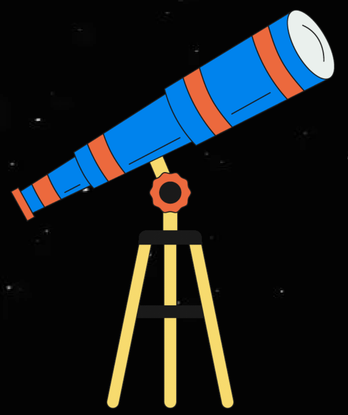
unsupervised learning

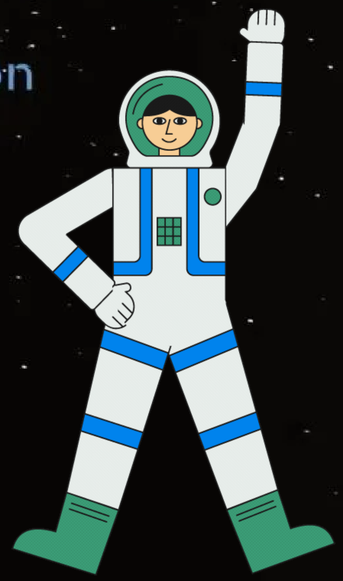
K-means clustering algorithm

TUNSA

Tunisian Space Association

الجمعية التونسية للفضاء

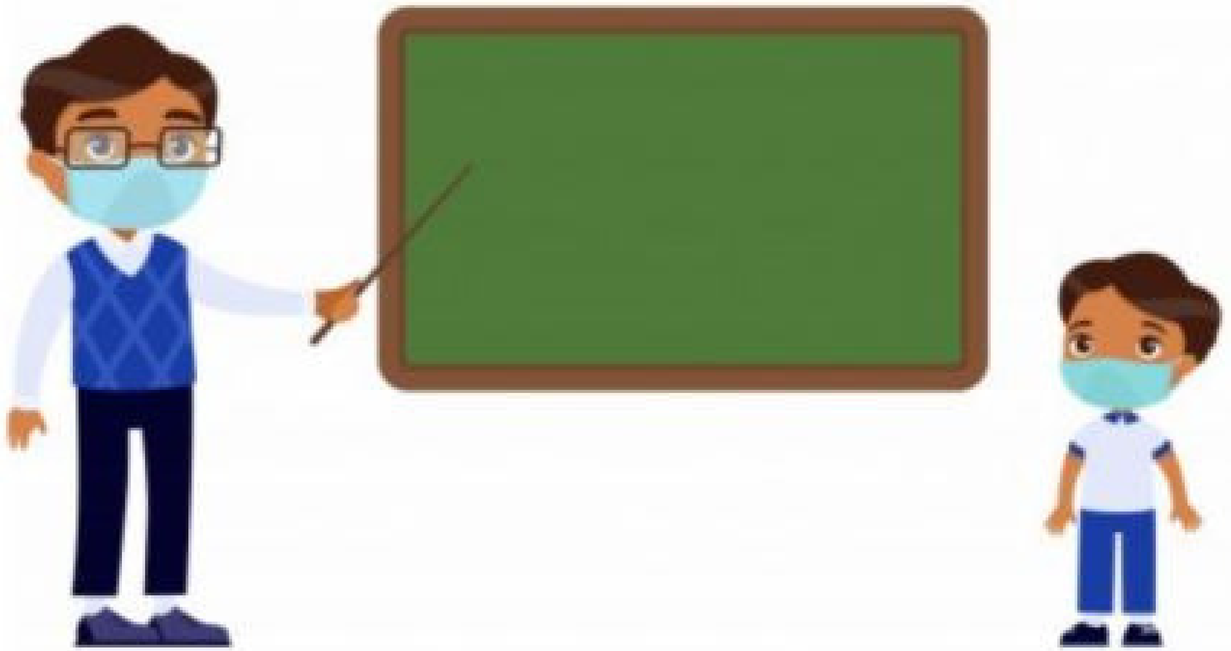




UNSUPERVISED LEARNING

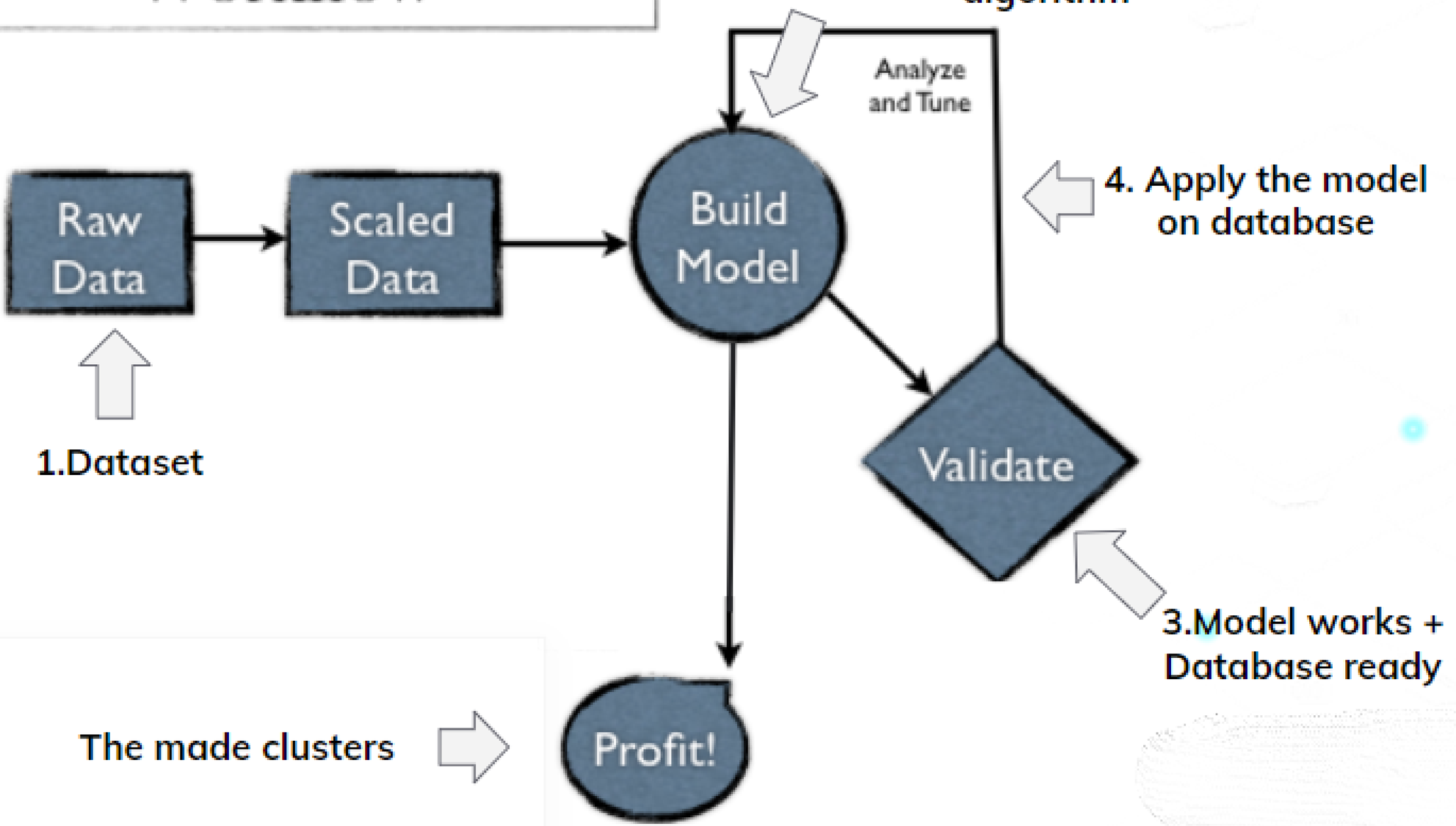
- Infer a function to describe/reveal hidden structure from unlabeled data.
- Unlike Supervised learning no supervision signal to evaluate a potential solution.
- Examples: K-means / Agglomerative Hierarchical

Supervised

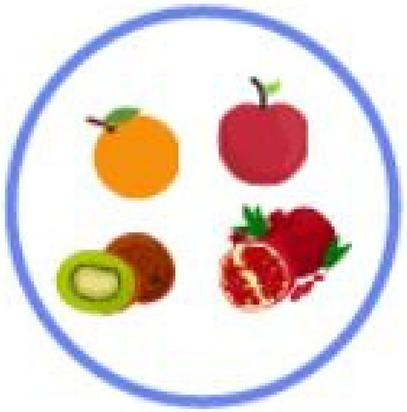


Unsupervised

Unsupervised Learning Workflow



Unsupervised Learning



Cluster 2

Cluster 1

Cluster 3

GOOGLE NEWS

Google Actualités

Rechercher des sujets, des lieux et des sources

- Pour vous
- Favoris
- Recherches enregistrées
- COVID-19
- France
- International
- Vos actualités locales
- Économie
- Sciences et technologies
- Divertissement
- Sports
- Santé

Comment les articles sont-ils classés ?

À la une

Voir plus : À la une

Actualités COVID-19 : Consultez les dernières actualités sur le coronavirus

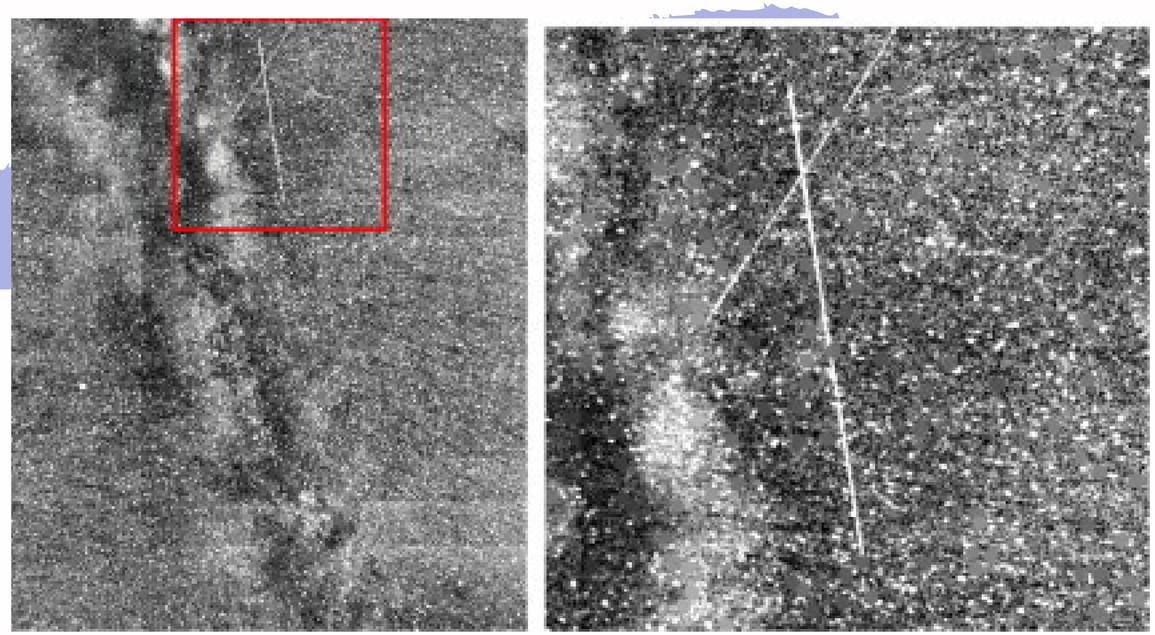
Covid-19 : regain de mobilisation contre le pass sanitaire et le projet de pass vaccinal, avec 105 000 mani...

Franceinfo · Il y a 1 heure

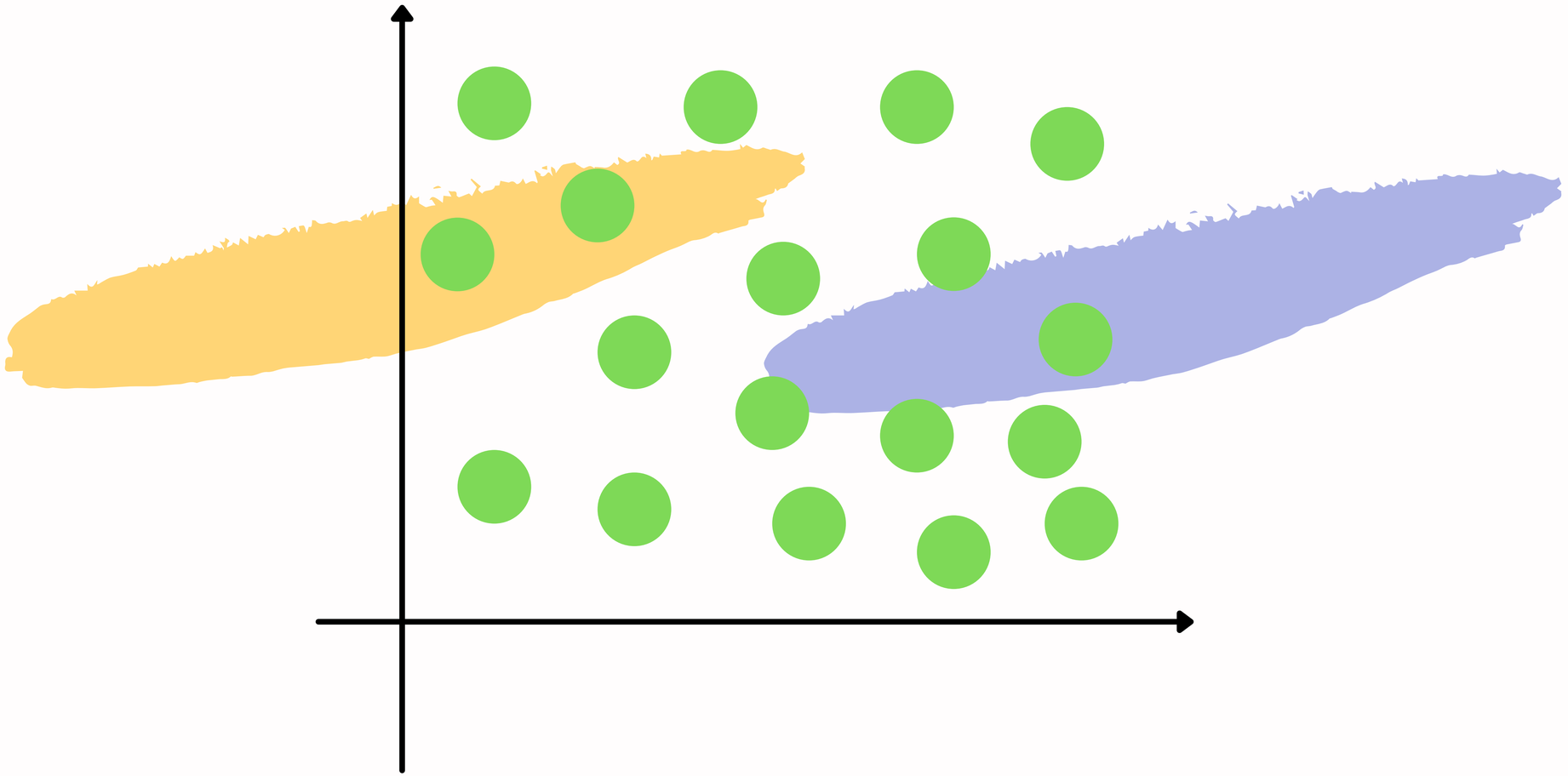
- «Ca n'a rien de sanitaire» : à Paris, des milliers de personnes contre le passe vaccinal
  Le Parisien · Il y a 4 heures
- DIRECT - Coronavirus : plus de 100.000 manifestants ont défilé contre le pass vaccinal dans plusieurs villes de France
  CNEWS · Il y a 15 heures
- Pass vaccinal: Une mobilisation en demi teinte des anti-pass après les propos de Macron
  LeHuffPost · Il y a 5 heures
- TÉMOIGNAGE. "Je ne me sens plus citoyenne", Margaux manifeste contre le pass vaccinal et nous explique pourquoi

Social networking analysis

unsupervised learning can help astronomers with automatic objects detection



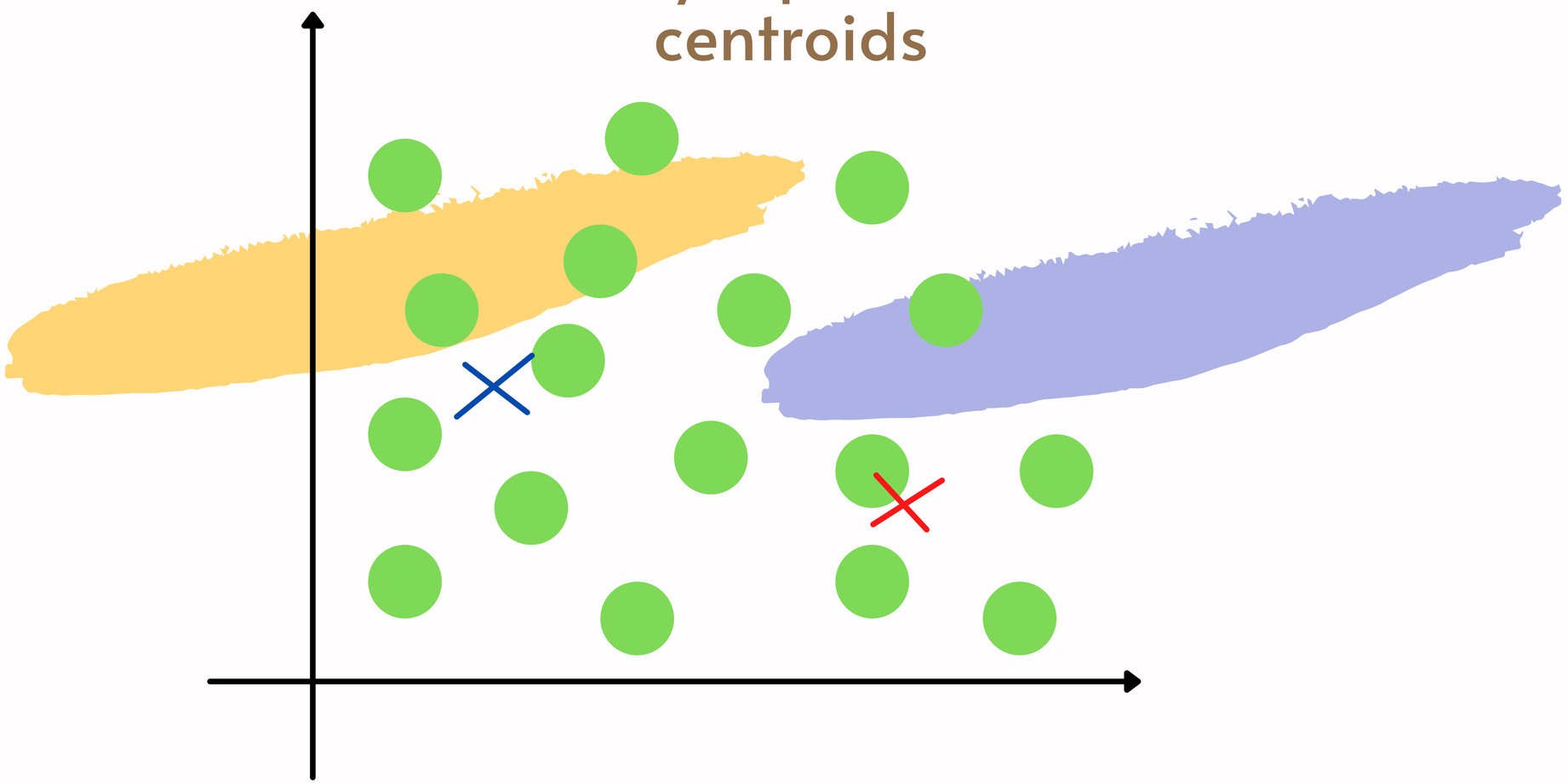
K-means clustering algorithm



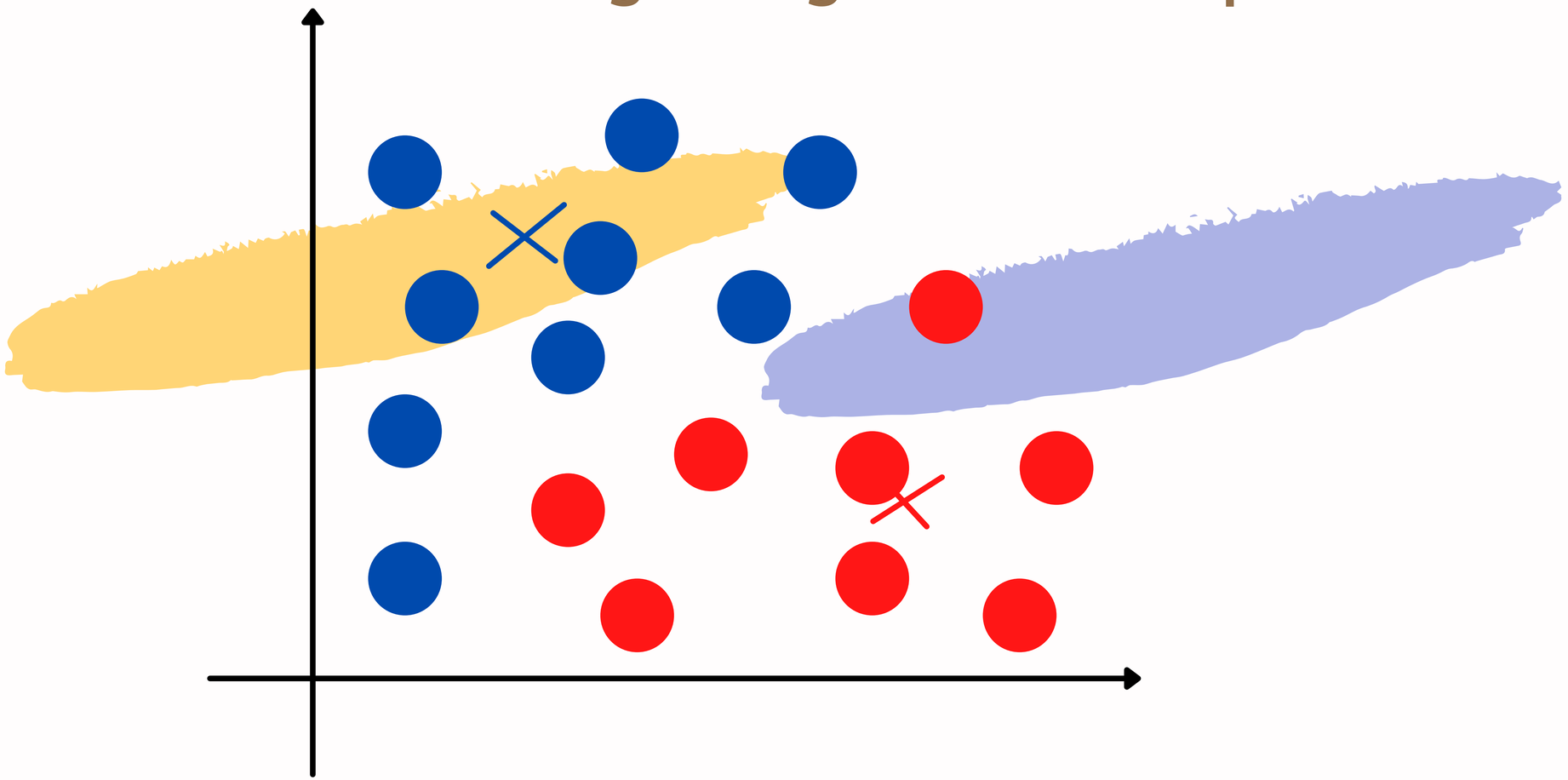
example on 2 clusters

1st step:

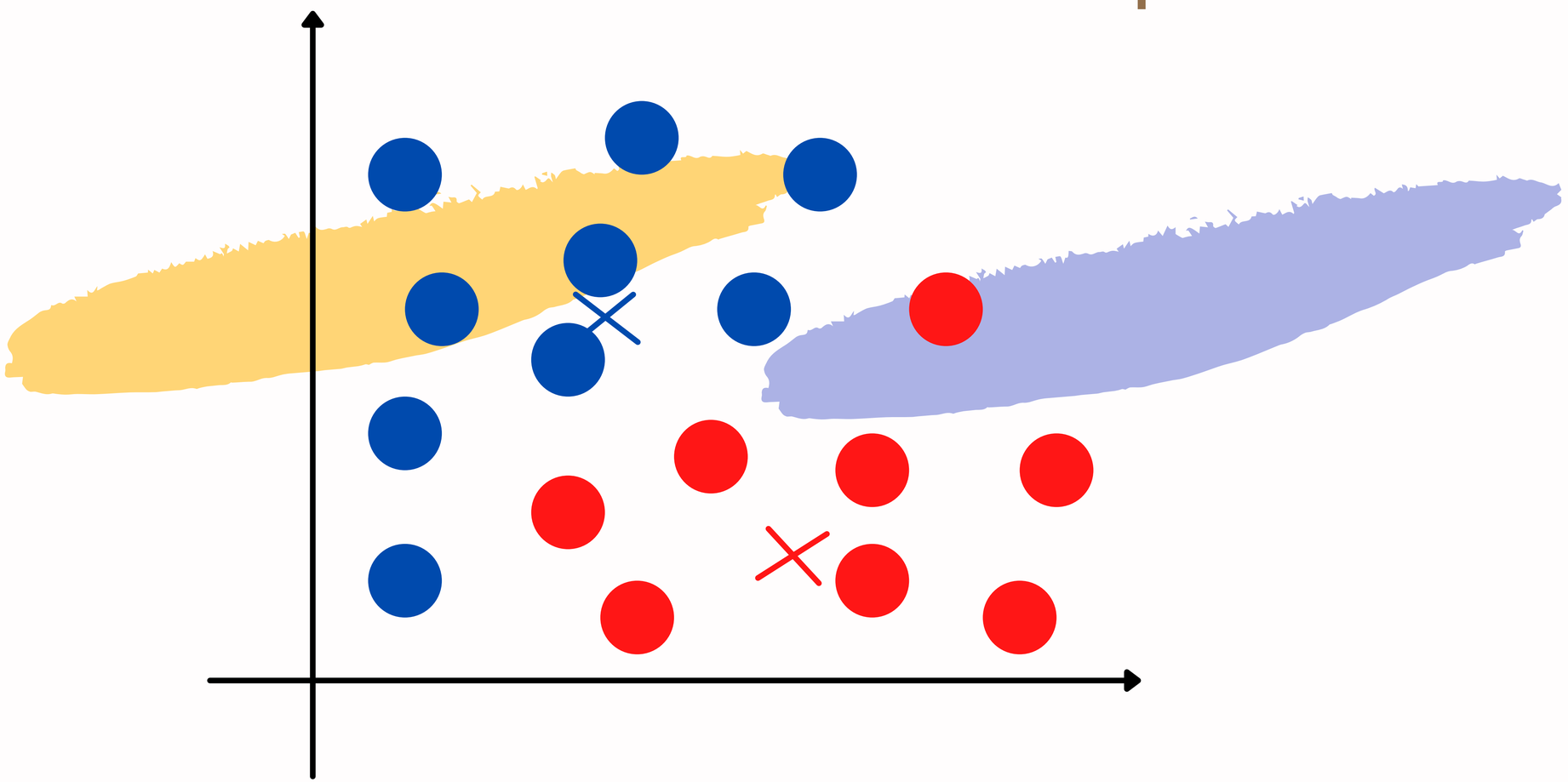
choose randomly 2 pts called the cluster centroids

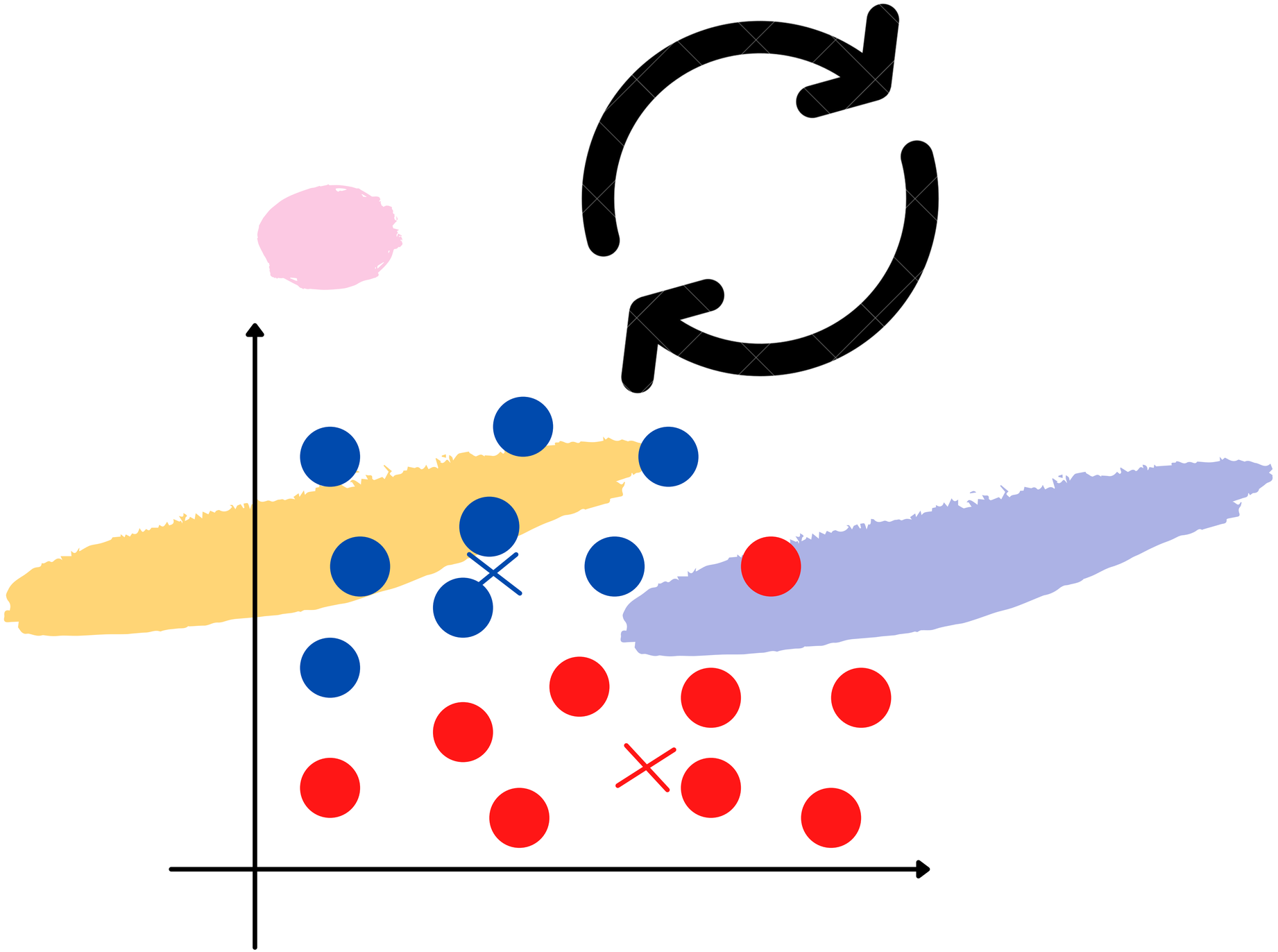


2nd step:
clustering assignement step

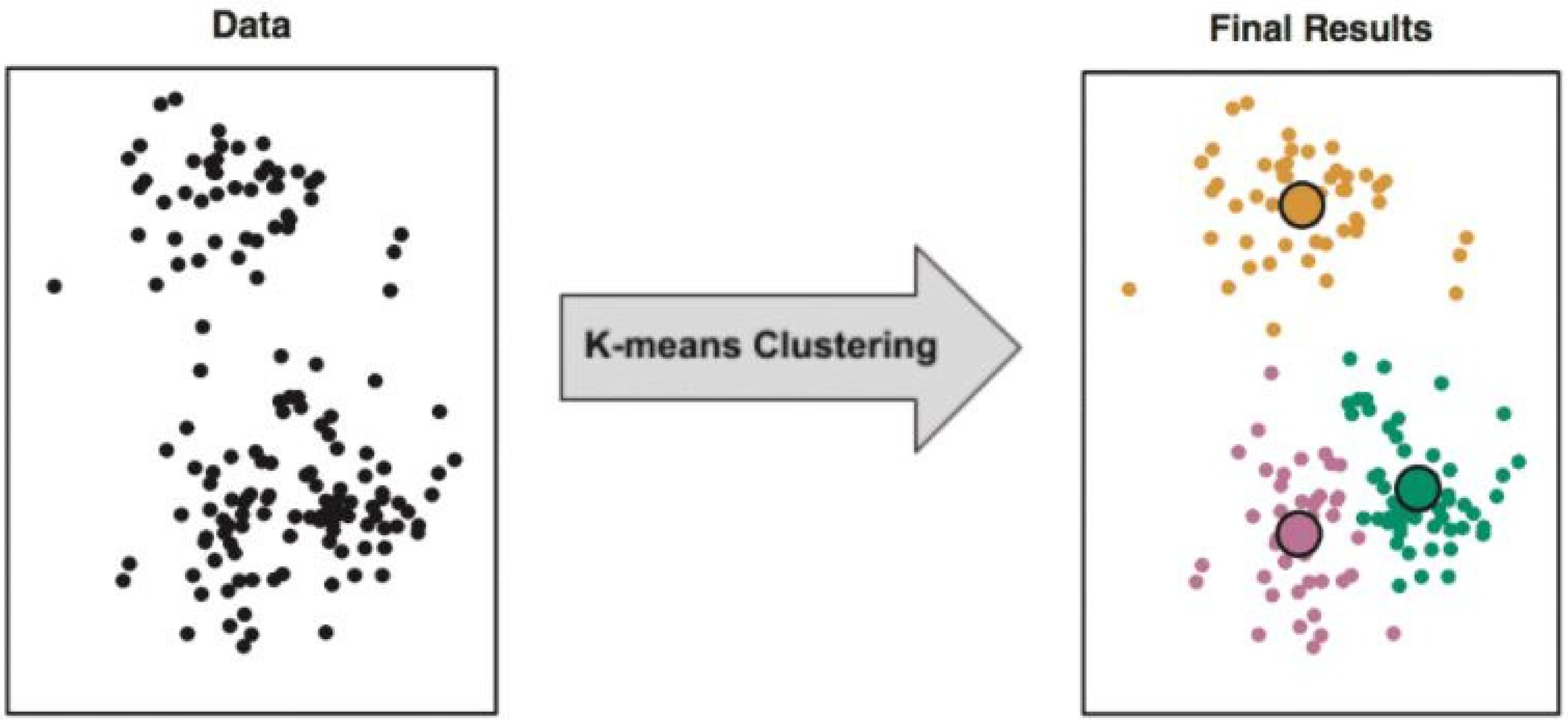


3rd step:
the move centroid step





Unsupervised Learning




Let's go deeper!

K-means algorithm

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$x^{(i)} \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)

K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat {

cluster assignement step

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

$\min_k \|x^{(i)} - \mu_k\|^2$

the move centroid step

for k = 1 to K

μ_k := average (mean) of points assigned to cluster k

$$\mu_2 = \frac{1}{3}\left[x^{(1)} + x^{(5)} + x^{(6)}\right]$$

Is the k-means algorithm guaranteed to converge?

but not necessarily to the global min

Solution: To run k-mean many times using different random initial values for the cluster centroids.

the distortion function $J(c,\mu) = \frac{1}{m}\sum_{i=1}^{m} \| x^{(i)} - \mu_c^{(i)} \|$

K-means algorithm → is exactly coordinate descent on J

optimization algorithm ←

↘ minimizing J

cluster assignement step:

minimizing J with respect to c(1),...c(m) while holding the cluster centroids fixed

the move centroid step:

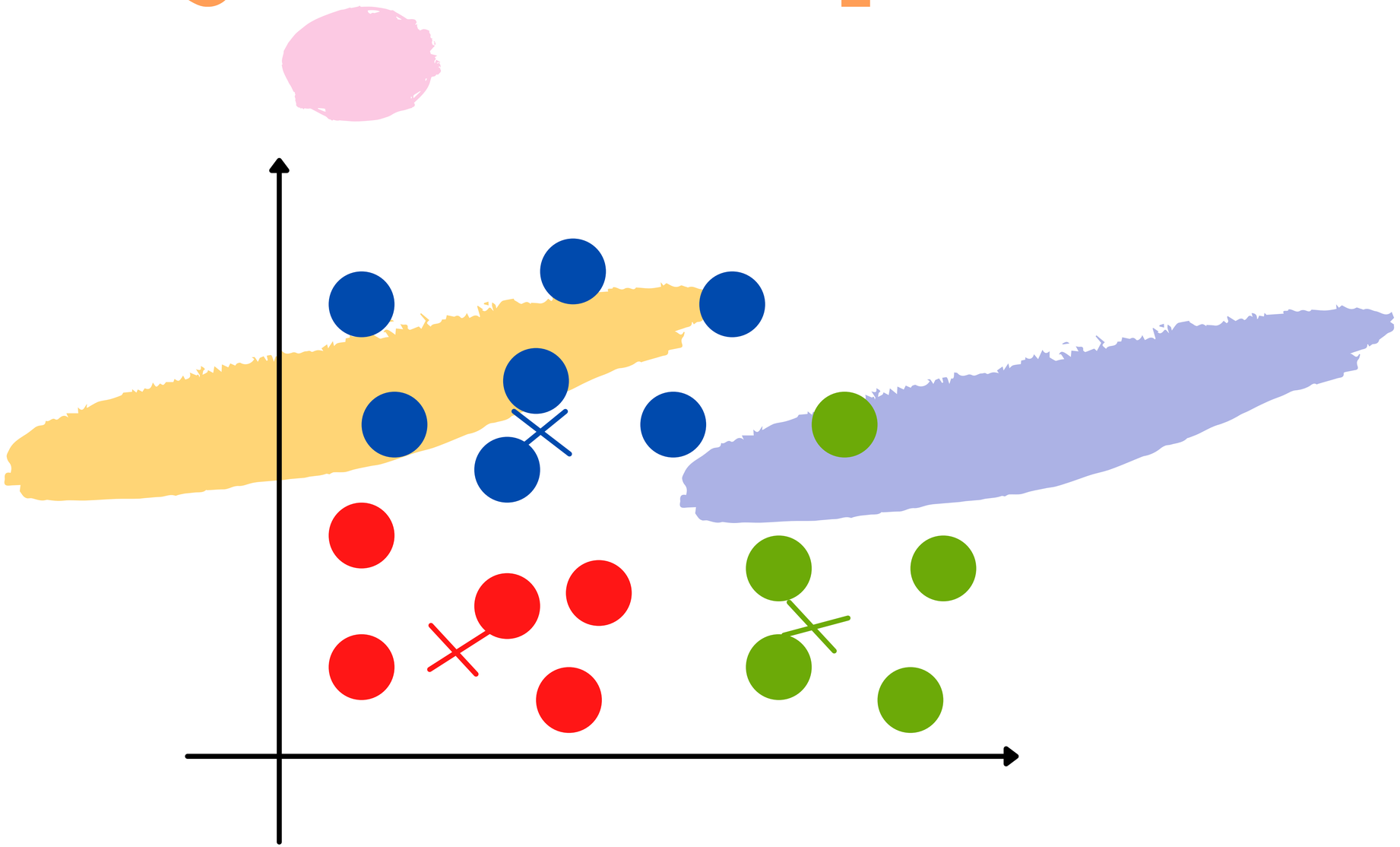
minimizing J with respect to the cluster centroids while holding c(1),...c(m) fixed

Random initialization

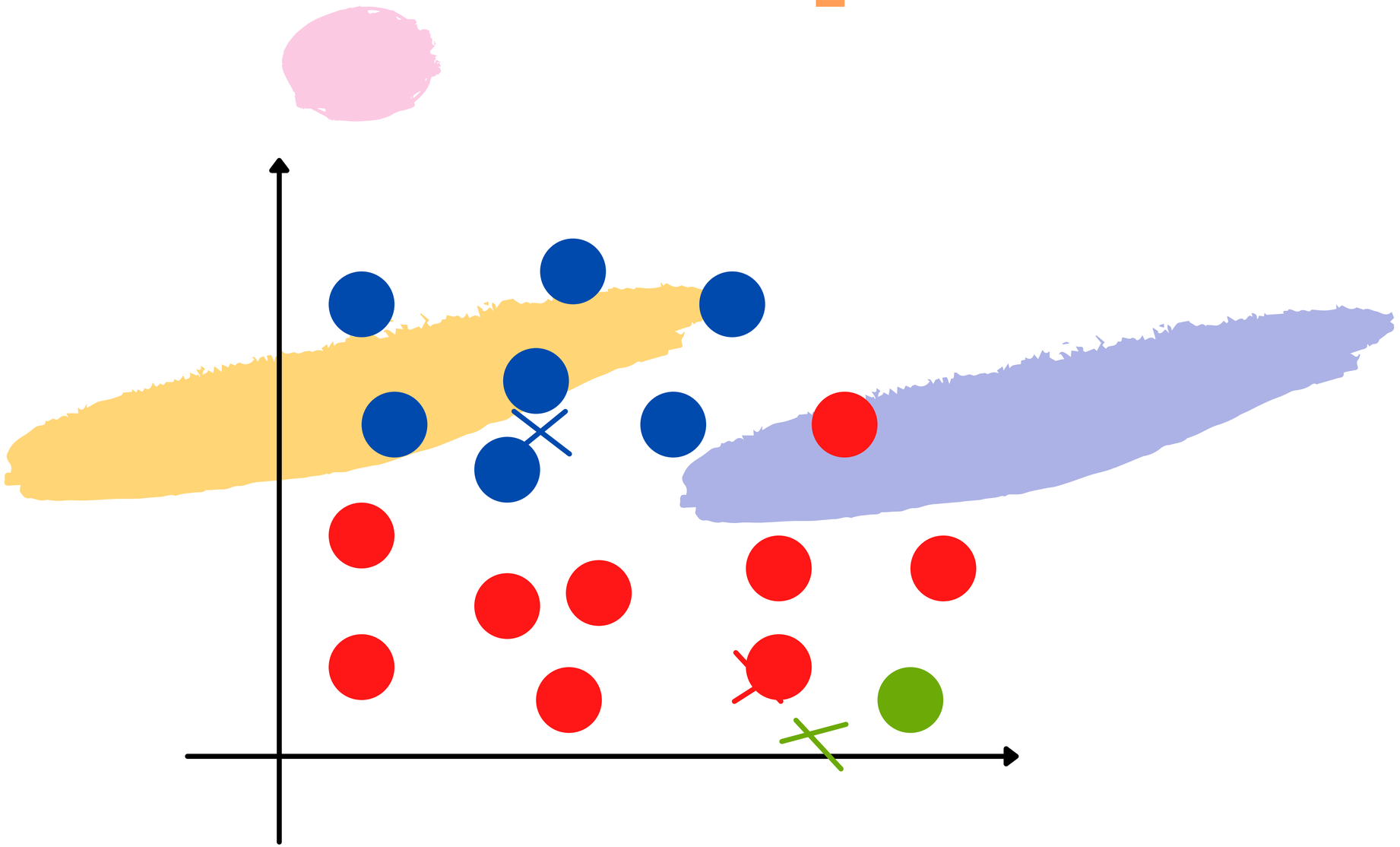
$$\mu_1, \cdots, \mu_K$$

- K<m : number of features
- Randomly choose K training examples & set clusters centroids equal to them

good local optima



bad local optima



SOLUTION

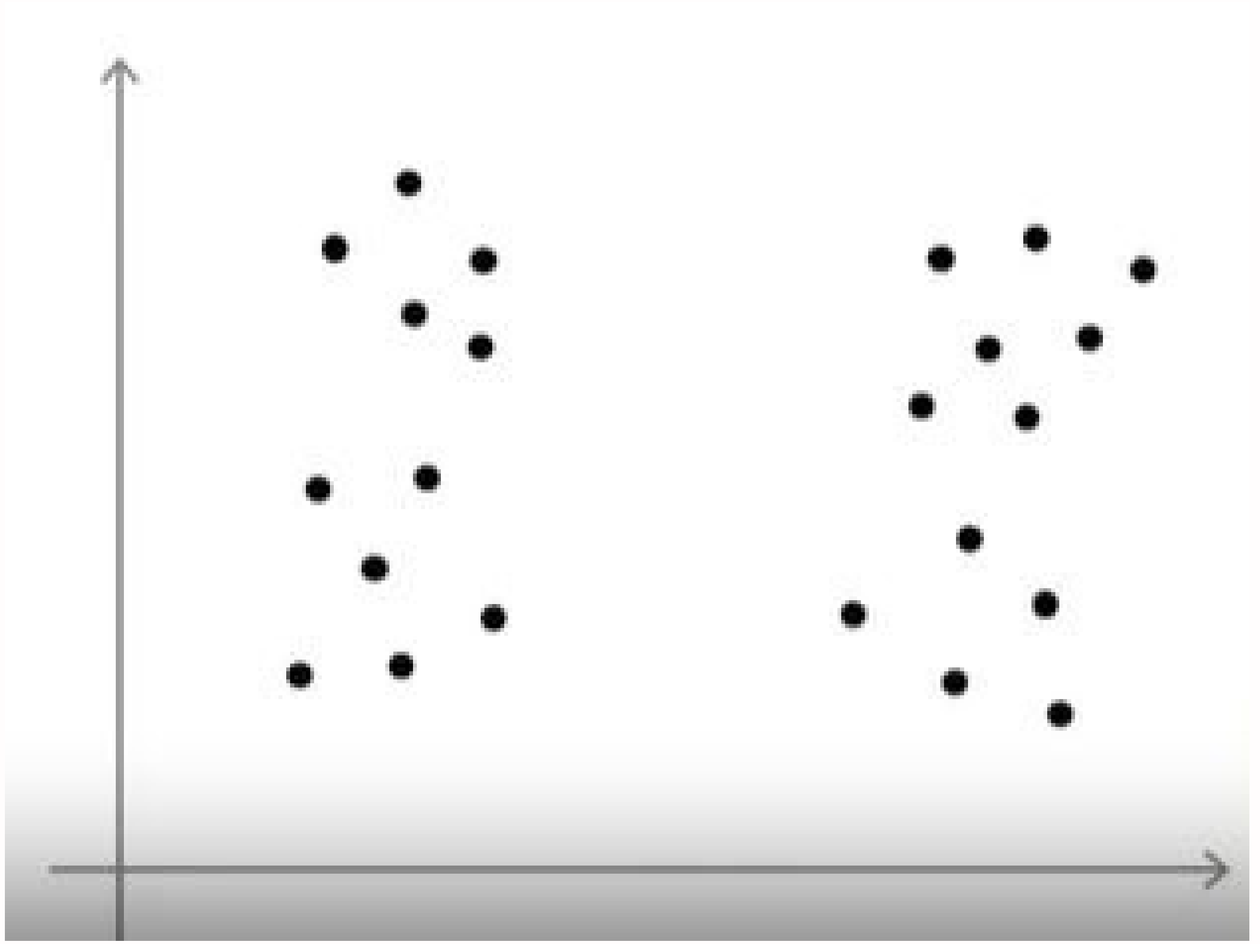
Random initialization

For i = 1 to 100 {

Randomly initialize K-means.
Run K-means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.
Compute cost function (distortion)
$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$
}

pick clustering that gave lowest cost J

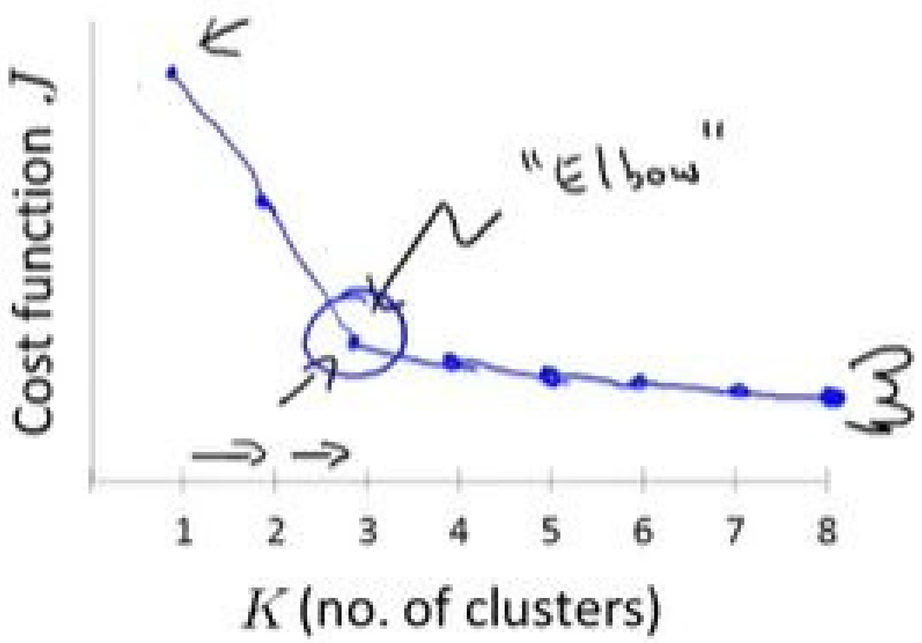
How to choose the number of cluster ?



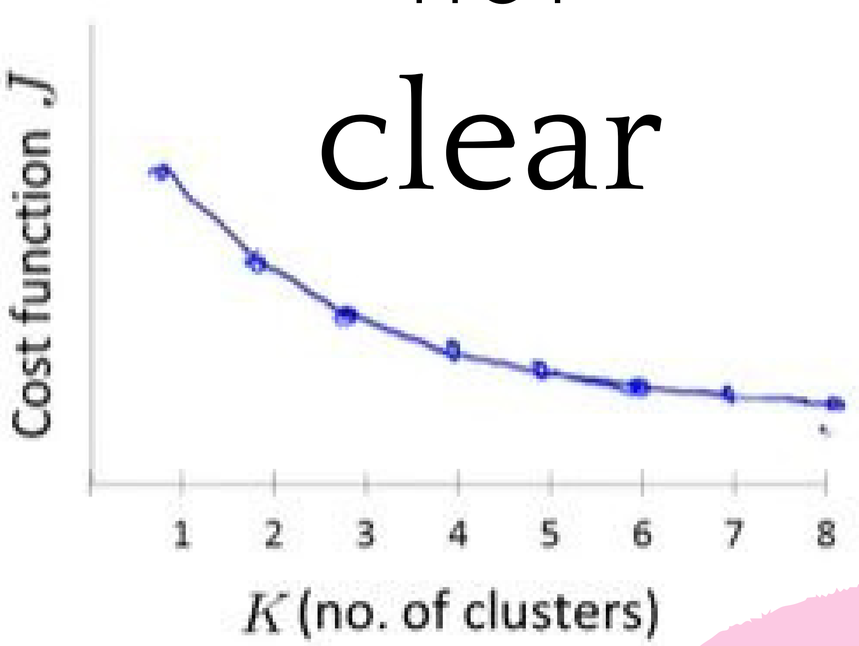
the most common method is the Elbow meth

Choosing the value of K

Elbow method:



not clear



Suppose you run k-means using k = 3 and k = 5. You find that the cost function J is much higher for k = 5 than for k = 3. What can you conclude?

In the run with k = 5, k-means got stuck in a bad local minimum. You should try re-running k-means with multiple random initializations.

choosing of K

Sometimes it depends on the later purpose

for example: T-shirt sizing

- xs,s,M,L,XL
- s,M,L

Let's practice!