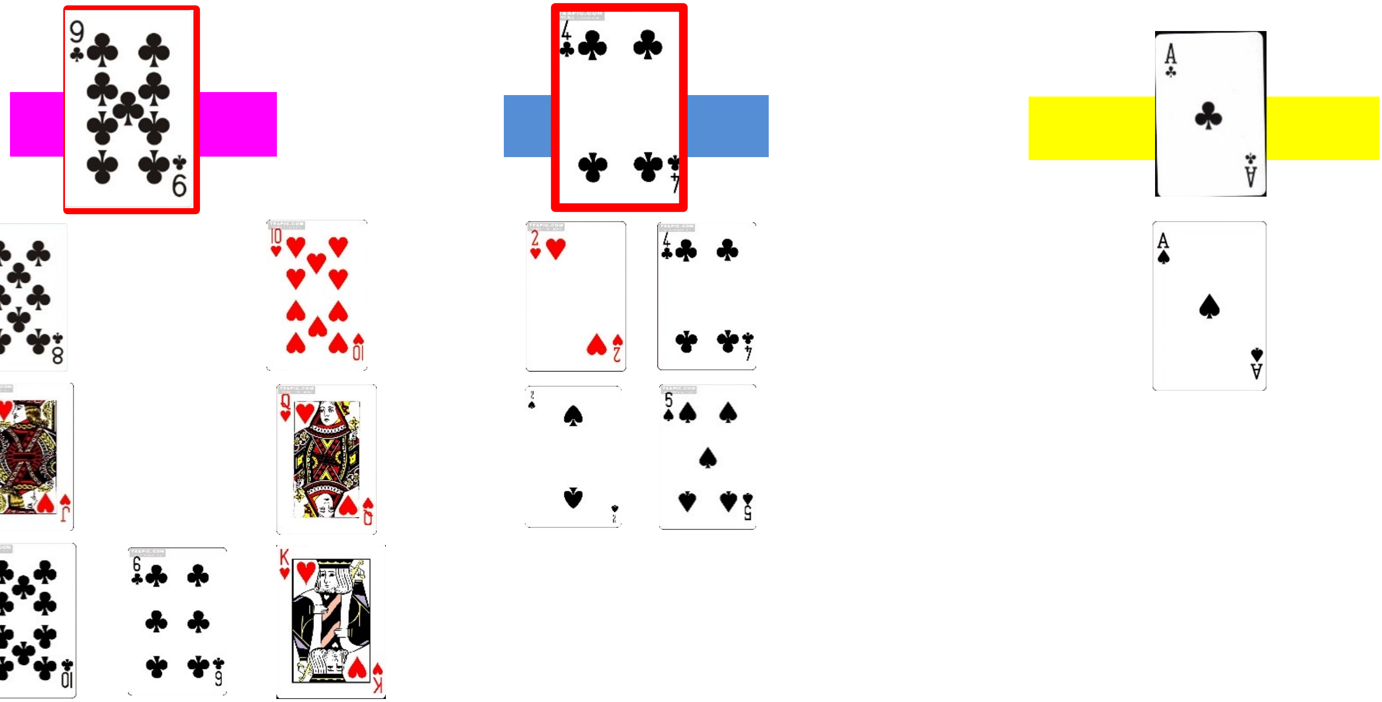
Step4: Repeat step2 and step3-Iteration2

- Update the cluster



Step4: Repeat step2 and step3- Iteration2

- Update the centroid



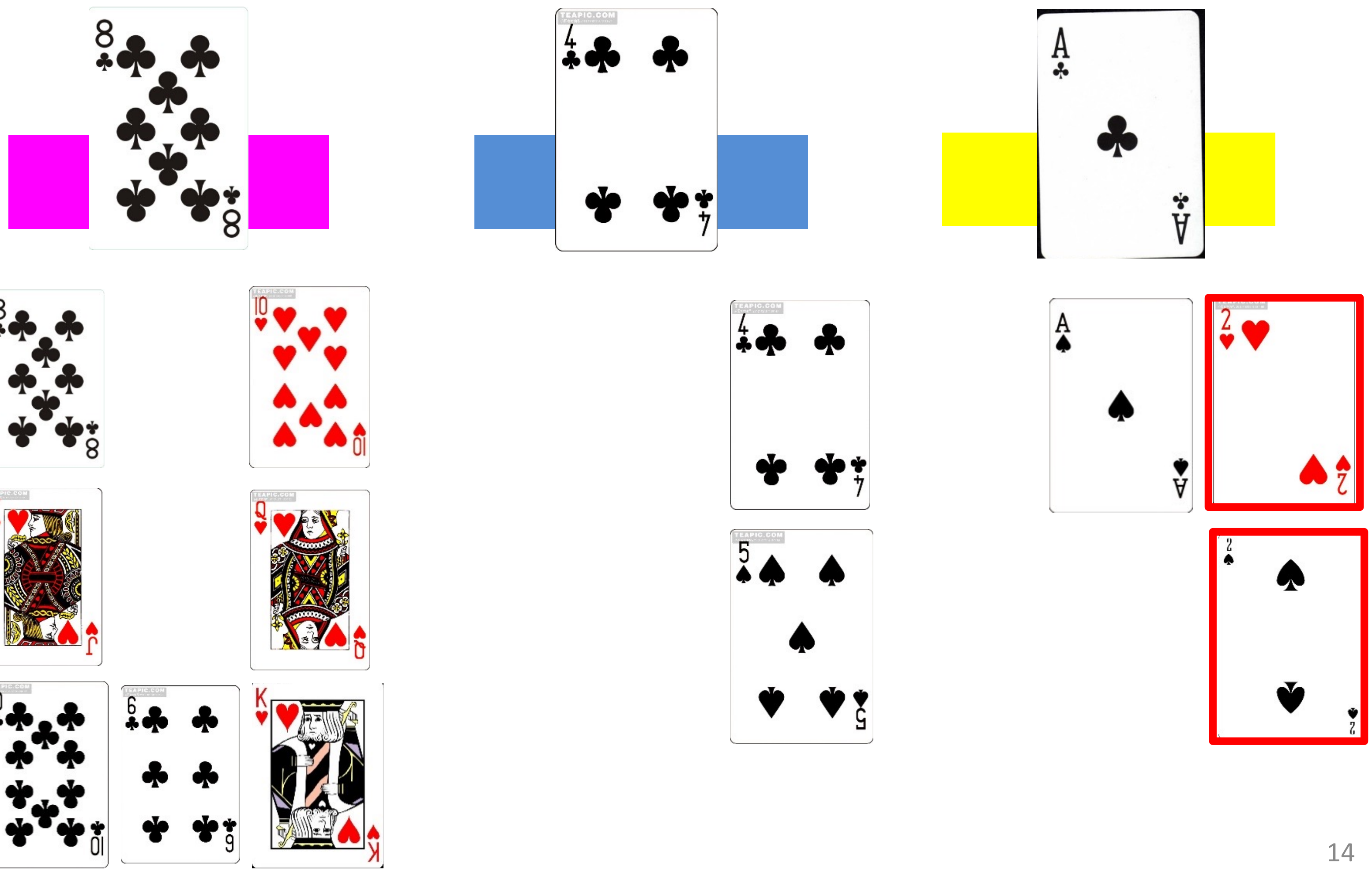
Pink Group:
8,10,11,12,10,6,13
Sum:70
# cards=9
Mean=70/9 => About 8

Blue Group:
2,2,4,5
Sum:13
# cards=4
Mean=13/4 => 4

Yellow Group:
1
Sum:1
# cards=1
Mean=1/1 => 1

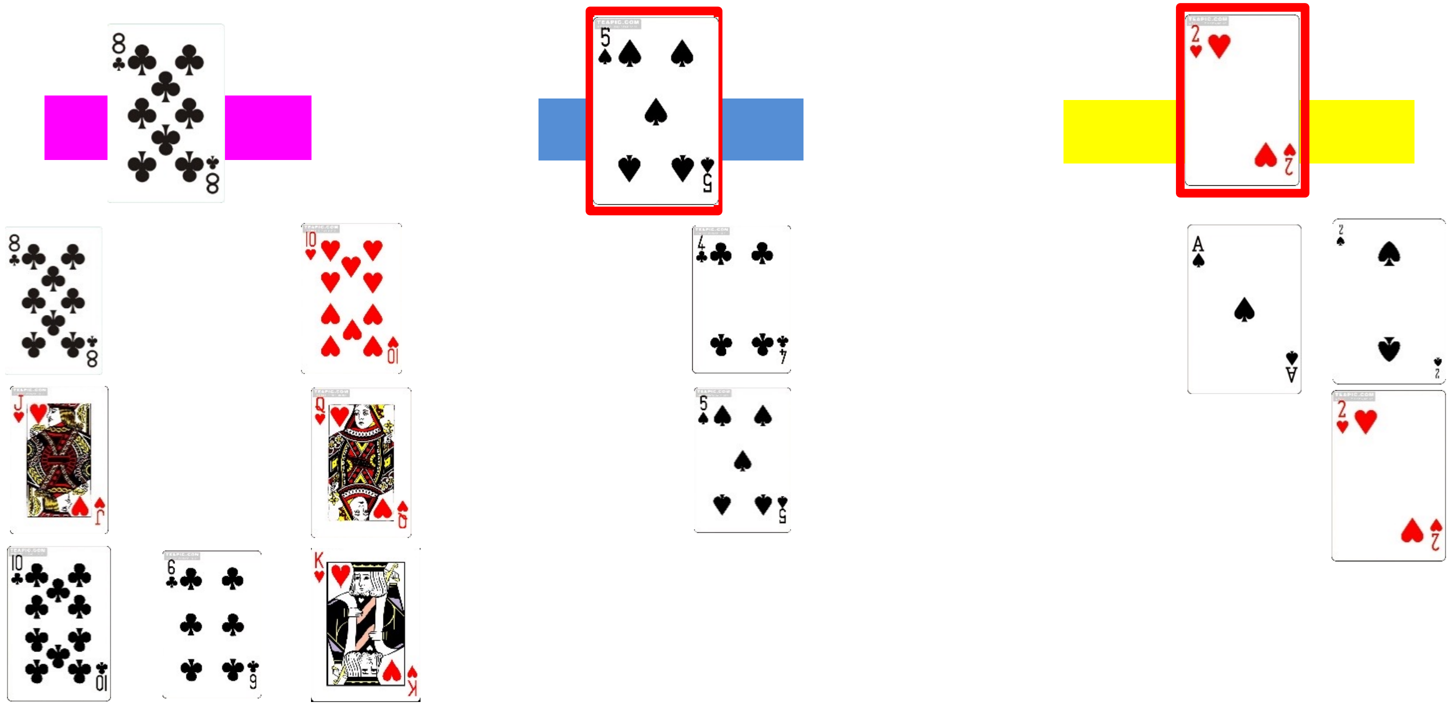
Step4: Repeat step2 and step3-Iteration3

- Update the cluster



Step4: Repeat step2 and step3-Iteration3

- Update the centroid



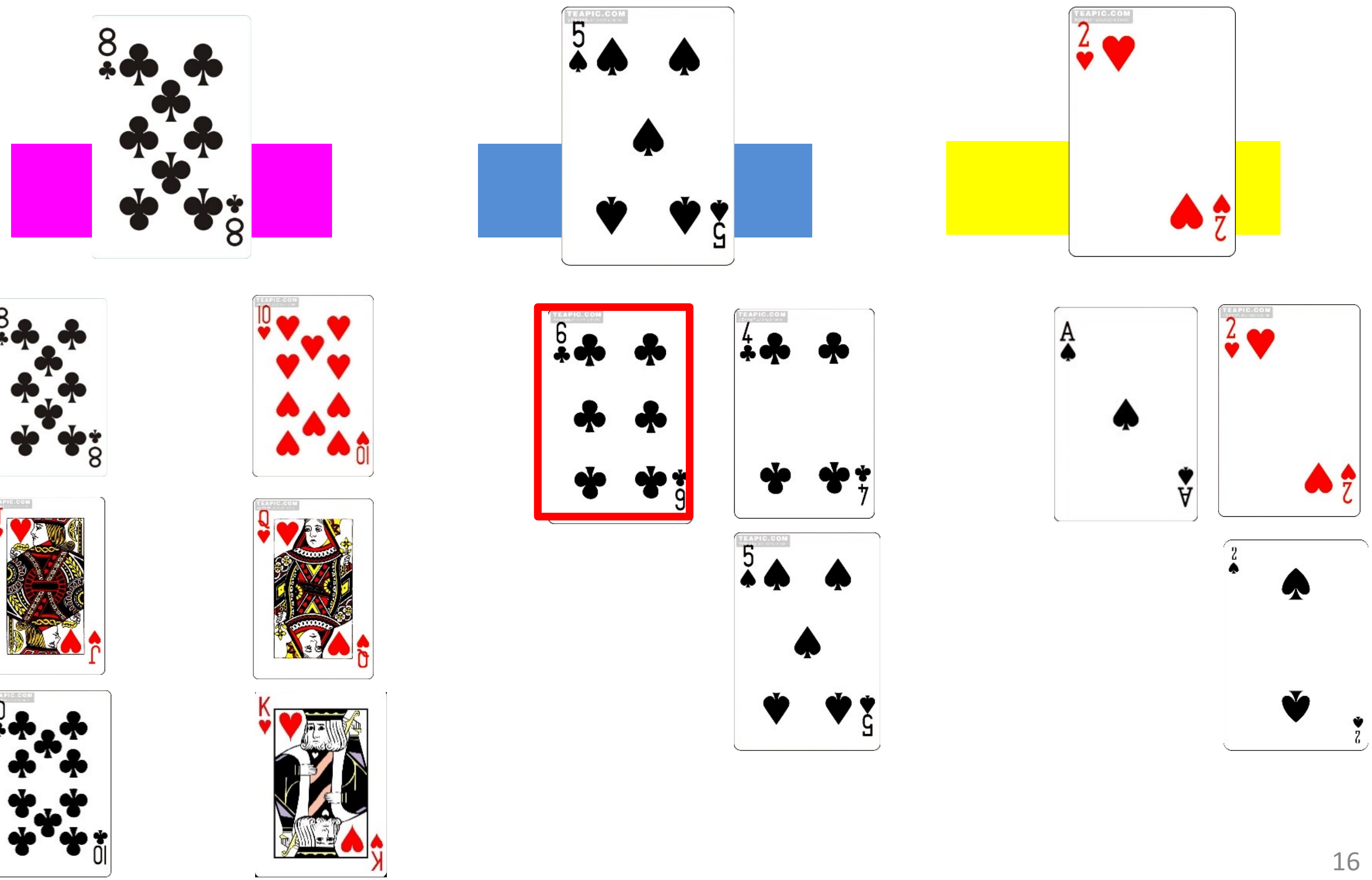
Pink Group:
8,10,11,12,10,6,13
Sum:70
# cards=9
Mean=70/9 => About 8

Blue Group:
4,5
Sum:9
# cards=2
Mean=9/2 => 5

Yellow Group:
1,2,2
Sum:5
# cards=3
Mean=5/3 => 2

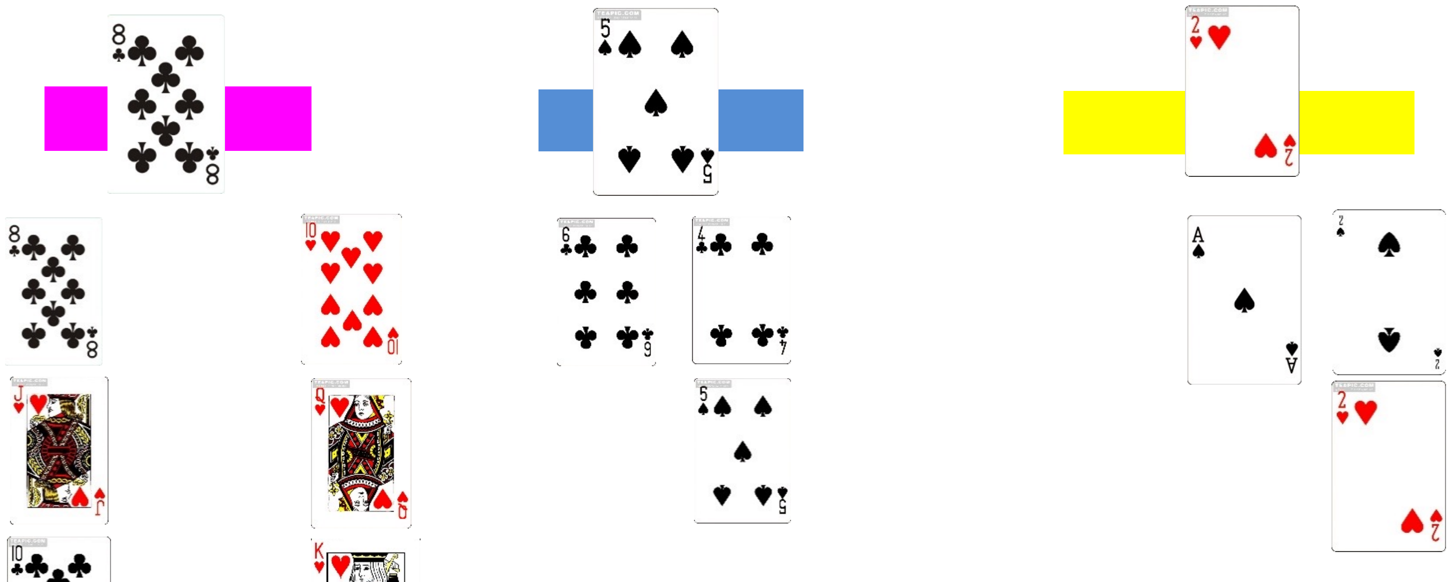
Step4: Repeat step2 and step3-Iteration4

- Update the cluster



Step4: Repeat step2 and step3-Iteration4

- Update the centroid



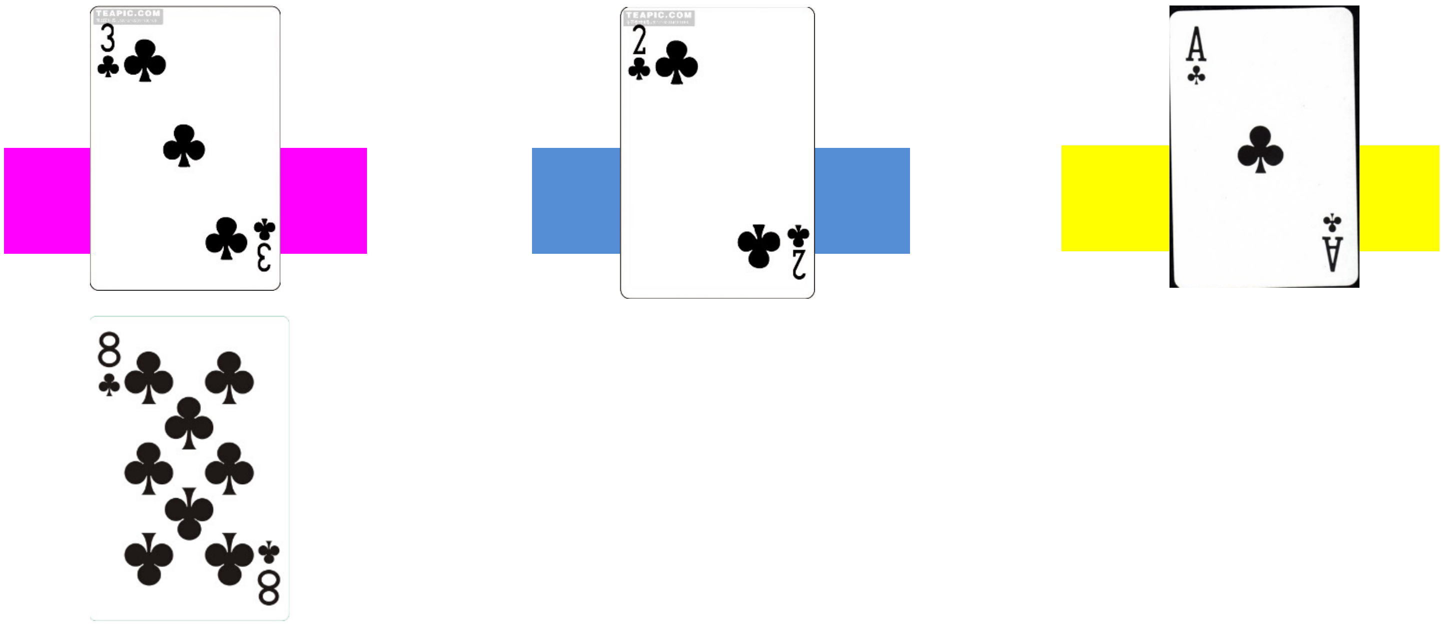
The center of every group does not change.!!!

Pink Group:
8,10,11,12,10,6,13
Sum:70
# cards=9
Mean=70/9 => About 8

Blue Group:
4,5,6
Sum:15
# cards=3
Mean=15/3 => 5

Yellow Group:
1,2,2
Sum:5
# cards=3
Mean=5/3 => 2

Distance Computation



The node is “8”
Find the closest centroid:
Current centorids:3,2,1
8-3=5 (Closest)   =>Calculate the distance
8-2=6
8-1=7

Distance Measure Method

- **Euclidean distance measure**:
  - Simplest
  - The Euclidean distance between point *p* and *q* in N-dimensional space is given as:

$$d=(p,q)=\sqrt{\sum_{i=1}^{N}(p_i - q_i)^2}$$

- **Cosine distance measure**:
  - Finds the cosine of angle between two vectors (vectors drawn from origin to the points.)

$$d = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

- **Manhattan distance measure**:
  - The sum of the absolute differences of the coordinates of two points.

$$d=(p,q) = |\sum_{i=1}^{N} p_i - q_i|$$

The Drawback of K-means

- The parameter of K-means:
  - Must decide the number of cluster in advance.
  - Different initial center will result in different cluster result.
- The center of K-means can be virtual node.

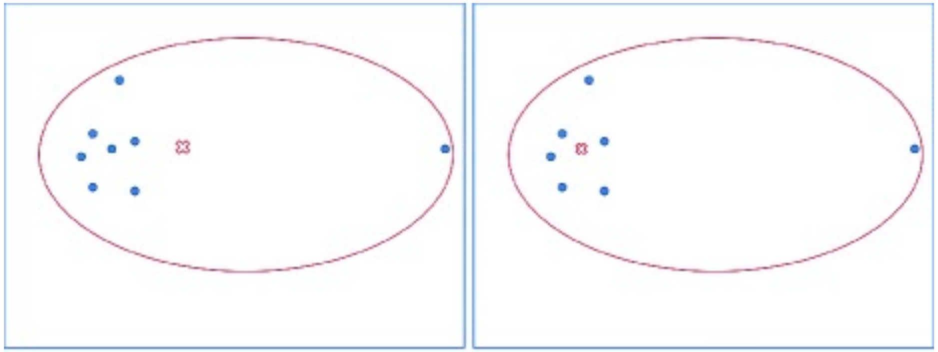
- Drawback:
- K-means cannot deal with category data.
- K-means is heavily affect by noise(離群值).
  - K-medoids

K-medoids

- Step1: Given n objects, initialize k cluster centers.
- Step2: Compute the distance of each object and cluster centers. Assign each object to its closest cluster center.
- Step3: Update the center for each cluster.
- Step4: Repeat 2 and 3 until no change in each cluster center.

- Same with K-means?
- Update the node which can make the sum of distance becomes minimum.

K-means vs. K-medoids



(a) Mean (b) Medoid

	K-means	K-mediods
Center	Virtual node	Real node
The method to update center	The mean of nodes in the cluster.	The node which can make the sum of distance be minimum.

Outline

- K-means
- K-medioids
- Hierarchical Clustering
- Density Based Clustering (DBSCAN)

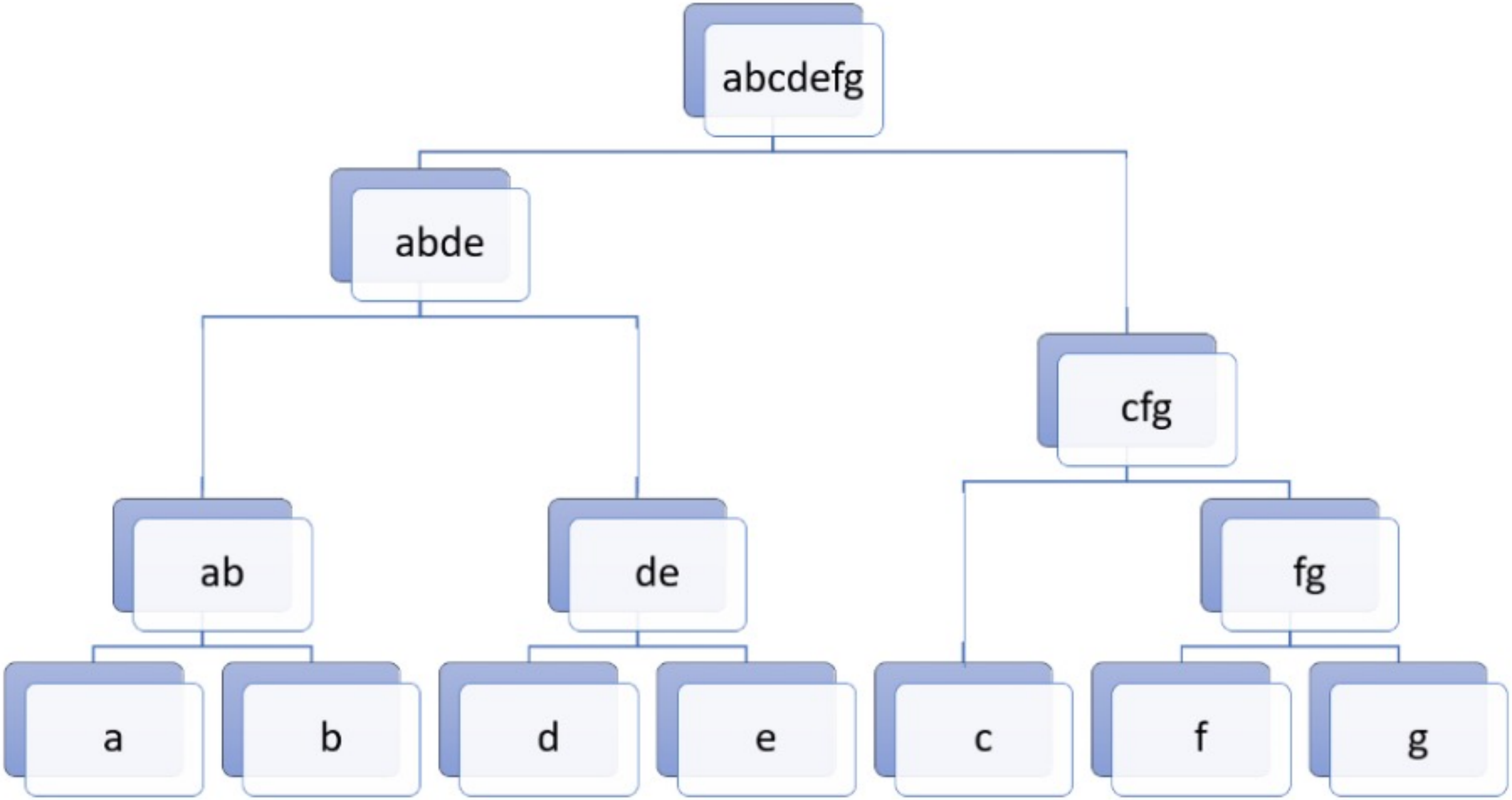
Hierarchical Clustering

- Hierarchical clustering (階層式分群法) is a hierarchical method which generate the clusters by iteratively (聚合) or divisive (分裂) data.

Agglomerative (1/2)

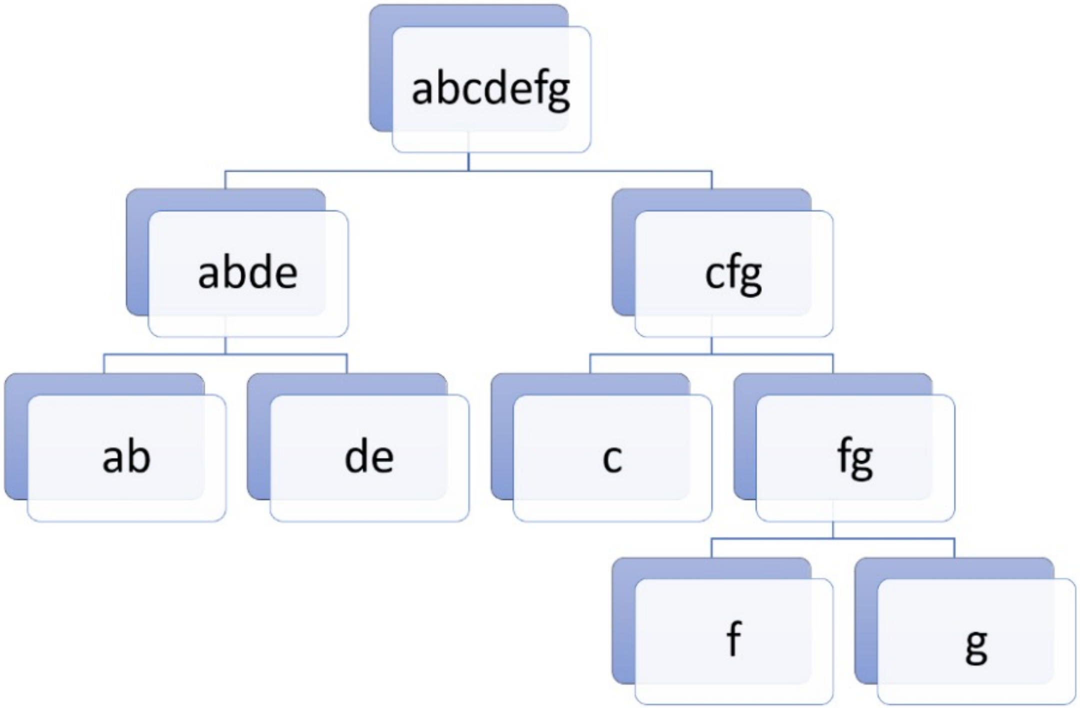
- It is a “bottom-up” method.
- Prepare basic components and iteratively combine the components to be a final solution.

Agglomerative (2/2)



Divisive

- It is a “top-down” method.
- See the whole picture of the problem and iteratively add the detail to make the solution clear.
- Regard the data as a cluster and iteratively divide the data.



Steps of Agglomerative

- Step1: Every node is a cluster.
- Step2: Scan all the nodes. Choose two nodes which are closest to be a cluster.
- Step4: Repeat 2 and 3 until all data becomes a cluster or achieve the x cluster.

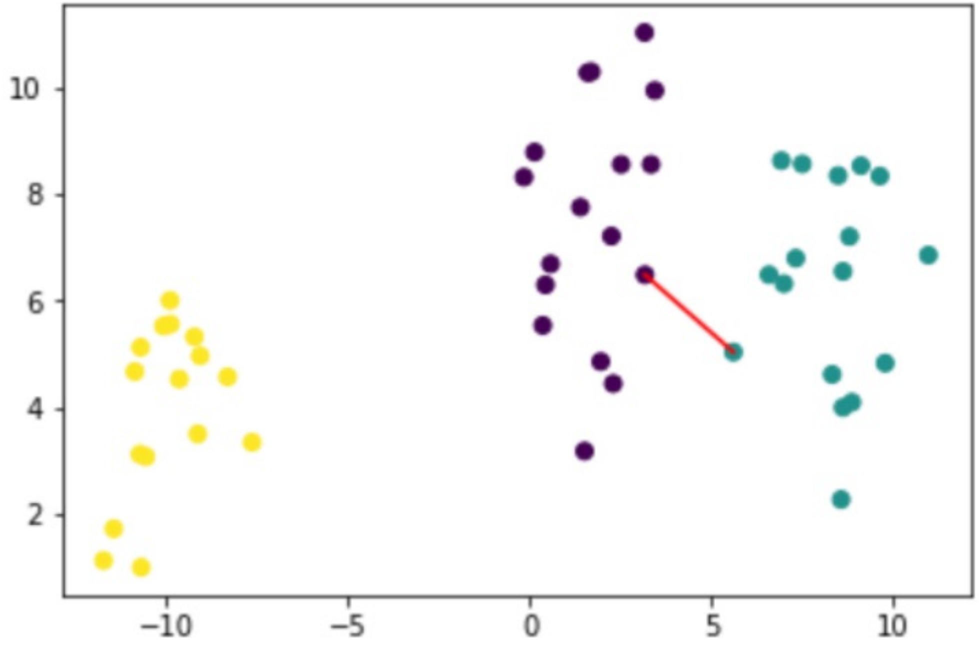
Distance of Two Clusters

- Single-linkage agglomerative algorithm (單一連結聚合演算法)
- Complete-linkage agglomerative algorithm (完整連結聚合演算法)
- Average-linkage agglomerative algorithm (平均連結聚合演算法)
- Centroid method (中心聚合演算法)
- Ward’s method (沃德法)

Single-linkage Agglomerative Algorithm

- The distance is defined as the distance between the closest points in the two clusters.

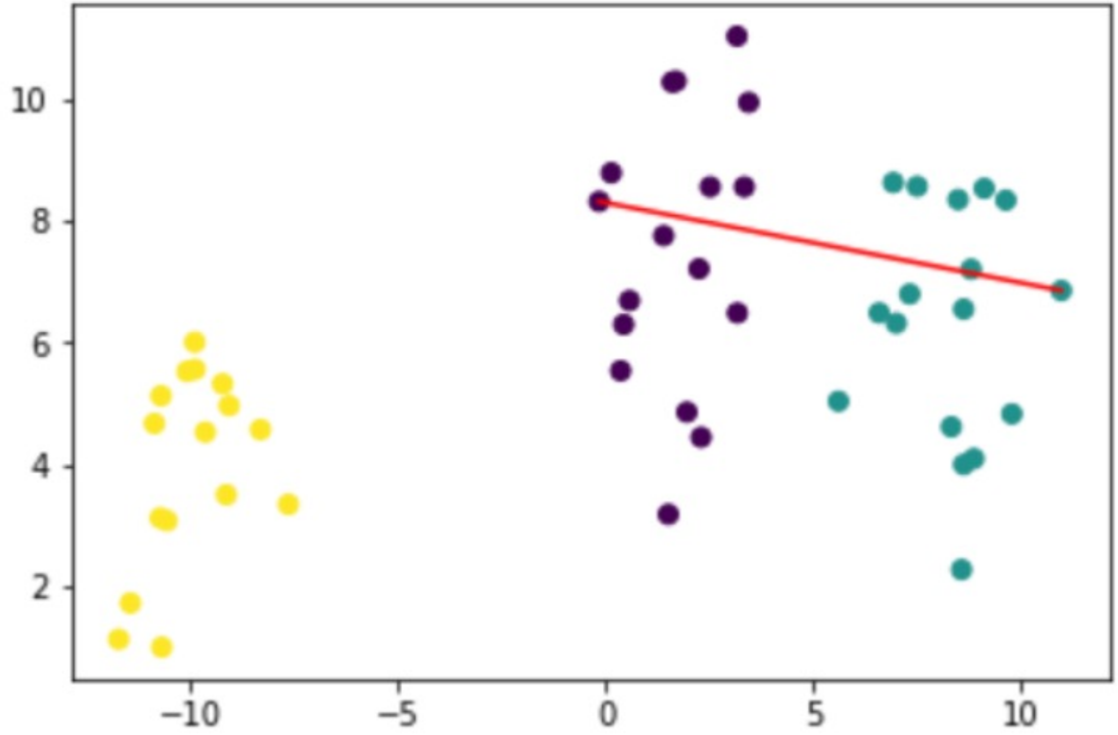
$$d(C_i, C_j) = \min_{a \in C_i,\, b \in C_j} d(a, b)$$



Complete-linkage Agglomerative Algorithm

- The distance is defined as the distance between the furthest points in the two clusters.

$$d(C_i, C_j) = \max_{a \in C_i,\, b \in C_j} d(a, b)$$



Average-linkage Agglomerative Algorithm

- The distance is defined as the mean of the sum of the distance between the points in the two clusters.

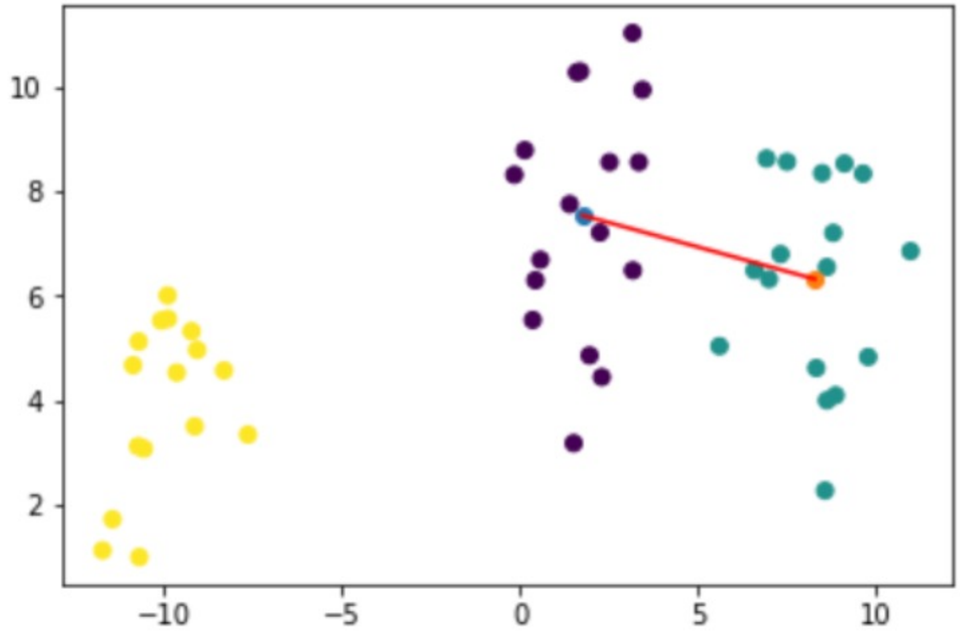
$$d(C_i, C_j) = \sum_{a \in C_i,\, b \in C_j} \frac{d(a, b)}{|C_i||C_j|}$$

Centroid Method

- The distance is defined as the distance between center points in the two clusters.

$$d(C_i, C_j) = \|\mu_{C_i}, \mu_{C_j}\|$$

mu_C指的是C集合中的平均值



紅色線的長度即為中心聚合算法的距離（藍色點為紫色資料點的中心點，橘色則為綠色資料點的中心點）

Ward's Method

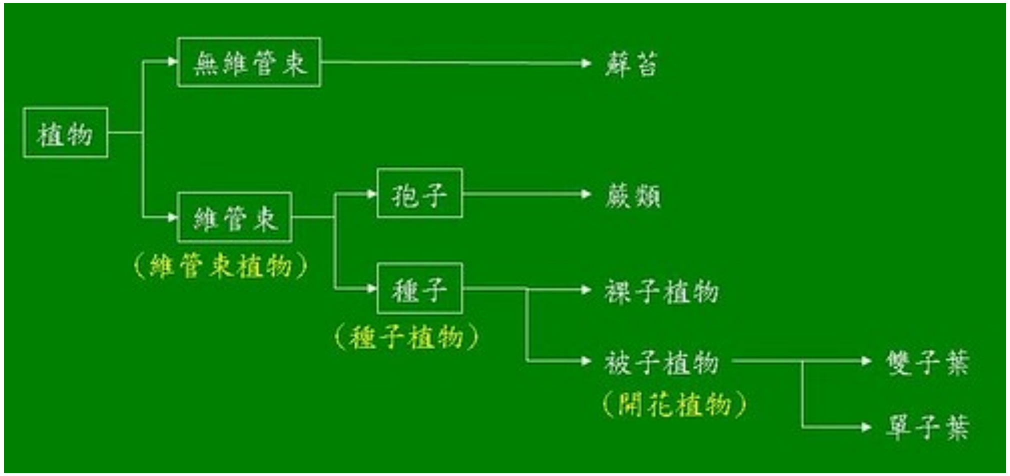
- The distance is defined as the sum of the square distance between every point and the new center point which is generated after two cluster merge.

$$d(C_i, C_j) = \sum_{a \in C_i \bigcup C_j} \|a - \mu_{C_i \bigcup C_j}\|$$

- The method can be regarded as finding the similarity of two clusters. Merging the clusters which have higher similarity.

Drawback of Hierarchical Clustering

- Define the distance measure of two clusters.
- Define the number of cluster.
- Suitable for biological clustering.

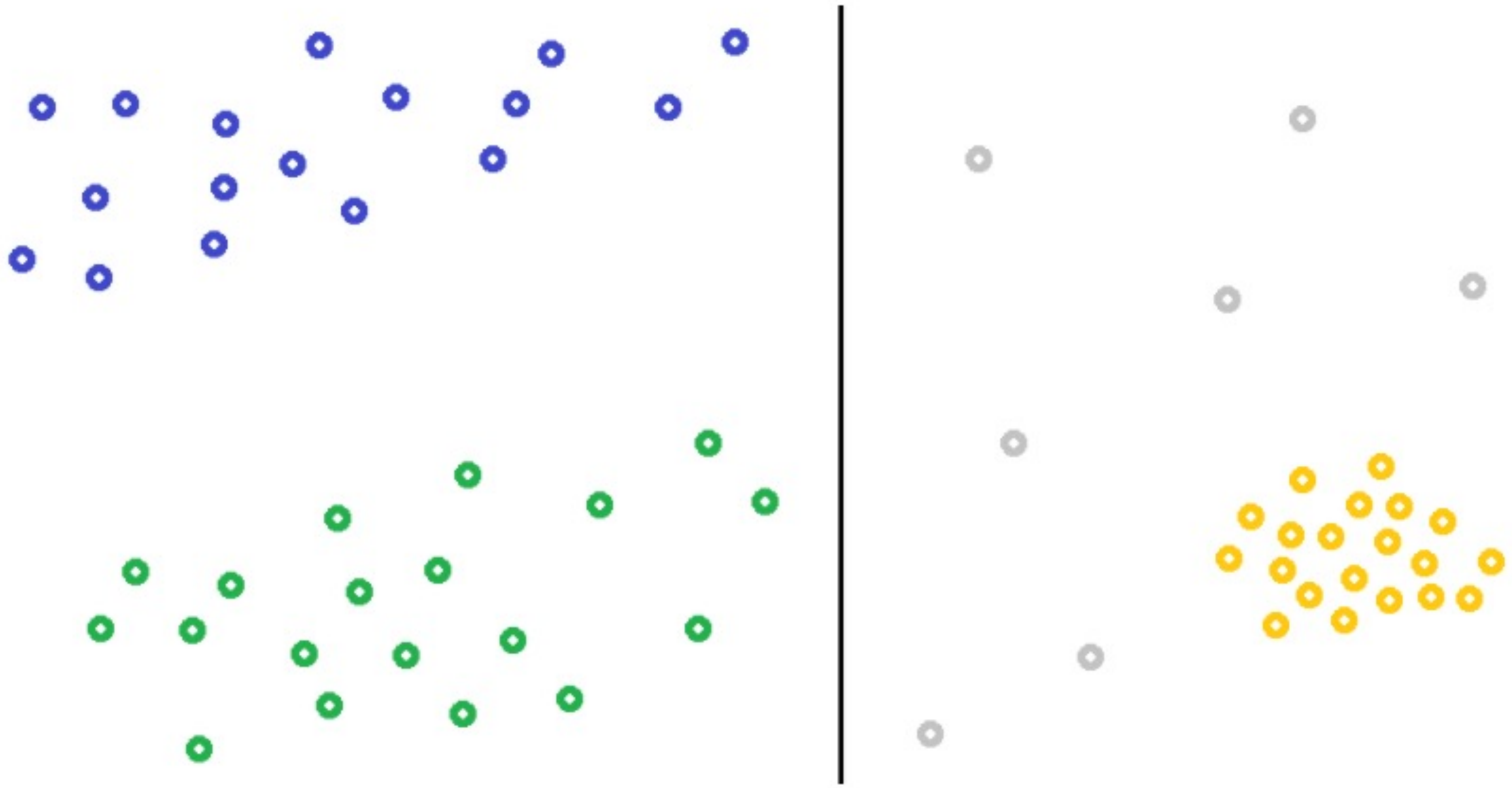


- Drawback:
- Hierarchical clustering needs much computation resource since the method has to scan every data in each iteration.

Density Based Clustering (DBSCAN)

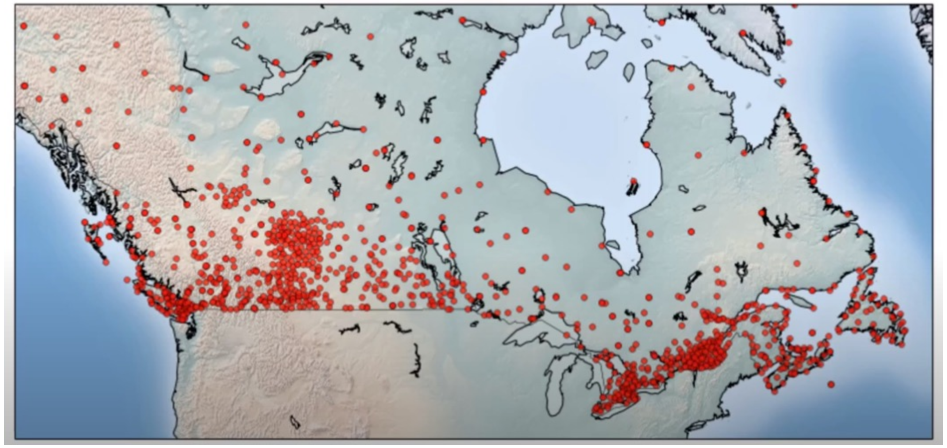
K-means can find good clusters!

K-means cannot find good clusters.

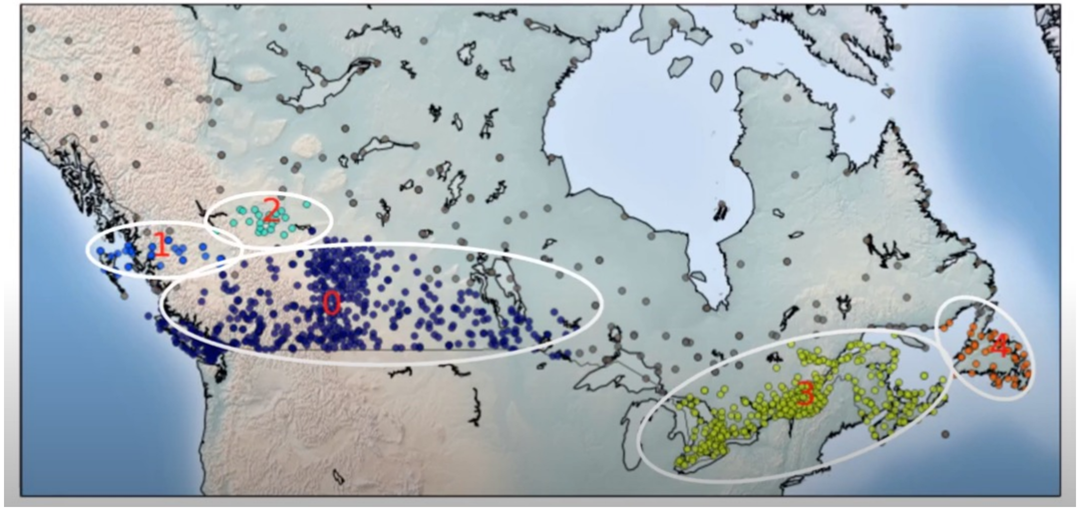


Example of Density Based Clustering

- The weather station of Canada.

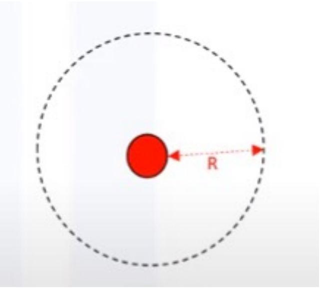


Use DBSCAN to find the cluster which show the same weather condition.

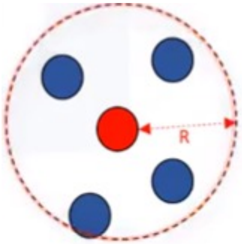


DBSCAN

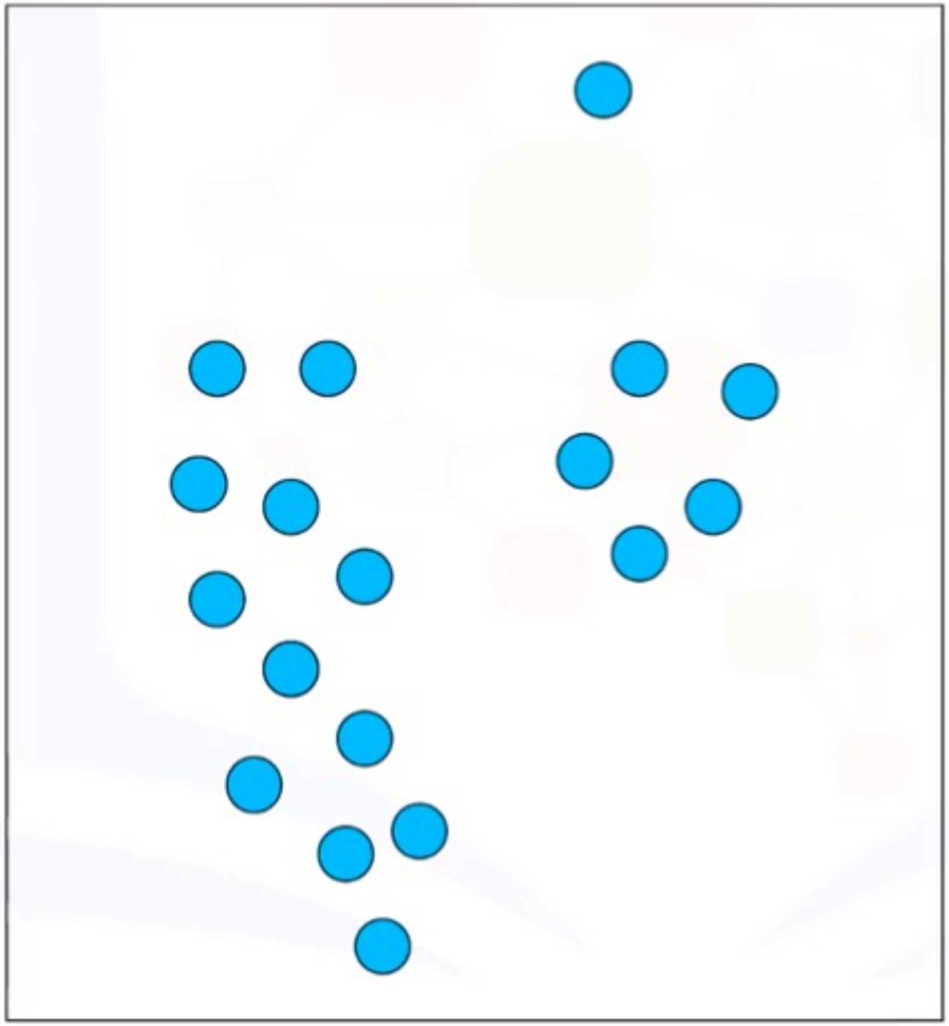
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - One of the most common clustering algorithms.
  - Works based on density of objects.



- R (Radius of neighborhood)
  - Radius (R) that if includes enough number of points within, we call it a dense area.



- M (Min number of neighbors)
  - The minimum number of data points we want in a neighborhood to define a cluster.



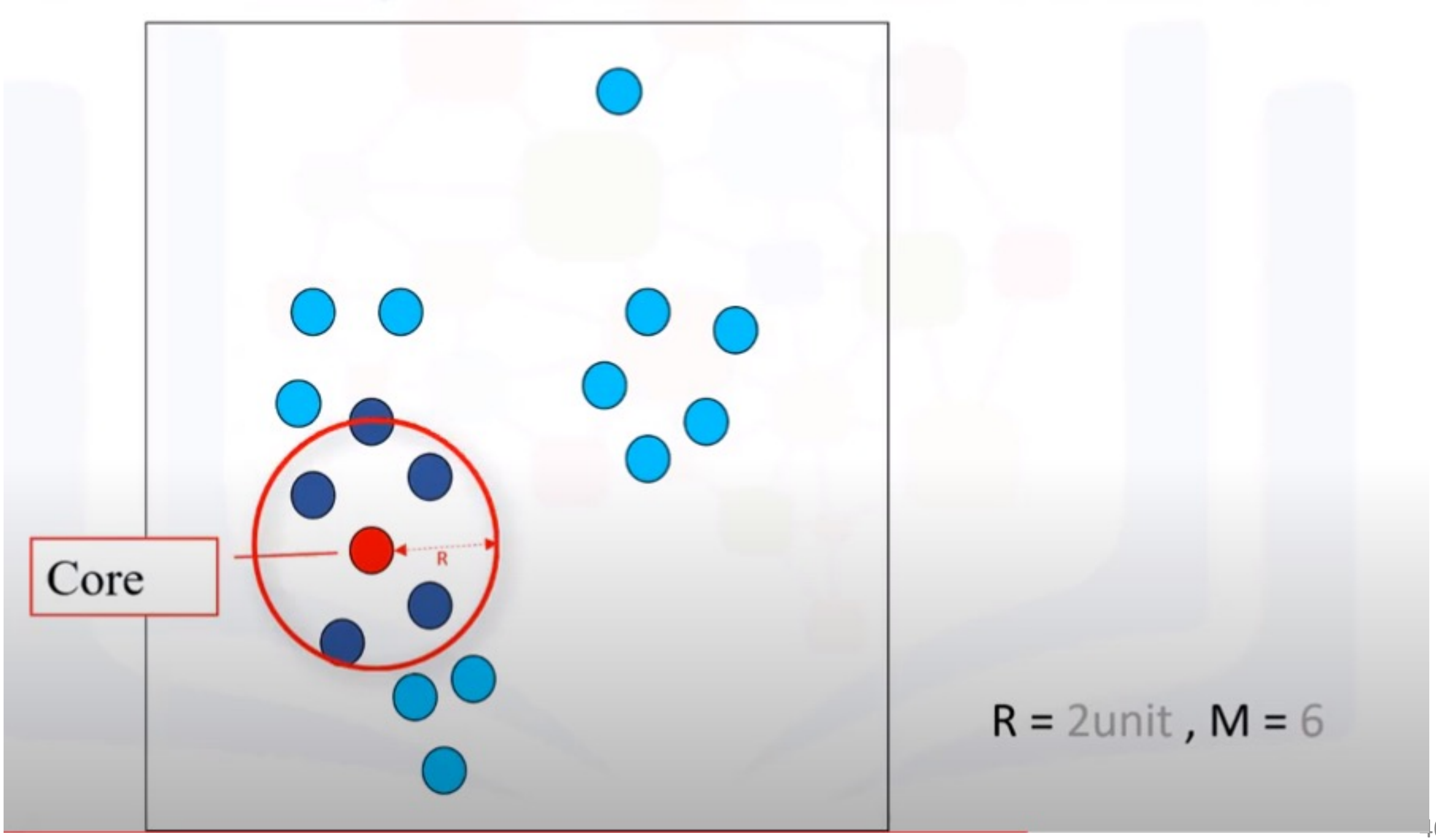
Each point is either:

- *core point*
- *border point*
- *outlier point*

R = 2unit , M = 6

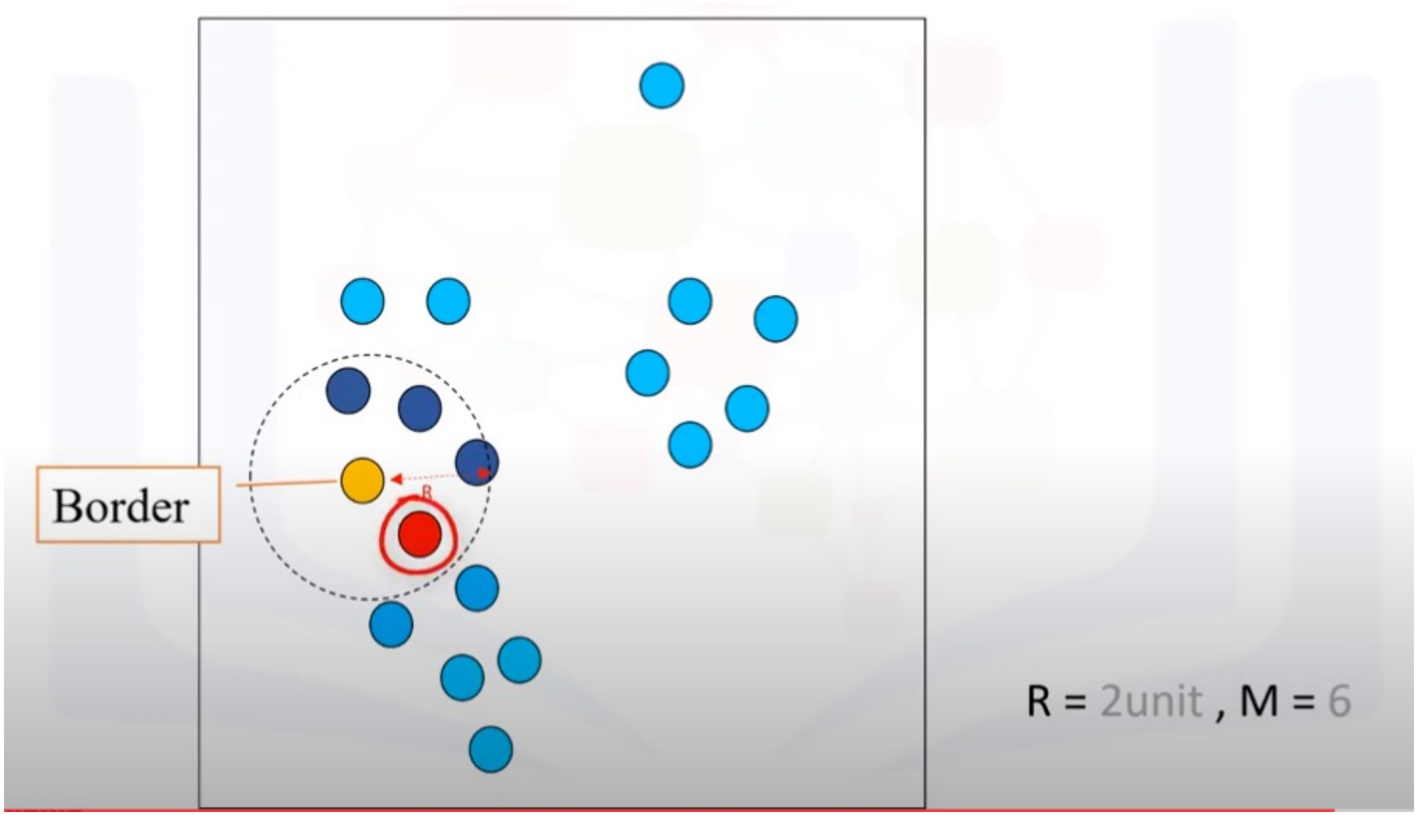
Core Point

- Core point: Within R neighborhood of the point, there are at least M points.



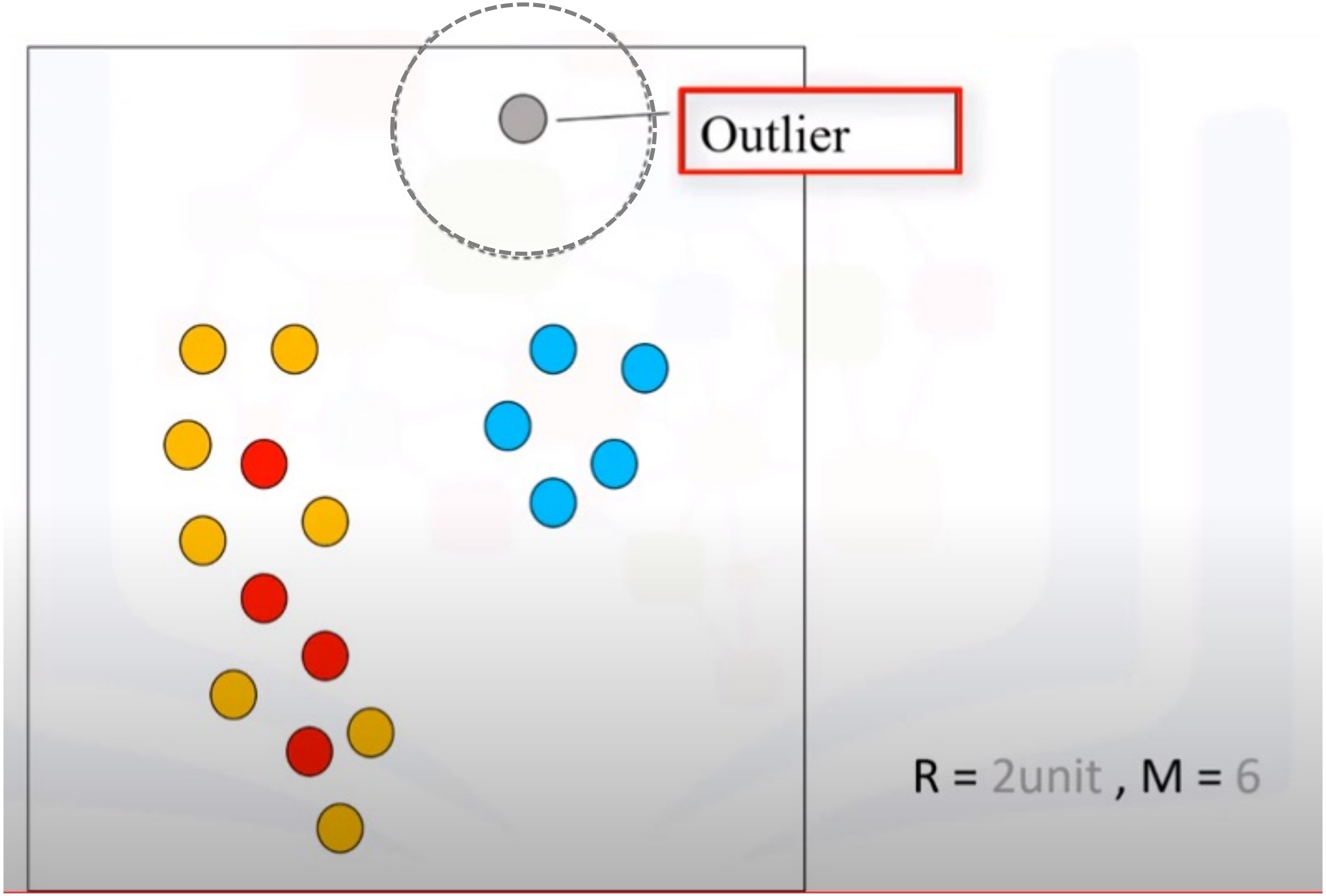
Border Point

- Border point: Its neighborhood contains at least M data point or it is reachable from some core points.
- Reachable: It is within R distance from the core point.



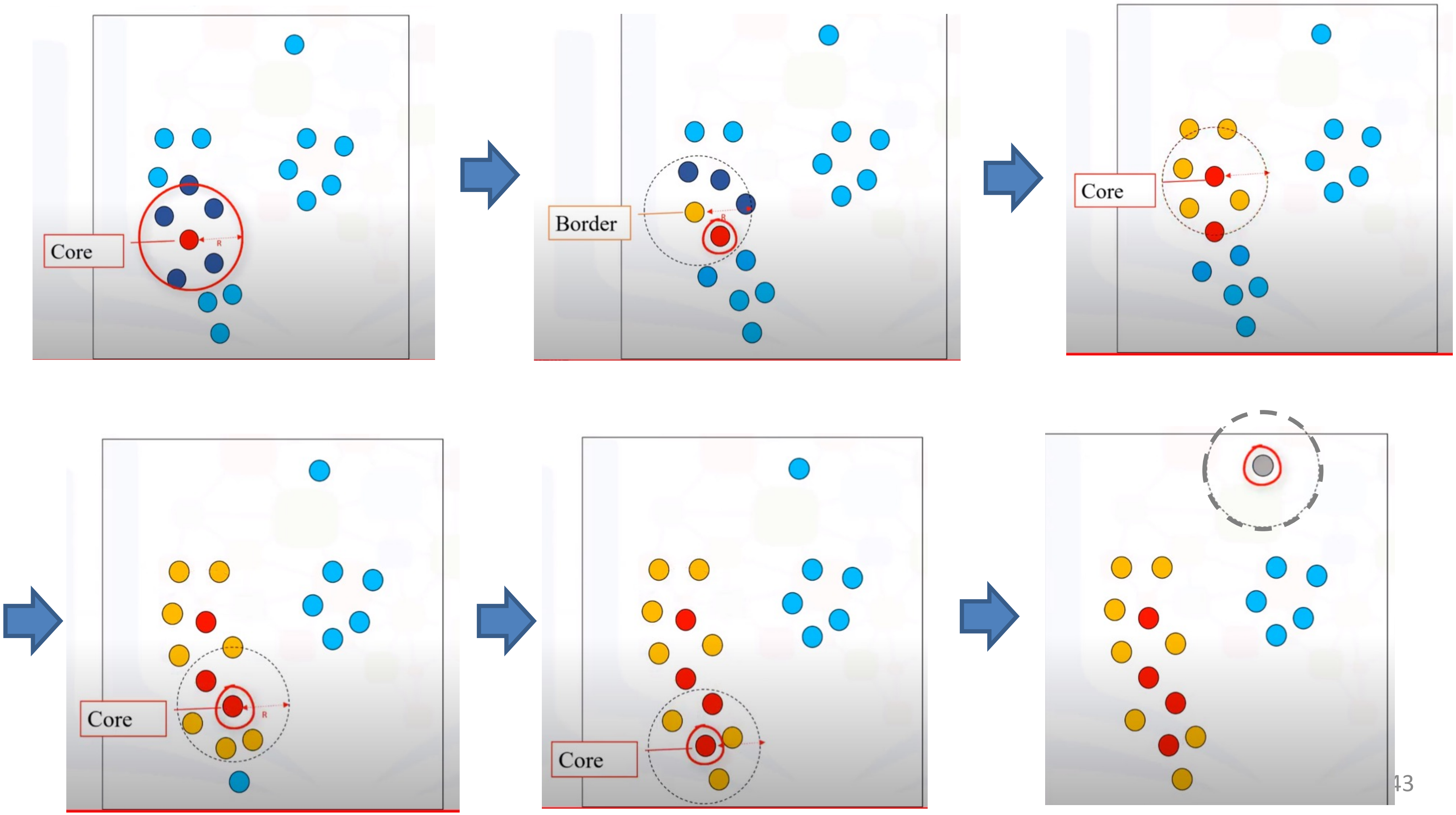
Outlier Point

- Not a core point nor a board point => outlier point

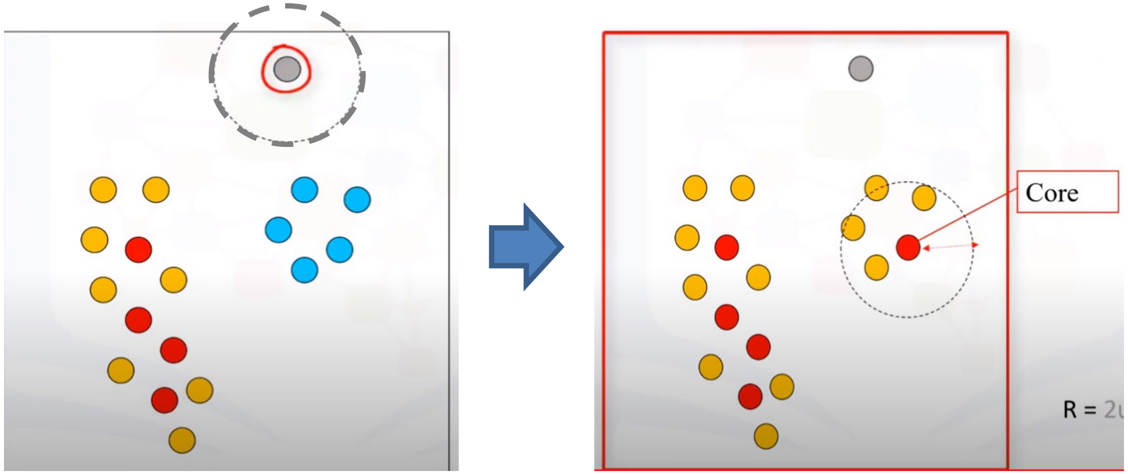


Step1 of DBSCAN (1/2)

- Step1: Label points.

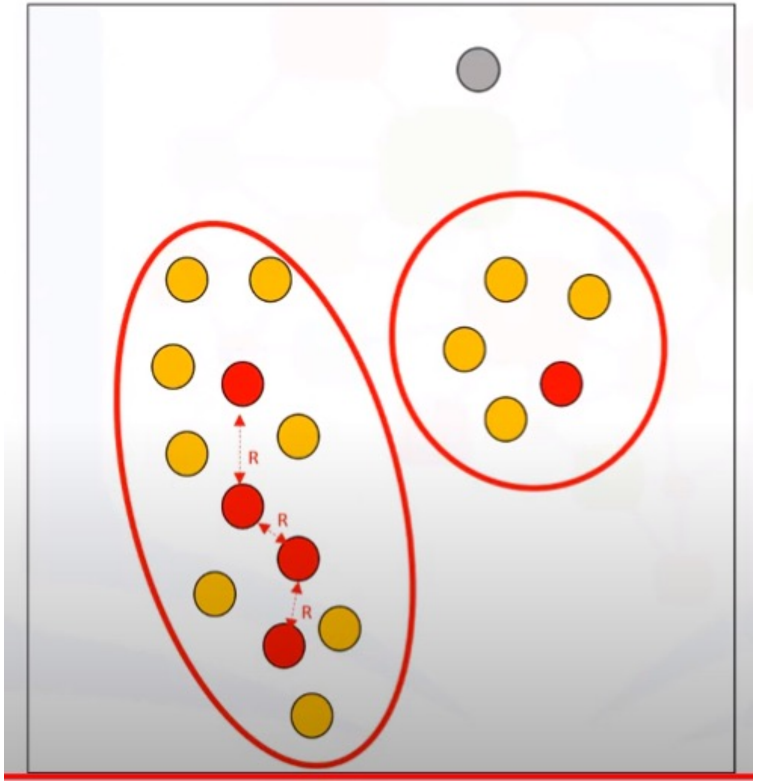


Step1 of DBSCAN (2/2)



Step2 of DBSCAN

- Step2: Connect Core Points that are neighbors and put them in the same cluster.



- Cluster is formed by at least one core point and all reachable border points.

Advantages of DBSCAN

- 1. Arbitrarily shaped clusters.
- 2. Robust to outliers.
- 3. Does not require specification of the number of clusters.