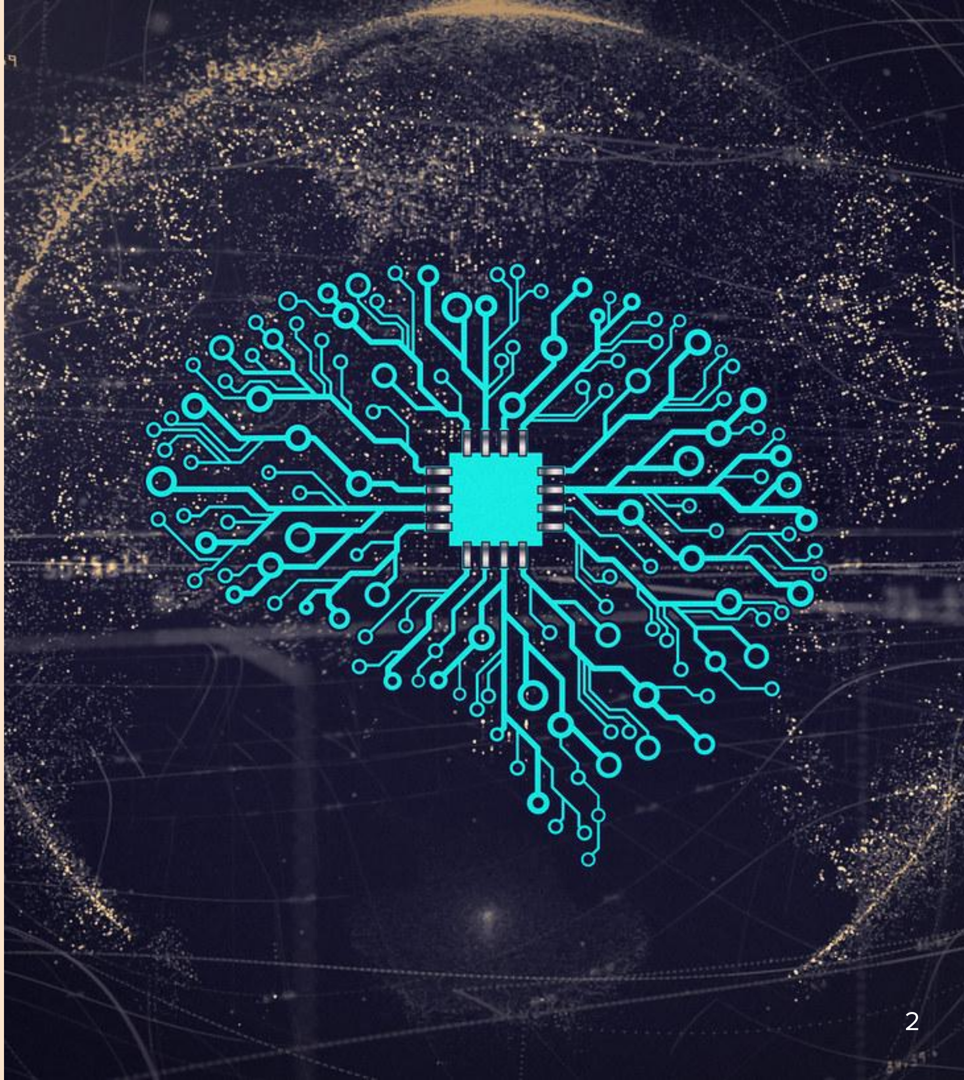
K-means Clustering

Dr. Paisit Khanarsa

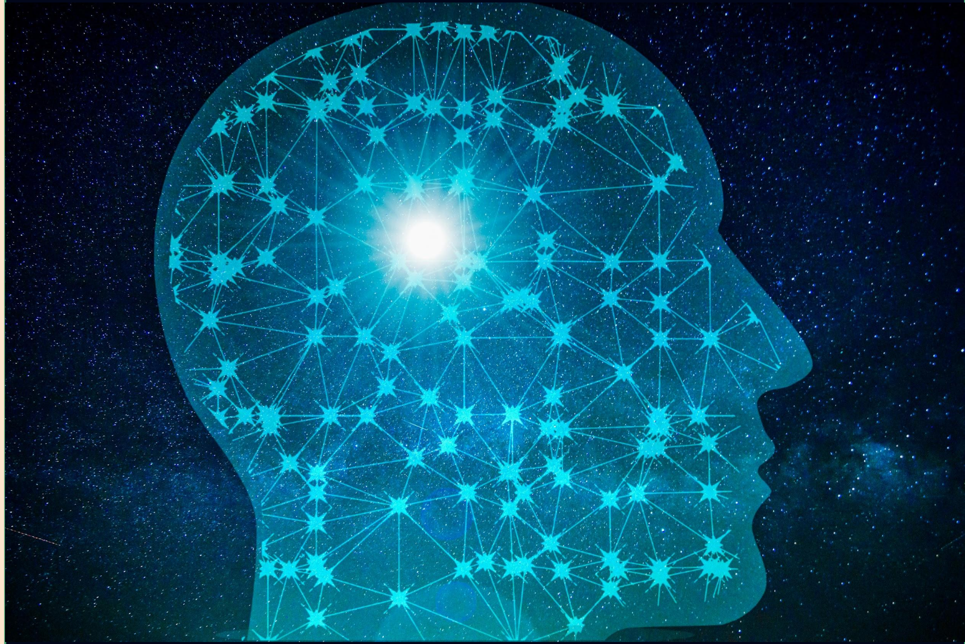
Fibo, Kmutt

Outline

- ❖ Unsupervised learning
- ❖ K-means clustering
- ❖ K-means clustering performance



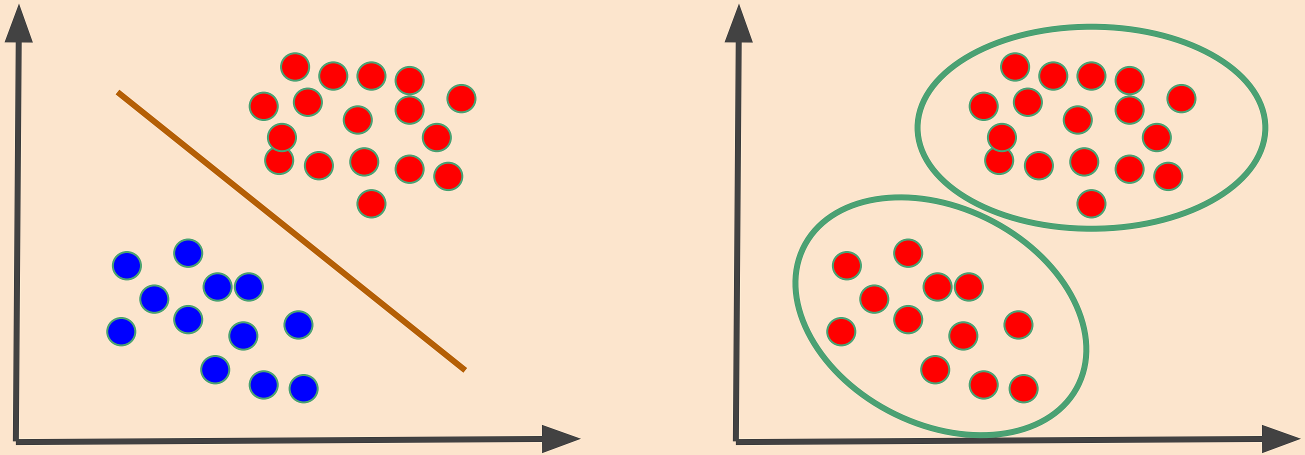
Unsupervised Learning



Unsupervised learning

In supervised learning, your data come with labels indicating what class corresponds to each sample.

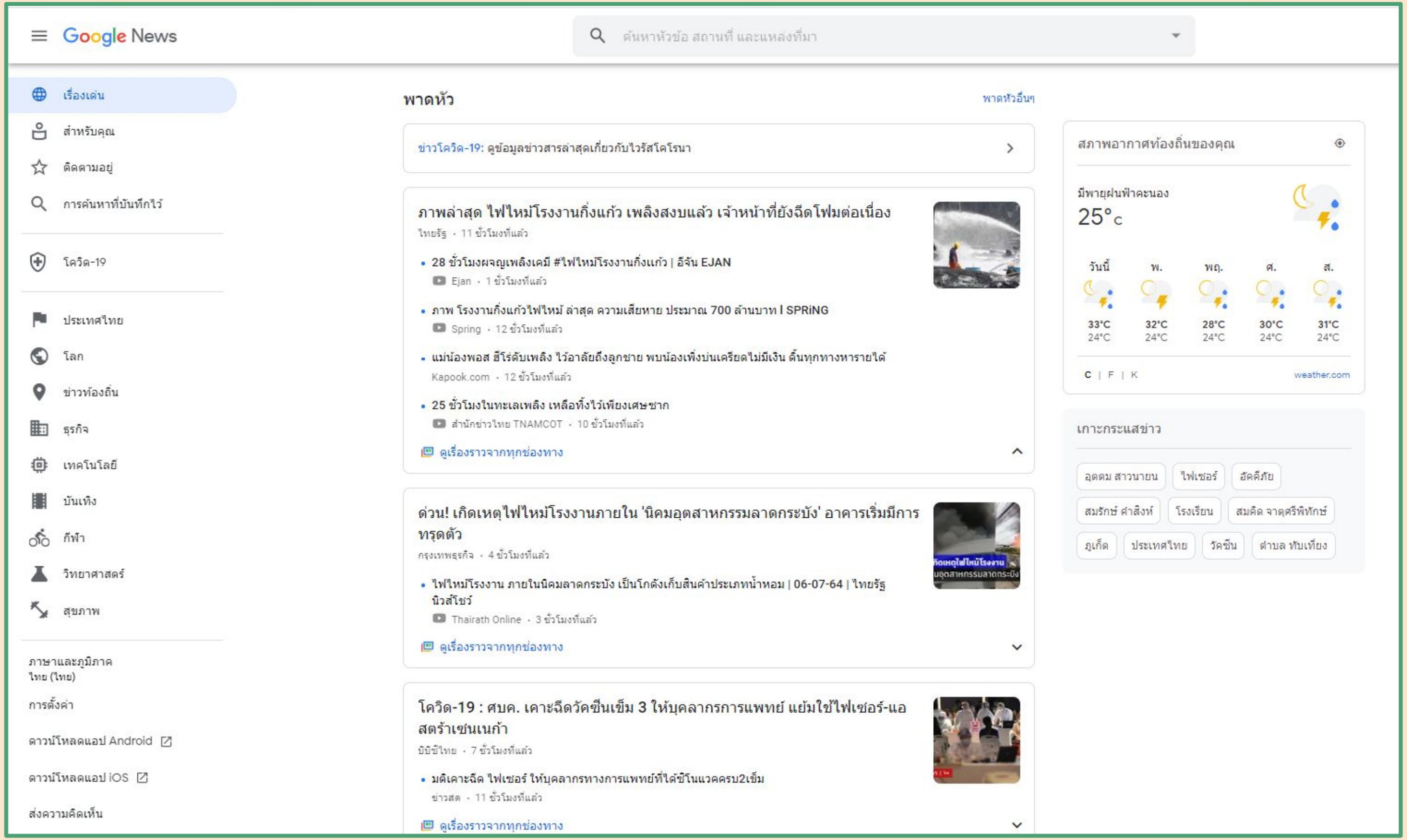
Sometimes, data do not come with categorical labels, but you can tell that there is a grouping structure



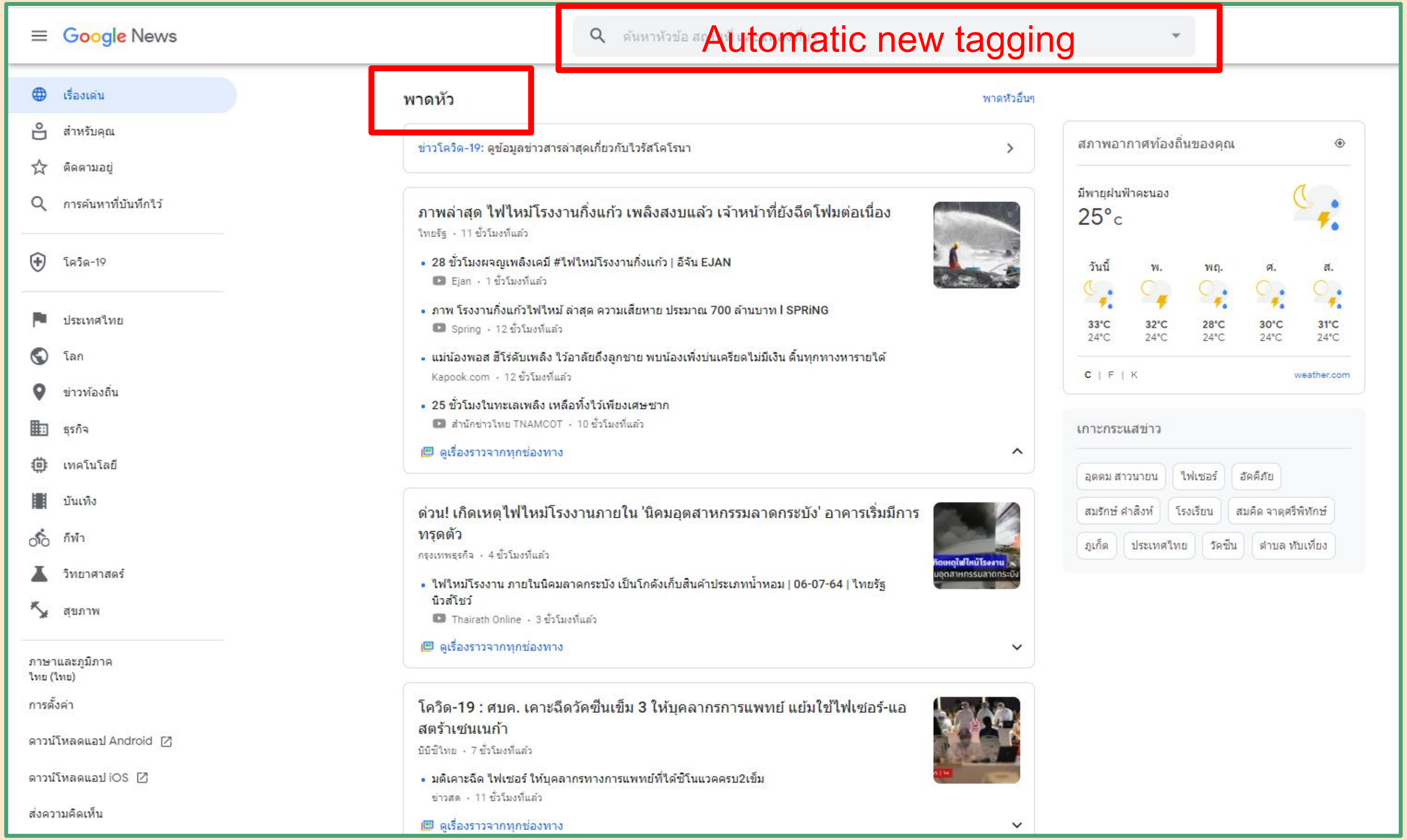
Two class

No class

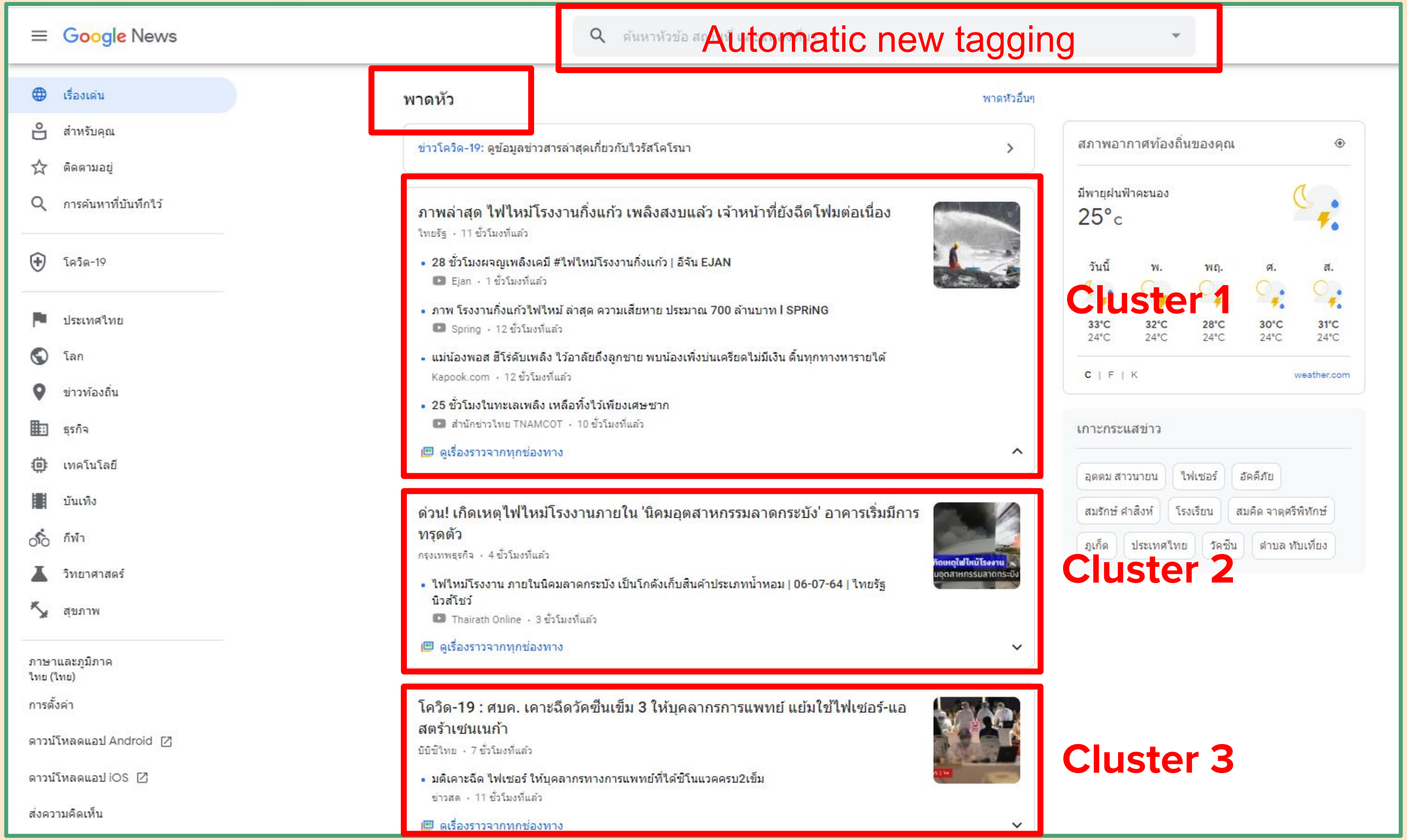
Example of clustering problems



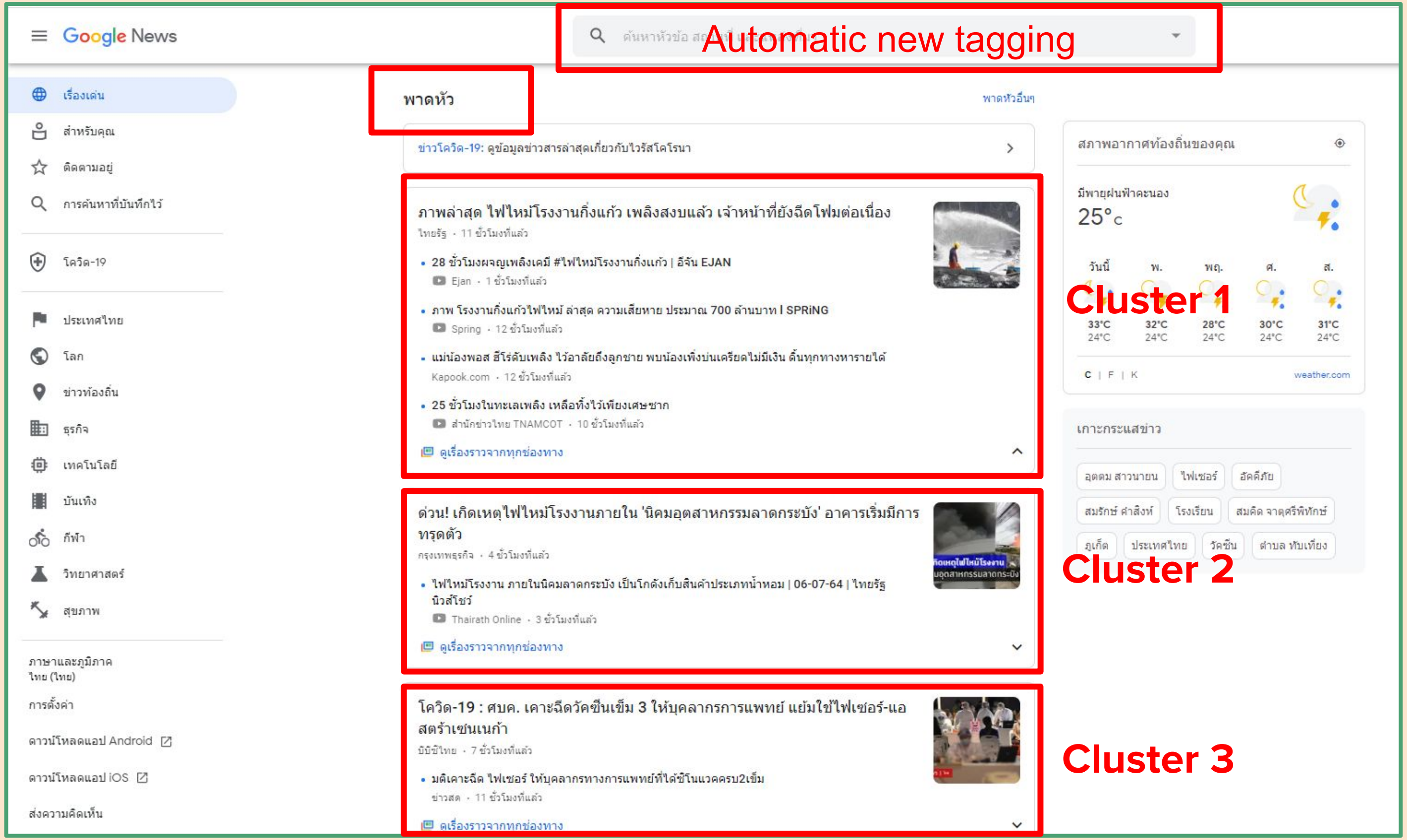
Example of clustering problems



Example clustering problems

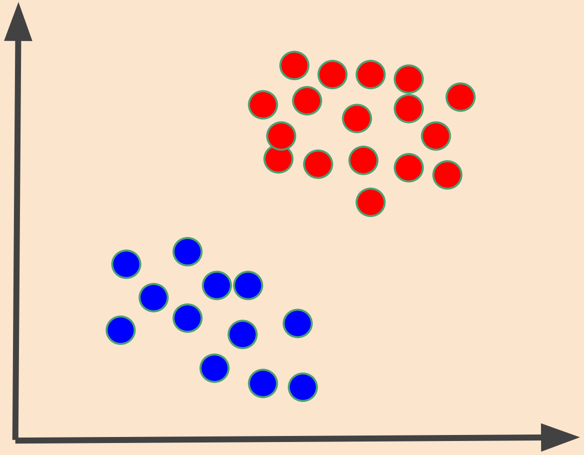


Example clustering problems

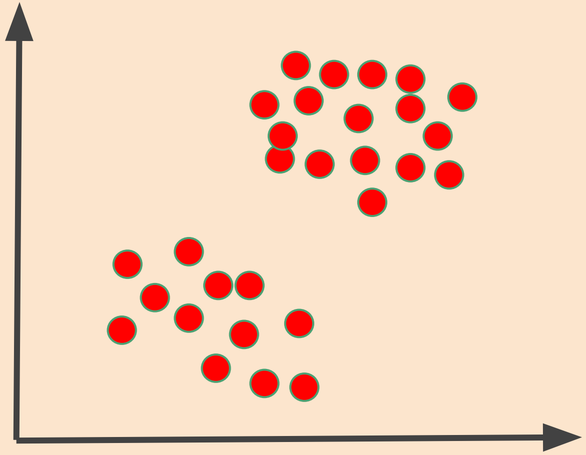


Impossible to label all news !!!

Supervised learning vs Unsupervised learning



Labeled



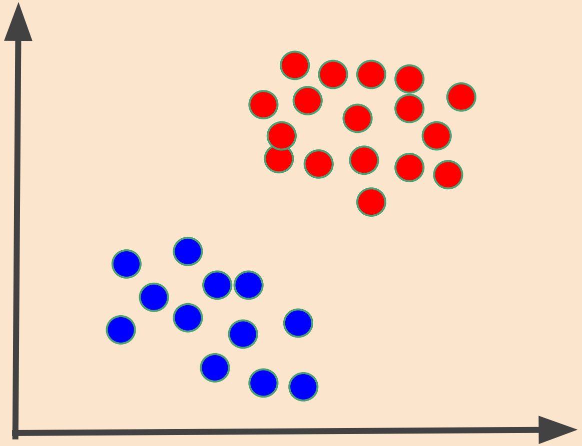
Not label

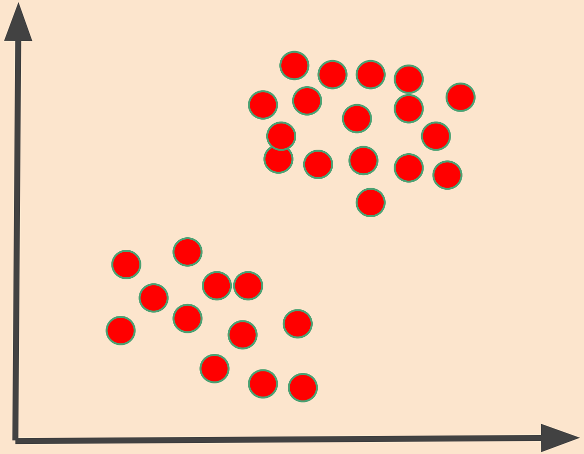
Cost function

Supervised learning → **Error**

Unsupervised learning → **Objective function**

Supervised learning vs Unsupervised learning

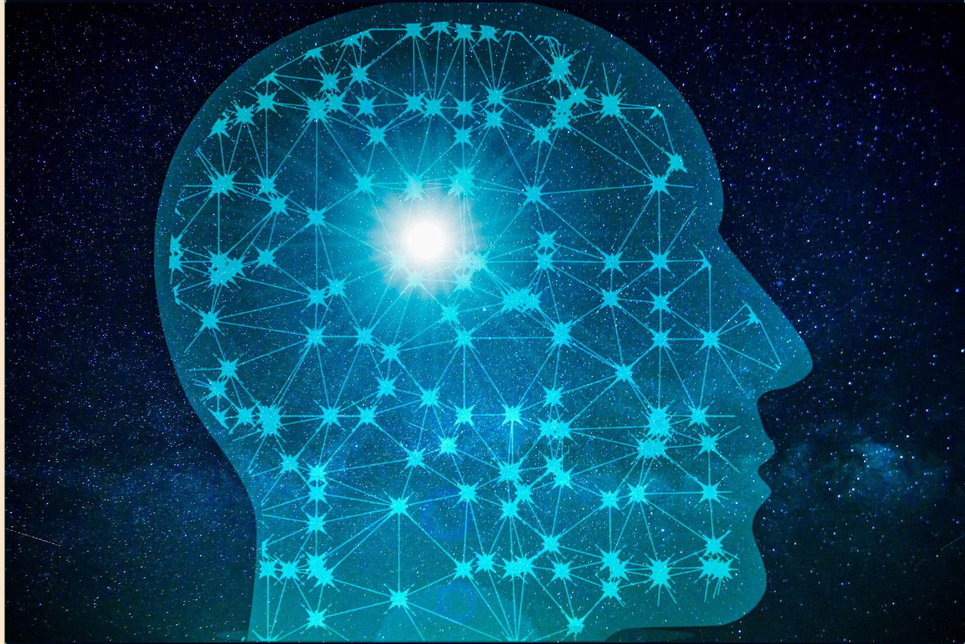




Training **Testing**

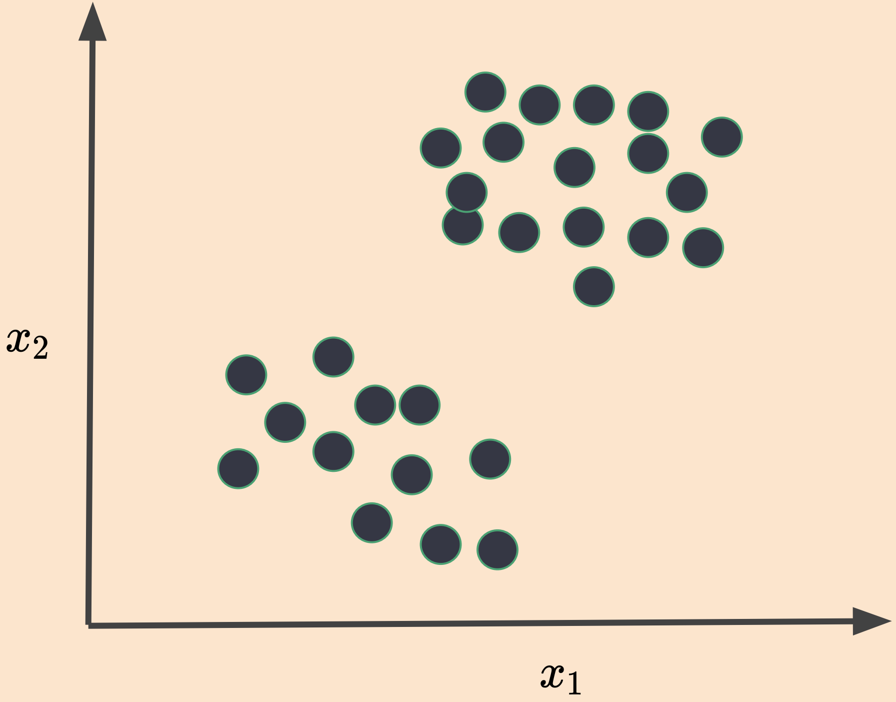
	X	Y	x’	y’	Predict	Cost function
Supervised	Yes	Yes	Yes	No	y’	Error
Unsupervised	Yes	No	Yes	No	Y,y’	Objective function

K-Means Clustering



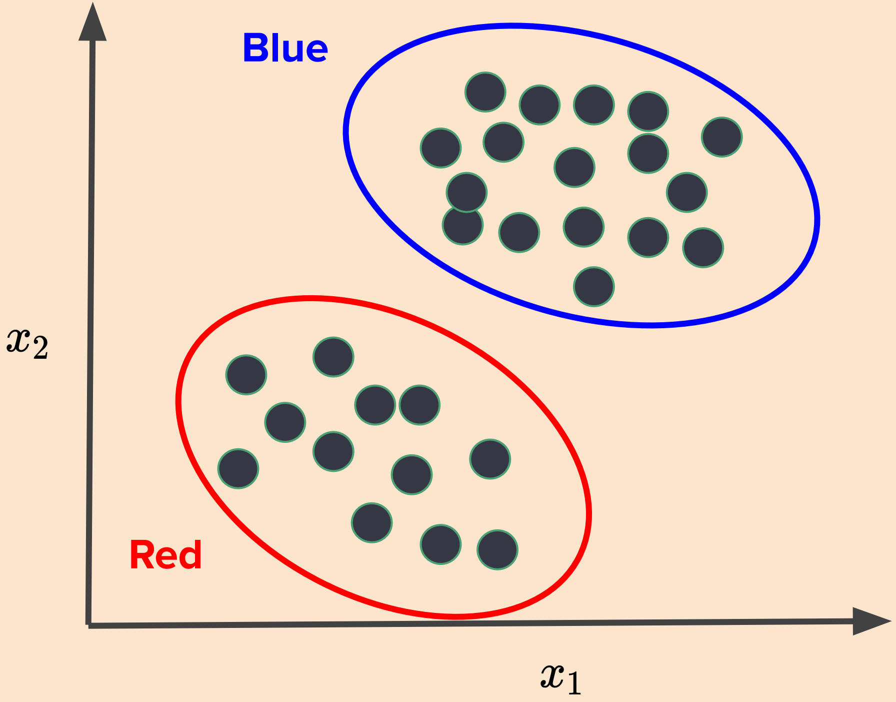
K-means clustering

- K-means clustering is the most popular clustering algorithm.
- K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (centroid).



K-means clustering

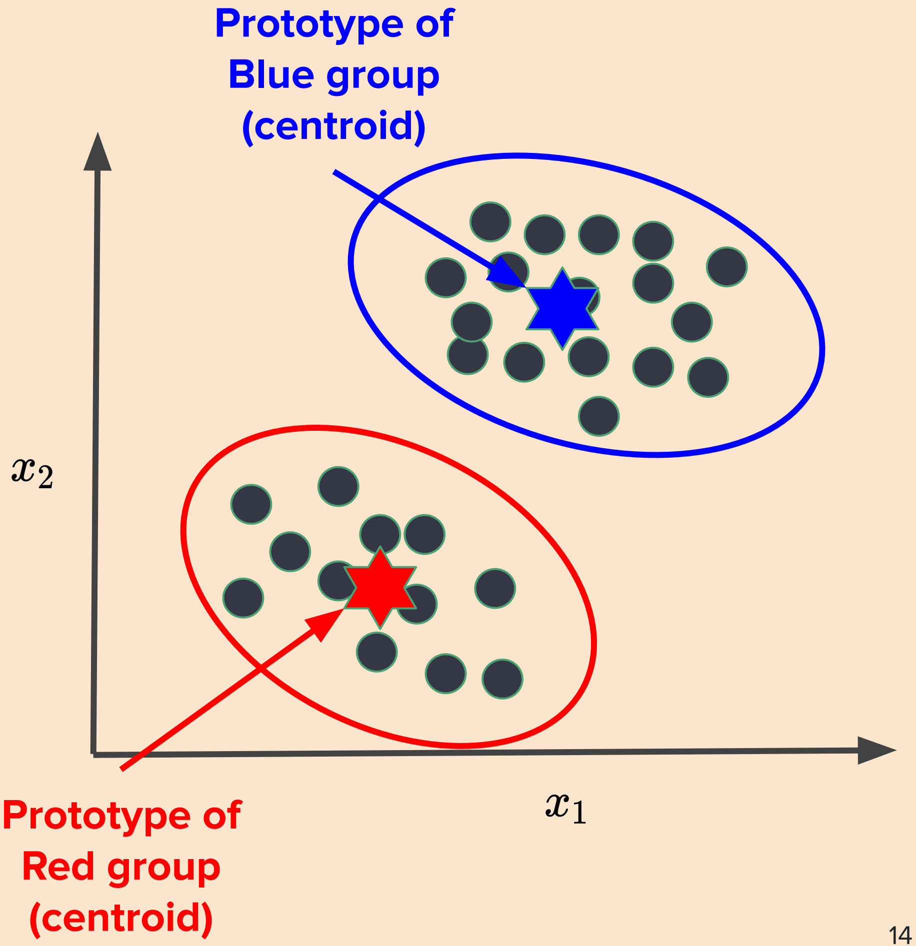
- K-means clustering is the most popular clustering algorithm.
- K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (centroid).



K-means clustering

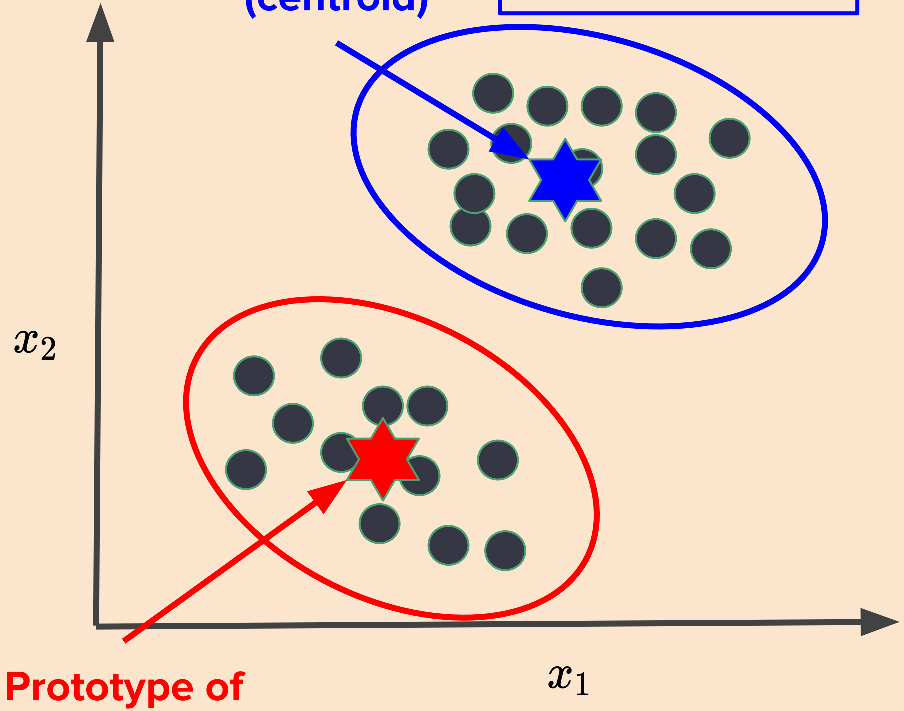
- K-means clustering is the most popular clustering algorithm.

- K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (centroid).



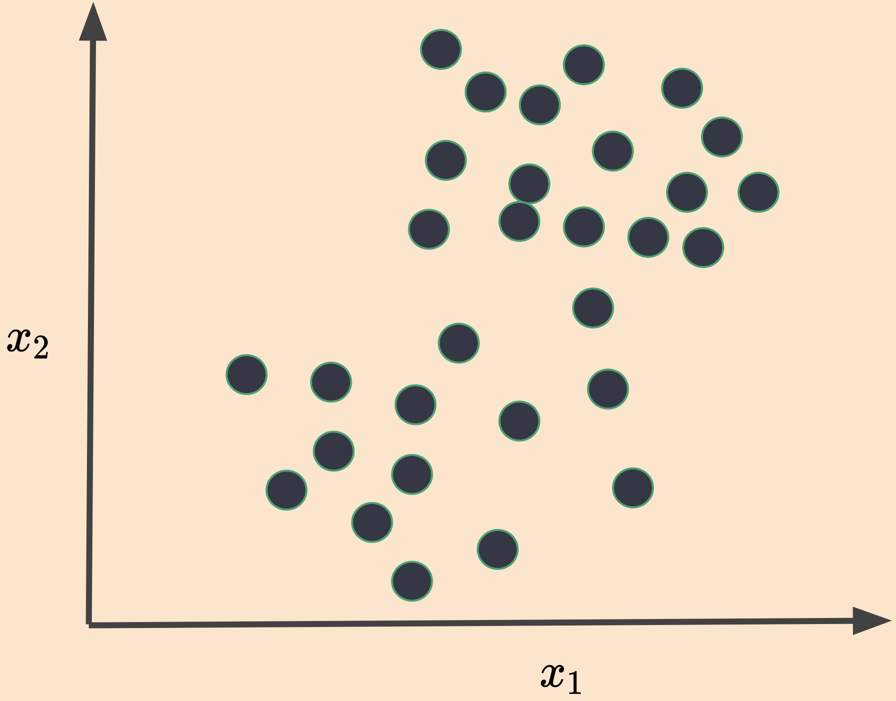
K-means clustering

- ❖ K-means clustering is the most popular clustering algorithm.
- ❖ K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (centroid).



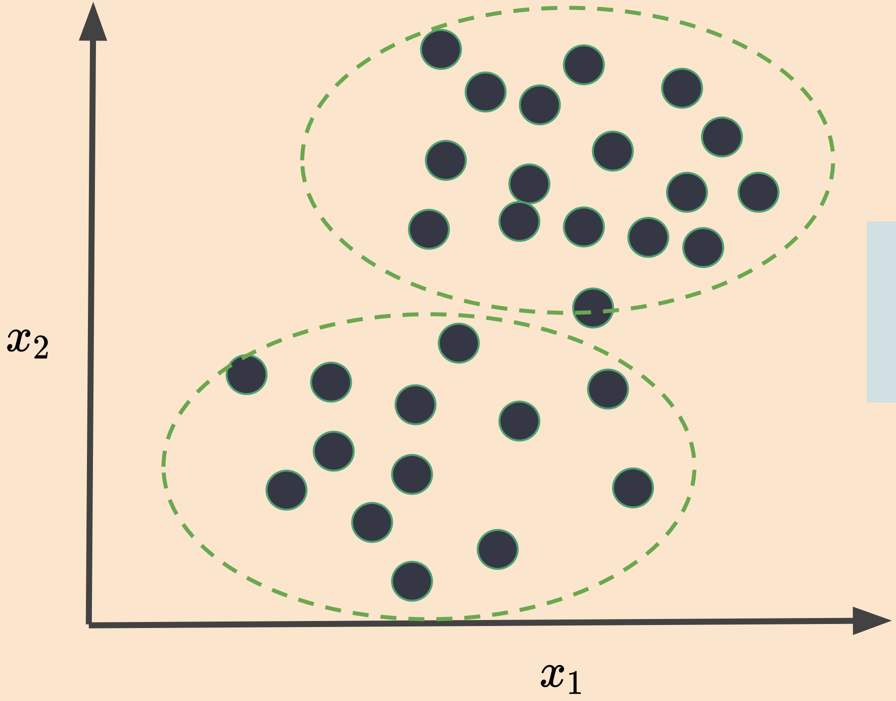
K-means algorithm

❖ **First step: Define k ;** we have a bunch of unlabeled data points. We decide that we are going to find **two clusters** in this data



K-means algorithm

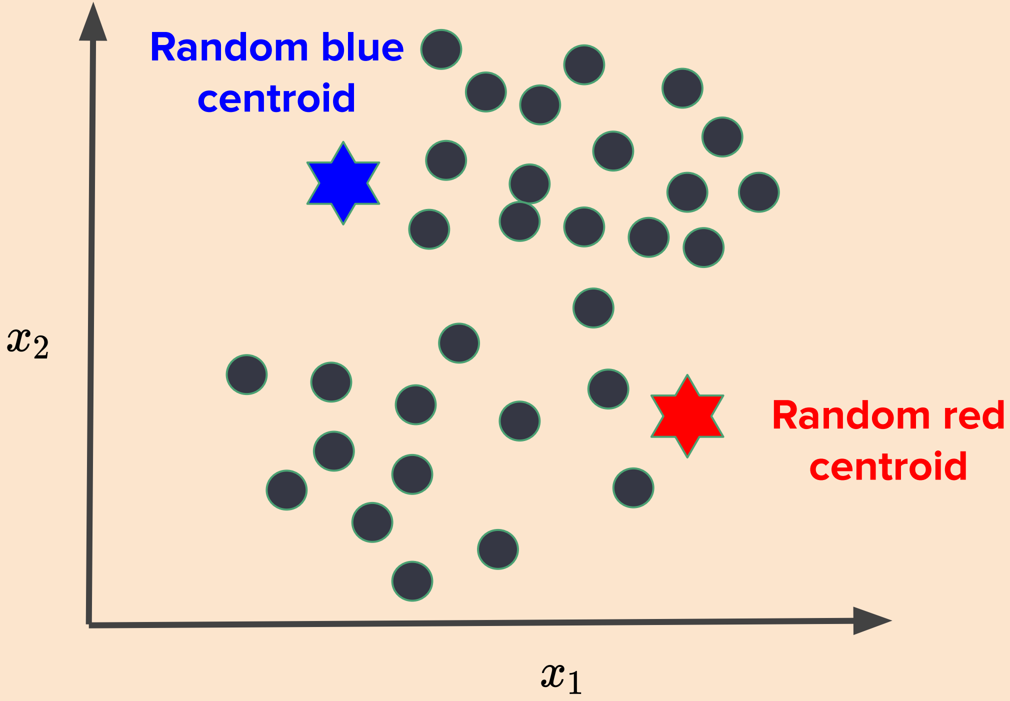
- **First step: Define k ;** we have a bunch of unlabeled data points. We decide that we are going to find **two clusters** in this data



Let's set k = 2 by humans

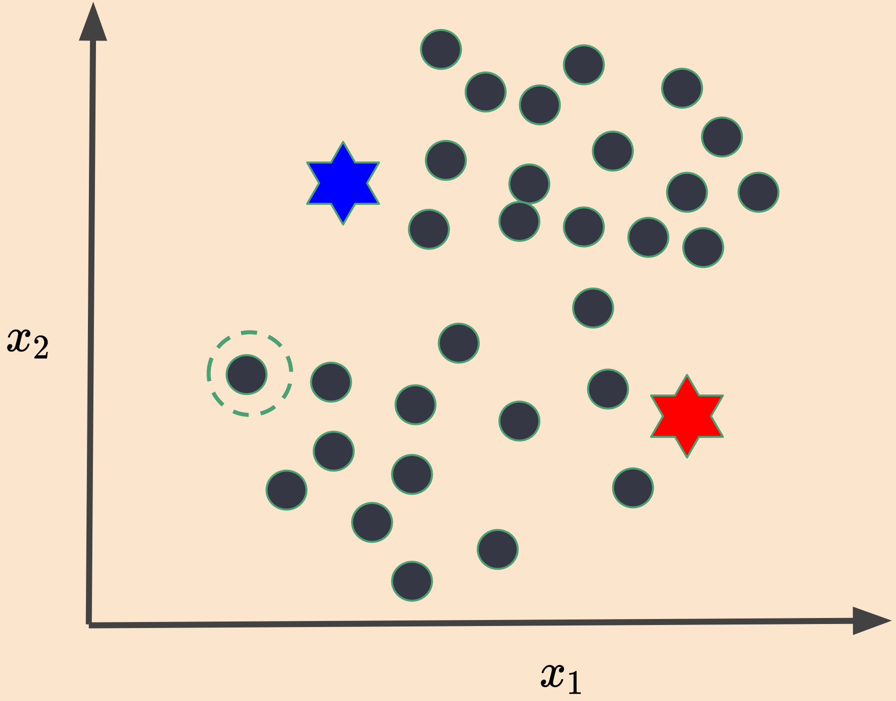
K-means algorithm

❖ **Second step: Random centroids ;** This step is to pick two random locations to be our cluster centroids.



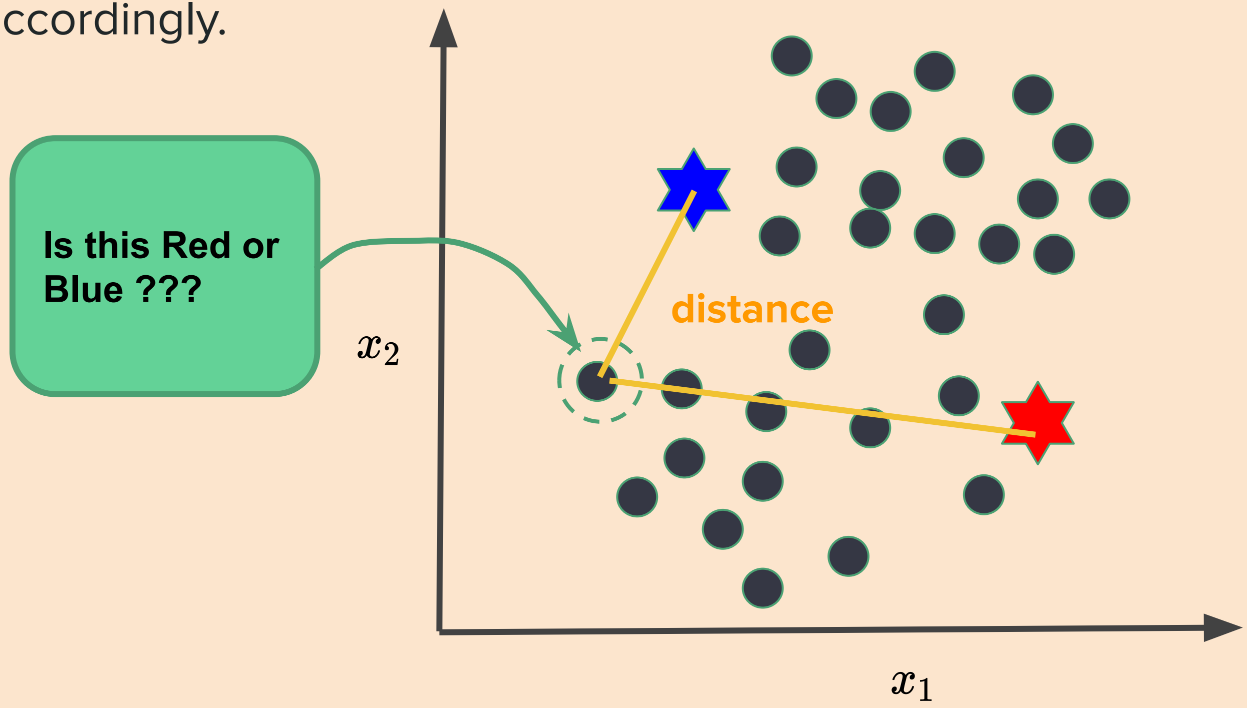
K-means algorithm

❖ **Third step: Cluster assignment ;** This step is to determine whether each dot in every sample is closer to red or blue centroid and label the sample to red or blue accordingly.



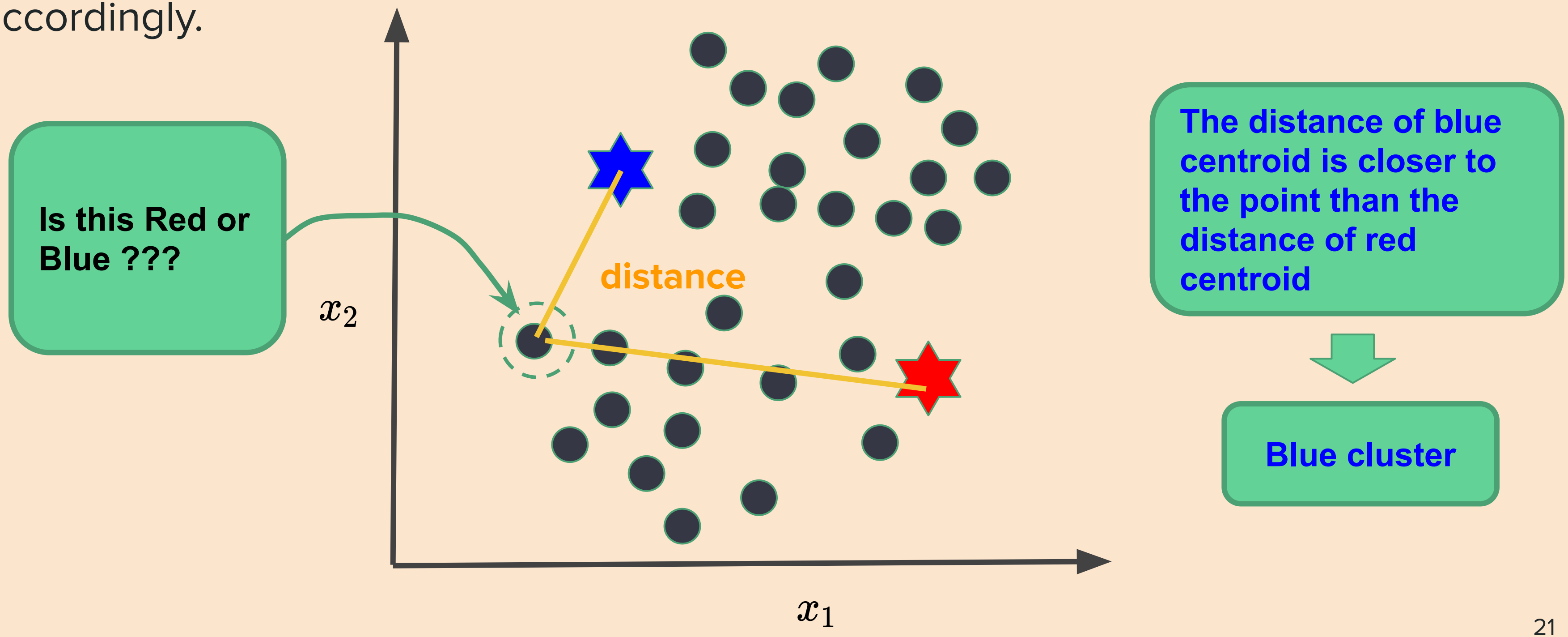
K-means algorithm

❖ **Third step: Cluster assignment** ; This step is to determine whether each dot in every sample is closer to red or blue centroid and label the sample to red or blue accordingly.



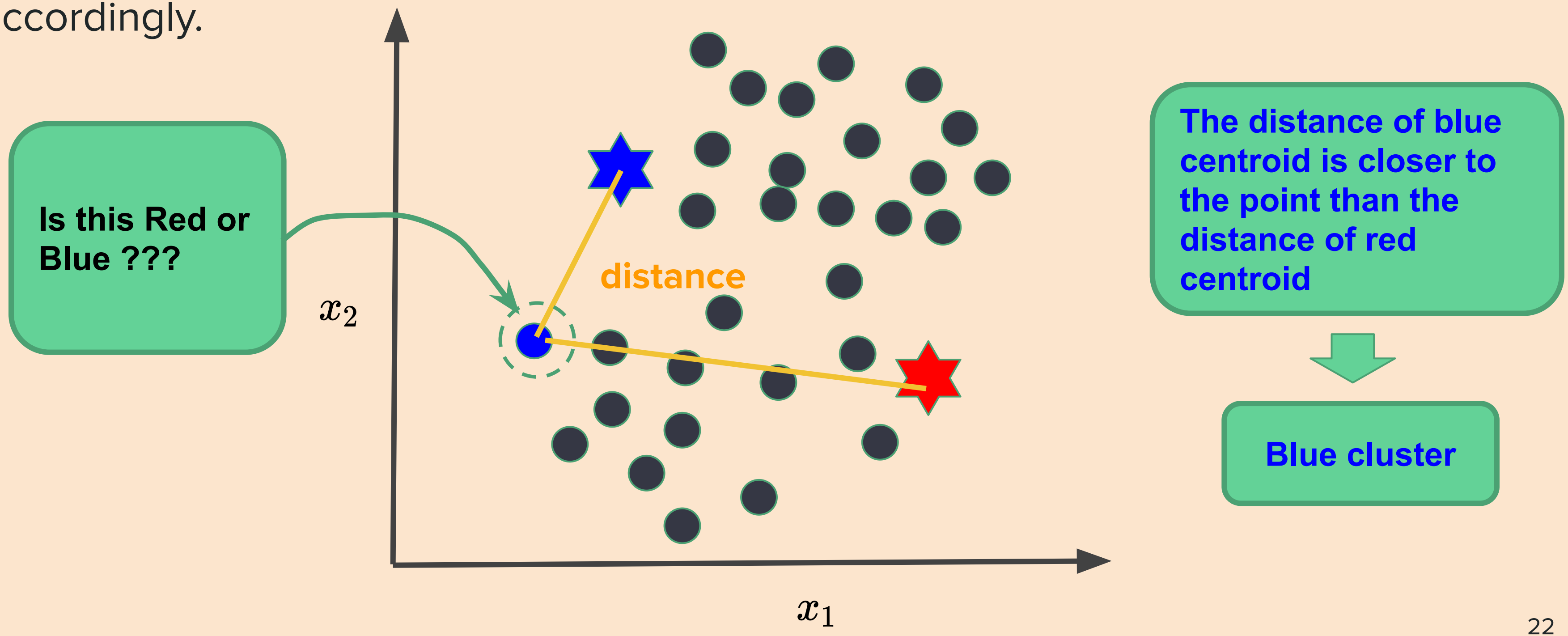
K-means algorithm

❖ **Third step: Cluster assignment ;** This step is to determine whether each dot in every sample is closer to red or blue centroid and label the sample to red or blue accordingly.



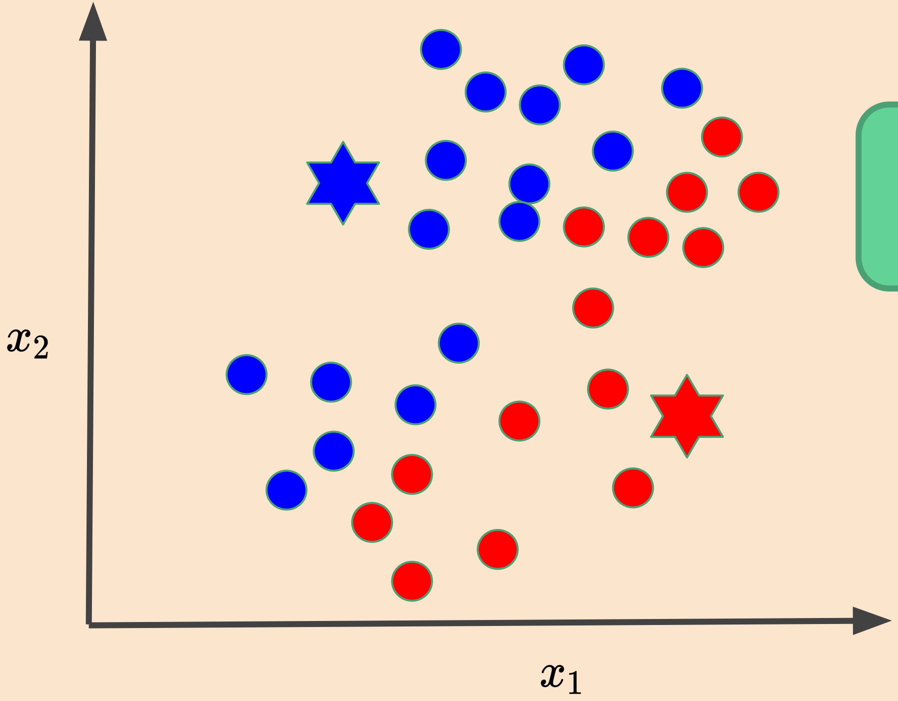
K-means algorithm

❖ **Third step: Cluster assignment ;** This step is to determine whether each dot in every sample is closer to red or blue centroid and label the sample to red or blue accordingly.



K-means algorithm

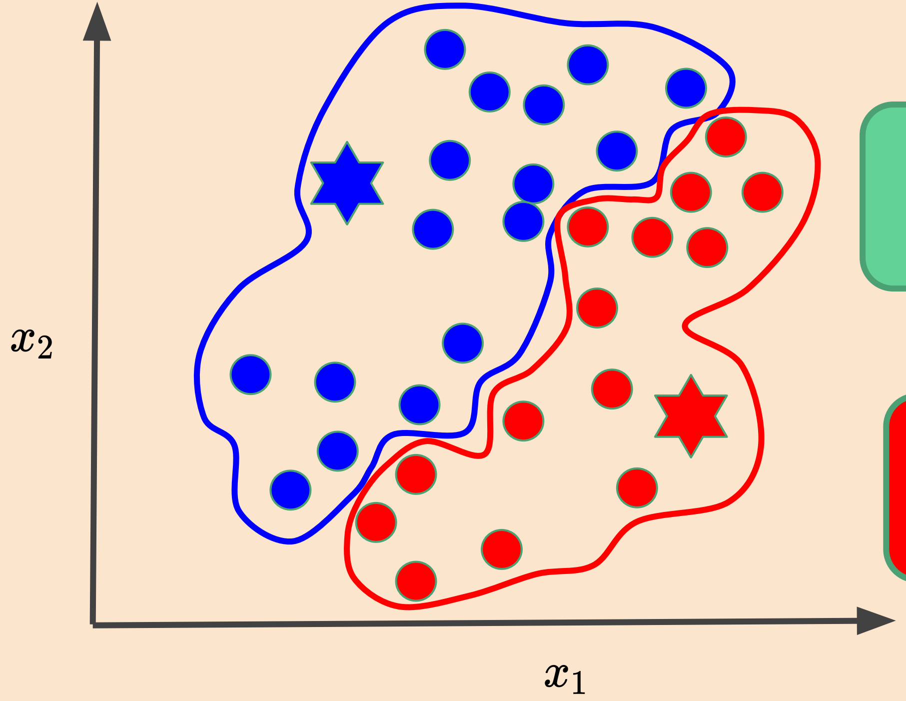
❖ **Third step: Cluster assignment** **;** This step is to determine whether each dot in every sample is closer to red or blue centroid and label the sample to red or blue accordingly.



To this step in every sample

K-means algorithm

❖ **Third step: Cluster assignment ;** This step is to determine whether each dot in every sample is closer to red or blue centroid and label the sample to red or blue accordingly.

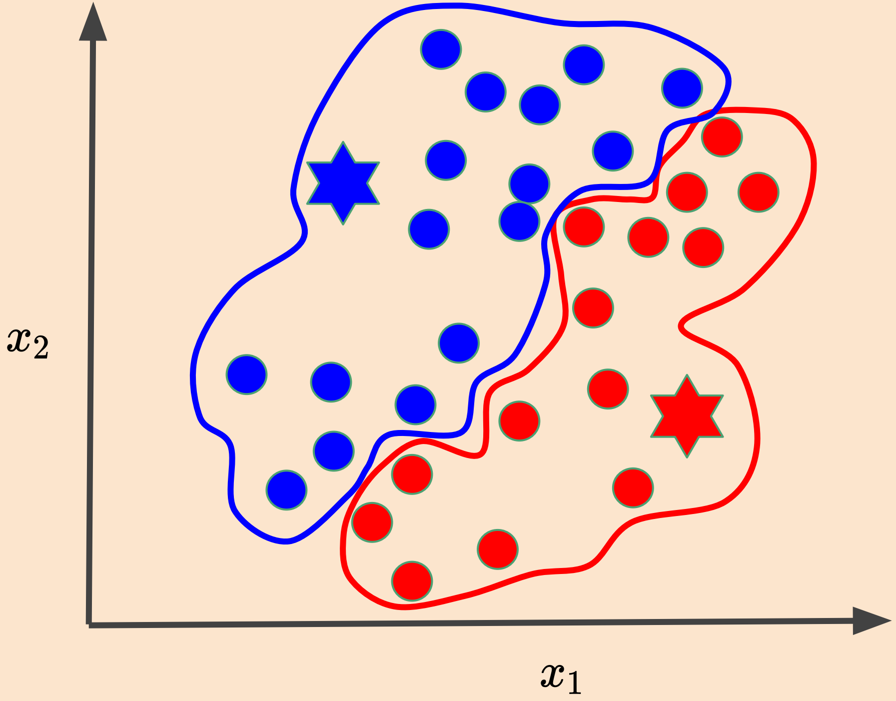


To this step in every sample

This is the bad centroids!!!

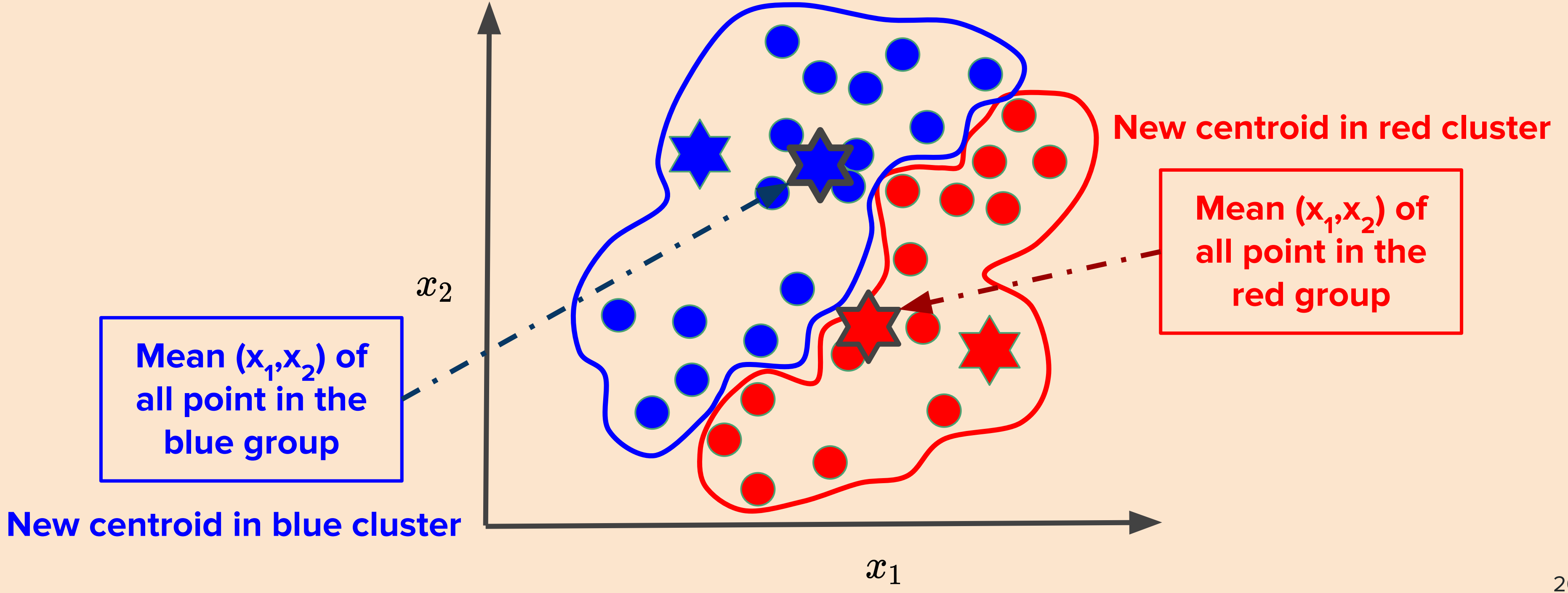
K-means algorithm

❖ **Fourth step: Centroid movement** ; This step is to move the red and blue centroids to the means of clusters.



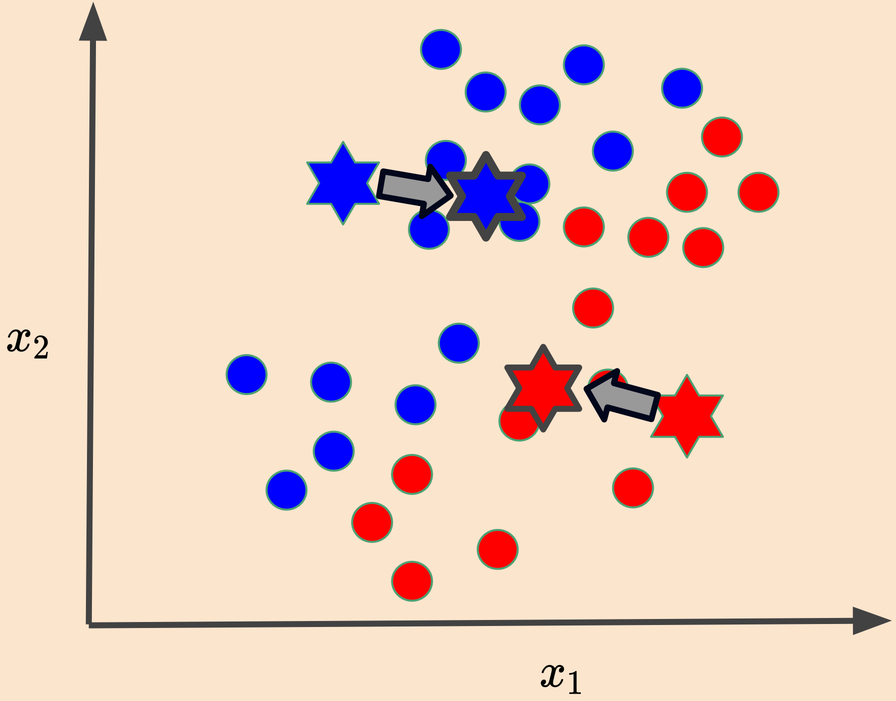
K-means algorithm

❖ **Fourth step: Centroid movement** ; This step is to move the red and blue centroids to the means of clusters.



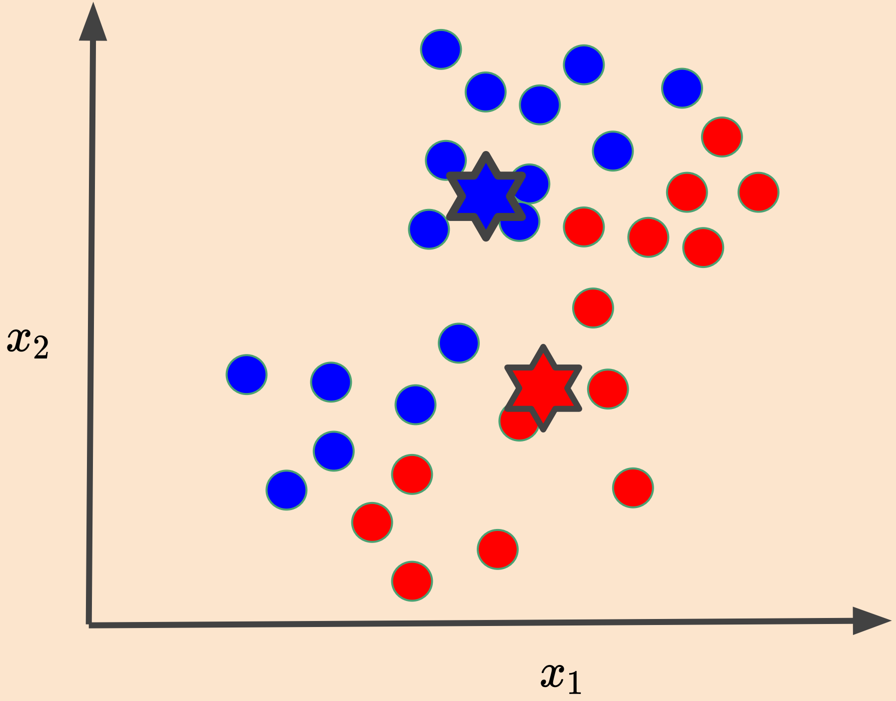
K-means algorithm

- **Fourth step: Centroid movement ;** This step is to move the red and blue centroids to the means of clusters.



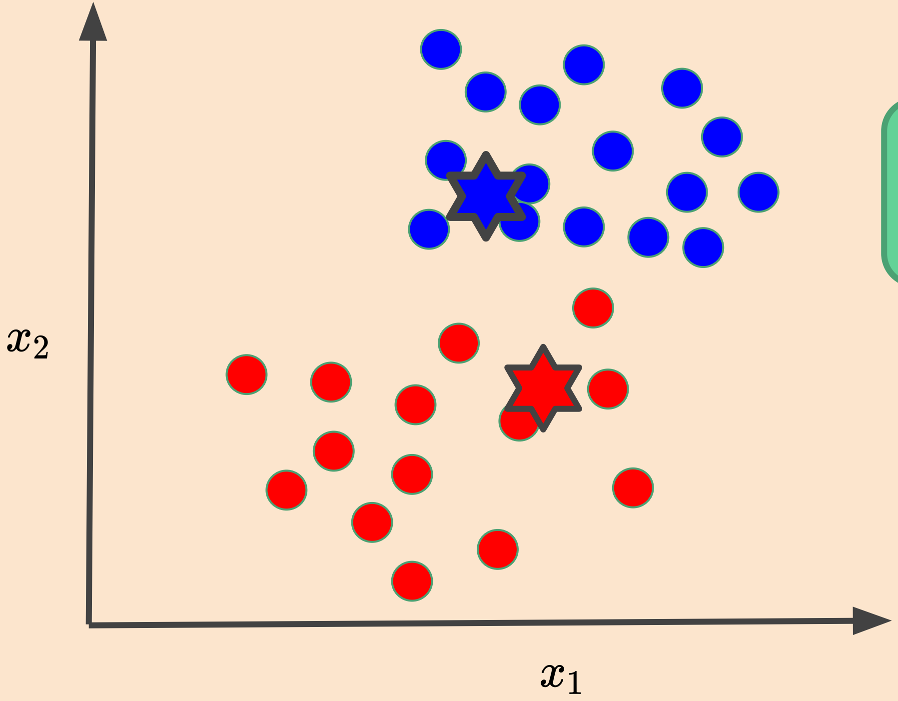
K-means algorithm

❖ **Fourth step: Centroid movement ;** This step is to move the red and blue centroids to the means of clusters.



K-means algorithm

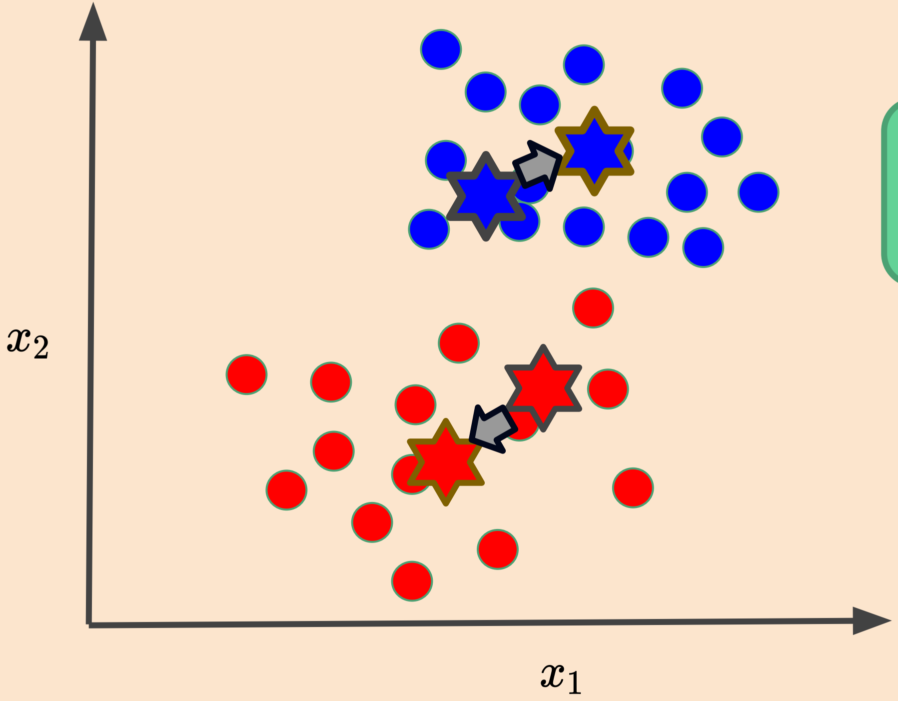
❖ **Fifth step: Repeat to third step and fourth step**



Third step: Cluster assignment

K-means algorithm

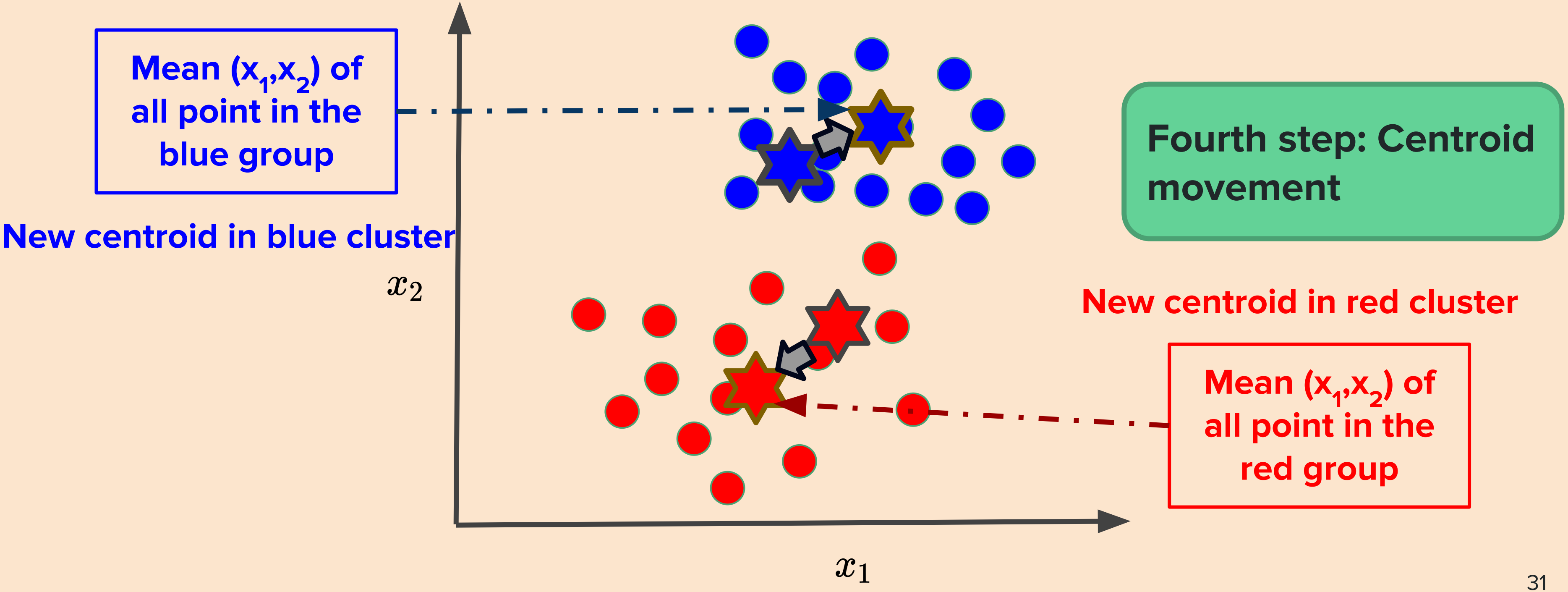
❖ **Fifth step: Repeat to third step and fourth step**



Fourth step: Centroid movement

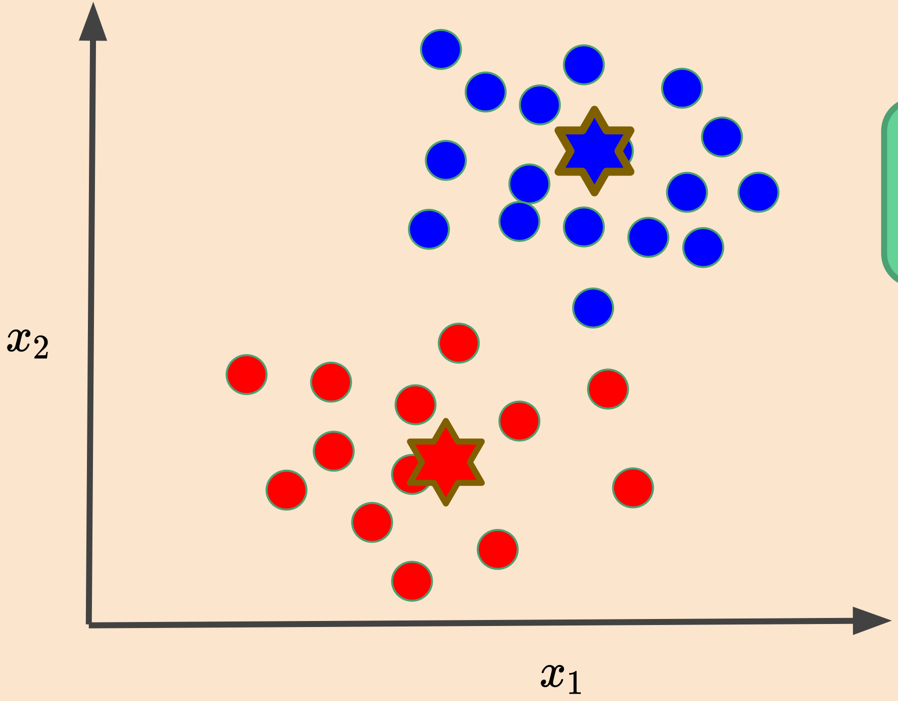
K-means algorithm

- **Fifth step: Repeat to third step and fourth step**



K-means algorithm

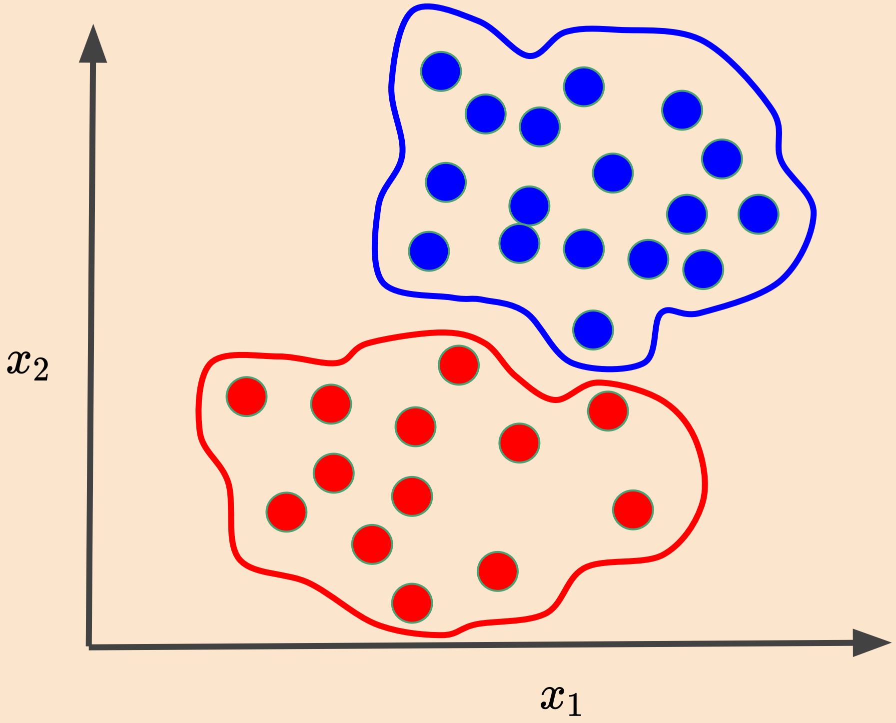
❖ **Fifth step: Repeat to third step and fourth step**



Third step: Cluster assignment

K-means algorithm

- **The algorithm converge to the solution when cluster centroids are not changed anymore.**

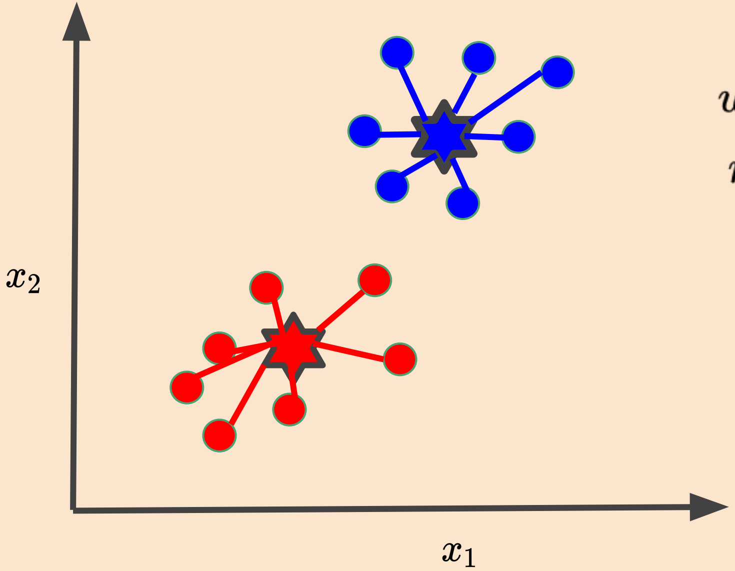


Objective function of K-means algorithm

- **The algorithm minimize the total distances between all data points and the centroids of the clusters they belong to.**

$$Objective\ function: find\ \bar{x}_{red}, \bar{x}_{blue}$$

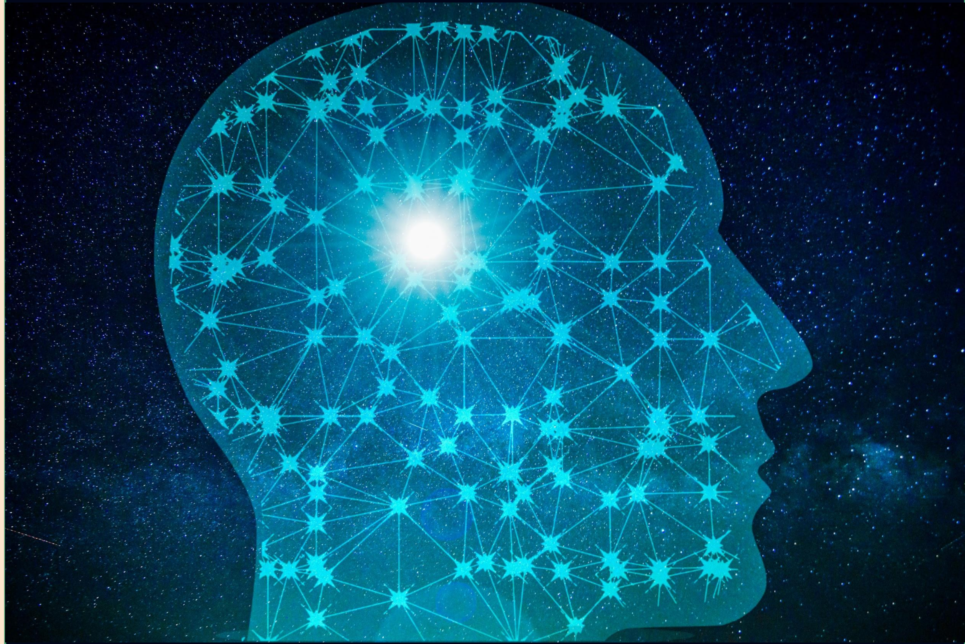
$$Minimize:\ z = \sum_{x_i}^{m_{red}} Distance(x_i, \bar{x}_{red}) + \sum_{x_i}^{m_{blue}} Distance(x_i, \bar{x}_{blue})$$



$where\ m_{red}$ is the set of points having $\bar{x}_{red}$ be the centroid

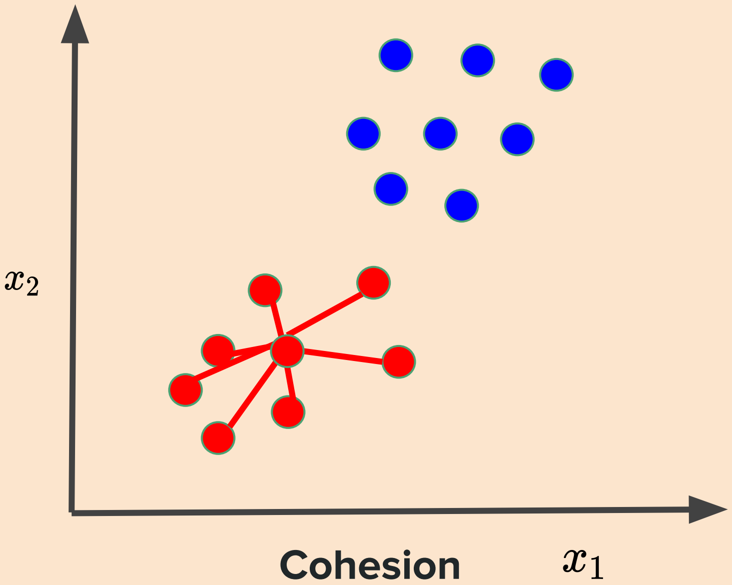
m_{blue} is the set of points having $\bar{x}_{blue}$ be the centroid

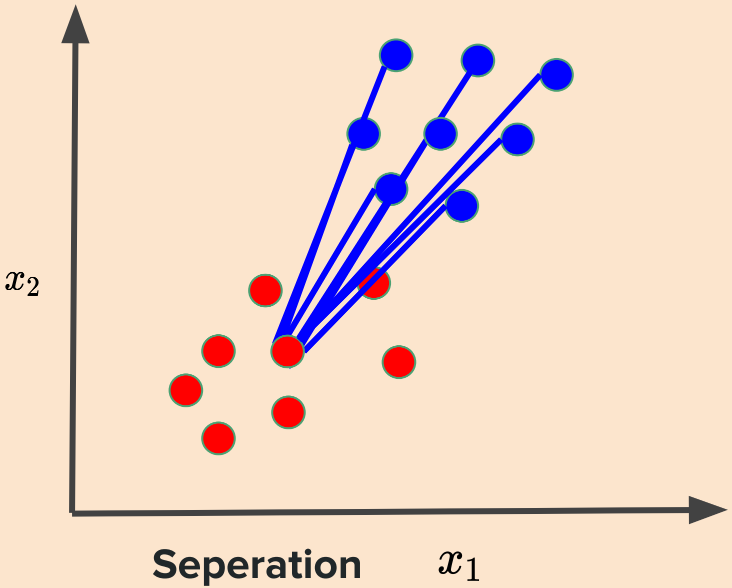
K-means clustering performance



Clustering performance

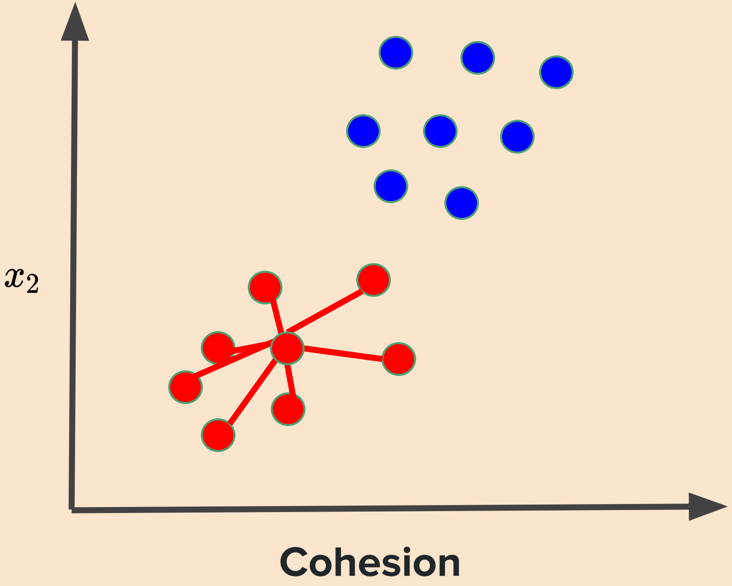
❖ The performance of clustering algorithm is often judged based on how well you algorithm separates out the data into several clusters.





Clustering performance

❖ The performance of clustering algorithm is often judged based on how well you algorithm separates out the data into several clusters.

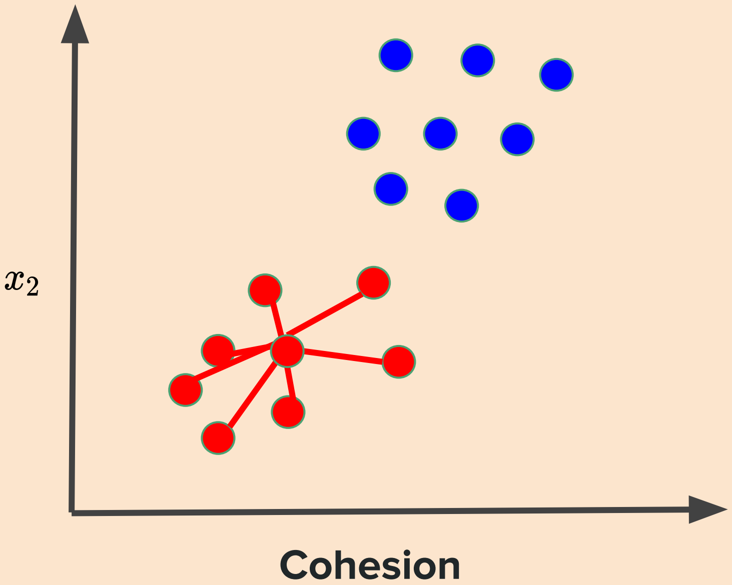


Cohesion: Distance of one data point to all data point in the sample group. (Within-Cluster-Sum-of-Squares (WCSS))

Show the performance within group distance

Clustering performance

❖ The performance of clustering algorithm is often judged based on how well you algorithm separates out the data into several clusters.



Cohesion: Distance of one data point to all data point in the sample group. (Within-Cluster-Sum-of-Squares (WCSS))

$$\text{WCSS} = \sum_{C_k}^{C_n} \left(\sum_{d_i \, in \, C_i}^{d_m} distance(d_i, C_k)^2 \right)$$

Where,

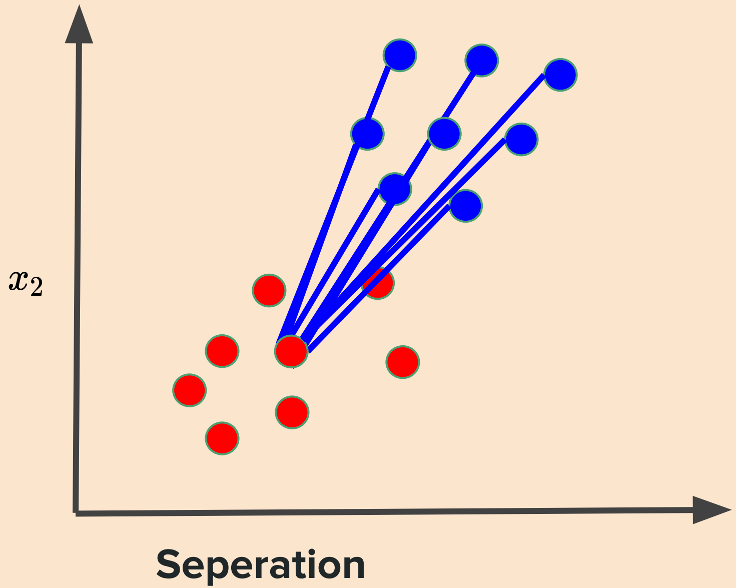
C is the cluster centroids and d is the data point in each Cluster.

Clustering performance

❖ The performance of clustering algorithm is often judged based on how well you algorithm separates out the data into several clusters.

Seperation: Distance of one data point to all data point in the different group.

Show the performance between group distance



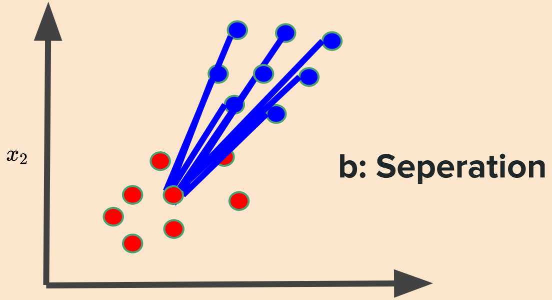
Silhouette coefficient

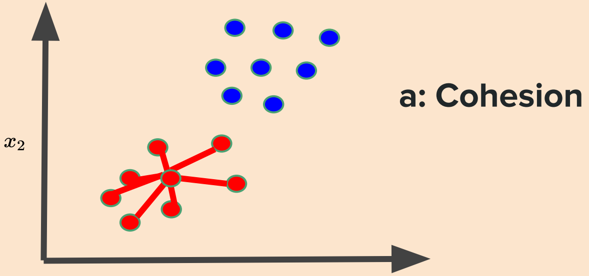
❖ Silhouette coefficient is defined by two separated scores.

$$s = \frac{b - a}{max(b,a)}$$

a : mean distance between a sample and all other points in the same class.

b : mean distance between a sample and all other points in the next nearest class.





Silhouette coefficient

❖ Silhouette coefficient is defined by two separated scores.

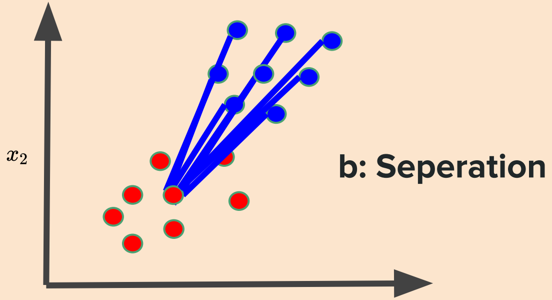
$$s = \frac{b-a}{max(b,a)}$$

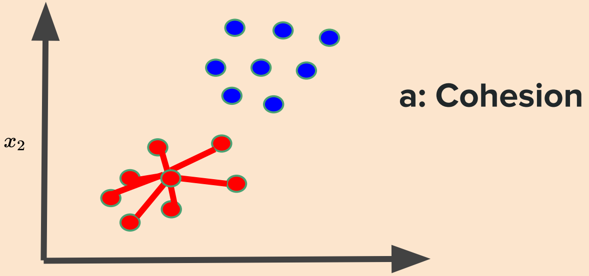
If b > a

$$s = 1 - \frac{a}{b}$$

a : mean distance between a sample and all other points in the same class.

b : mean distance between a sample and all other points in the next nearest class.





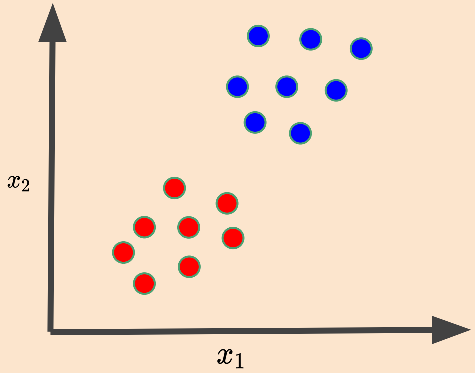
Silhouette coefficient

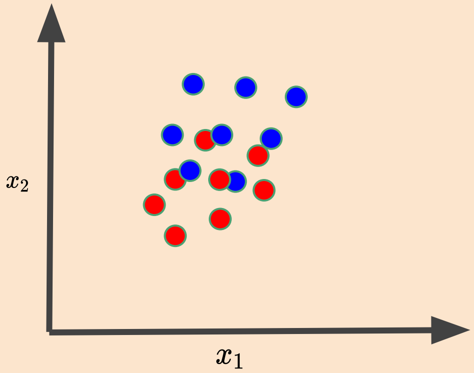
❖ Silhouette coefficient is defined by two separated scores.

$$s = \frac{b-a}{max(b,a)} \xrightarrow{\text{If b > a}} s = 1 - \frac{a}{b}$$

If $b >> a$: S is close to 1.

If $b \sim a$: S is close to 0.





Silhouette coefficient

❖ Silhouette coefficient is defined by two separated scores.

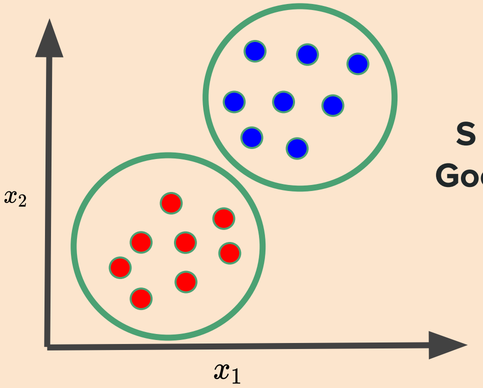
$$s = \frac{b-a}{max(b,a)}$$

If b > a

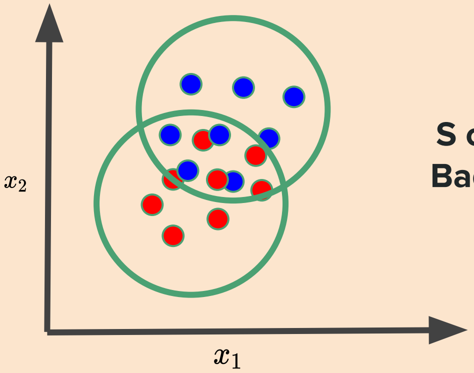
$$s = 1 - \frac{a}{b}$$

If $b >> a$: S is close to 1.

If $b \sim a$: S is close to 0.



S close to 1
Good clusters

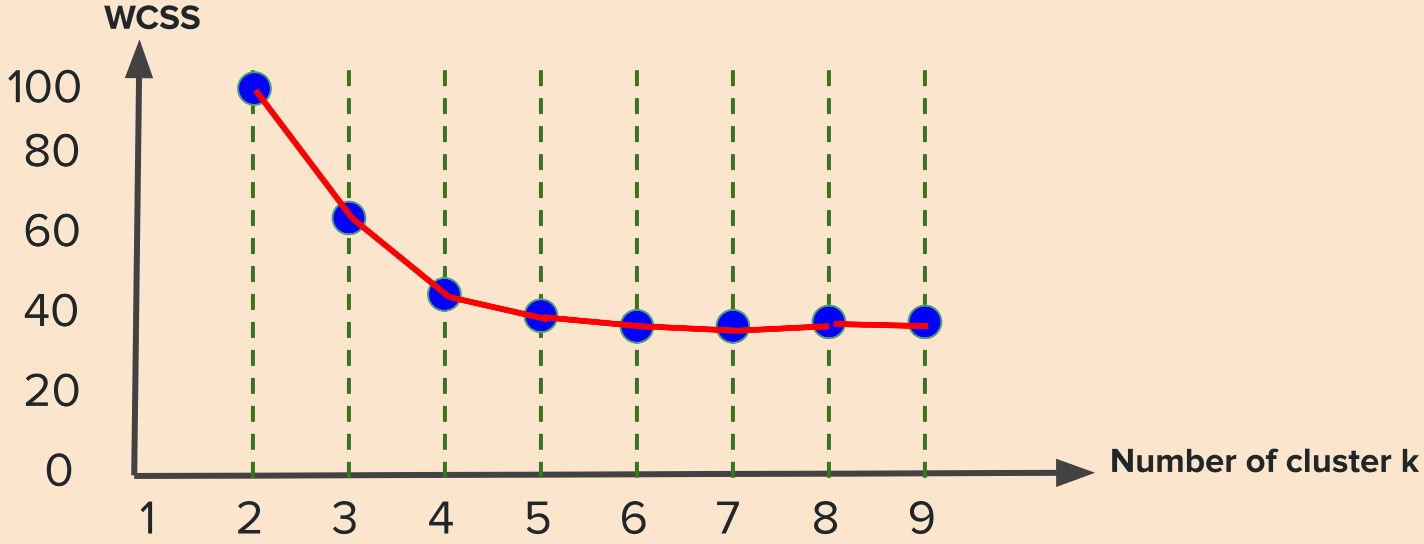


S close to 0
Bad clusters

Choosing the right K

❖ Using **elbow method**

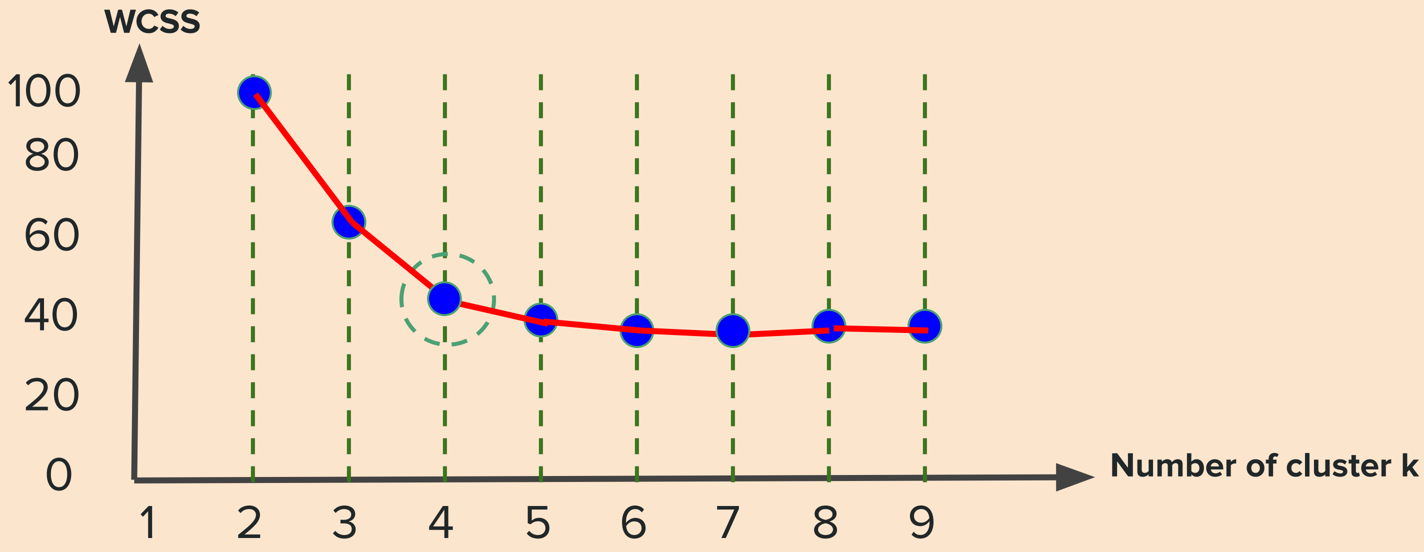
- You compute performance metrics such as Silhouette coefficient or WCSS, while varying k.
- Pick the k at the elbow point. At this point, more clusters do not necessarily mean higher performance.



Choosing the right K

❖ Using **elbow method**

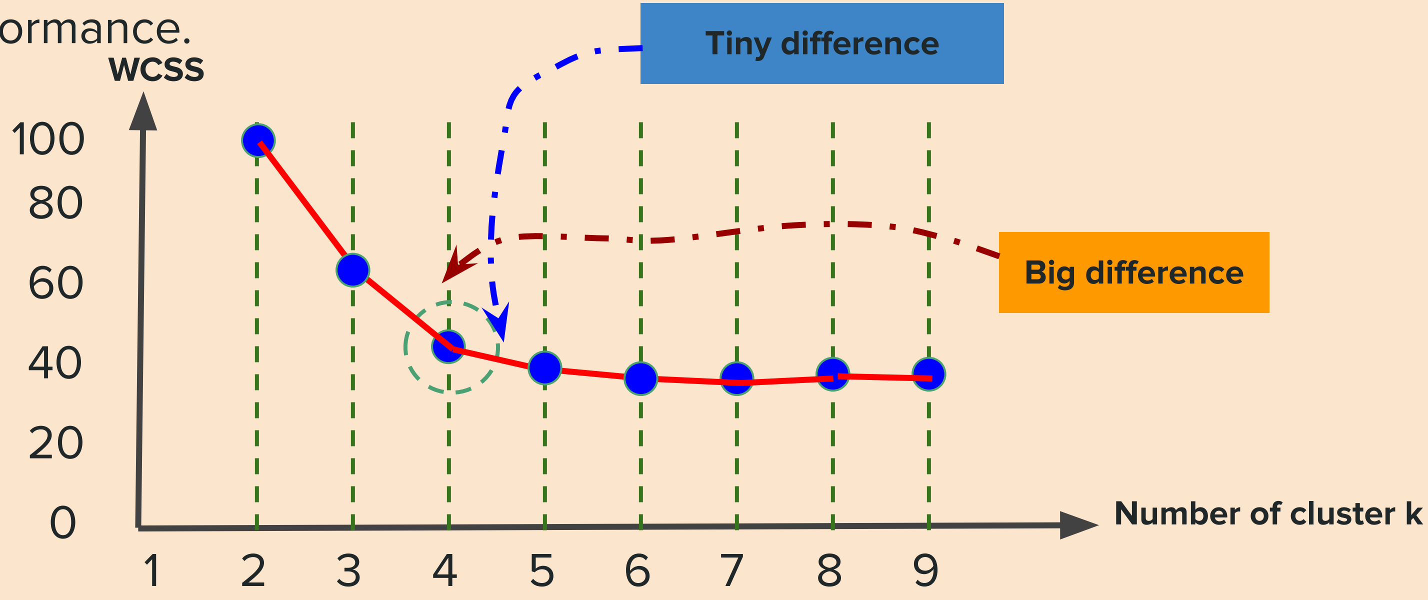
- You compute performance metrics such as Silhouette coefficient or WCSS, while varying k.
- Pick the k at the elbow point. At this point, more clusters do not necessarily mean higher performance.



Choosing the right K

❖ Using **elbow method**

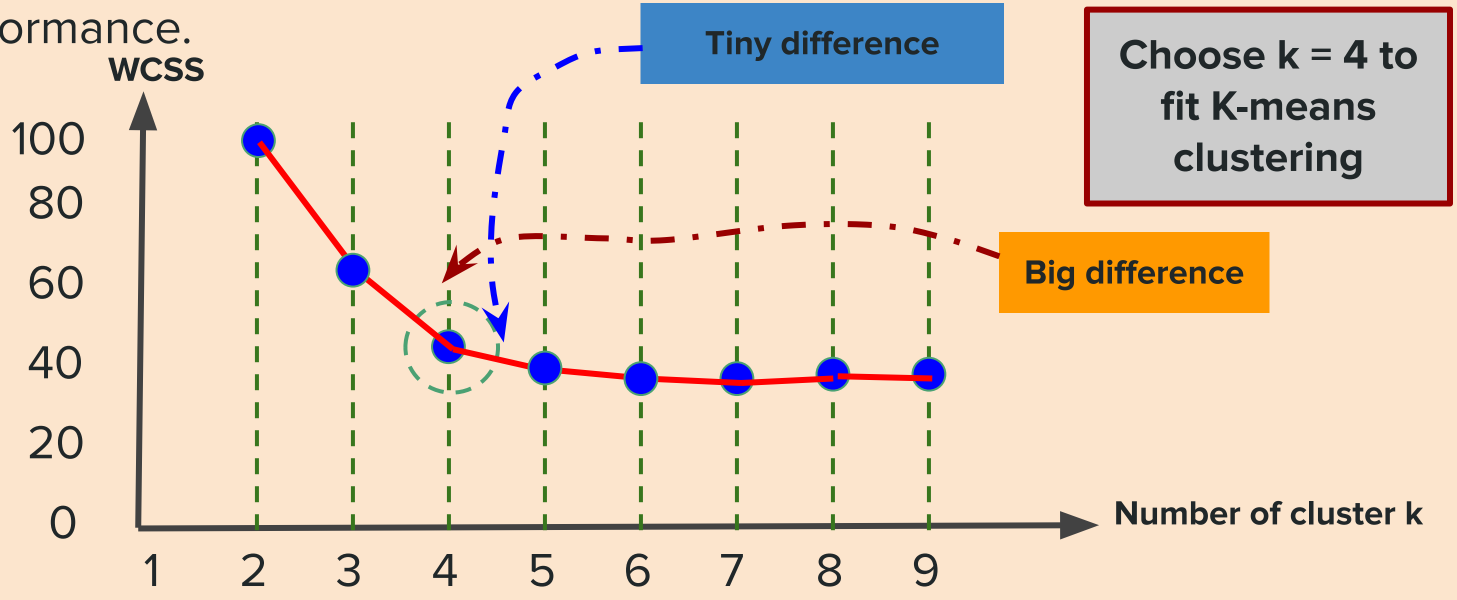
- You compute performance metrics such as Silhouette coefficient or WCSS, while varying k.
- Pick the k at the elbow point. At this point, more clusters do not necessarily mean higher performance.



Choosing the right K

❖ Using **elbow method**

- You compute performance metrics such as Silhouette coefficient or WCSS, while varying k.
- Pick the k at the elbow point. At this point, more clusters do not necessarily mean higher performance.



Good luck 😉


Uppsala universitet 1852–1916
Det goda universitetet. Festskrift till Anders Hallberg
Studenten, Staden och Sanningen
MITT LIV
UPPSALA UNIVERSITE