

# K-means Clustering

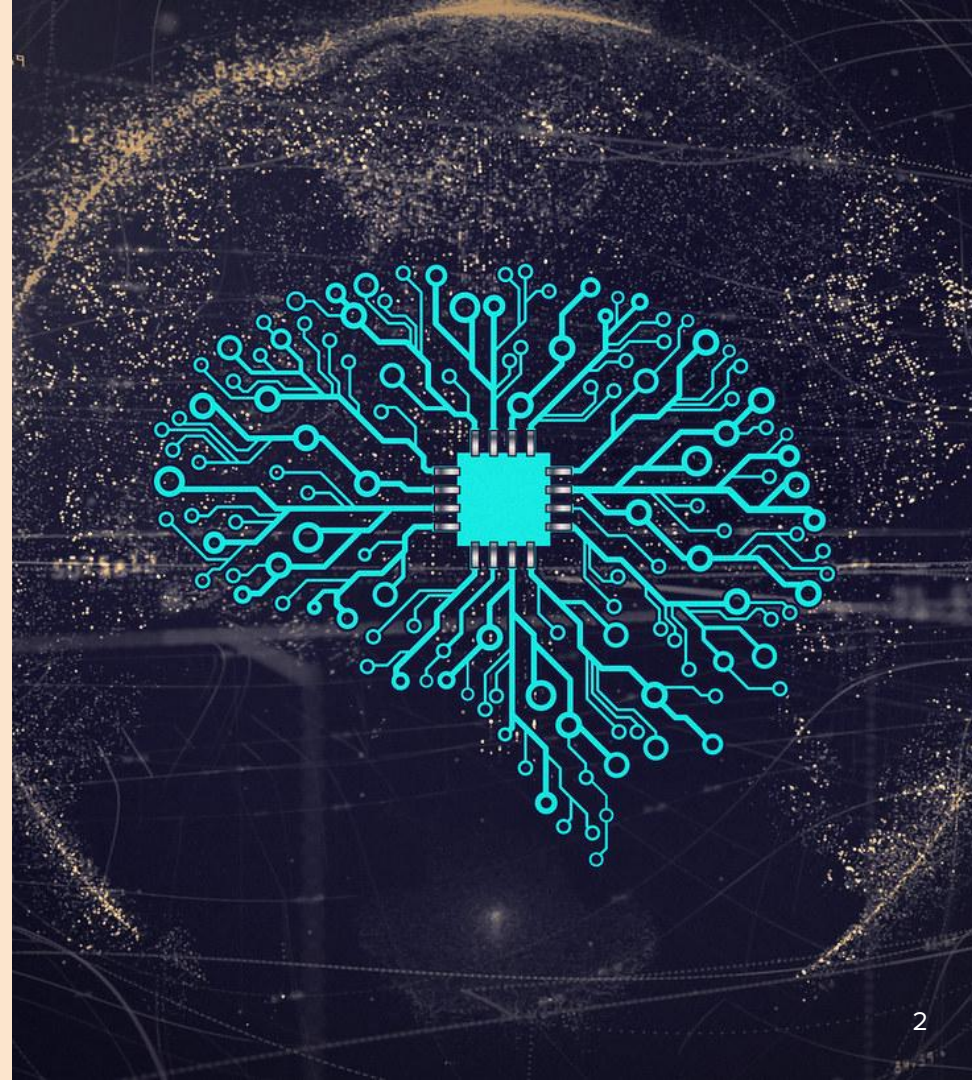
---

Dr. Paisit Khanarsa

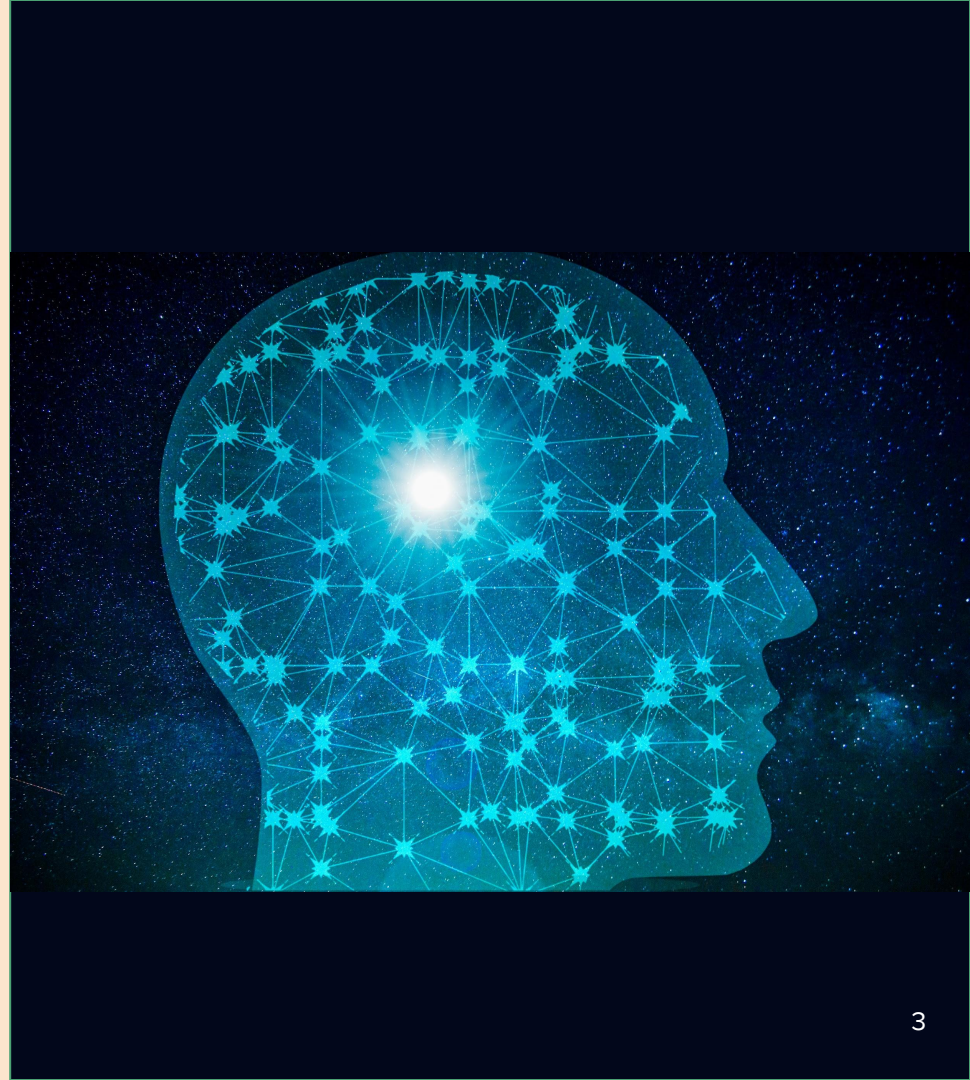
Fibo, Kmutt

# Outline

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



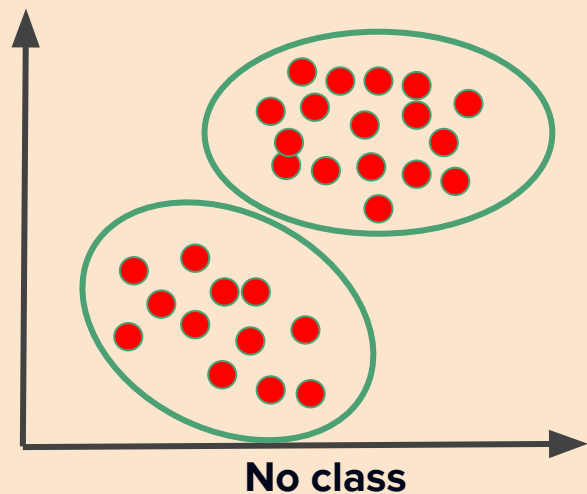
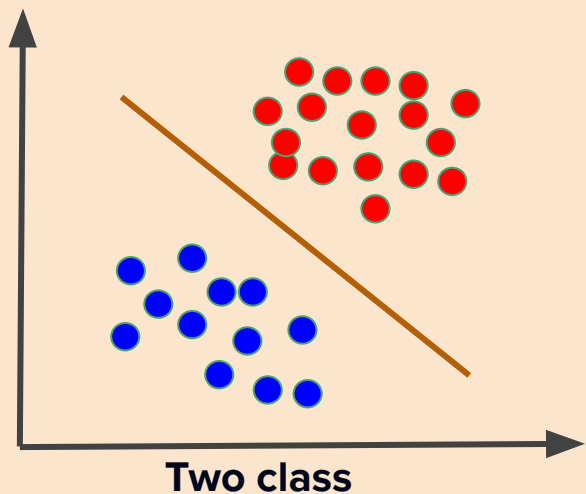
# Unsupervised Learning



# 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





# Example of clustering problems

The screenshot displays the Google News interface in Thai. The top navigation bar includes the Google News logo and a search bar. The left sidebar contains a list of categories and filters, such as 'เรื่องเด่น' (Featured), 'สำหรับคุณ' (For you), 'ติดตามอยู่' (Following), 'การค้นหาที่บันทึกไว้' (Saved searches), 'โควิด-19' (COVID-19), 'ประเทศไทย' (Thailand), 'โลก' (World), 'ข่าวท้องถิ่น' (Local news), 'ธุรกิจ' (Business), 'เทคโนโลยี' (Technology), 'บันเทิง' (Entertainment), 'กีฬา' (Sports), 'วิทยาศาสตร์' (Science), 'สุขภาพ' (Health), 'ภาษาและภูมิภาคไทย (ไทย)' (Language and Region: Thailand), 'การตั้งค่า' (Settings), 'ดาวน์โหลดแอป Android' (Download Android app), 'ดาวน์โหลดแอป iOS' (Download iOS app), and 'ส่งความคิดเห็น' (Send feedback).

The main content area is titled 'พาดหัว' (Headlines) and features several news stories:

- ข่าวโควิด-19: ดูข้อมูลข่าวสารล่าสุดเกี่ยวกับไวรัสโคโรนา** (COVID-19 News: View the latest news about the coronavirus). This section includes three articles:
  - ภาพล่าสุด ไฟไหม้โรงงานกึ่งแก้ว เหล็กสงบแล้ว เจ้าหน้าที่ยังฉีดไฟต่อเนื่อง** (Latest photo: Fire at a semi-glass factory, iron is calm, firefighters still spraying). It mentions a fire at a glass factory in Ejan, 11 minutes ago.
  - ภาพ โรงงานกึ่งแก้วไฟไหม้ ล่าสุด ความเสียหาย ประมาณ 700 ล้านบาท | SPRING** (Photo: Glass factory fire, latest damage estimated at 700 million baht). It mentions a fire at a glass factory in Spring, 12 minutes ago.
  - แม่ป้องพลอส ฮีโร่ดับเพลิง 'ไวอาลีย์' ลึกซึ้ง ขยาย พบน้องเหิงบนเคียวไม่มีเงิน คืนทุกทางหารายได้** (Mother of Pollos, fire hero 'Viawalee' is deep, expanding, found Nong Heng on a sword with no money, returning everything, earning money). It mentions a fire at a glass factory in Kapook.com, 12 minutes ago.
  - 25 ชั่วโมงในทะเลเพลิง เหลือทิ้งไว้เพียงเศษซาก** (25 hours in a sea of fire, only a few scraps left). It mentions a fire at a glass factory in Thairath Online, 10 minutes ago.
- ด่วน! เกิดเหตุไฟไหม้โรงงานภายใน 'นิคมอุตสาหกรรมลาดกระบัง' อาคารเริ่มมีการทรุดตัว** (Urgent! Fire at a factory inside 'Ladkrabang Industrial Estate' building is starting to collapse). It mentions a fire at a glass factory in Kongsamut, 4 minutes ago.
- โควิด-19 : ศบค. เคาะฉีดวัคซีนเข็ม 3 ให้บุคลากรทางการแพทย์ แยมใช้ไฟเซอร์-แอสตราเซนเนกา** (COVID-19 : SCMC. Vaccinate 3rd dose for healthcare workers, use Pfizer-AstraZeneca). It mentions a vaccine in Nidchi, 7 minutes ago.

On the right side, there is a weather widget titled 'สภาพอากาศท้องถิ่นของคุณ' (Local weather conditions) showing a temperature of 25°C and a 5-day forecast. Below the weather widget is a 'เกาะกระแสข่าว' (News flow) section with buttons for various topics like 'ลดนม สวามาน', 'ไฟเซอร์', 'ฮัดสิดีย', 'สมรักษ์ คำสิงห์', 'โรงเรียน', 'สมคิด จาตุศรีพิทักษ์', 'ภูเก็ต', 'ประเทศไทย', 'วัดจีน', 'ศาลา ทับเที่ยง'.

# Example of clustering problems

Google News

ค้นหาหัวข้อ... Automatic new tagging

พาดหัว

ข่าวโควิด-19: คู่มือลดข่าวสารล่าสุดเกี่ยวกับไวรัสโคโรนา

ภาพล่าสุด ไฟไหม้โรงงานกึ่งแก้ว เพลิงสงบแล้ว เจ้าหน้าที่ยังฉีดไฟต่อเนื่อง

ไทยรัฐ · 11 ชั่วโมงที่แล้ว

- 28 ชั่วโมงผจญเพลิงเคมี่ #ไฟไหม้โรงงานกึ่งแก้ว | อีจัน EJAN  
Ejan · 1 ชั่วโมงที่แล้ว
- ภาพ โรงงานกึ่งแก้วไฟไหม้ ล่าสุด ความเสียหาย ประมาณ 700 ล้านบาท | SPRING  
Spring · 12 ชั่วโมงที่แล้ว
- แม่กองพล อีโรสดับเพลิง ใว้อาสัยถึงลูกชาย พบน้องหมีบนเคียวไม่มีเงิน ดันทุทางหารายได้  
Kapook.com · 12 ชั่วโมงที่แล้ว
- 25 ชั่วโมงทะเลเพลิง เกลือทิ้งไว้เพียงเศษซาก  
สำนักข่าวไทย TNAMCOT · 10 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

ด่วน! เกิดเหตุไฟไหม้โรงงานภายใน 'นิคมอุตสาหกรรมลาดกระบัง' อาคารเริ่มมีการทรุดตัว

กรุงเทพธุรกิจ · 4 ชั่วโมงที่แล้ว

- ไฟไหม้โรงงาน ภายในนิคมลาดกระบัง เป็นโศกภัยเกินคำประเทน้ำหมอม | 06-07-64 | ไทยรัฐนิวส์โชว์  
Thairath Online · 3 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

โควิด-19 : ศบค. เคาะฉีดวัคซีนเข็ม 3 ให้บุคลากรทางการแพทย์ แยมใช้ไฟเซอร์-แอสตราเซนเนกา

มติชนออนไลน์ · 7 ชั่วโมงที่แล้ว

- มติเคาะฉีด ไฟเซอร์ ให้บุคลากรทางการแพทย์ที่ได้ขึ้นแวคครบ2เข็ม  
ข่าวสด · 11 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

สภาพอากาศท้องถิ่นของคุณ

มีพายุฝนฟ้าคะนอง  
25°C

วันนี้	พ.	พ.	ศ.	ส.
33°C 24°C	32°C 24°C	28°C 24°C	30°C 24°C	31°C 24°C

C | F | K

เกาะกระแสข่าว

ลดตม สาขานาน | ไฟเซอร์ | ฮัดดี้

สมรักษ์ คำสิงห์ | โรงเรียน | สมคิด จาตุศรีพิทักษ์

ภูเก็ต | ประเทศไทย | วัคซีน | คำลวด ทัพเรือ

# Example clustering problems

**Automatic new tagging**

**พาดหัว**

**Cluster 1**

ข่าวโควิด-19: คู่มือลดข่าวสารล่าสุดเกี่ยวกับไวรัสโคโรนา

ภาพล่าสุด ไฟไหม้โรงงานกิ่งแก้ว เพชรสังขมแล้ว เจ้าหน้าที่ยังฉีดไฟต่อเนื่อง  
ไทยรัฐ · 11 ชั่วโมงที่แล้ว

- 28 ชั่วโมงก่อนเผยเคสเคมี่ #ไฟไหม้โรงงานกิ่งแก้ว | อีจัน EJAN  
Ejan · 1 ชั่วโมงที่แล้ว
- ภาพ โรงงานกิ่งแก้วไฟไหม้ ล่าสุด ความเสียหาย ประมาณ 700 ล้านบาท | SPRING  
Spring · 12 ชั่วโมงที่แล้ว
- แม่ป้องพอส ฮีโร่ดับเพลิง ใจอาสาเสี่ยงลูกชาย พบน้องหมีบ้านเครือคิดไม่มีเงิน ดันทุกทางหารายได้  
Kapook.com · 12 ชั่วโมงที่แล้ว
- 25 ชั่วโมงในทะเลเพลิง เหลือทั้งไวไฟเพียงเศษซาก  
สำนักข่าวไทย TNAMCOT · 10 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

**Cluster 2**

ด่วน! เกิดเหตุไฟไหม้โรงงานภายใน นิคมอุตสาหกรรมลาดกระบัง อาคารเริ่มมีการ  
ทรุดตัว  
กรุงเทพธุรกิจ · 4 ชั่วโมงที่แล้ว

- ไฟไหม้โรงงาน ภายในนิคมลาดกระบัง เป็นโศกภัยเกินสิ้นค่าประเภทน้ำหอม | 06-07-64 | ไทยรัฐ  
นิวส์โฮว์  
Thairath Online · 3 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

**Cluster 3**

โควิด-19 : ศบค. เคาะฉีดวัคซีนเข็ม 3 ให้อนุคลากรแพทย์ แย้มใช้ไฟเซอร์-แอสตราเซนเนกา  
บีบีซีไทย · 7 ชั่วโมงที่แล้ว

- มดีเคาะฉีด ไฟเซอร์ ให้อนุคลากรทางการแพทย์ที่ได้ขึ้นแวคครบ2เข็ม  
ข่าวสด · 11 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

สภาพอากาศท้องถิ่นของคุณ

มีพายุฝนฟ้าคะนอง  
25°C

วันนี้	พ.	พฤ.	ศ.	ส.
33°C 24°C	32°C 24°C	28°C 24°C	30°C 24°C	31°C 24°C

C | F | K

เกาะกระแสข่าว

ลดตาม สำนวนาน ไฟเซอร์ ฮัดดี้

สมรักษ์ คำสิงห์ โรงเรียน สมคิด จาตุศรีพิทักษ์

ภูเก็ต ประเทศไทย วัคซีน ค่าลด ภาษีเงิน

# Example clustering problems

Google News

ค้นหาหัวข้อ... Automatic new tagging

พาดหัว

ข่าวโควิด-19: คู่มือคนข่าวสารล่าสุดเกี่ยวกับไวรัสโคโรนา

ภาพล่าสุด ไฟไหม้โรงงานกิ่งแก้ว เพลิงสงบแล้ว เจ้าหน้าที่ยังคงฉีดน้ำต่อเนื่อง

ไทยรัฐ · 11 ชั่วโมงที่แล้ว

- 28 ชั่วโมงก่อนเผยเคมีย์ #ไฟไหม้โรงงานกิ่งแก้ว | อีจัน EJAN  
Ejan · 1 ชั่วโมงที่แล้ว
- ภาพ โรงงานกิ่งแก้วไฟไหม้ ล่าสุด ความเสียหาย ประมาณ 700 ล้านบาท | SPRING  
Spring · 12 ชั่วโมงที่แล้ว
- แม่ป้องกันพลส ธีรธัมมสิงห์ โฉมใหม่ถึงลูกชาย พบน้องหญิงบนเตียงไม่มีเงิน ดันทุทางหายไค้  
Kapook.com · 12 ชั่วโมงที่แล้ว
- 25 ชั่วโมงก่อนเผยเคมีย์ เพลิงไหม้โรงงานกิ่งแก้ว ไฟไหม้โรงงานกิ่งแก้ว  
สำนักข่าวไทย THAMCOT · 10 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

ด่วน! เกิดเหตุไฟไหม้โรงงานภายใน 'นิคมอุตสาหกรรมลาดกระบัง' อาคารเริ่มมีการทรุดตัว

กรุงเทพธุรกิจ · 4 ชั่วโมงที่แล้ว

- ไฟไหม้โรงงาน ภายในนิคมลาดกระบัง เป็นโศกตึงเกินสิ้นคำประเภทน้ำหอม | 06-07-64 | ไทยรัฐนิวส์โชว์  
Thairath Online · 3 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

โควิด-19 : ศบค. เคาะฉีดวัคซีนเข็ม 3 ให้บุคลากรทางการแพทย์ แยมใช้ไฟเซอร์-แอสตราเซนเนกา

บีบีซีไทย · 7 ชั่วโมงที่แล้ว

- มติเคาะฉีด ไฟเซอร์ ให้บุคลากรทางการแพทย์ที่ได้ฉีดในแคว้น 2 เข็ม  
ข่าวสด · 11 ชั่วโมงที่แล้ว

ดูเรื่องราวจากทุกช่องทาง

สภาพอากาศท้องถิ่นของคุณ

มีพายุฝนฟ้าคะนอง

25°C

วันนี้ พ. พุธ. ศ. ส.

33°C 32°C 28°C 30°C 31°C  
24°C 24°C 24°C 24°C 24°C

C | F | K

weath.com

เกาะกระแสข่าว

ลดตม สาขานาน ไฟเซอร์ อดคีตย

สมรักษ์ คำสิงห์ โรงเรียน สมคิด จาคศิริทิพย์

ภูเก็ต ประเทศไทย รัตนัน คำดล ทัพเรือ

Cluster 1

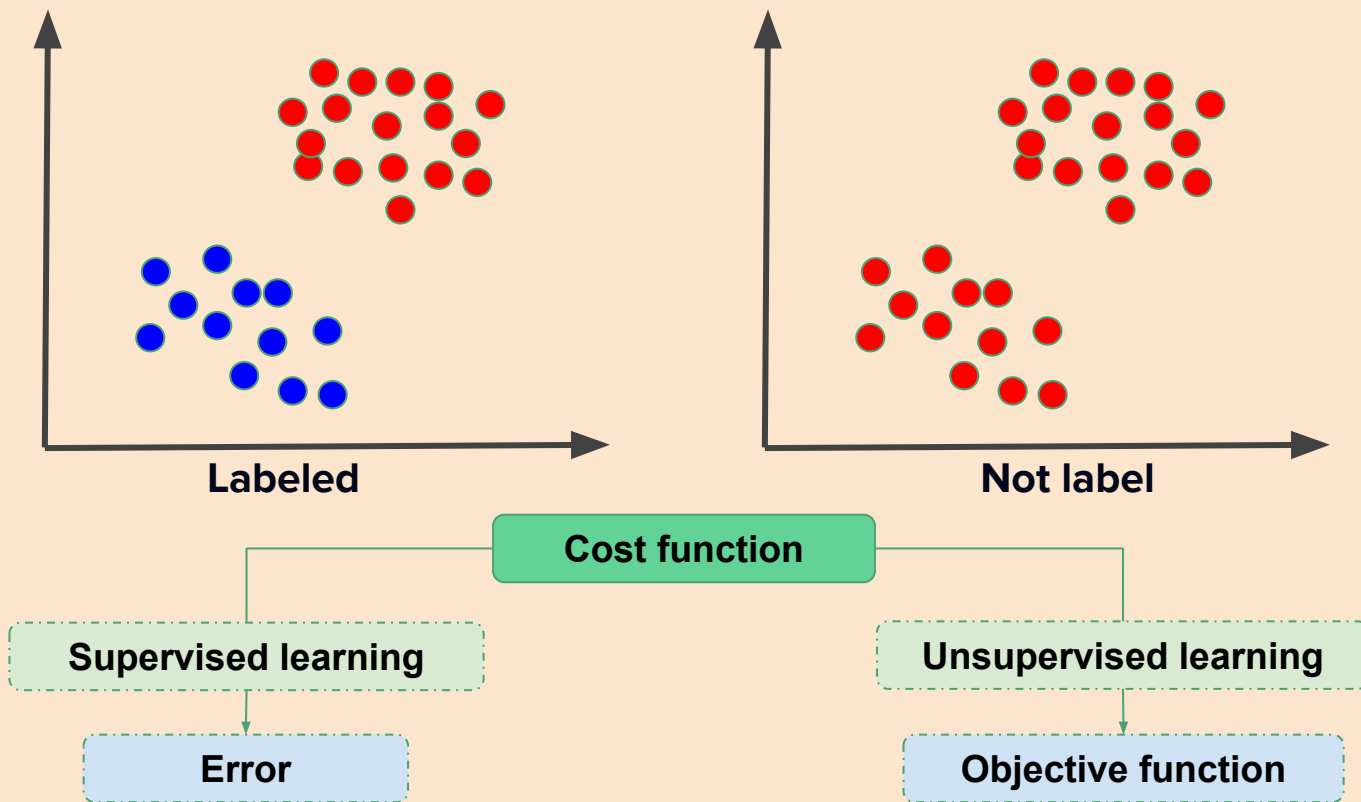
Cluster 2

Cluster 3

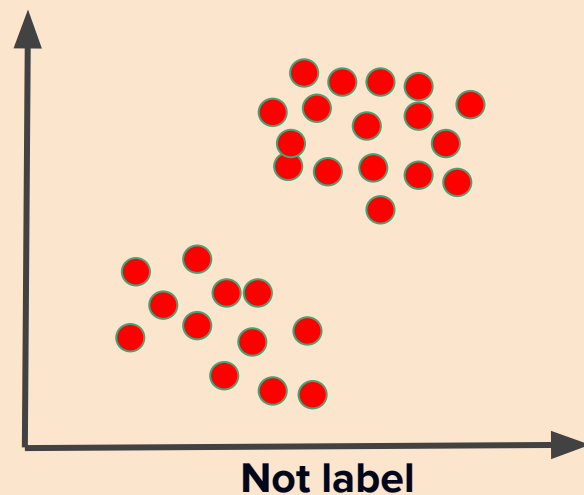
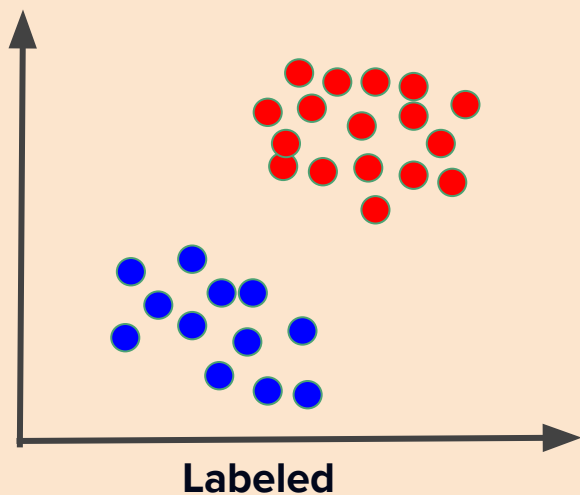
Impossible to label all news !!!



# Supervised learning vs Unsupervised learning



# Supervised learning vs Unsupervised learning

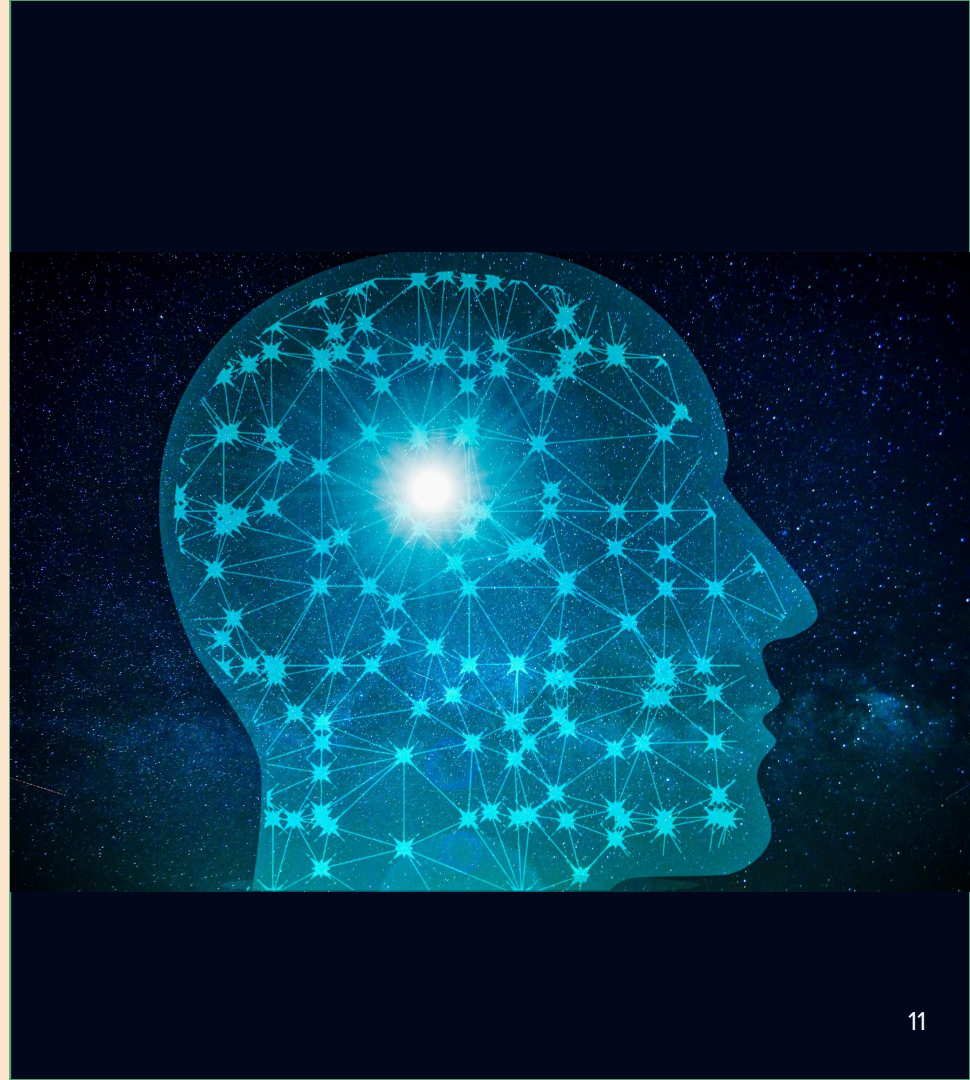


Training

Testing

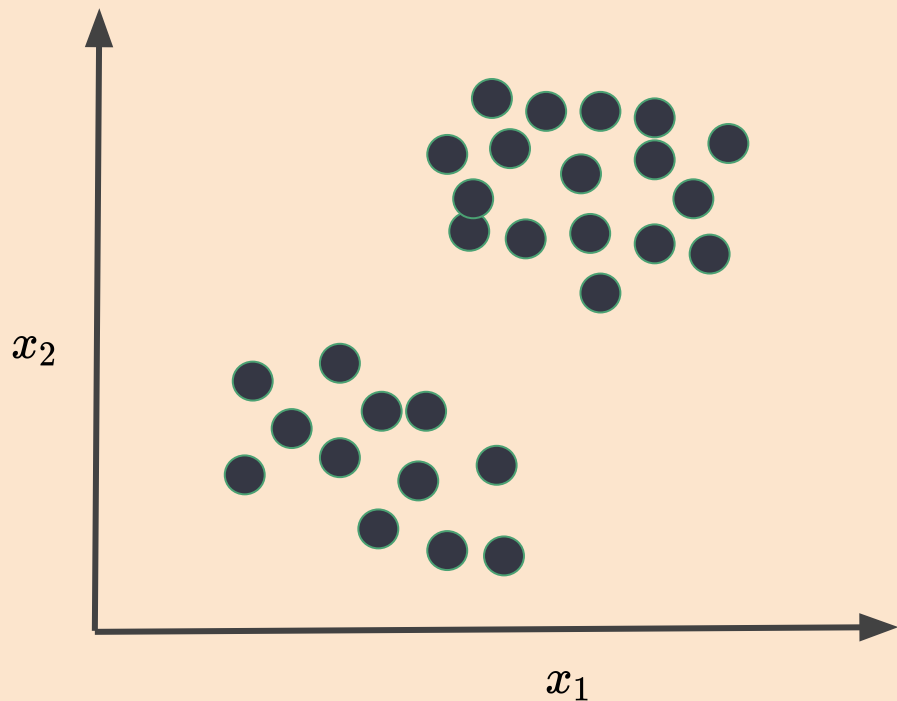
	X	Y	$x'$	$y'$	Predict	Cost function
Supervised	Yes	Yes	Yes	No	$y'$	Error
Unsupervised	Yes	No	Yes	No	$Y, y'$	Objective function

# K-Means Clustering



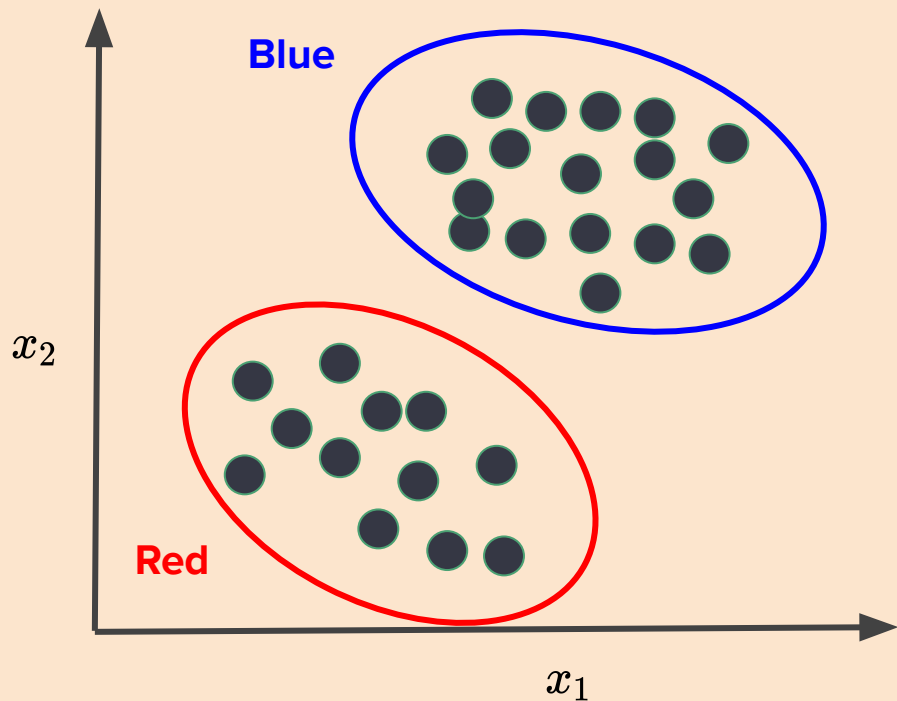
# K-means clustering

- ❖ 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

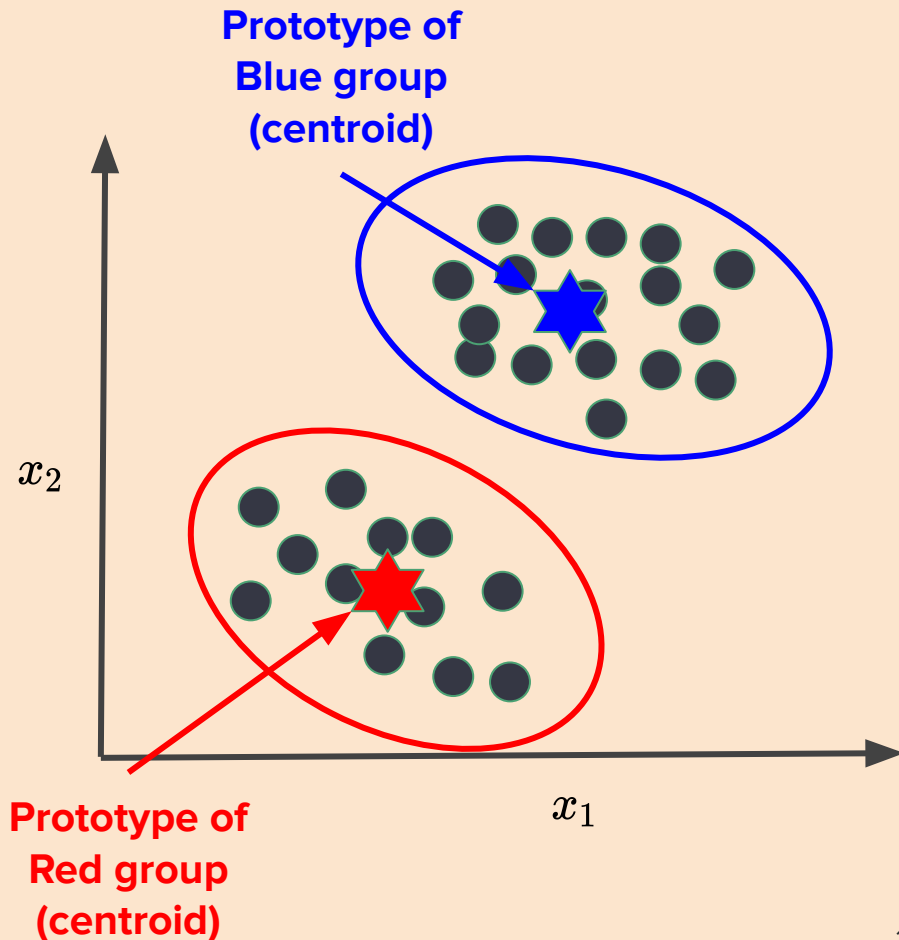
- ❖ K-means clustering is the most popular clustering algorithm.
- ❖ K-means clustering aims to partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (centroid).





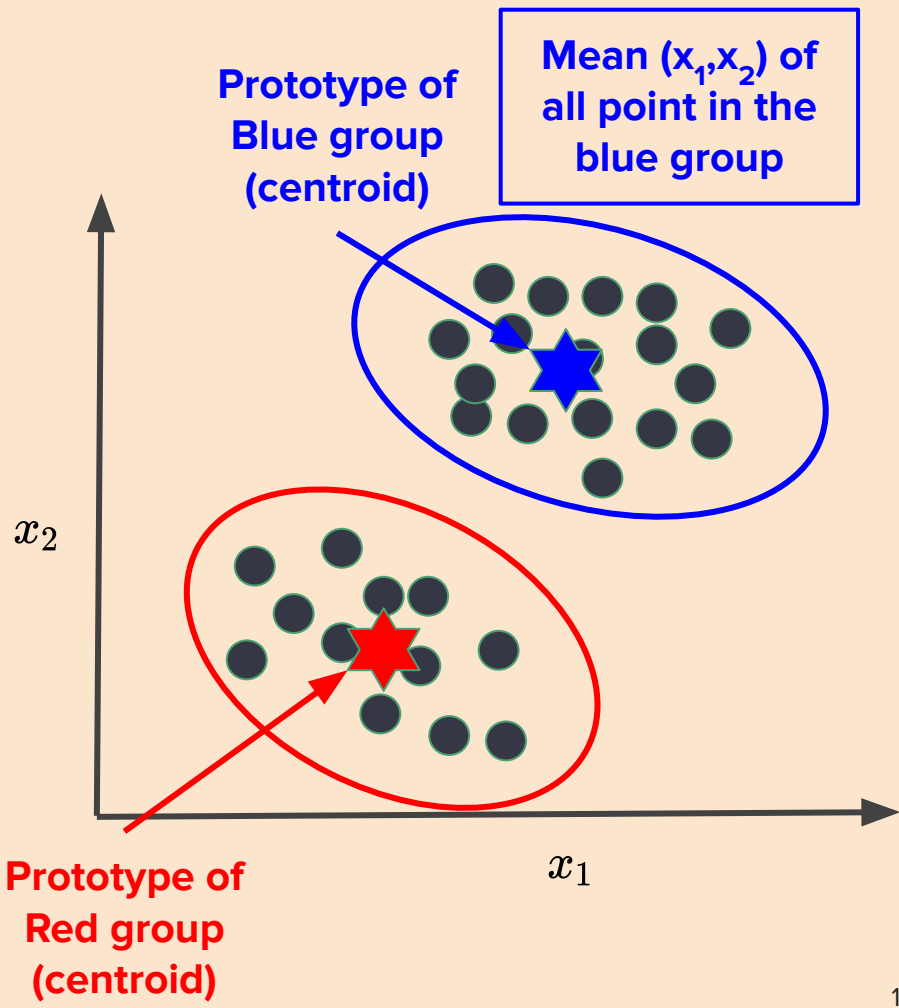
# K-means clustering

- ❖ 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).



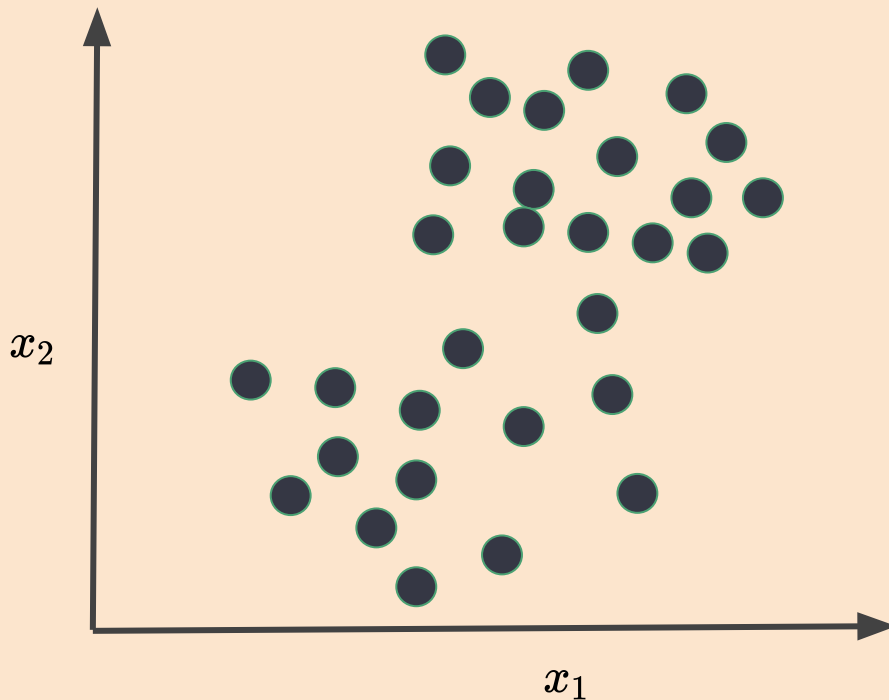
# K-means clustering

- ❖ 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).



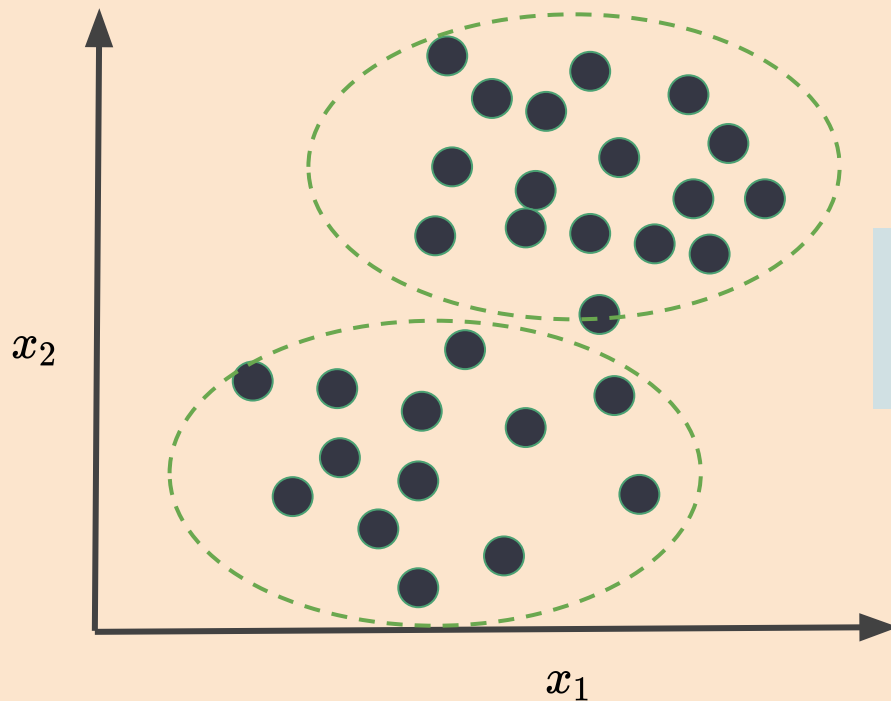
# K-means algorithm

- ❖ **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

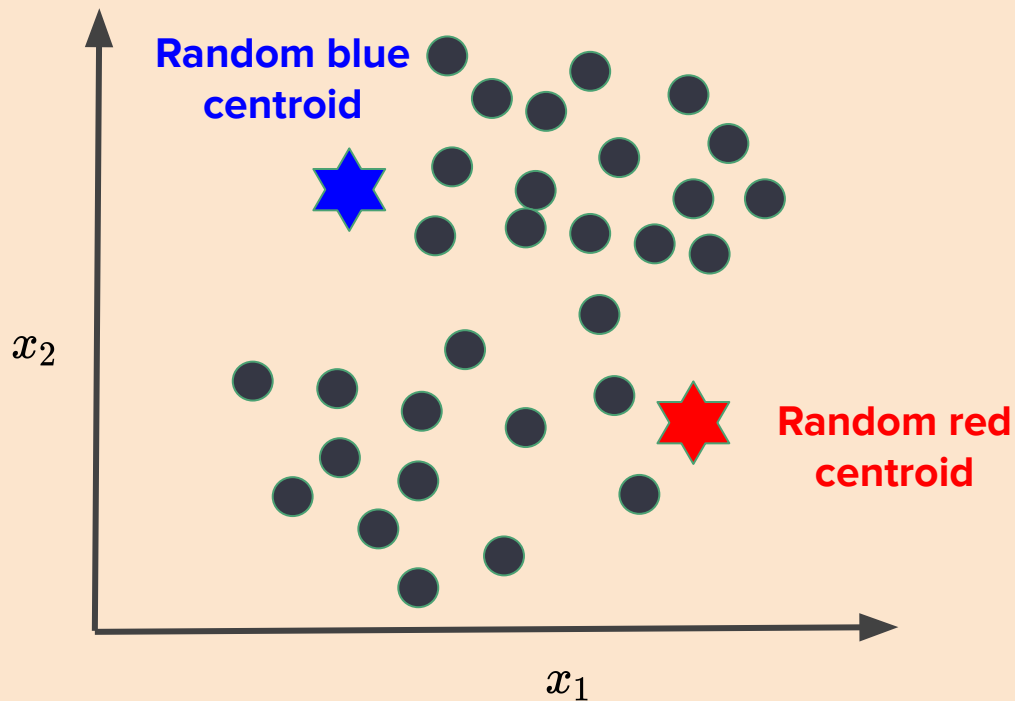
- ❖ **First step: Define  $k$**  ; we have a bunch of unlabeled data points. We decide that we are going to find **two clusters** in this data



**Let's set  $k = 2$   
by humans**

# K-means algorithm

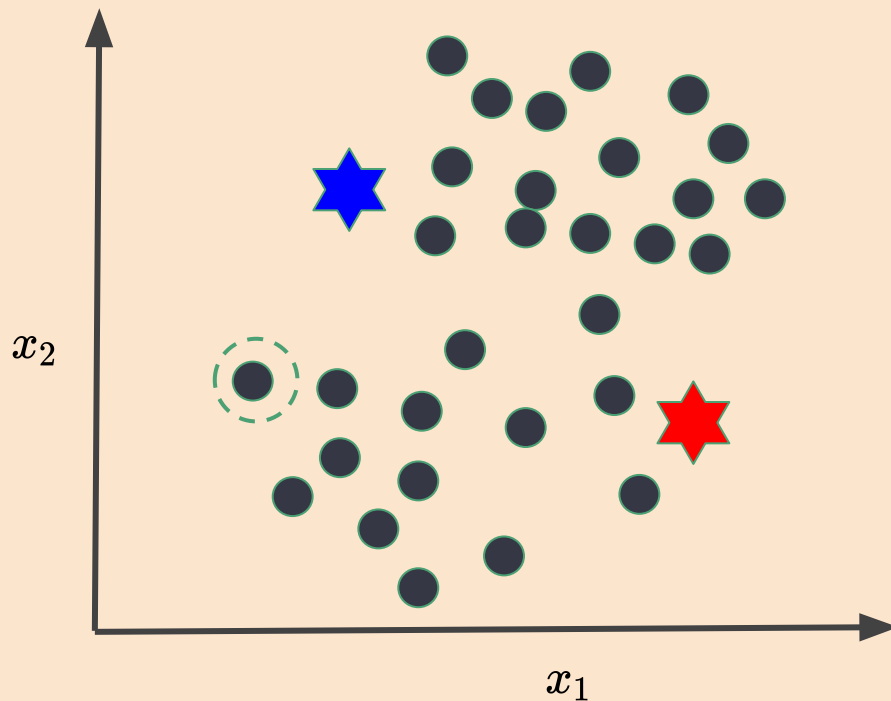
- ❖ **Second step: Random centroids** ; This step is to pick two random locations to be our cluster centroids.





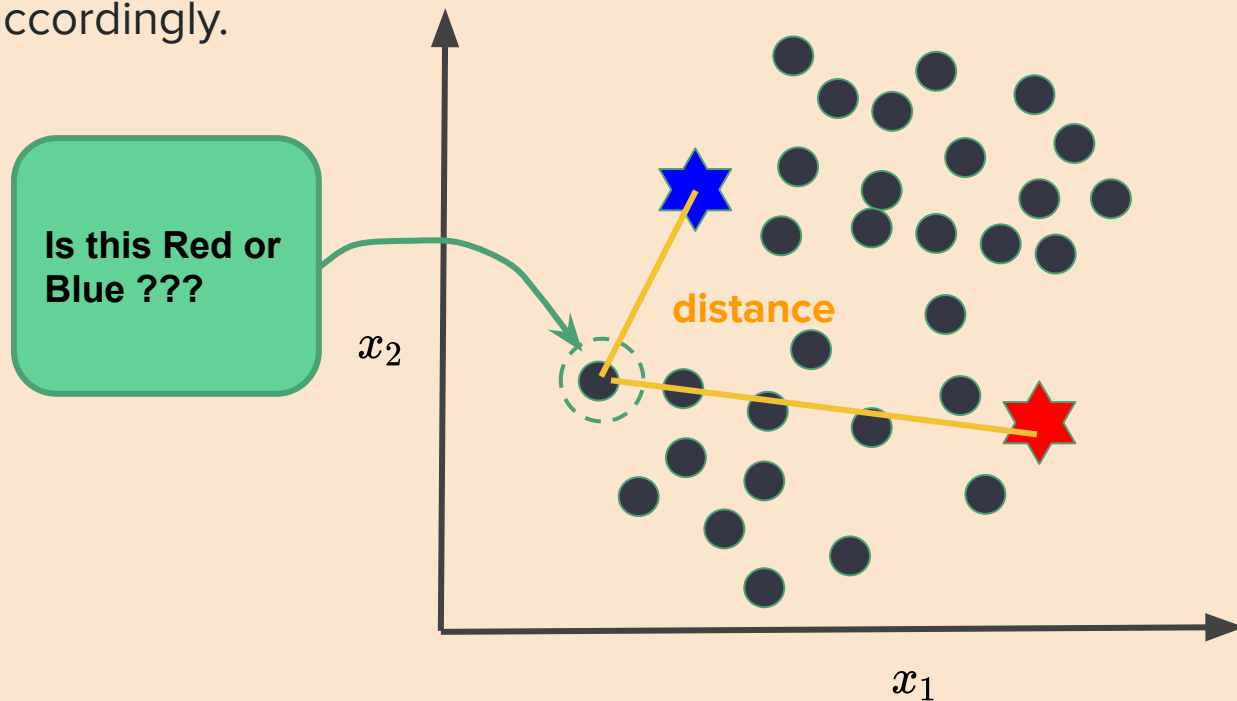
# K-means algorithm

- ❖ **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.



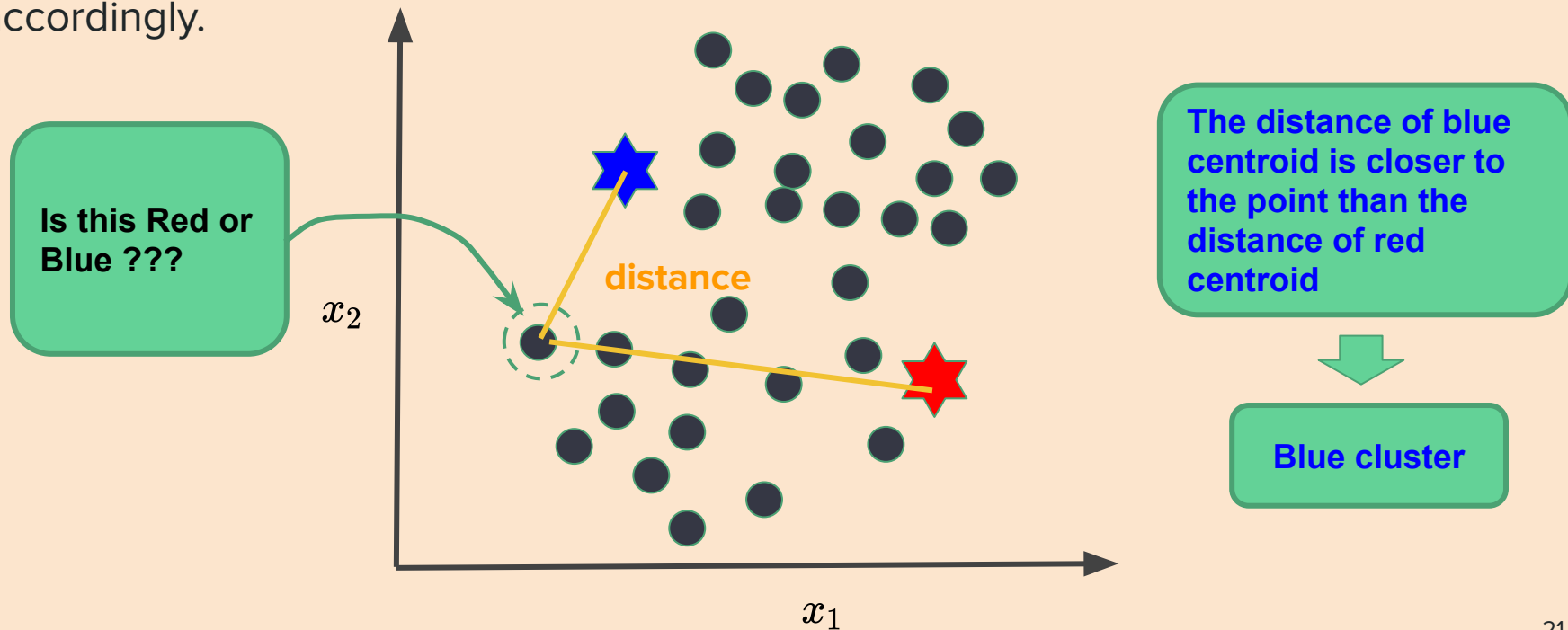
# K-means algorithm

- ❖ **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.



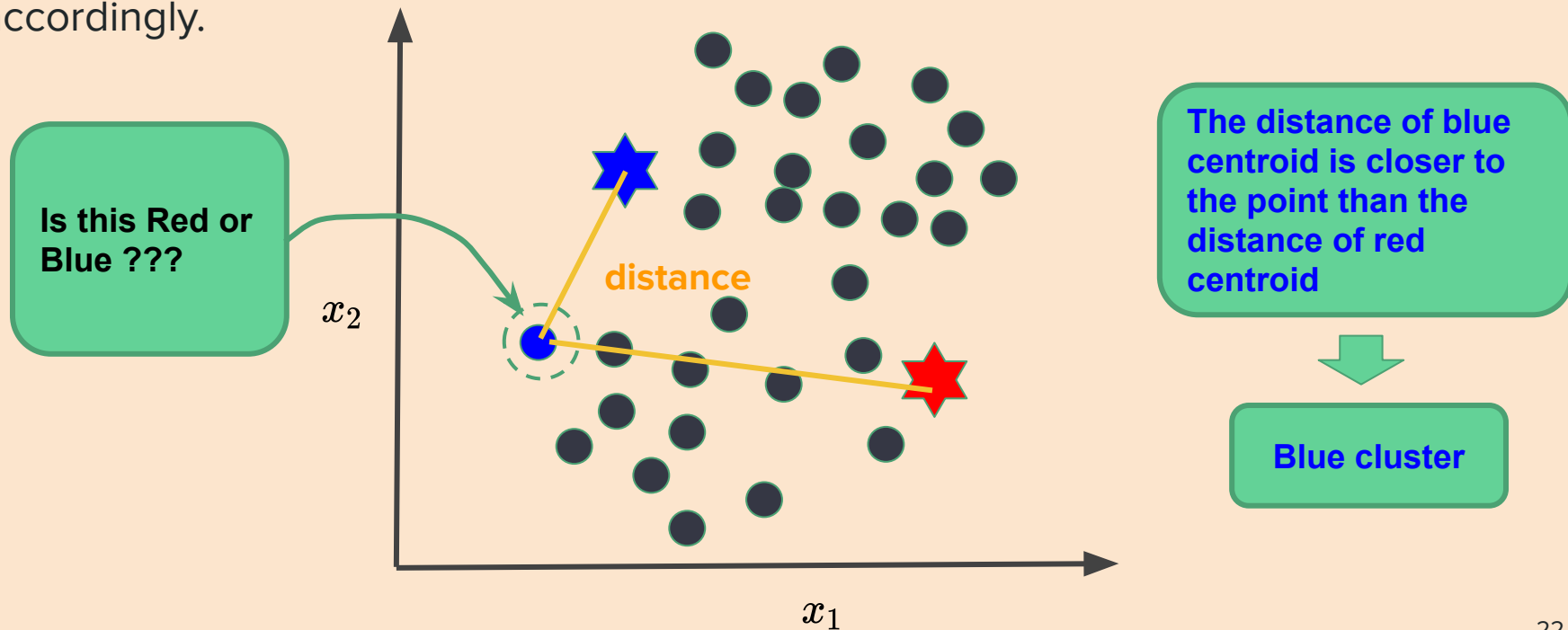
# K-means algorithm

- ❖ **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.



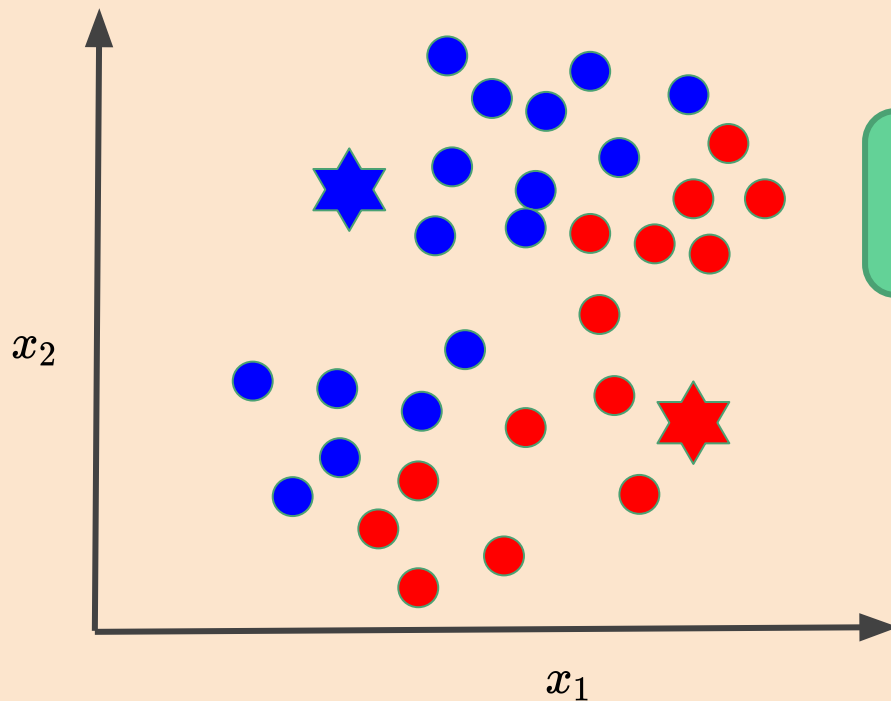
# K-means algorithm

- ❖ **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.

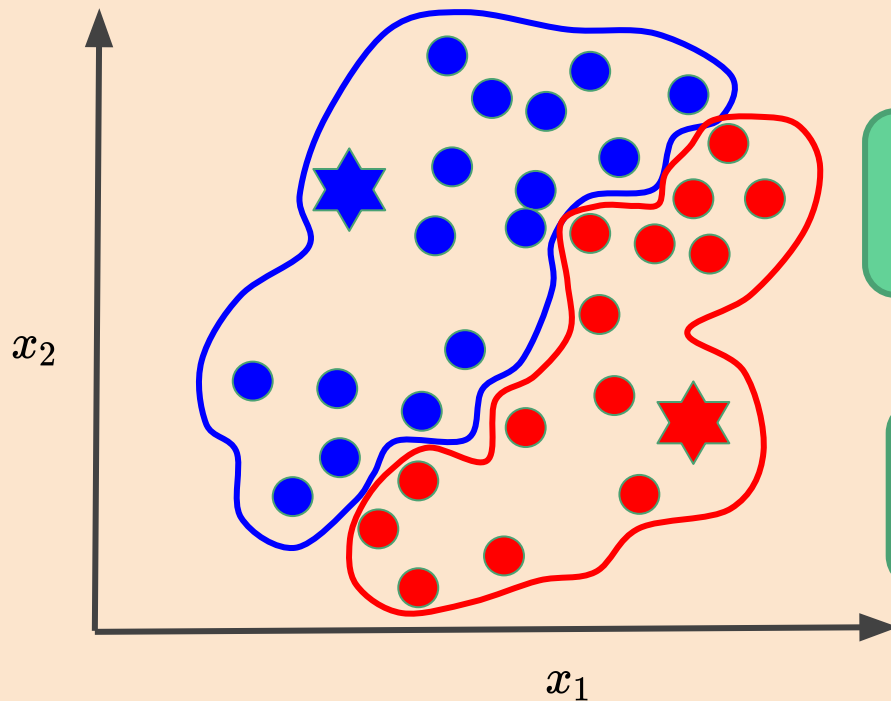


To this step in every sample



# 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.

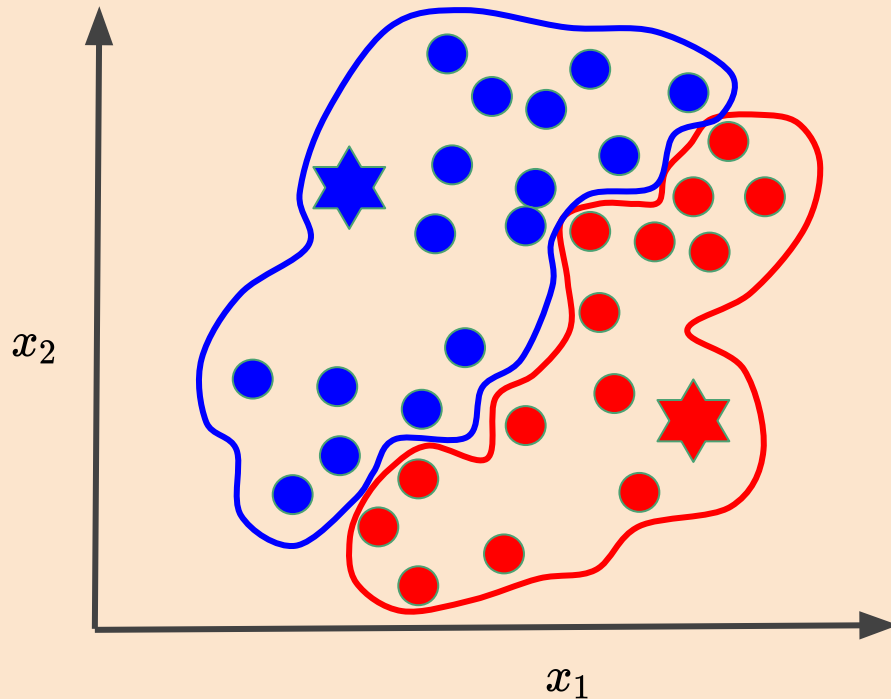


To this step in every sample

This is the bad centroids!!!

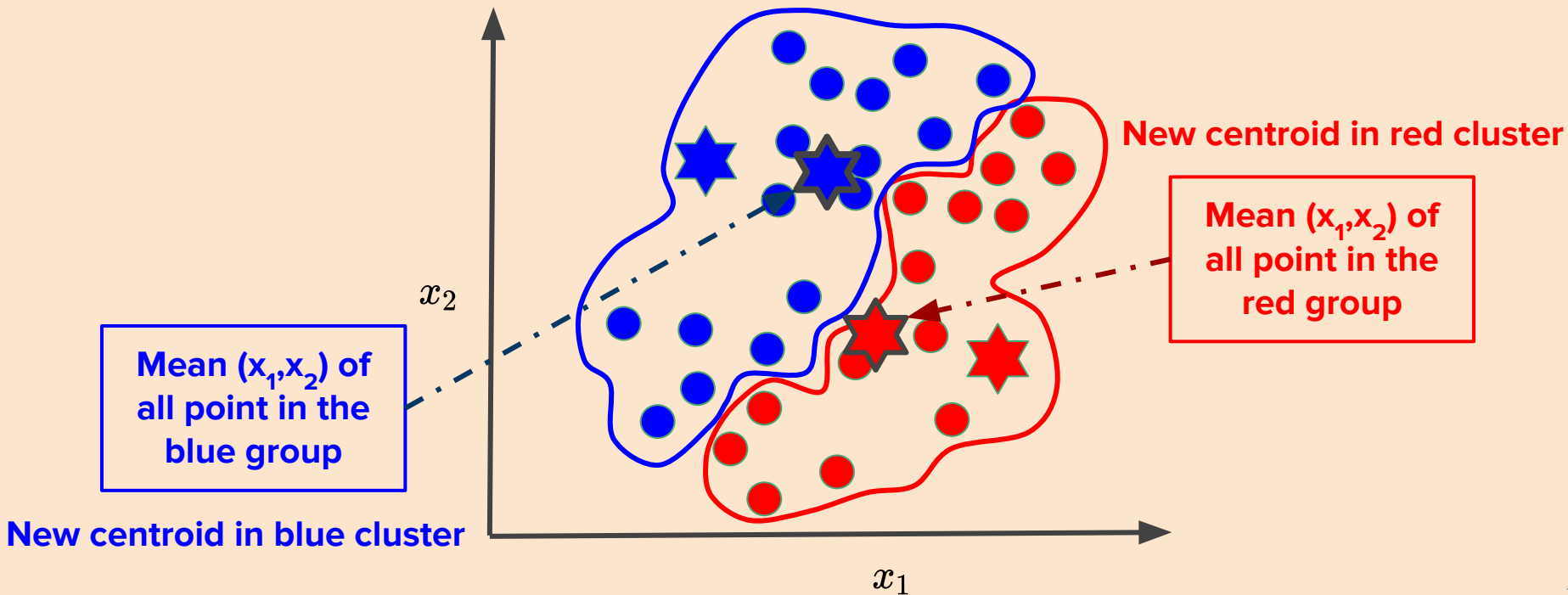
# K-means algorithm

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



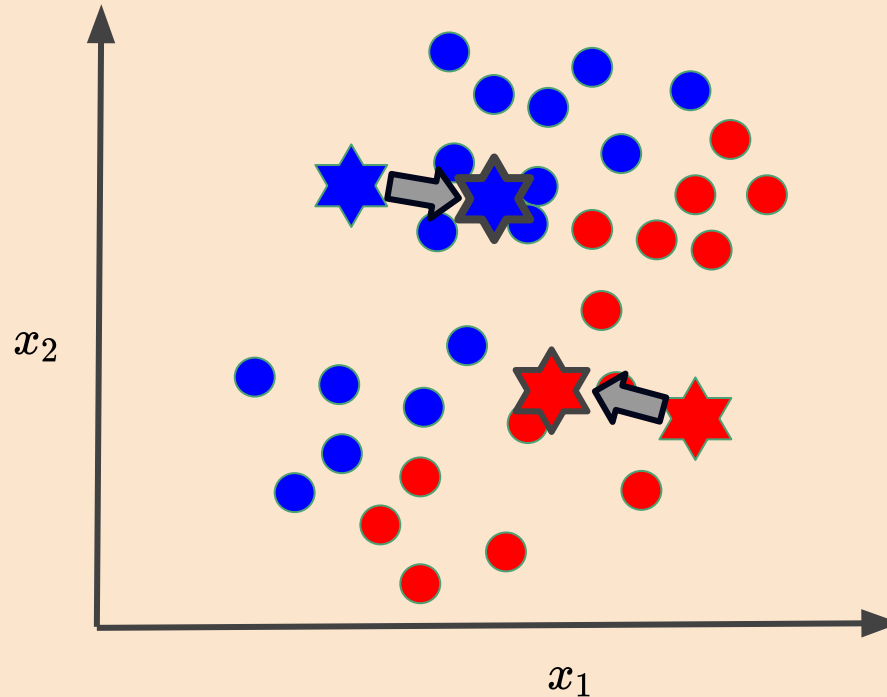
# K-means algorithm

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



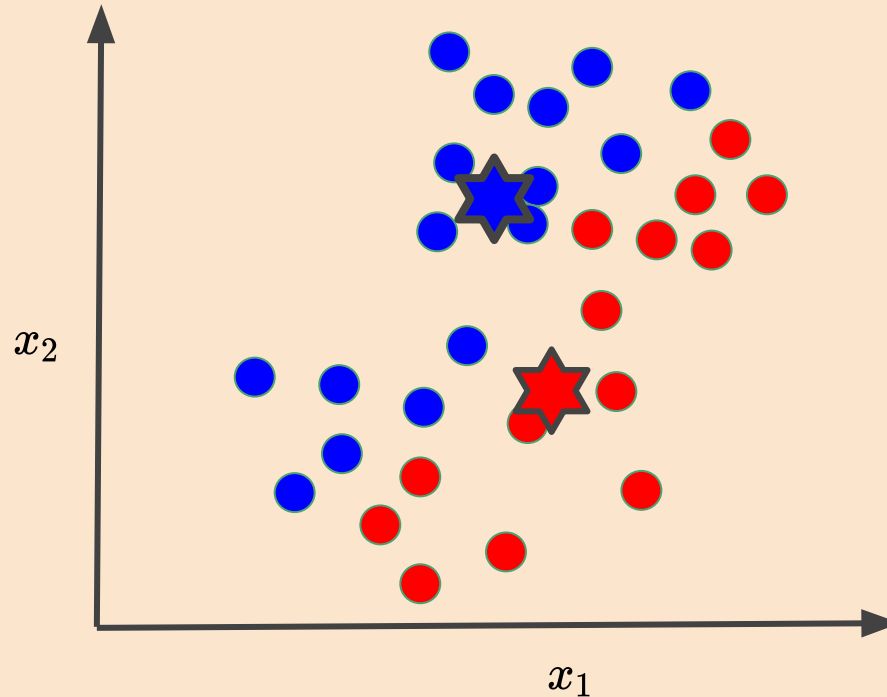
# K-means algorithm

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



# K-means algorithm

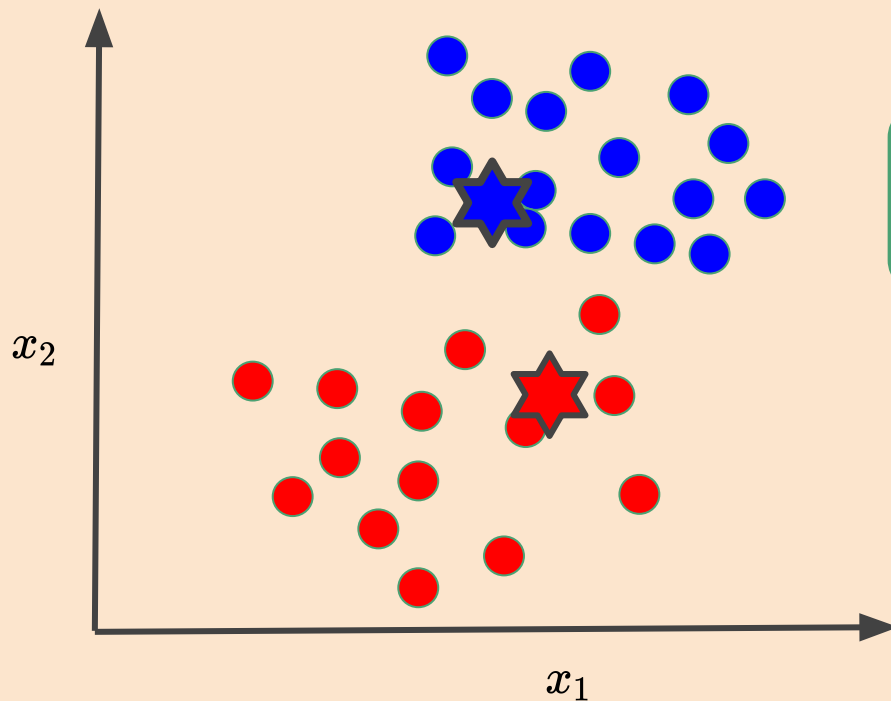
- ❖ **Fourth step: Centroid movement** ; This step is to move the red and blue centroids to the means of clusters.





# K-means algorithm

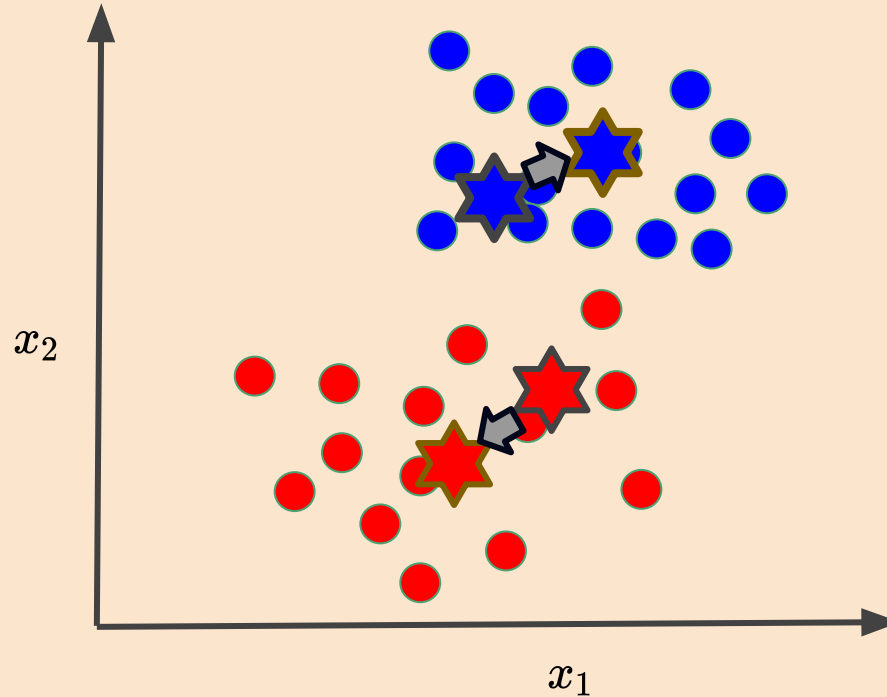
- ❖ Fifth step: Repeat to third step and fourth step



**Third step: Cluster assignment**

# K-means algorithm

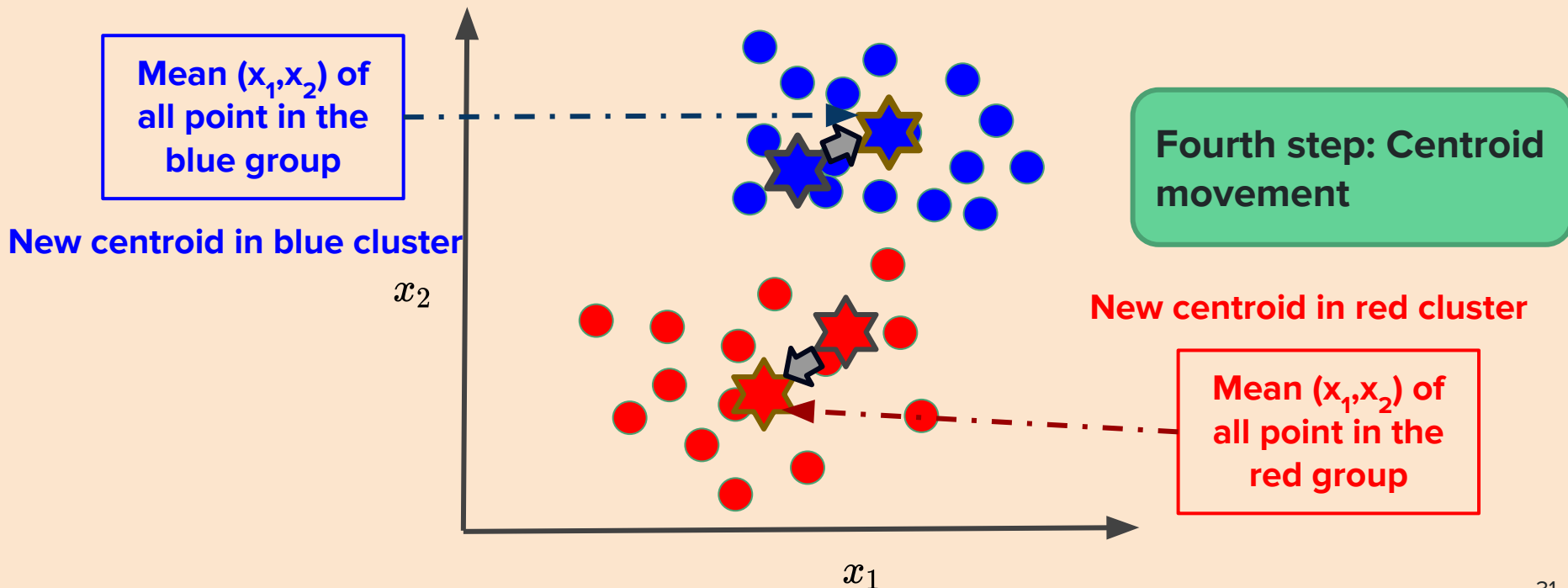
- ❖ Fifth step: Repeat to third step and fourth step



**Fourth step: Centroid movement**

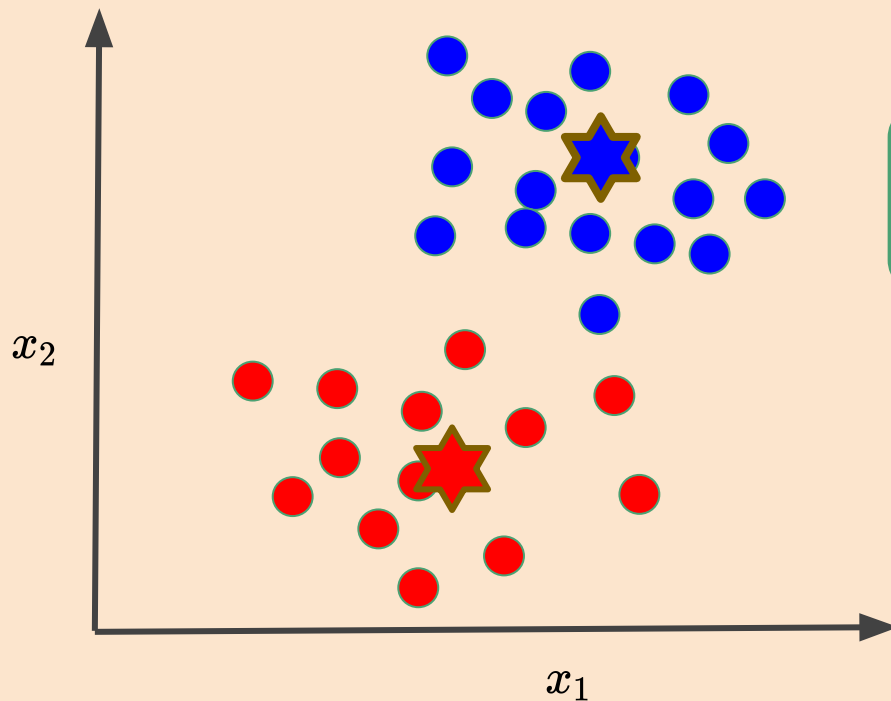
# K-means algorithm

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



# K-means algorithm

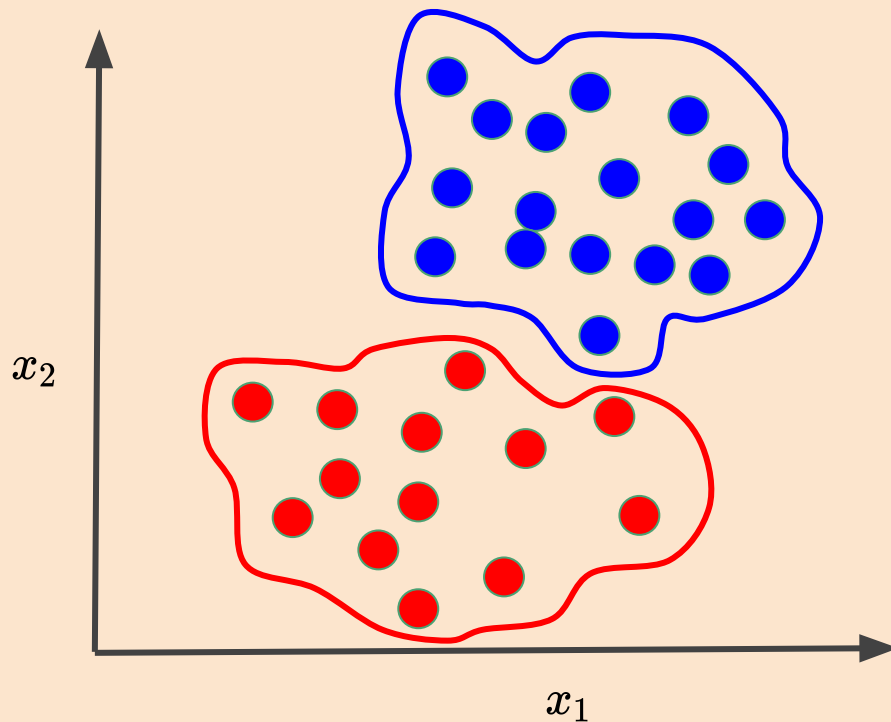
- ❖ Fifth step: Repeat to third step and fourth step



**Third step: Cluster assignment**

# K-means algorithm

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

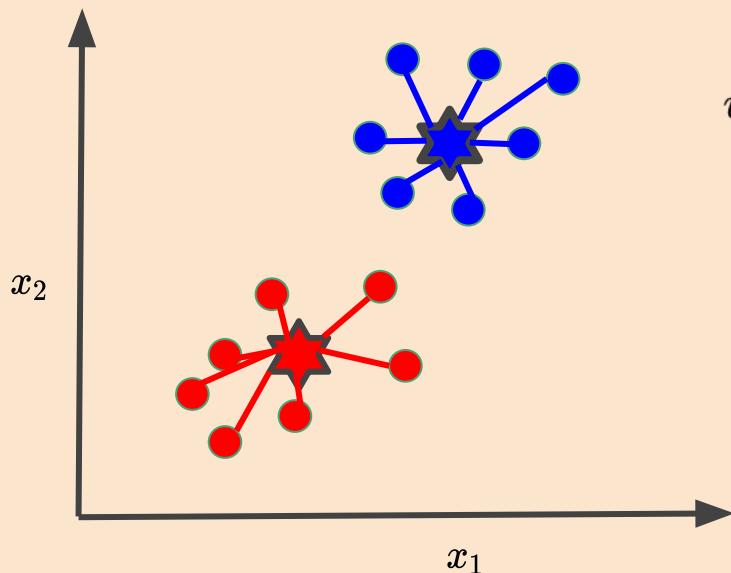


# Objective function of K-means algorithm

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

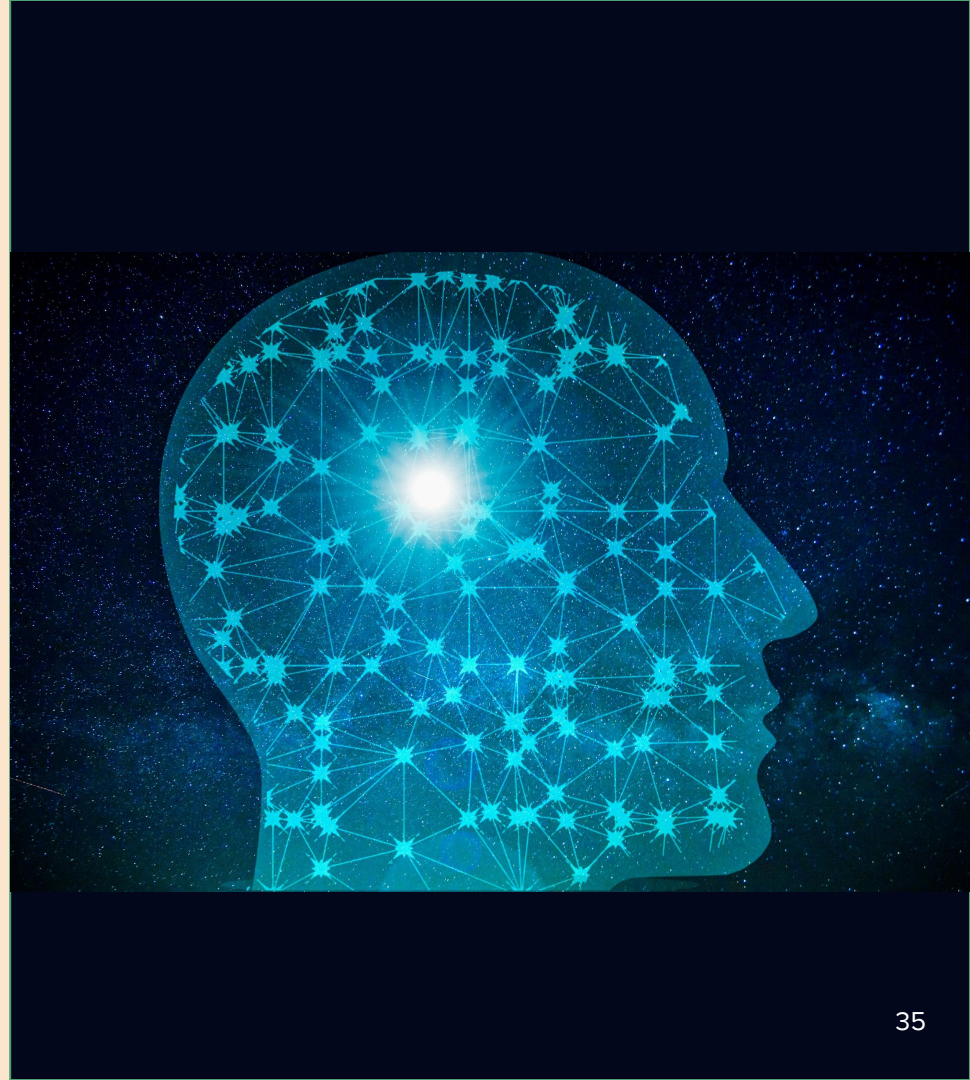
*Objective function : find  $\bar{x}_{red}, \bar{x}_{blue}$*

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



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

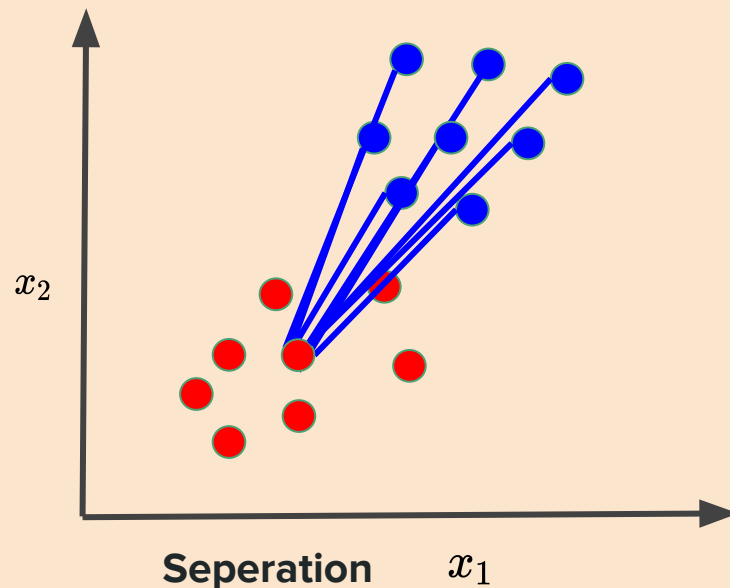
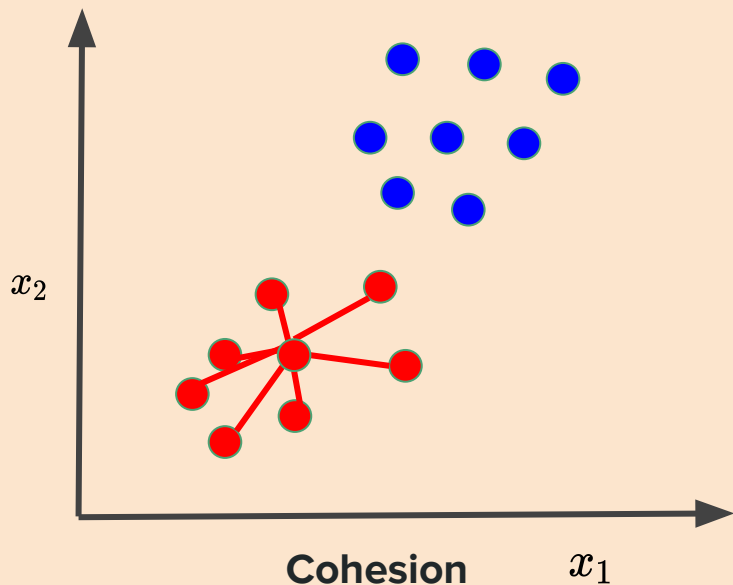
# K-means clustering performance





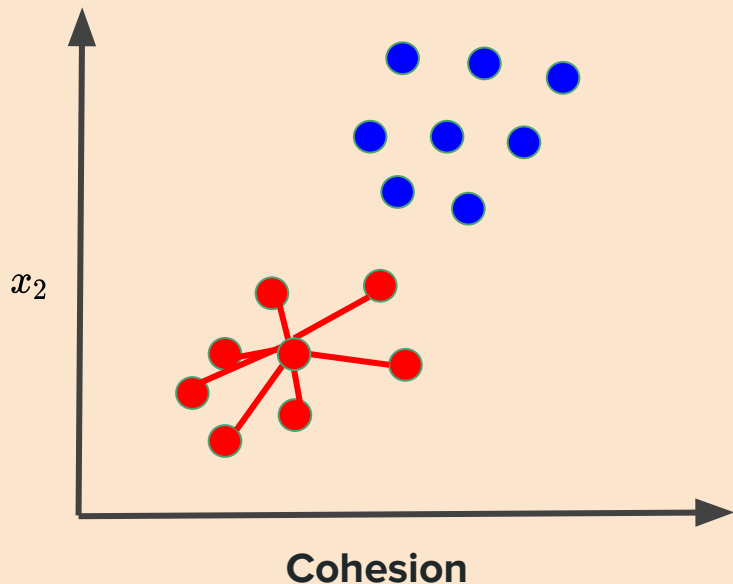
# Clustering performance

- ❖ 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.



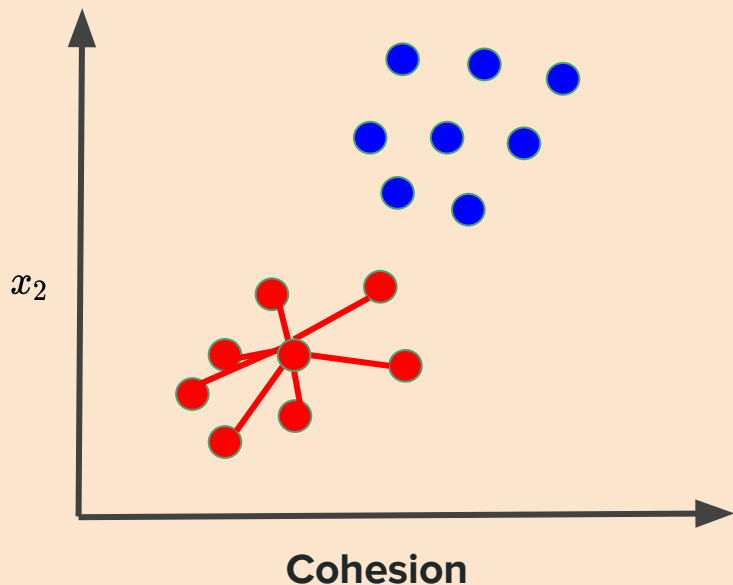
**Cohesion:** Distance of one data point to all data point in the sample group. (Within-Cluster-Sum-of-Squares (WCSS))



**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))



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

Where,

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

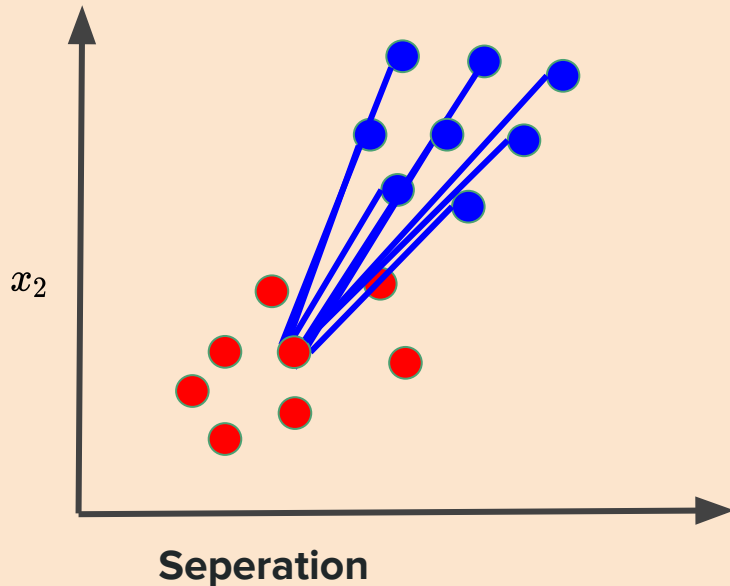
# Clustering performance

- ❖ 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**



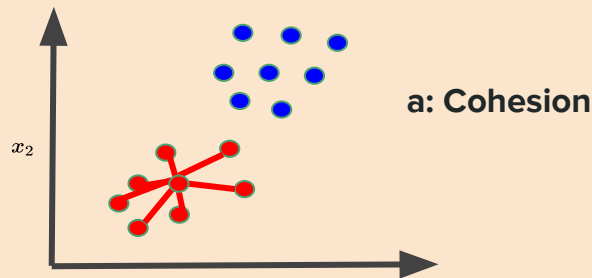
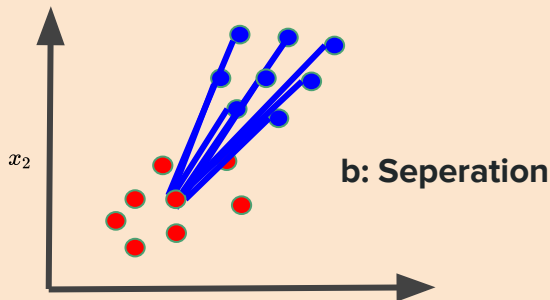
# Silhouette coefficient

- ❖ 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.

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

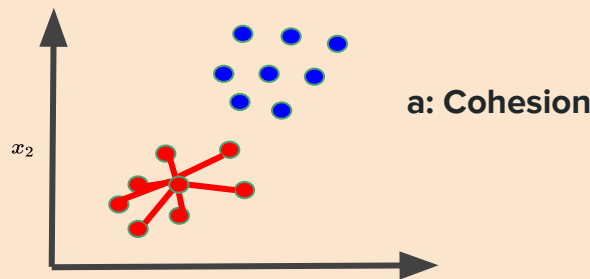
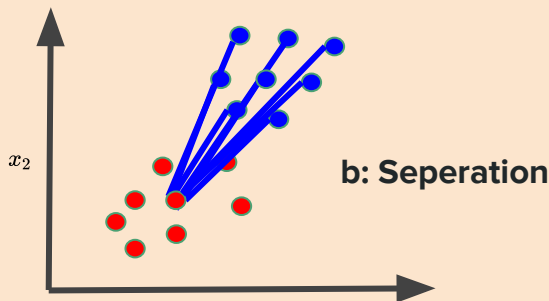
If  $b > a$



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

$a$  : 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.

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

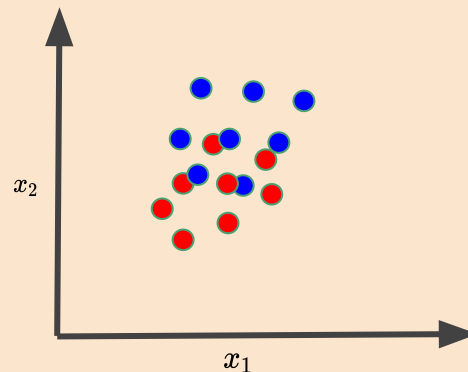
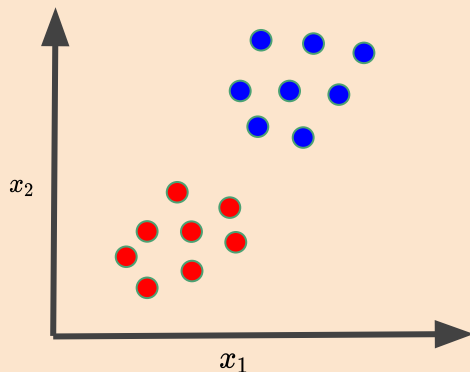
If  $b > a$



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

If  $b \gg a$  :  $S$  is close to 1.

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



# Silhouette coefficient

- ❖ Silhouette coefficient is defined by two separated scores.

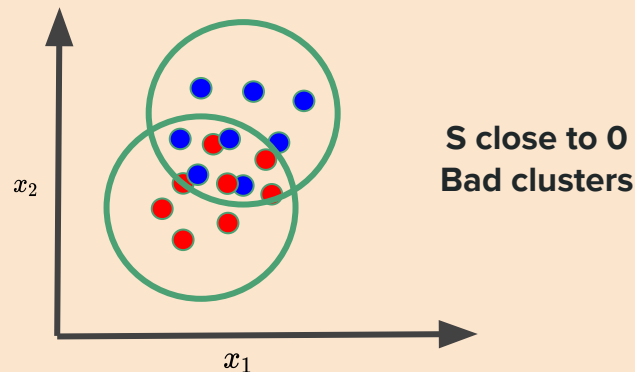
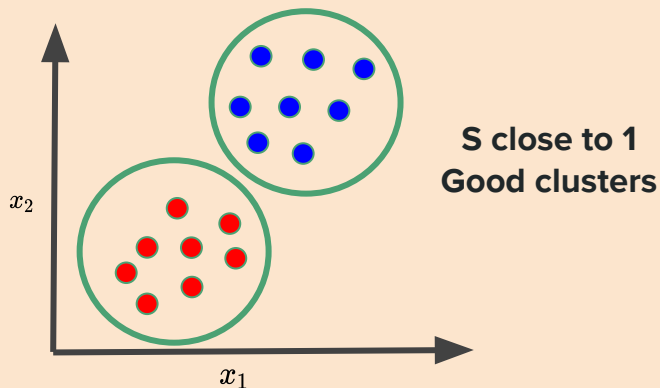
$$s = \frac{b-a}{\max(b,a)}$$

If  $b > a$   $\longrightarrow$

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

If  $b \gg a$  :  $S$  is close to 1.

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

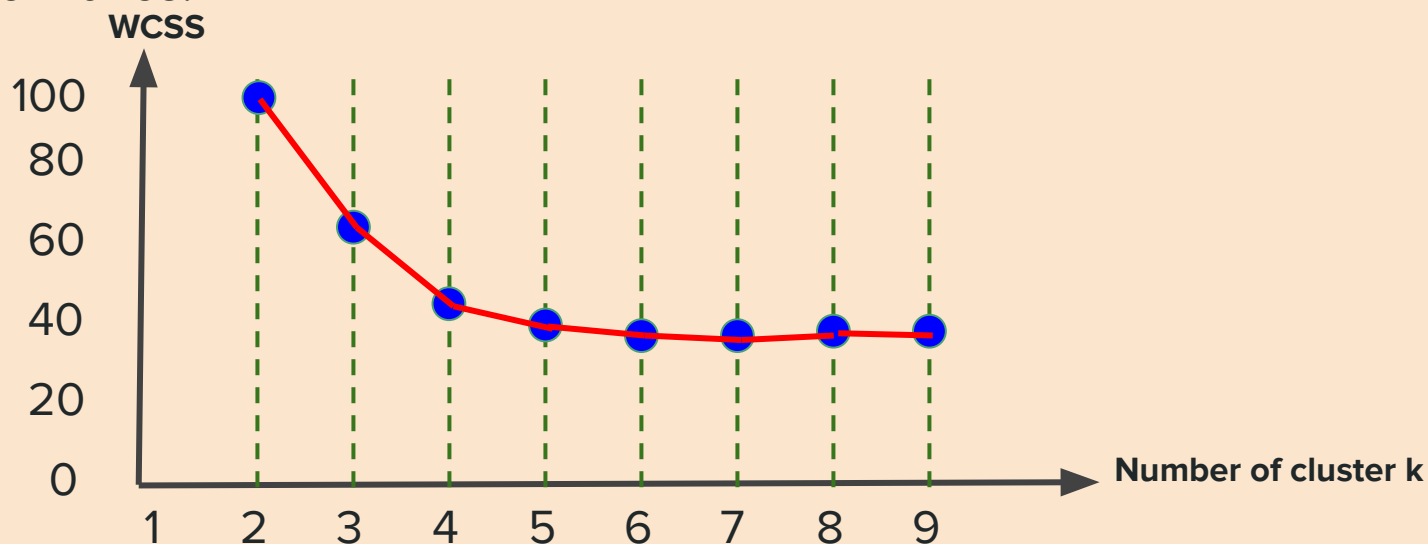




# Choosing the right K

## ❖ Using **elbow method**

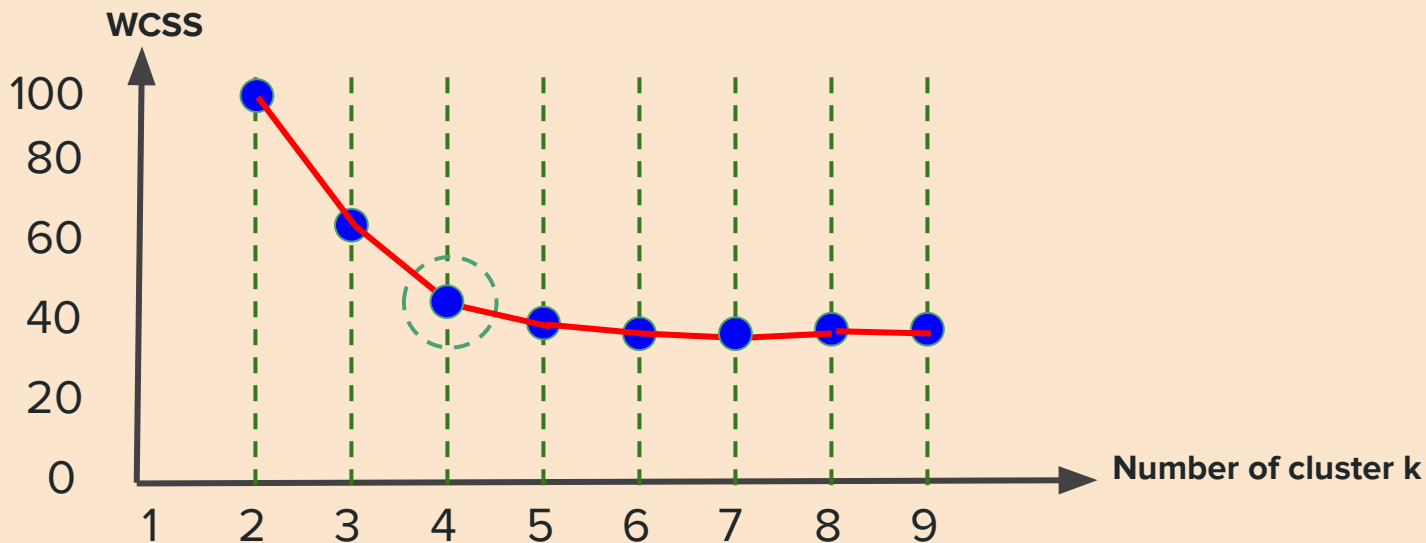
- You compute performance metrics such as Silhouette coefficient or WCSS, while varying  $k$ .
- Pick the  $k$  at the elbow point. At this point, more clusters do not necessarily mean higher performance.



# Choosing the right K

## ❖ Using **elbow method**

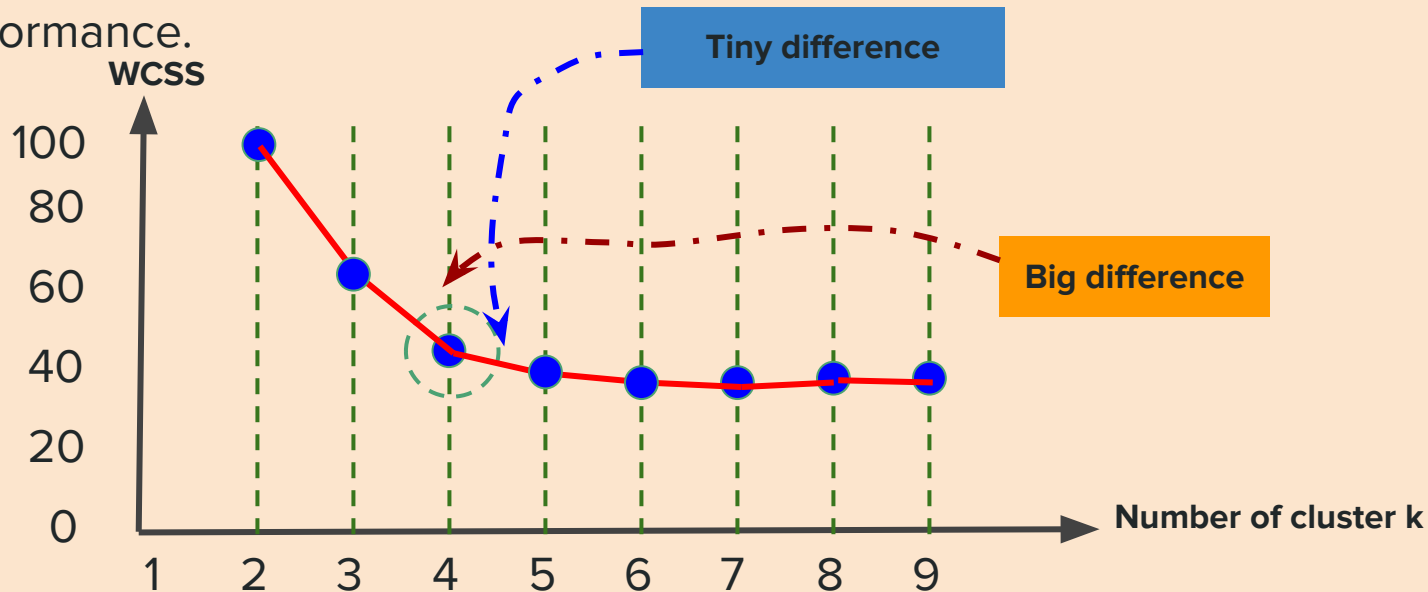
- You compute performance metrics such as Silhouette coefficient or WCSS, while varying  $k$ .
- Pick the  $k$  at the elbow point. At this point, more clusters do not necessarily mean higher performance.



# Choosing the right K

## ❖ Using **elbow method**

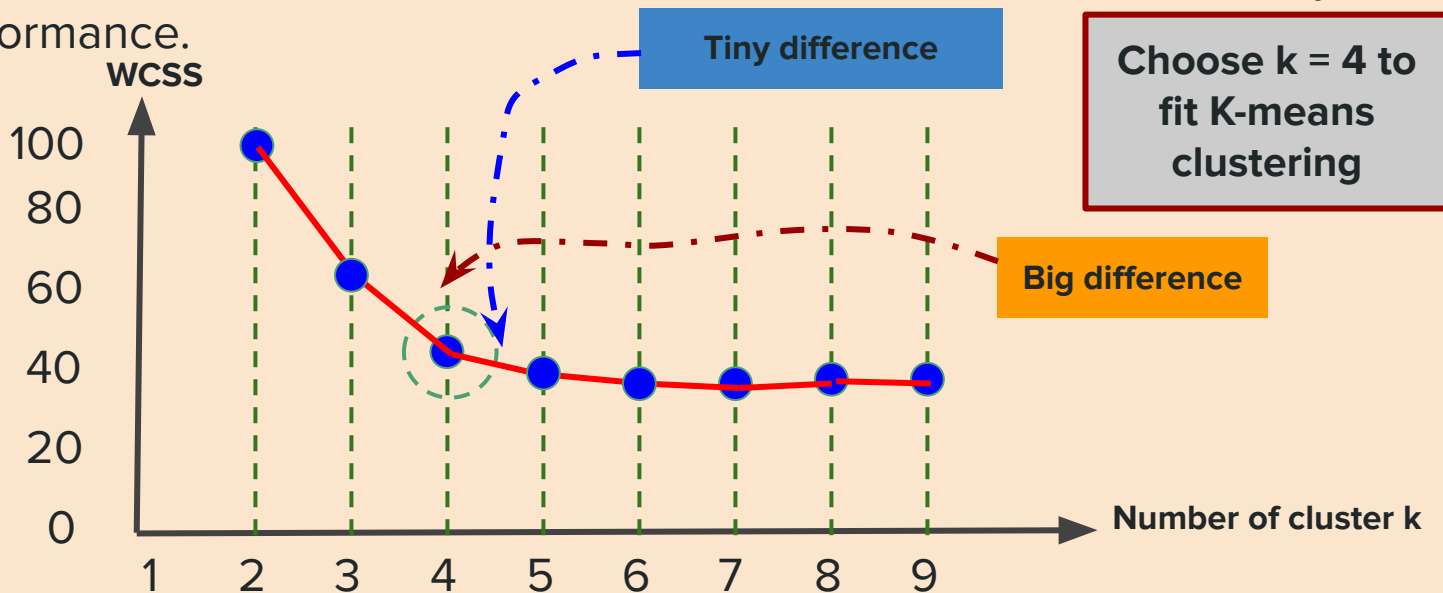
- You compute performance metrics such as Silhouette coefficient or WCSS, while varying  $k$ .
- Pick the  $k$  at the elbow point. At this point, more clusters do not necessarily mean higher performance.



# 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

