K Means Clustering

Jony Sugianto
jony@evolvemachinelearners.com
0812-13086659
github.com/jonysugianto

What is Clustering?

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

What is clustering?

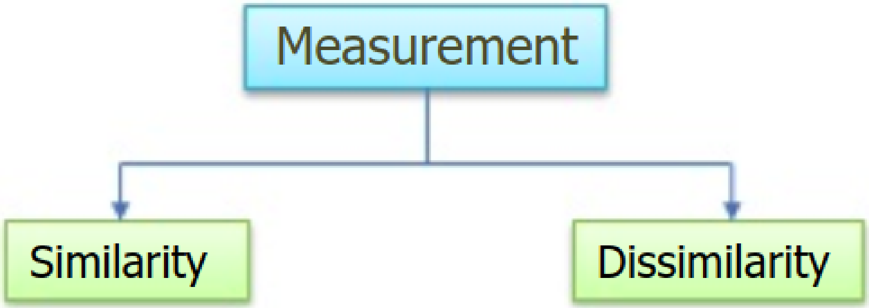
Organizing data into clusters such that there is:

- High intra-cluster similarity
- Low inter-cluster similkarity
- Informally, finding natural groupings among objects

Why clustering?

- Organizing data into clusters shows internal structure of the data
  - Clusterung genes
- Sometimes the partitioning is the goal
  - Market segmentation
- Prepare for other AI techniques
  - Summarize news
- Discovery in data
  - Underlying rues, reoccuring patterns, topics

Similarity/Dissimilarity Measurement

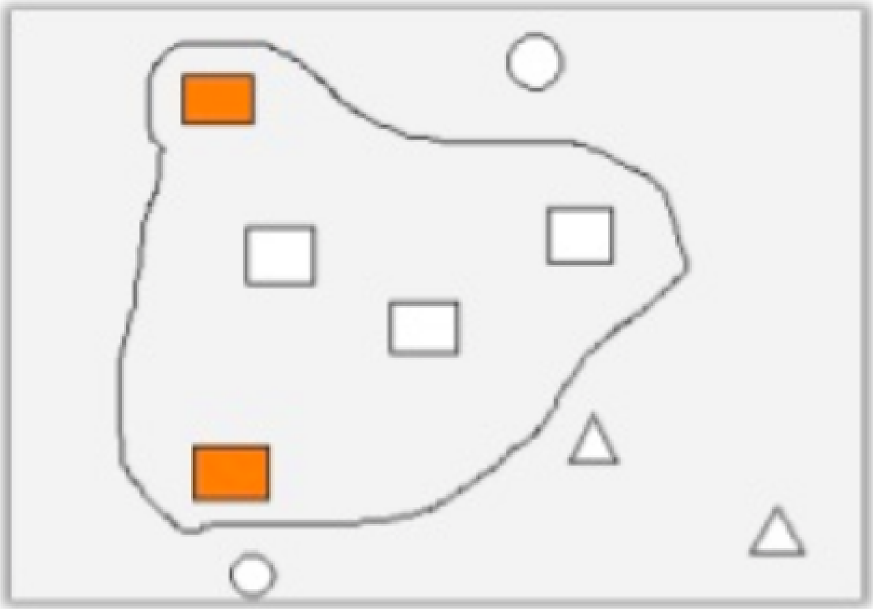


To achieve Clustering, a similarity/dissimilarity measure must be determined so as to cluster the data points based either on :

1. Similarity in the data or
2. Dissimilarity in the data

The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data.

Similarity Measurement

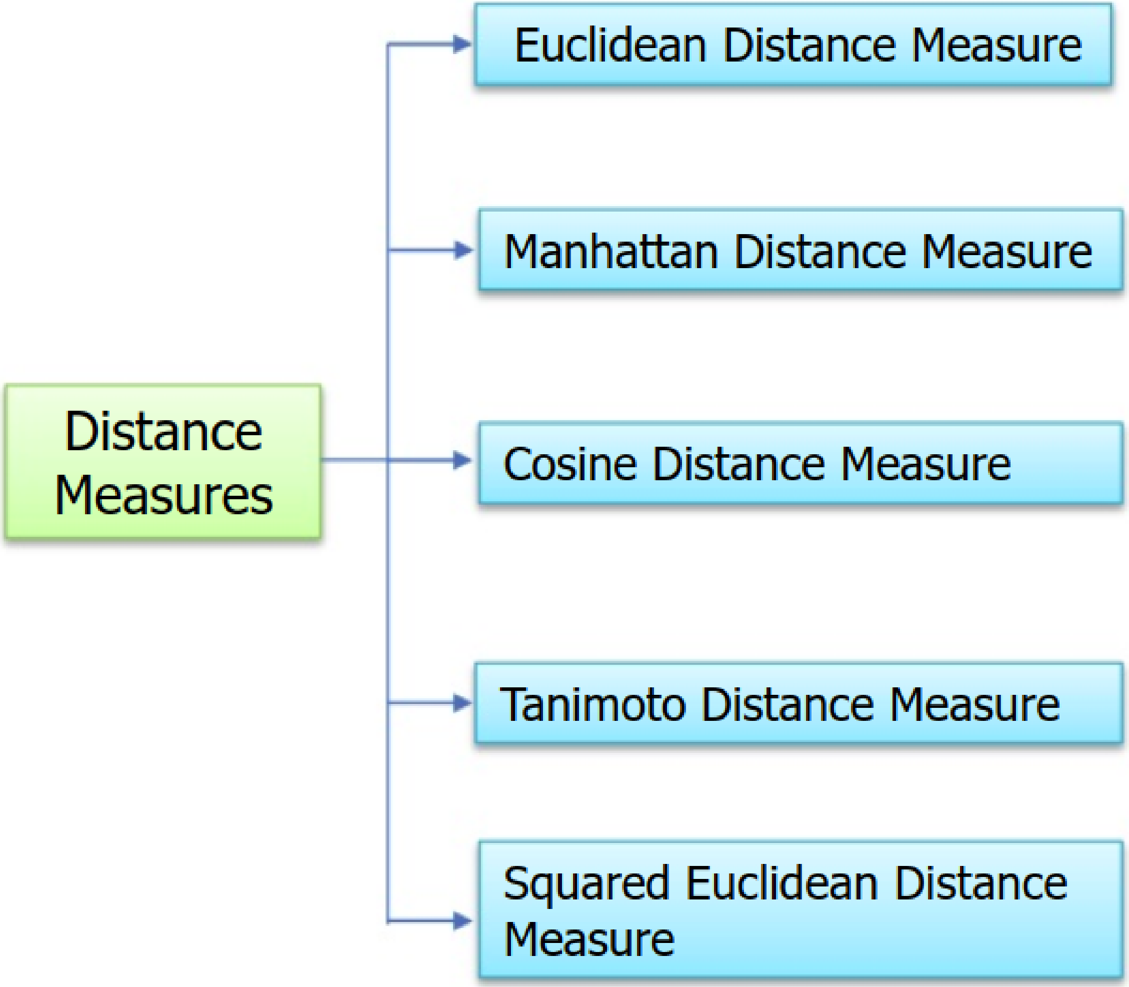


Similarity measures the degree to which a pair of objects are alike.

Concerning structural patterns represented as strings or sequences of symbols, the concept of pattern resemblance has typically been viewed from three main perspectives:

- Similarity as matching, according to which patterns are seen as different viewpoints, possible instantiations or noisy versions of the same object;
- Structural resemblance, based on the similarity of their composition rules and primitives;
- Content-based similarity.

Dissimilarity Measurement



Distance Measures

Euclidean Distance Measure

Manhattan Distance Measure

Cosine Distance Measure

Tanimoto Distance Measure

Squared Euclidean Distance Measure

Similarity can also be measured in terms of the placing of data points.

By finding the distance between the data points , the distance/difference of the point to the cluster can be found.

Difference between Euclidean and Manhattan

From this image we can say that, The Euclidean distance measure gives 5.65 as the distance between (2, 2) and (6, 6) whereas the Manhattan distance is 8.0

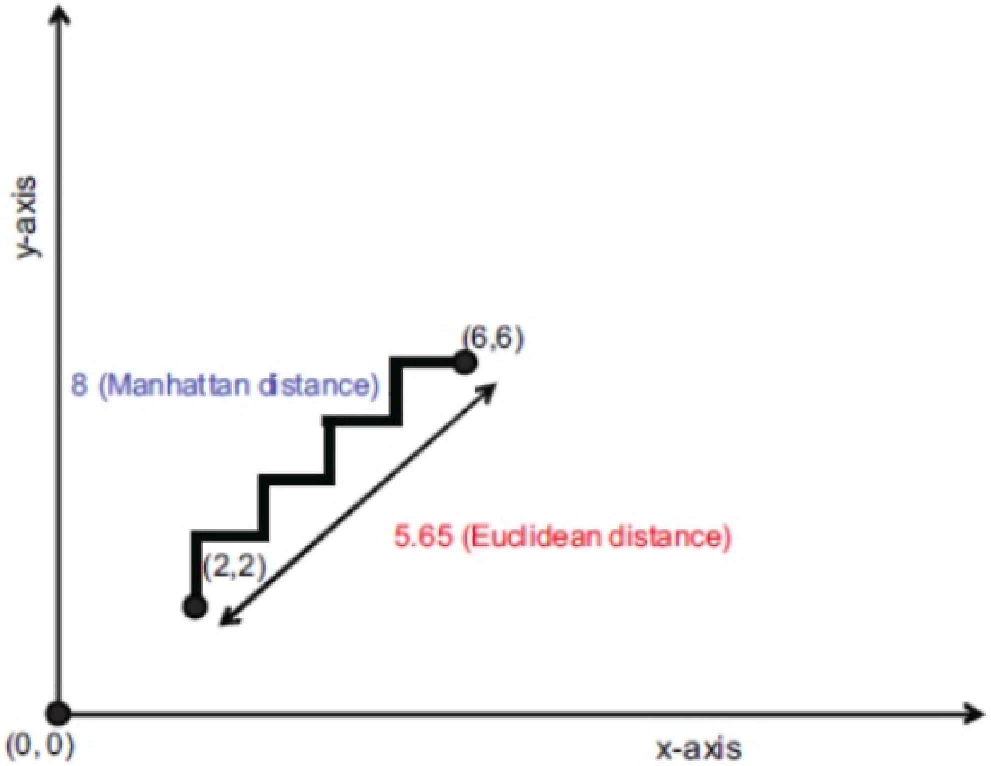
Mathematically, Euclidean distance between two n-dimensional vectors

(a$_1$, a$_2$, … , an) and (b$_1$,b$_2$,…,b$_n$) is:

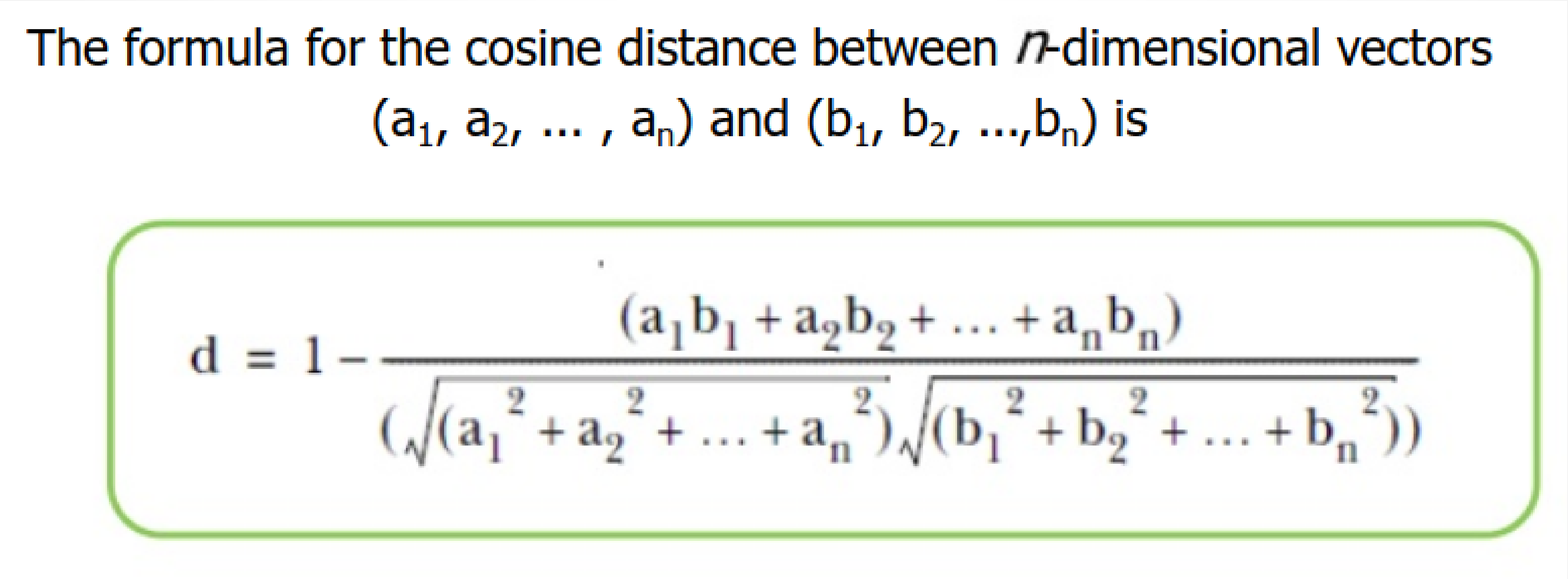
$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Manhattan distance between two n-dimensional vectors

$$d = |a1 - b1| + |a2 - b2| + \ldots + |an - bn|$$



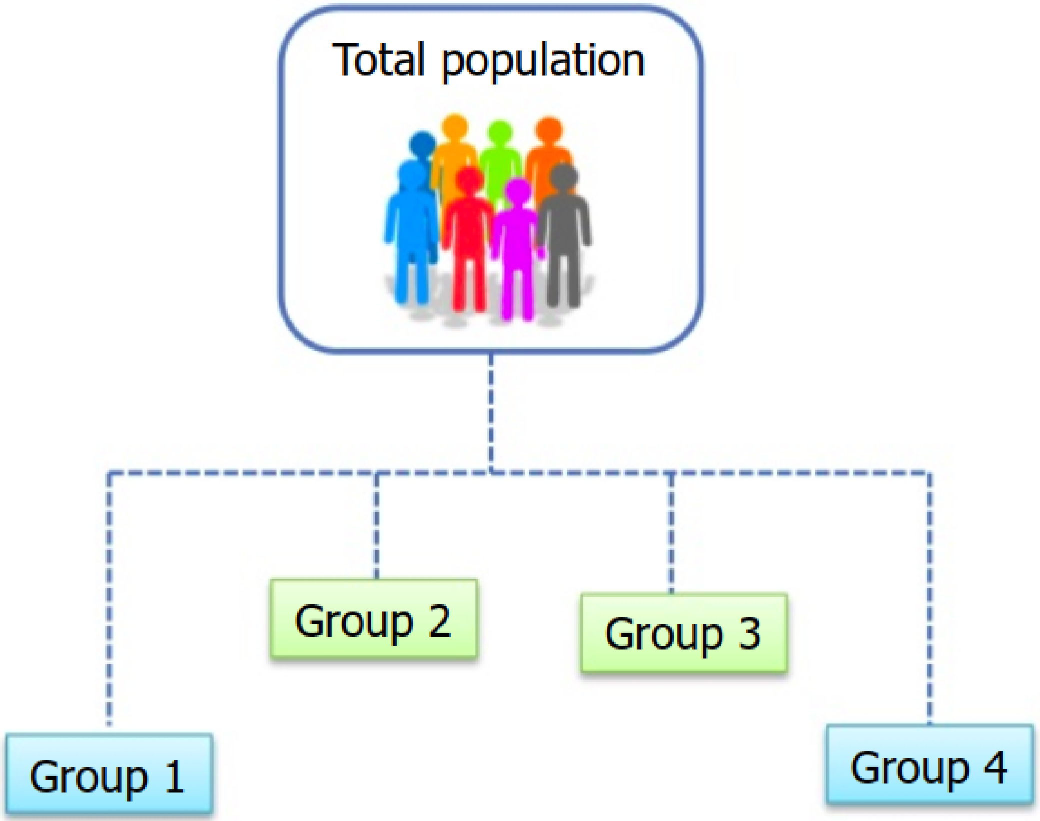
Cosine Distance Measure



The formula for the cosine distance between *n*-dimensional vectors $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ is

$$d = 1 - \frac{(a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}{(\sqrt{(a_1^{2} + a_2^{2} + \ldots + a_n^{2})}\sqrt{(b_1^{2} + b_2^{2} + \ldots + b_n^{2})})}$$

The idea behind K-means Clustering



Total population

Group 2

Group 3

Group 1

Group 4

- The process by which objects are classified into a number of groups so that they are as much dissimilar as possible from one group to another group, but as much similar as possible within each group.

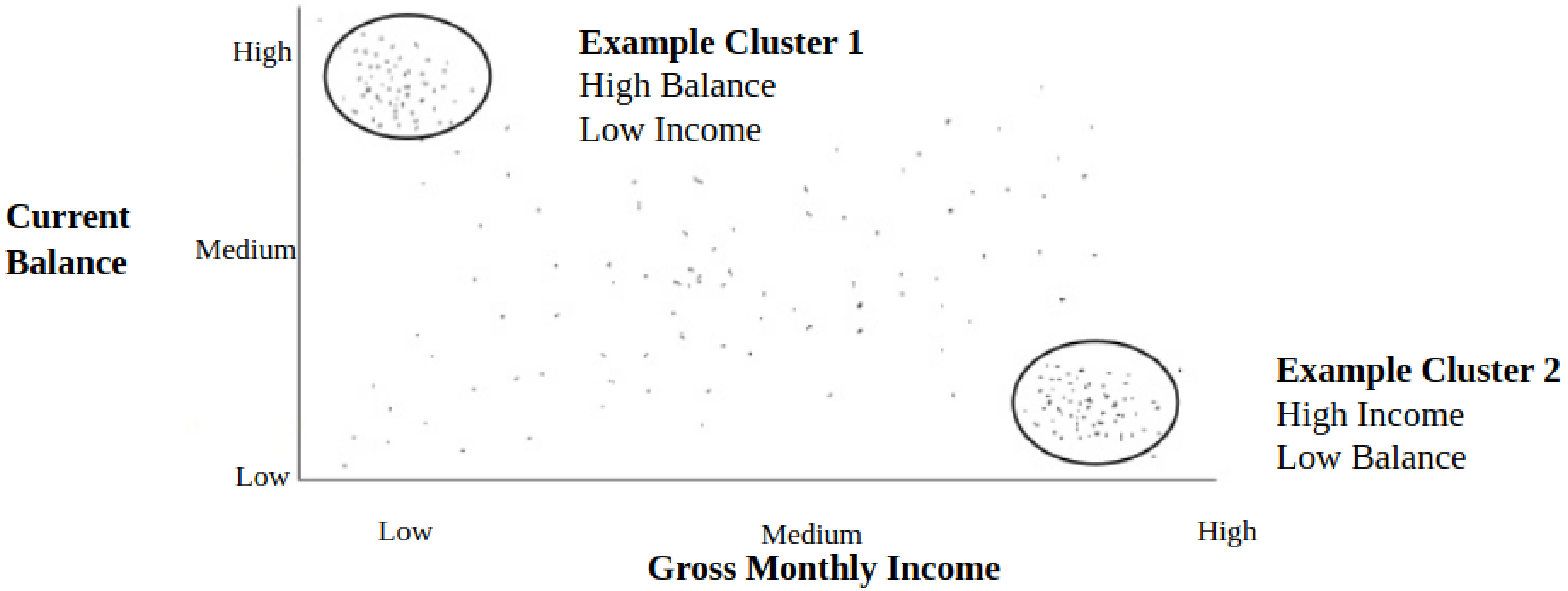
- The objects in group 1 should be as similar as possible.

- But there should be much difference between an object in group 1 and group 2.

- The attributes of the objects are allowed to determine which objects should be grouped together.

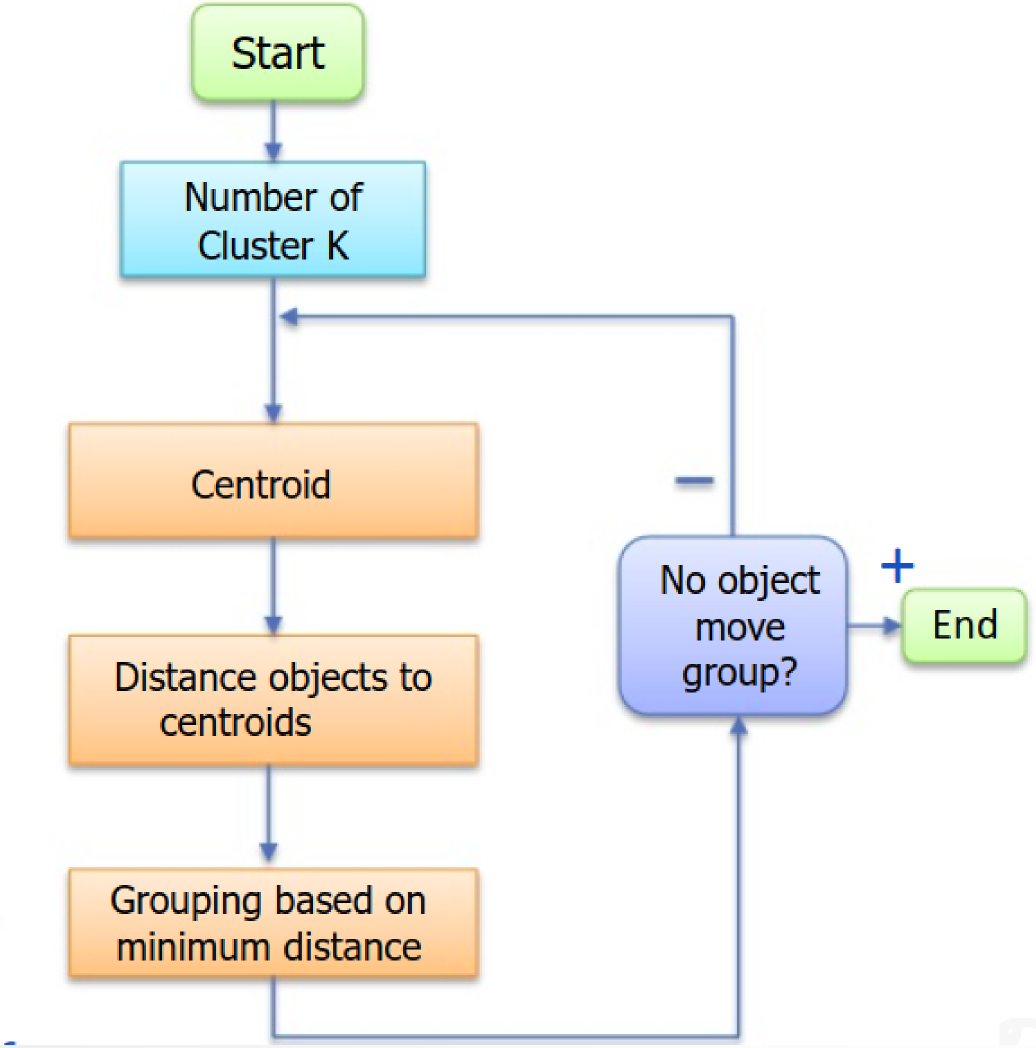
The idea behind K-means Clustering

Basic concepts of Cluster Analysis using two variables



- Cluster 1 and Cluster 2 are being differentiated by Income and Current Balance.
- The objects in Cluster 1 have similar characteristics (High Income and Low balance), on the other hand the objects in Cluster 2 have the same characteristic (High Balance and Low Income).
- But there are much differences between an object in Cluster 1 and an object in Cluster 2.

Process Flow of K-means Clustering


Start
Number of Cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
−
+
End

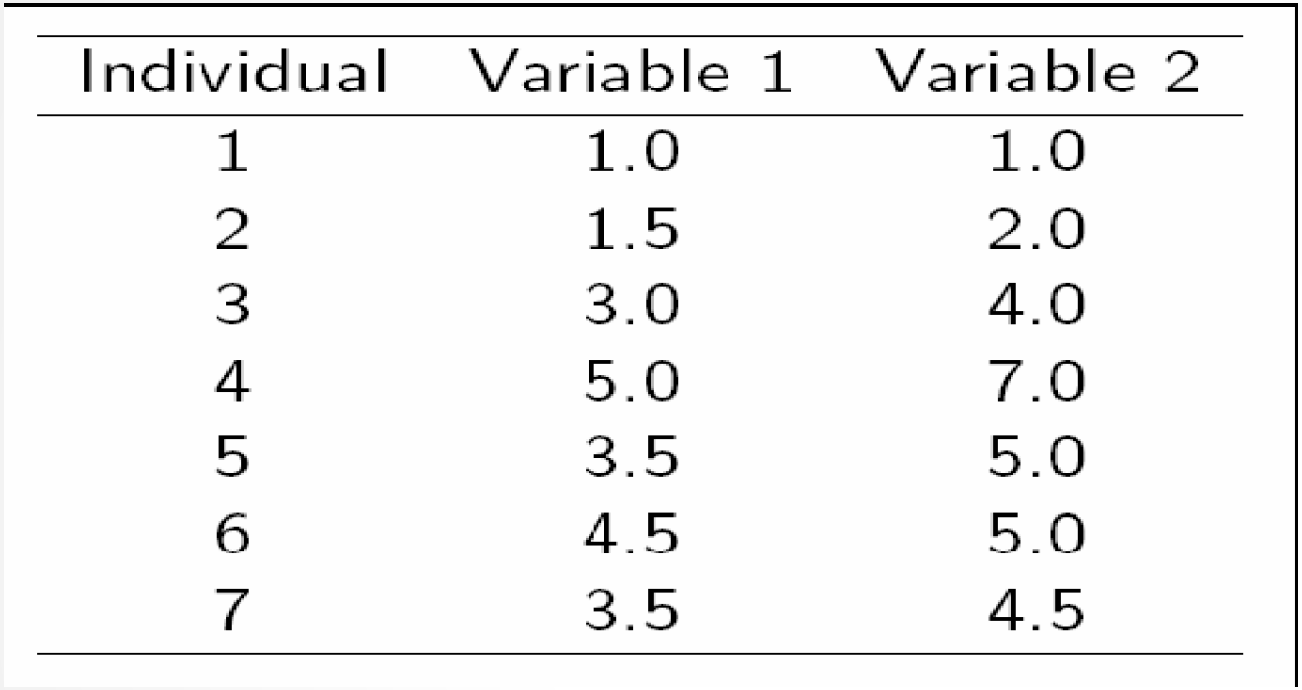
Iterate until *stable* (cluster centers converge):

1. Determine the centroid coordinate.

2. Determine the distance of each object to the centroids.

3. Group the object based on minimum distance (find the closest centroid)

A Simple example showing the implementation of k-means algorithm (using K=2)

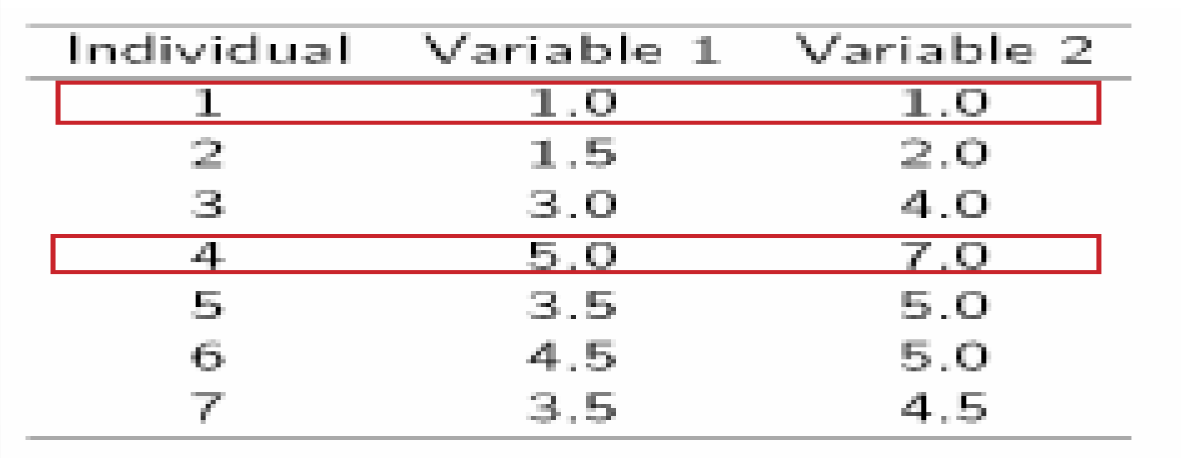


Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

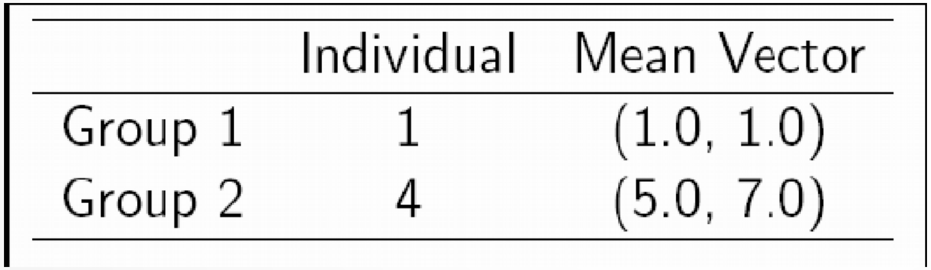
Step 1:

Initialization: Randomly we choose following two centroids (k=2) for two clusters.

In this case the 2 centroid are: m1=(1.0,1.0) and m2=(5.0,7.0).



Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5



	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Step 2:

- Thus, we obtain two clusters containing:

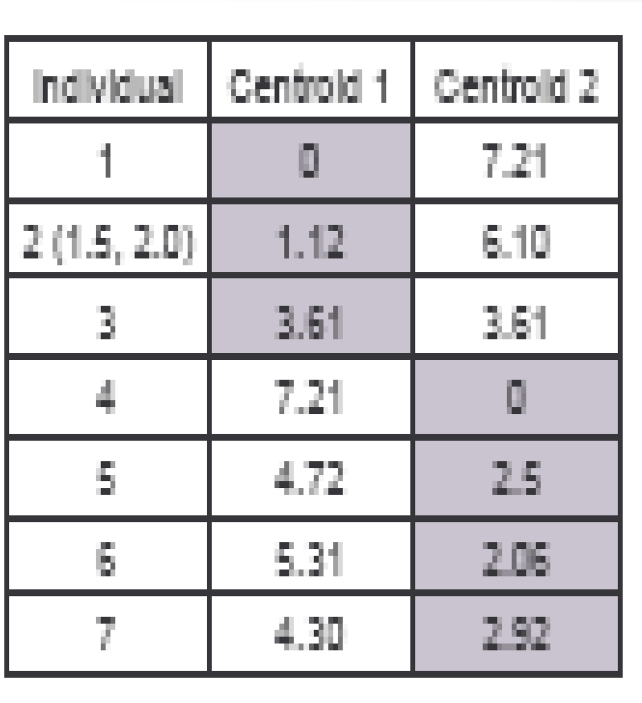
  {1,2,3} and {4,5,6,7}.

- Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$



Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

$$d(m_1, 2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} = 1.12$$

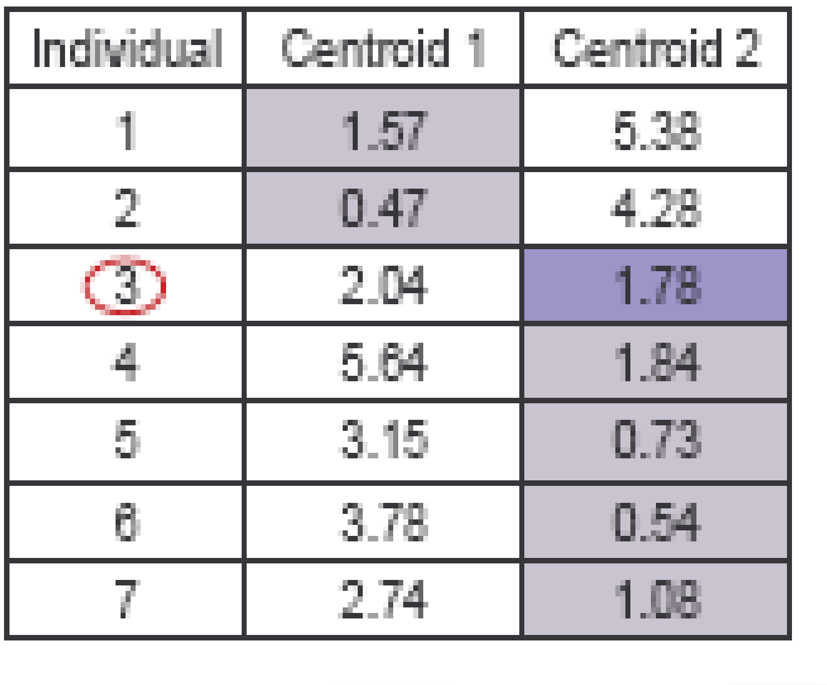
$$d(m_2, 2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

Step 3:

- Now using these centroids we compute the Euclidean distance of each object, as shown in table.

- Therefore, the new clusters are: {1,2} and {**3**,4,5,6,7}

- Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)



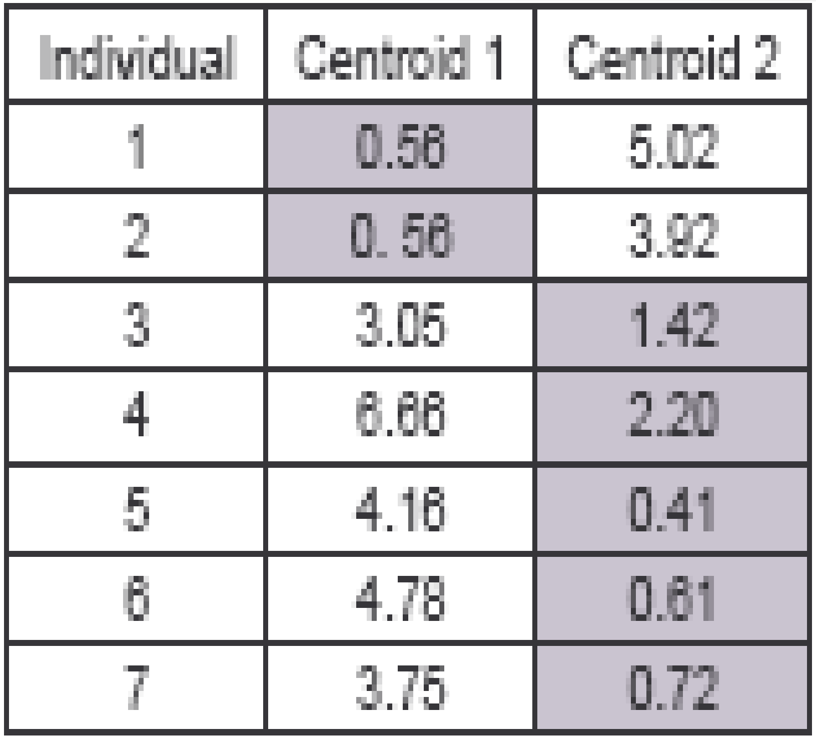
Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

- Step 4 :

  The clusters obtained are:

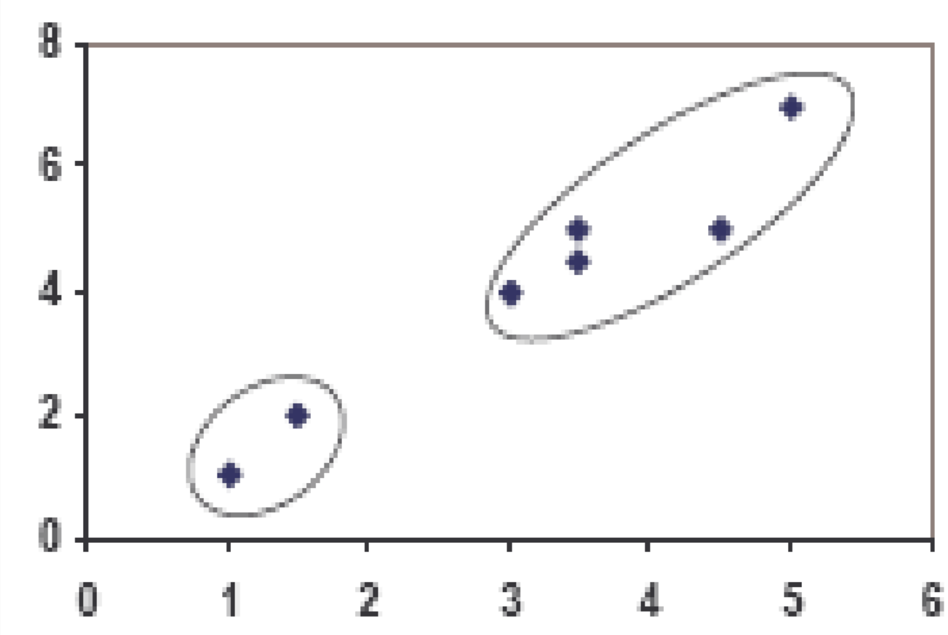
  {1,2} and {3,4,5,6,7}

- Therefore, there is no change in the cluster.
- Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.

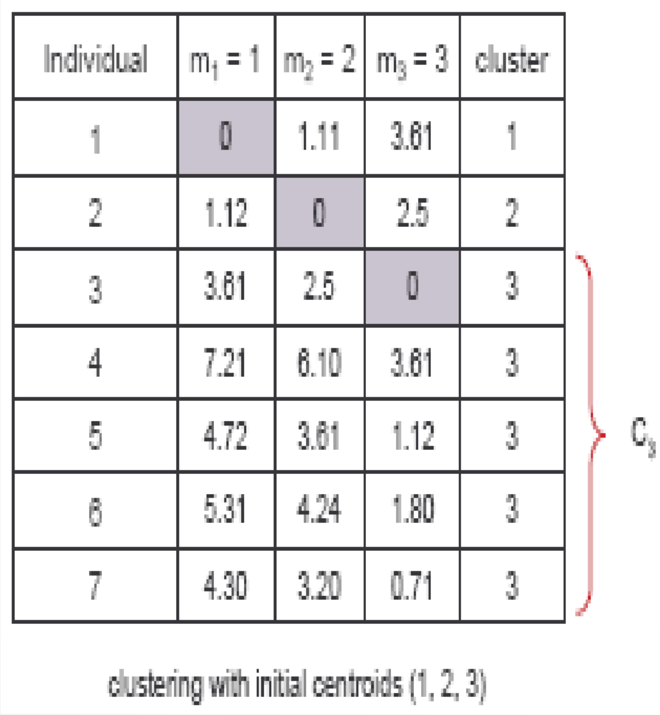


Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

PLOT



(with K=3)

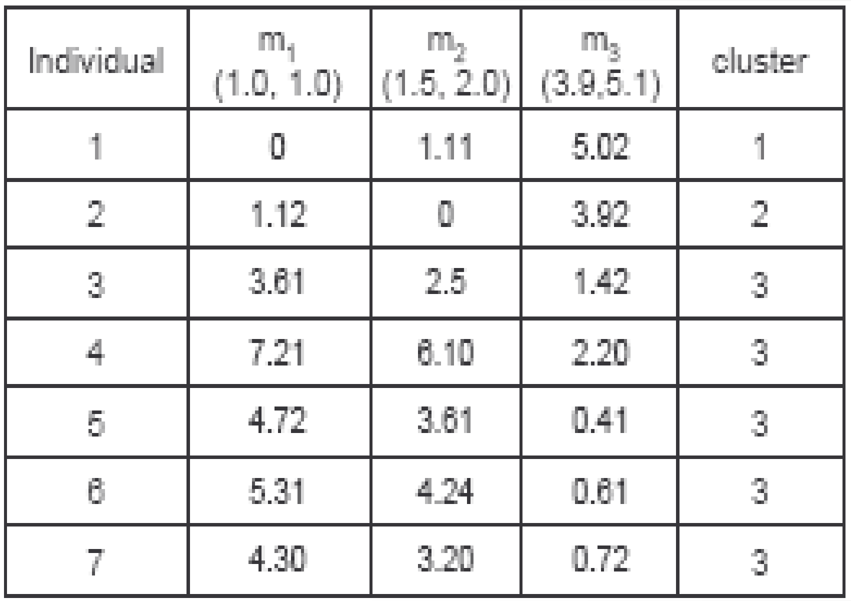


Individual	$m_1 = 1$	$m_2 = 2$	$m_3 = 3$	cluster
1	0	1.11	3.61	1
2	1.12	0	2.5	2
3	3.61	2.5	0	3
4	7.21	6.10	3.61	3
5	4.72	3.61	1.12	3
6	5.31	4.24	1.80	3
7	4.30	3.20	0.71	3

C_3

clustering with initial centroids (1, 2, 3)

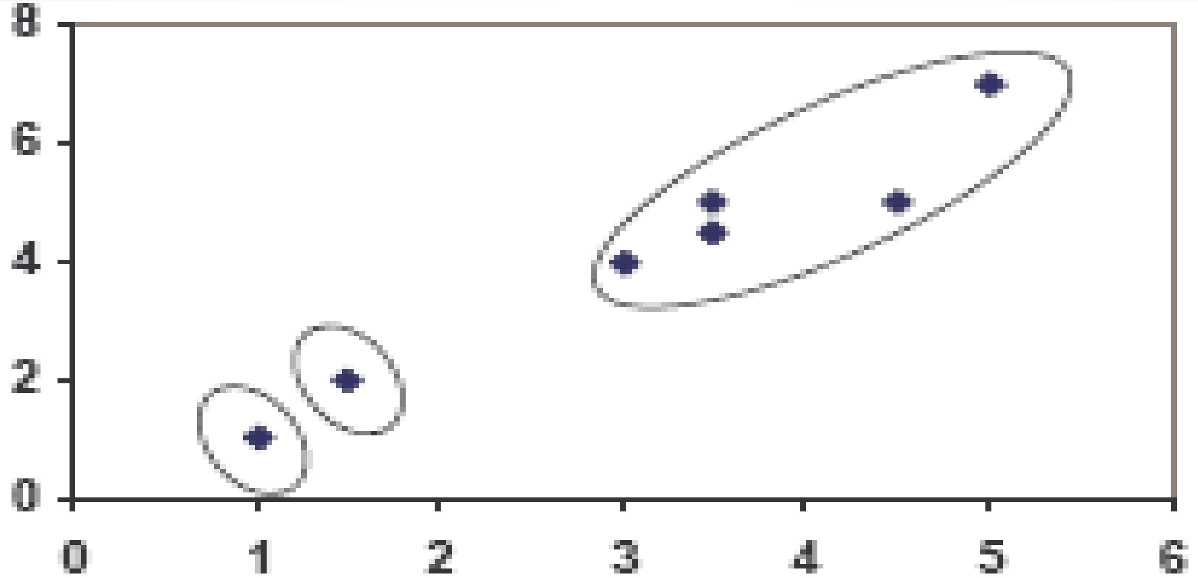
Step 1



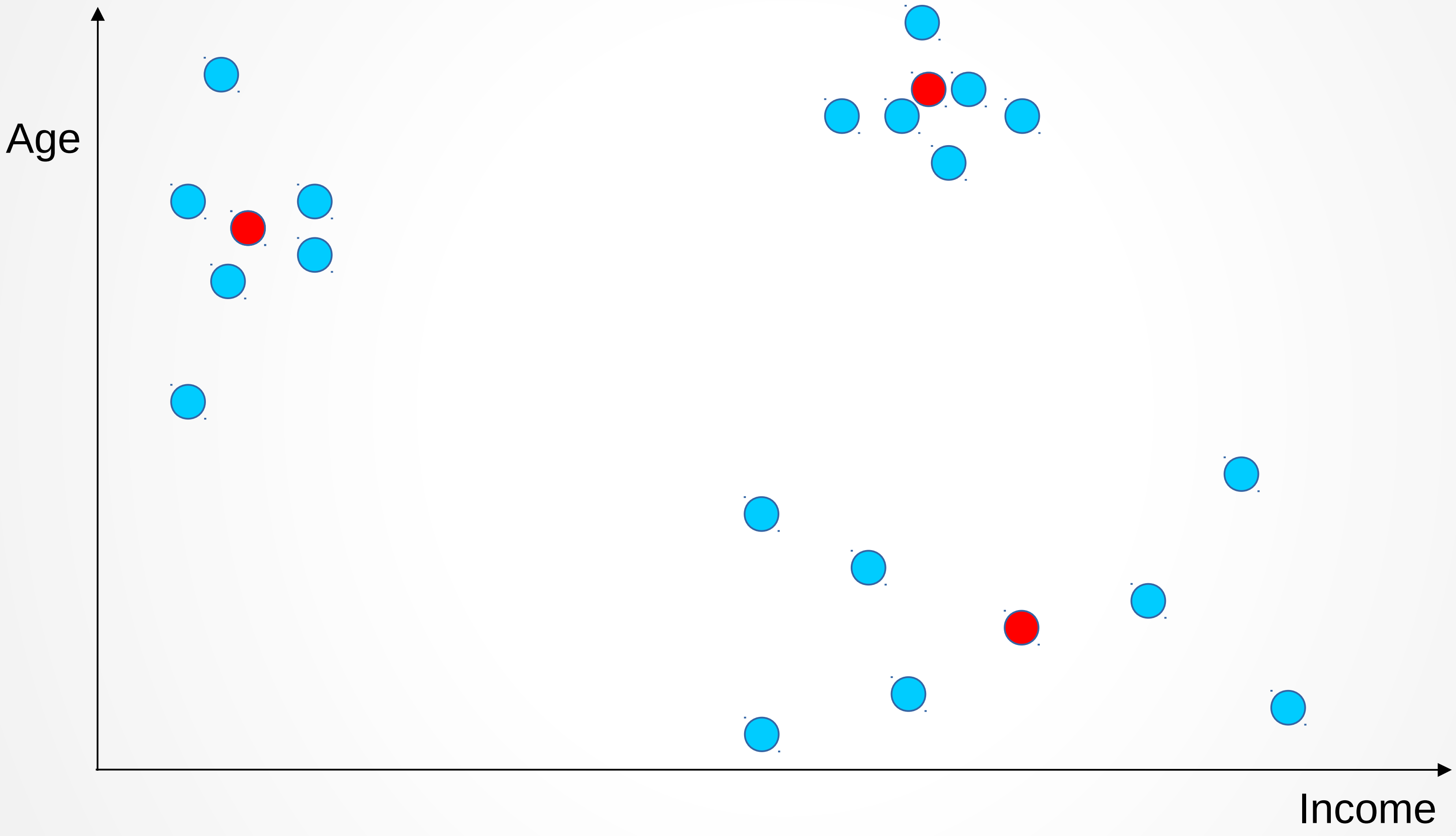
Individual	m_1 (1.0, 1.0)	m_2 (1.5, 2.0)	m_3 (3.9,5.1)	cluster
1	0	1.11	5.02	1
2	1.12	0	3.92	2
3	3.61	2.5	1.42	3
4	7.21	6.10	2.20	3
5	4.72	3.61	0.41	3
6	5.31	4.24	0.61	3
7	4.30	3.20	0.72	3

Step 2

PLOT

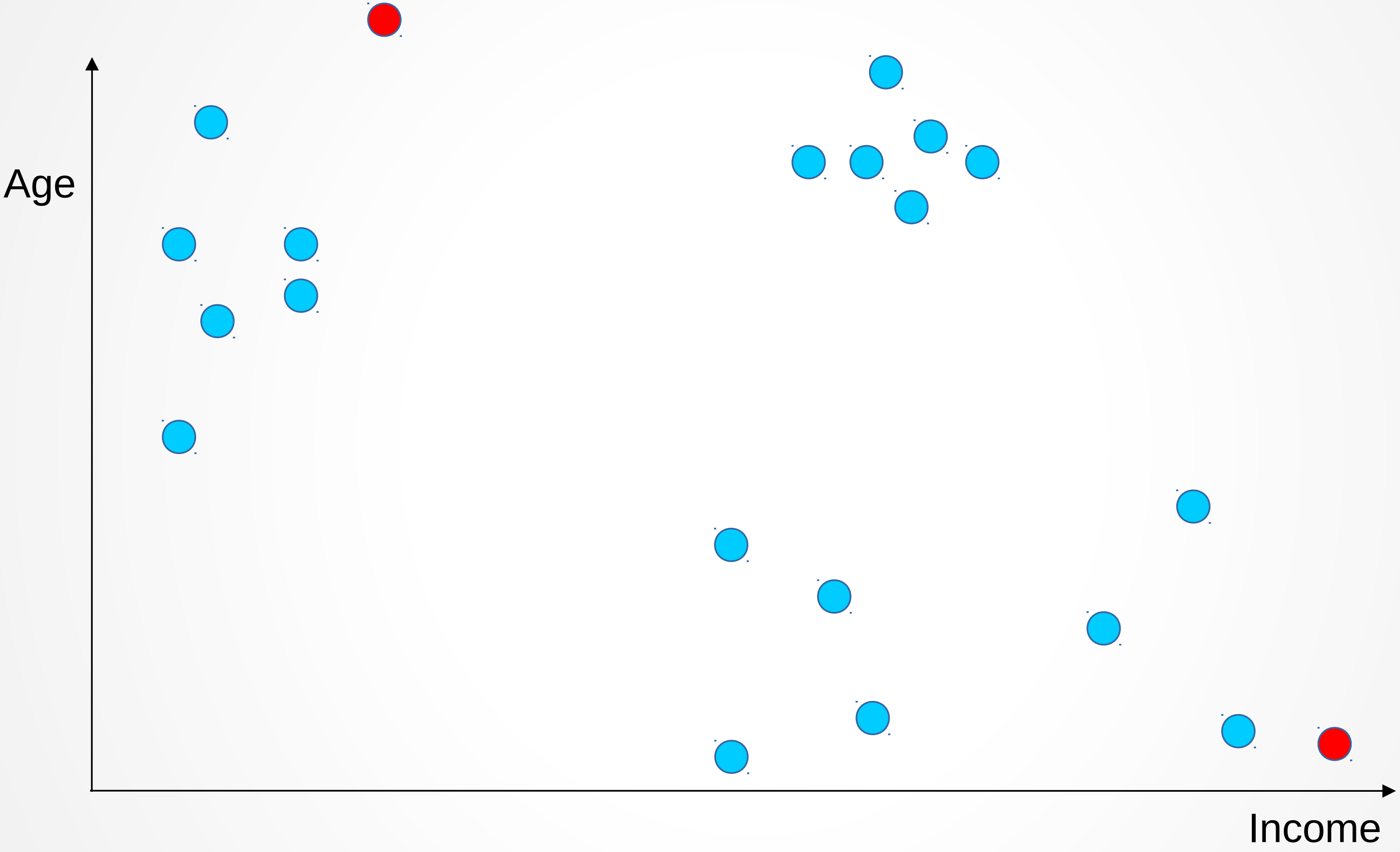


Performance Measurement Clustering Algorithms



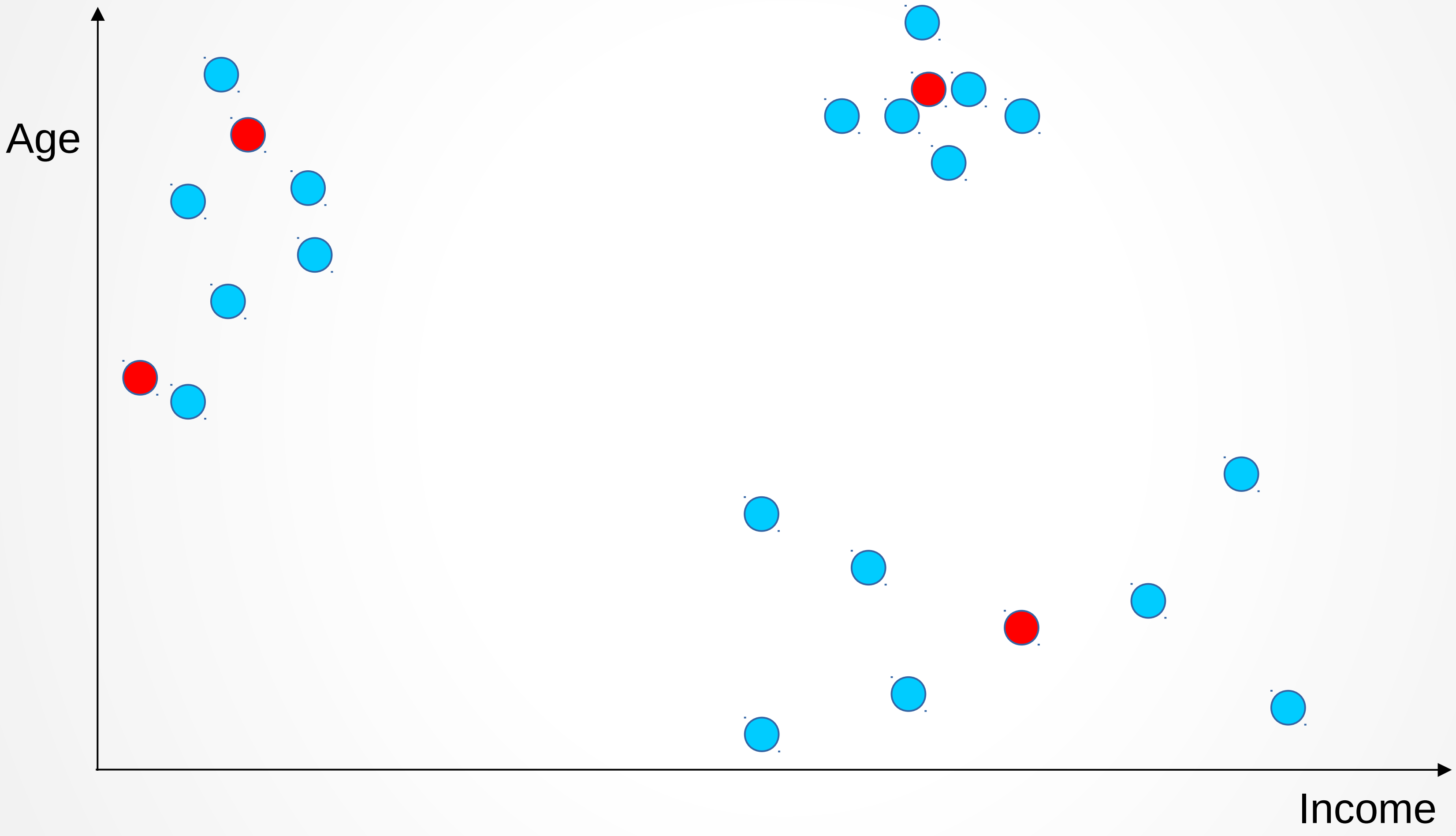
Good clustering
Distance(🔴, 🔴) => big
Distance(🔴, 🔵) => small

Performance Measurement Clustering Algorithms



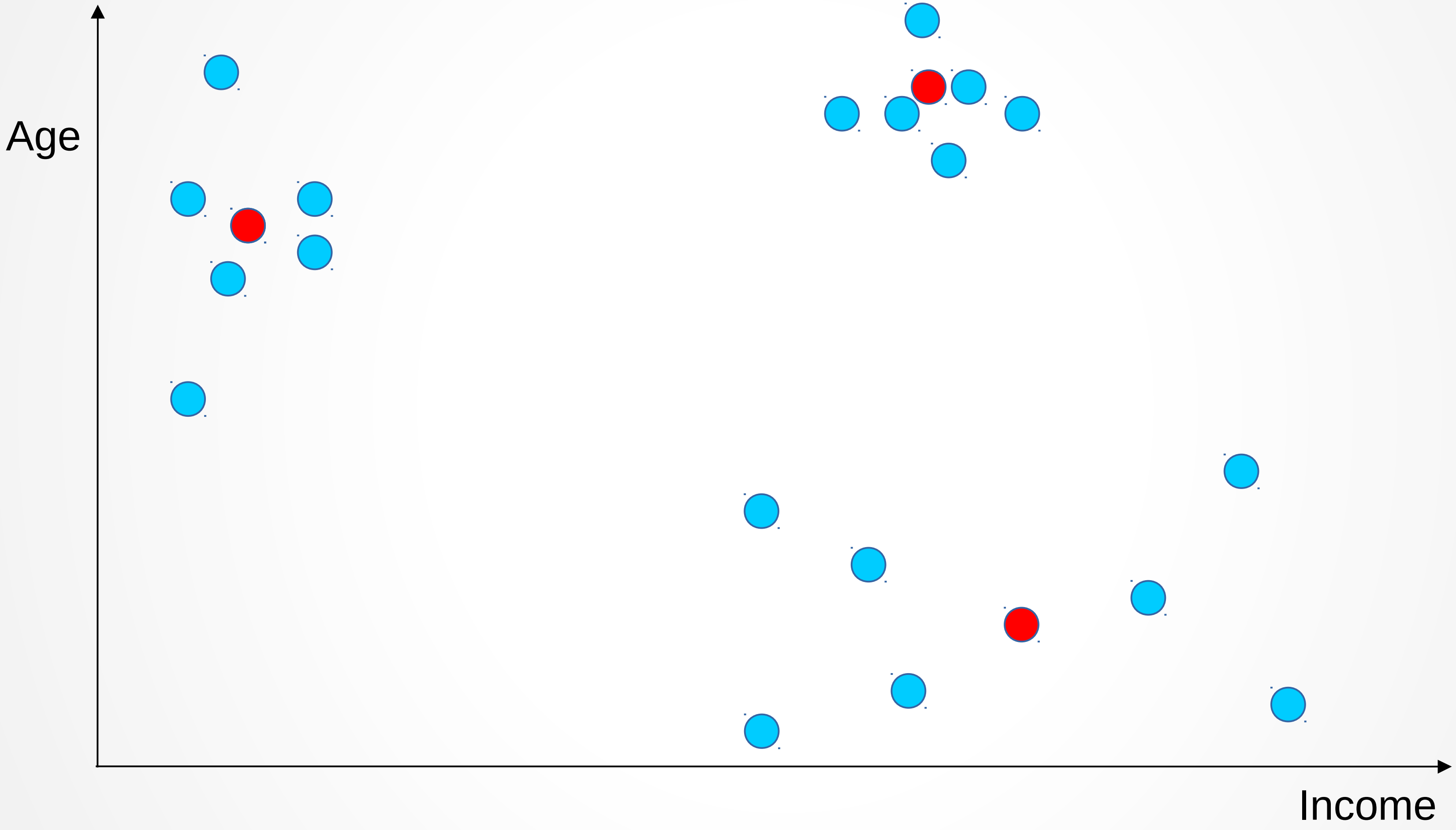
Bad clustering
Distance(🔴, 🔴) => big
Distance(🔴, 🔵) => big

Performance Measurement Clustering Algorithms



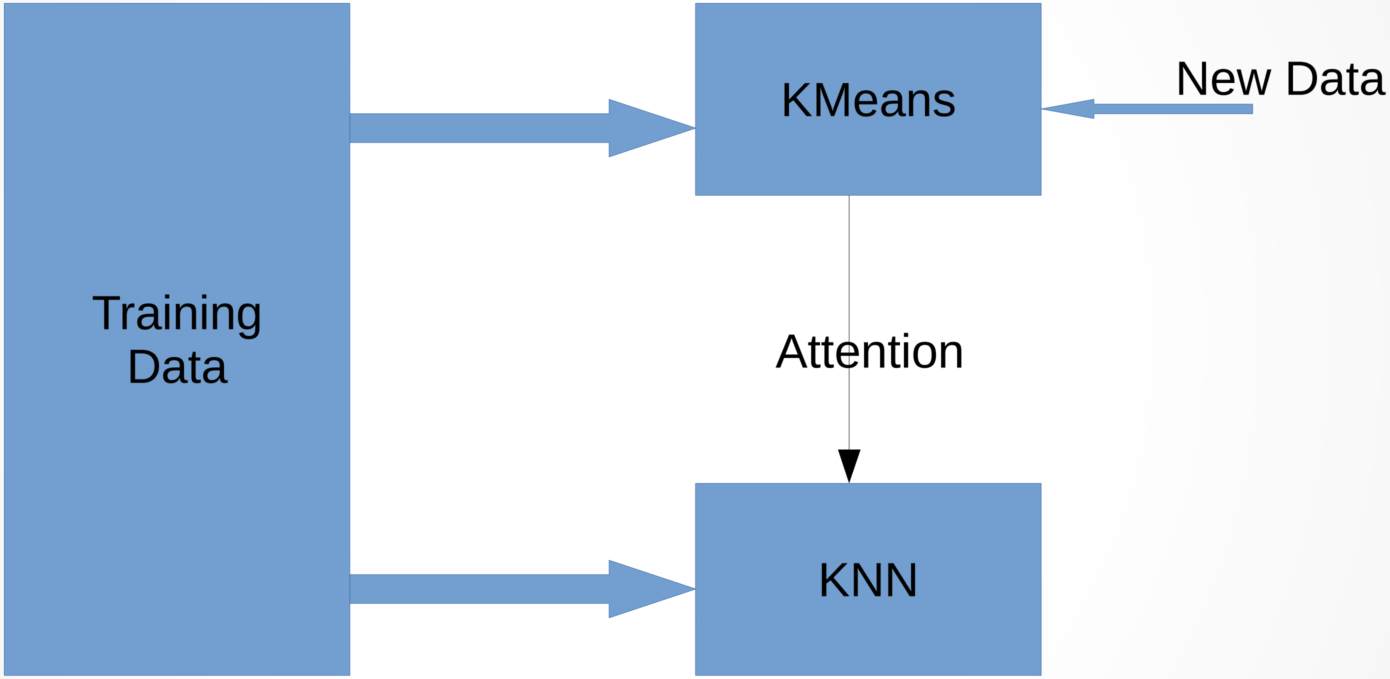
bad clustering
Distance(🔴, 🔴) => small
Distance(🔴, 🔵) => small

Dunn index



Dunn Index= Min(Distance(🔴, 🔴))/ Max(Distance(🔴,🔵))
Large dunn index means that compact and well separeted cluster exist

Tandem KNN and Kmeans



Tandem KNN and Kmeans
How To Calculate the Complexity

Tandem KNN and Kmeans
How To Improve Accuracy with cluster Global Labels