Unsupervised Learning :
Requires minimum hamn supervision
does not require any knowledge of HUMAN-LABELLED DATA

$X \in \mathbb{R}^K$

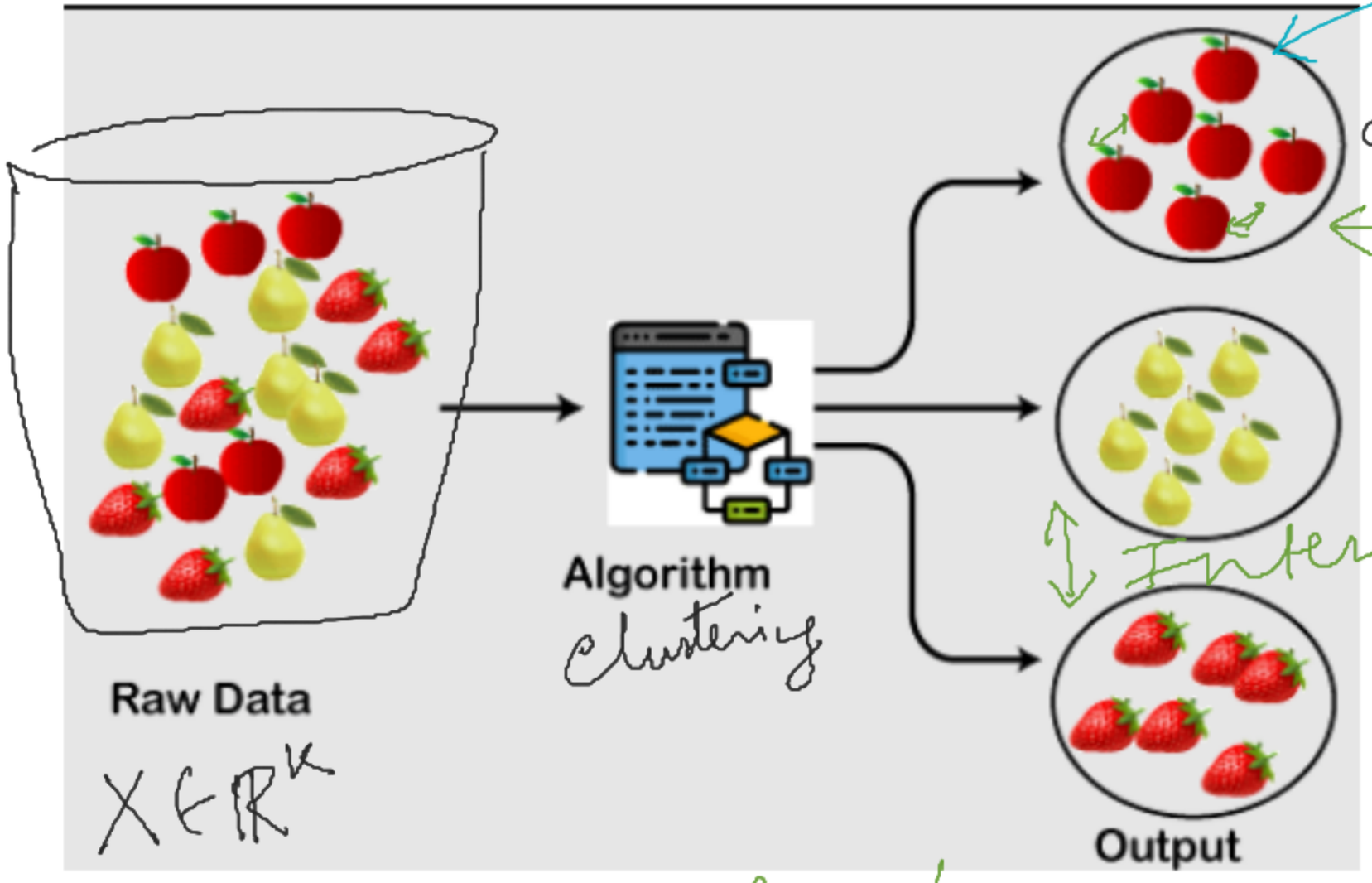
2 different types of unsupervised learning:
- clustering,
- dimensionality reduction

$Y \in \mathbb{R}^K$

Examples:
- recommendation system: to recommend customers:
  online shopping purchases
  movie recommendations
  .....

$$d(X_1, Y_1) = \sqrt{(x_1^1 - y_1^1)^2 + (x_1^2 - y_1^2)^2 + \dots + (x_1^K - y_1^K)^2}$$

$d(\vec{x}, \vec{y})$



Intra cluster within cluster

cluster - 1 Similar objects

cluster - 2

Inter cluster objects are dissimilar

cluster - 3

$X \in \mathbb{R}^K$

Intra cluster dist ↓

Inter cluster dist ↑

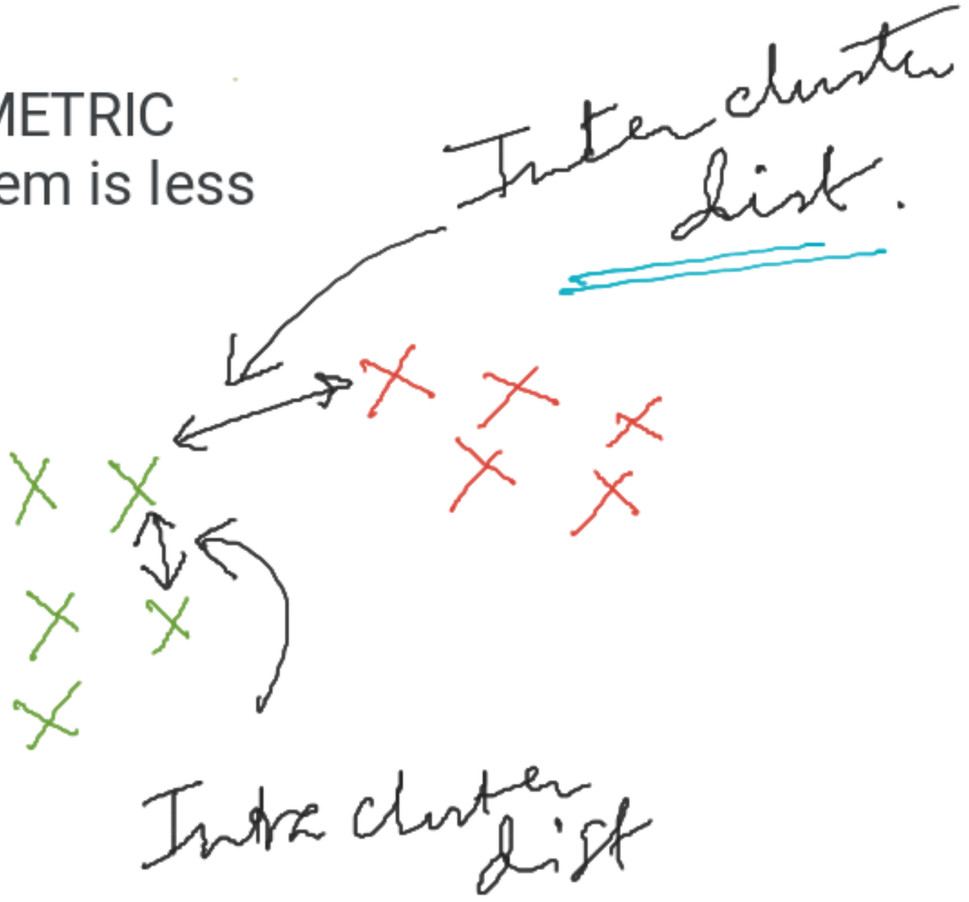
Clustering is an unsupervised learning technique

Objective of clustering:
to discover overall distribution patterns
correlations among data distribution

Given the features of sample data to be clustered, objectives are:

- to put similar samples in the same cluster
- the similarity b/w 2 samples is measured by DISTANCE METRIC
- the similarity of 2 samples is more, if the distance b/w them is less

A clustering alog is considered to be good, if:
INTER-cluster distance b/w different clusters is more
INTRA-cluster distance of the same cluster is less



Inter-cluster dist. is the dis. b/w 2 sample data belonging to 2 diff. clusters

Different measures:

SLD

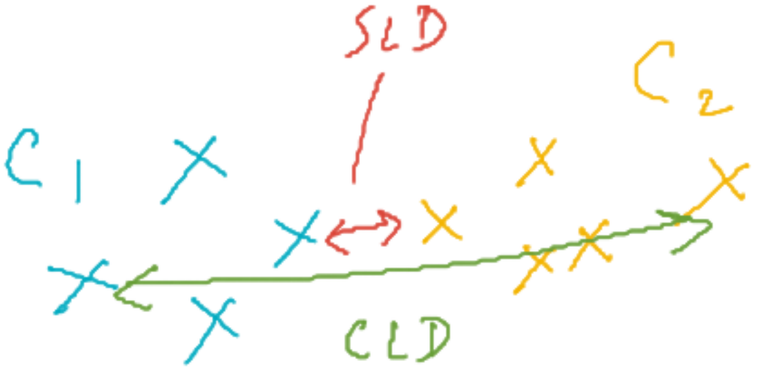
Single linkage distance: The minimum dist. b/w 2 sample data belonging to 2 diff. clusters

$$S(C_1, C_2) = \min_{\substack{x \in C_1, \\ y \in C_2}} d(x, y)$$

CLD

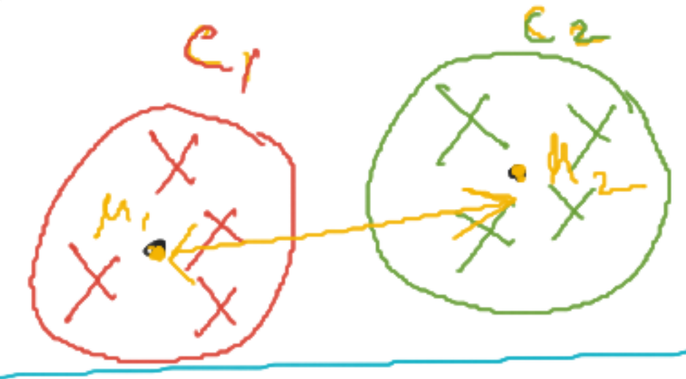
Complete linkage distance: The maximum dist. b/w 2 sample data belonging to 2 diff. clusters

$$S(C_1, C_2) = \max_{\substack{x \in C_1, \\ y \in C_2}} d(x, y)$$



Average linkage dist ... Avg. ....

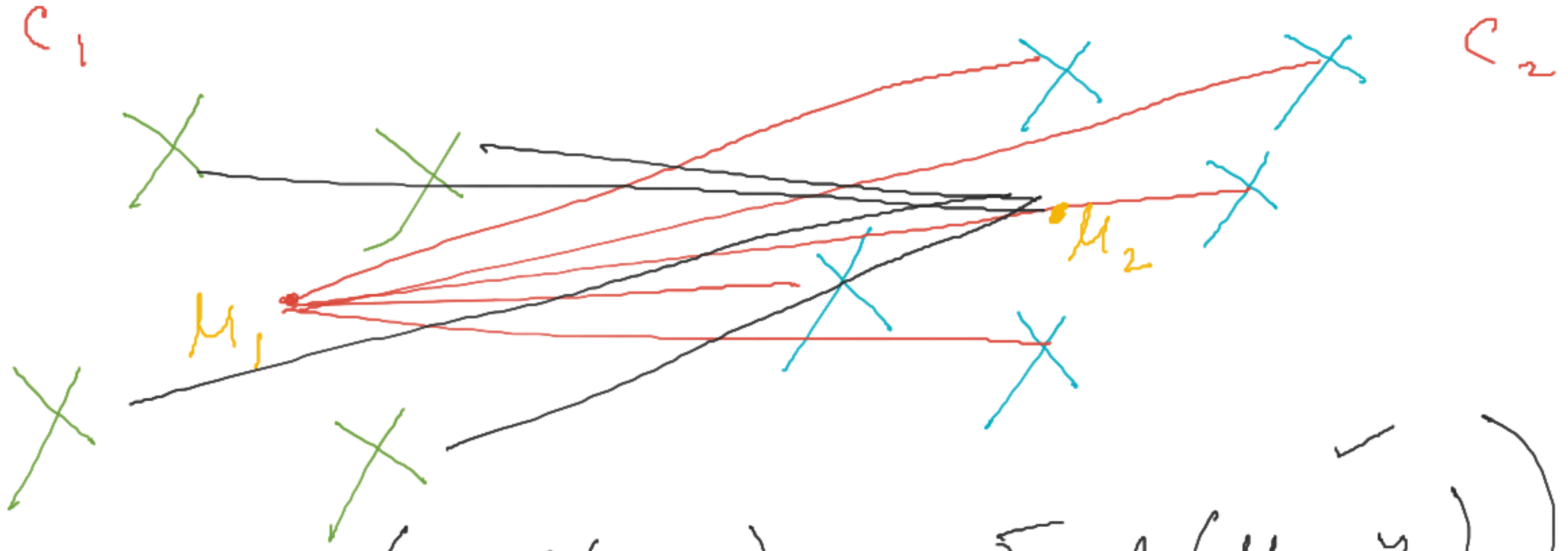
$$S(C_1, C_2) = \frac{1}{|C_1| \times |C_2|} \sum_{\substack{x \in C_1, \\ y \in C_2}} d(x, y)$$



Centroid linkage distance: The dist. b/w 2 centroids belonging to 2 diff. clusters

$$S(C_1, C_2) = d(\mu_1, \mu_2)$$

where, $\mu_1 = \frac{1}{|C_1|} \sum_{x \in C_1} x$ | $\mu_2 = \frac{1}{|C_2|} \sum_{y \in C_2} y$



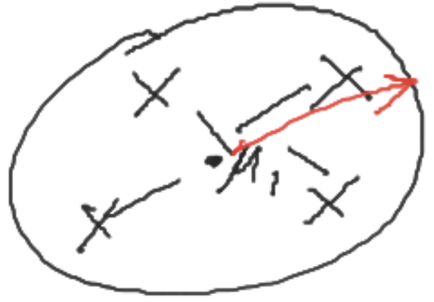
$$S(C_1, C_2) = \frac{1}{|C_1| + |C_2|} \left(\sum_{x \in C_1} d(x, \mu_2) + \sum_{y \in C_2} d(\mu_1, y) \right)$$

Average centroid linkage distance

Intra-cluster dist: dist. b/w 2 data belonging to the SAME cluster

Complete Diameter Dist

$$\Delta(C_1) = \max_{\substack{(x,y)\in C_1 \\ x \neq y}} d(x,y)$$



Centroid diameter dist

$$\Delta(C_1) = 2 \max_{x \in C_1} d(x, M_1)$$

Avg. diameter dist

$$\Delta(C_1) = \frac{\sum_{\substack{(x,y)\in C_1 \\ x \neq y}} d(x,y)}{}$$

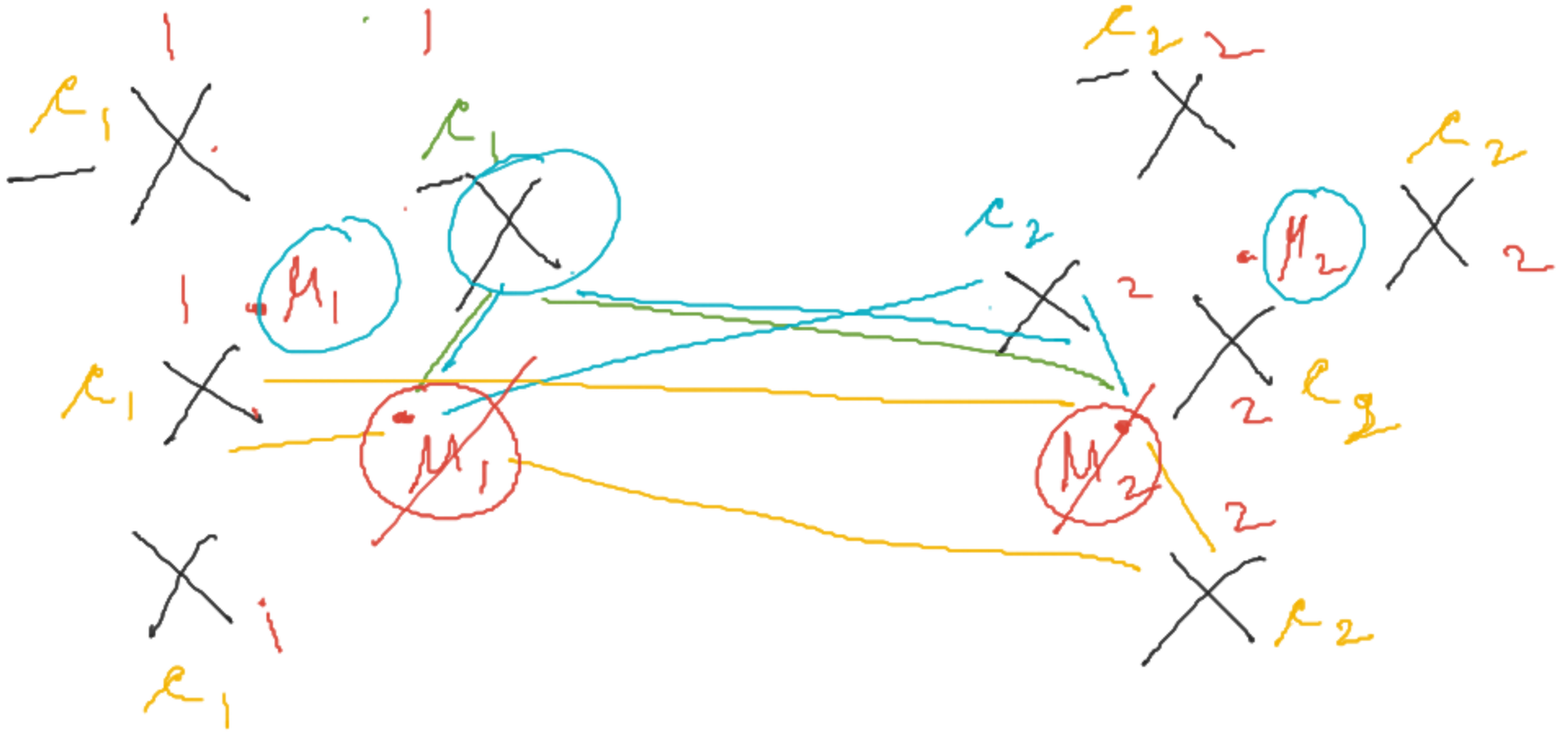
A good clustering algo:

A) more complete diameter dist
B) less complete linkage dist
C) less avg. diameter dist
D) more centroid diameter dist

K-means

i/p: no. of cluster (k)

k = 2



k = 2

K-means : i/p : ① Training data : $X = \{x_1, x_2, \cdots, x_n\}$

where $\forall x_i \in \mathbb{R}^m$

② K no. of clusters $\{C_1, C_2, \cdots, C_K\}$

Algo : Randomly initialize 'K' cluster centroids

$\mu_1, \mu_2, \cdots, \mu_K \in \mathbb{R}^m$

Repeat (until terminate) ← no update in centroid

{ for all $x_i \in X$:

$c_i = \underset{j \in \{1,2,\cdots,K\}}{\operatorname{argmin}} \| x_i - \mu_j \|^2$ // cluster assignment step

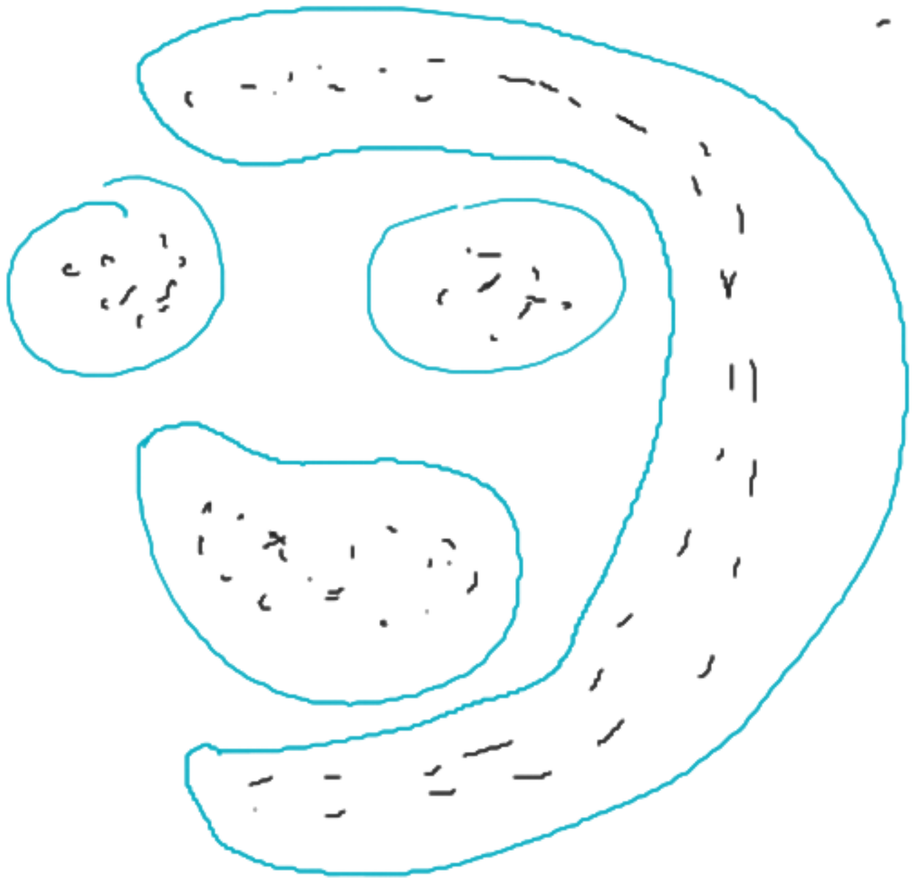
for all $\mu_i \in \{\mu_1, \mu_2, \cdots, \mu_K\}$

μ_i = update the value from newly generated clusters // centroid update

}

Random initialization of centroid

choosing "K"



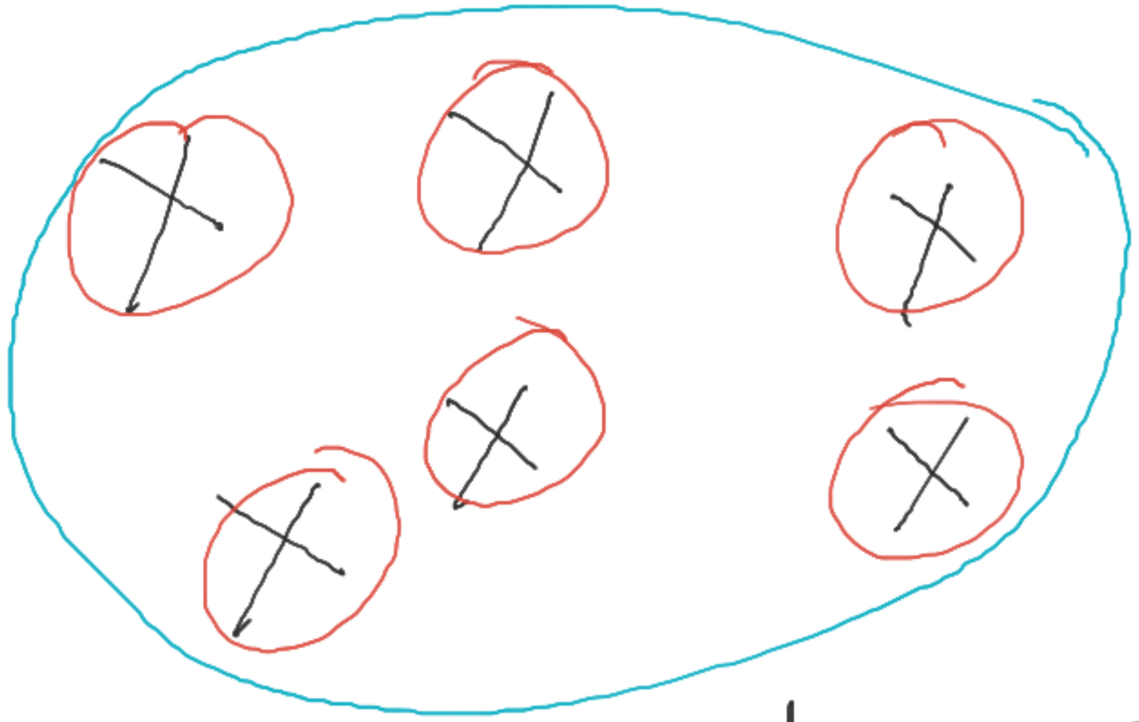
Globular data X

outlier X

→ (DBSCAN)

$n = 100$

$K \leq \sqrt{n} = 10$



k-mode

min # cluster = 1

max # cluster = n = 6

https://www.sciencedirect.com/science/article/abs/pii/S0167865509002323

k-means

o/p: $\forall x_i \in X, (x_i, r_i)$; $r_i \in \{C_1, C_2, \cdots, C_k\}$ is the cluster index

objective $\mu_i \in \mathbb{R}^m$ is the centroid of $C_i \in \{C_1, C_2, \cdots C_k\}$

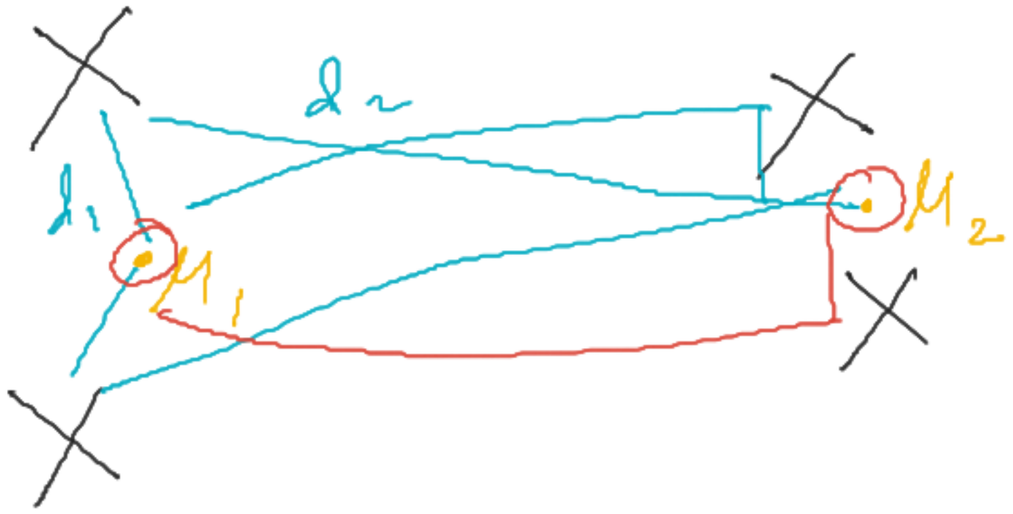
optimization objective

minimize $J(r_1, r_2, \cdots, r_n, \mu_1, \mu_2, \cdots, \mu_k)$

n + k parameters

$$J(r_1, \cdots, r_n, \mu_1, \cdots, \mu_k) = \frac{1}{n} \sum_{i=1}^{n} ||x_i - \mu_{r_i}||^2$$

$\mu_{r_i} \in \{\mu_1, \cdots \mu_k\}$ is the centroid of the cluster to which x_i belongs



$$J = \frac{1}{8}\left(d_1 + d_2 + d_3 + d_4 + d_5 + d_6 + d_7 + d_8 \right)$$

Suppose x_1 belongs to cluster 3, which are true

$c_1 = 3$

x_1

✓ A) $c_1 = 3$

B) $c_3 = 1$

C) $C_1 = 3$ ✗

D) $C_3 = 1$ ✗