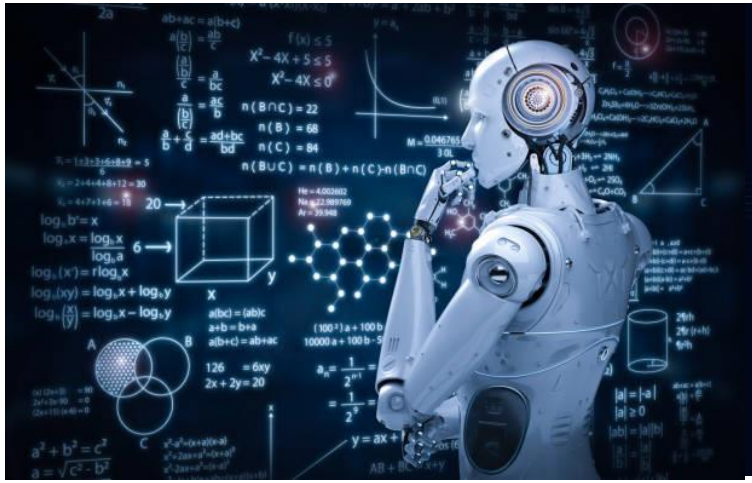
Welcome…

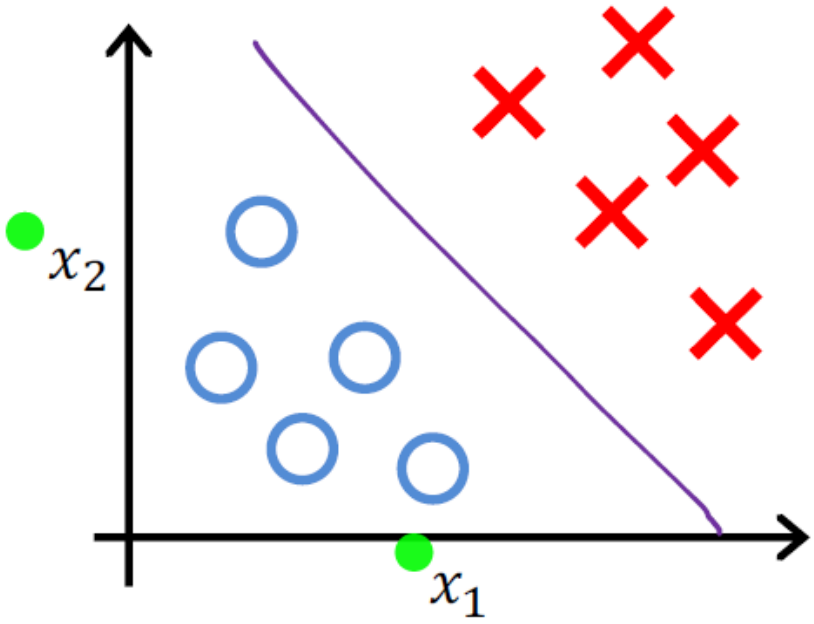
Unsupervised Learning: Clustering

CS 797Q
Fall 2024

10/30/2024

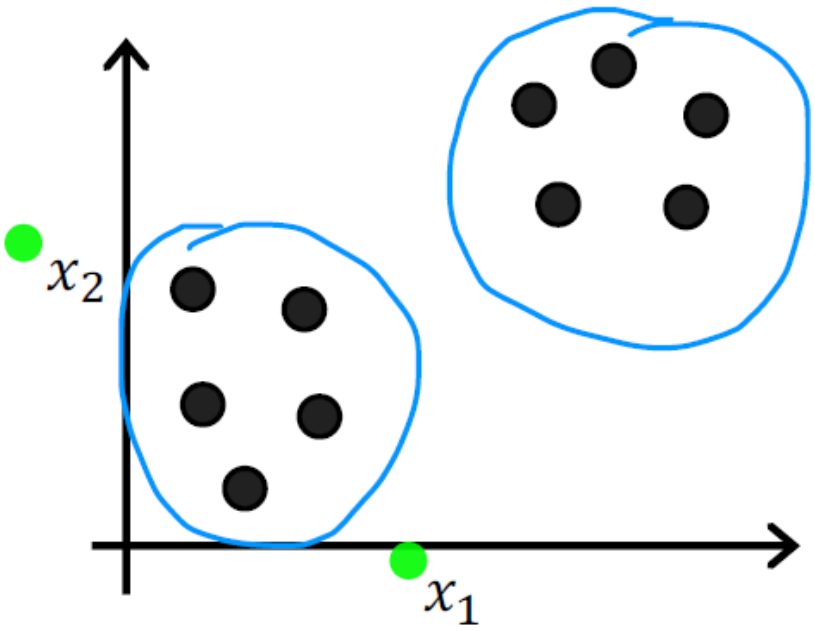


Supervised learning



Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$?

Unsupervised learning



Clustering

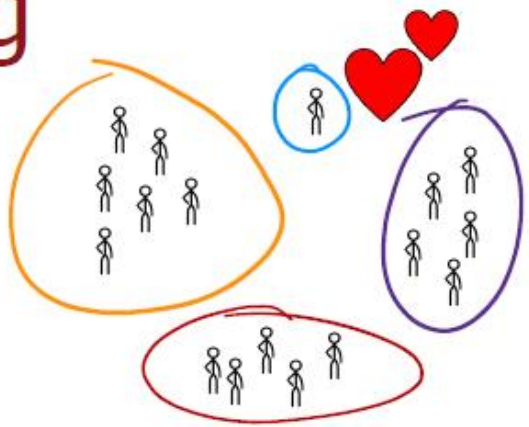
Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

Application of Clustering

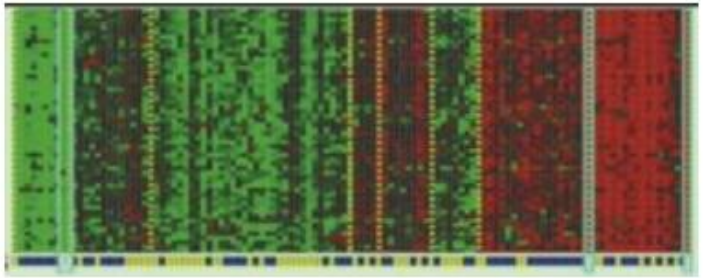
Applications of clustering

Grouping similar news

- Growing skills
- Develop career
- Stay updated with AI, understand how it affects your field of work



Market segmentation



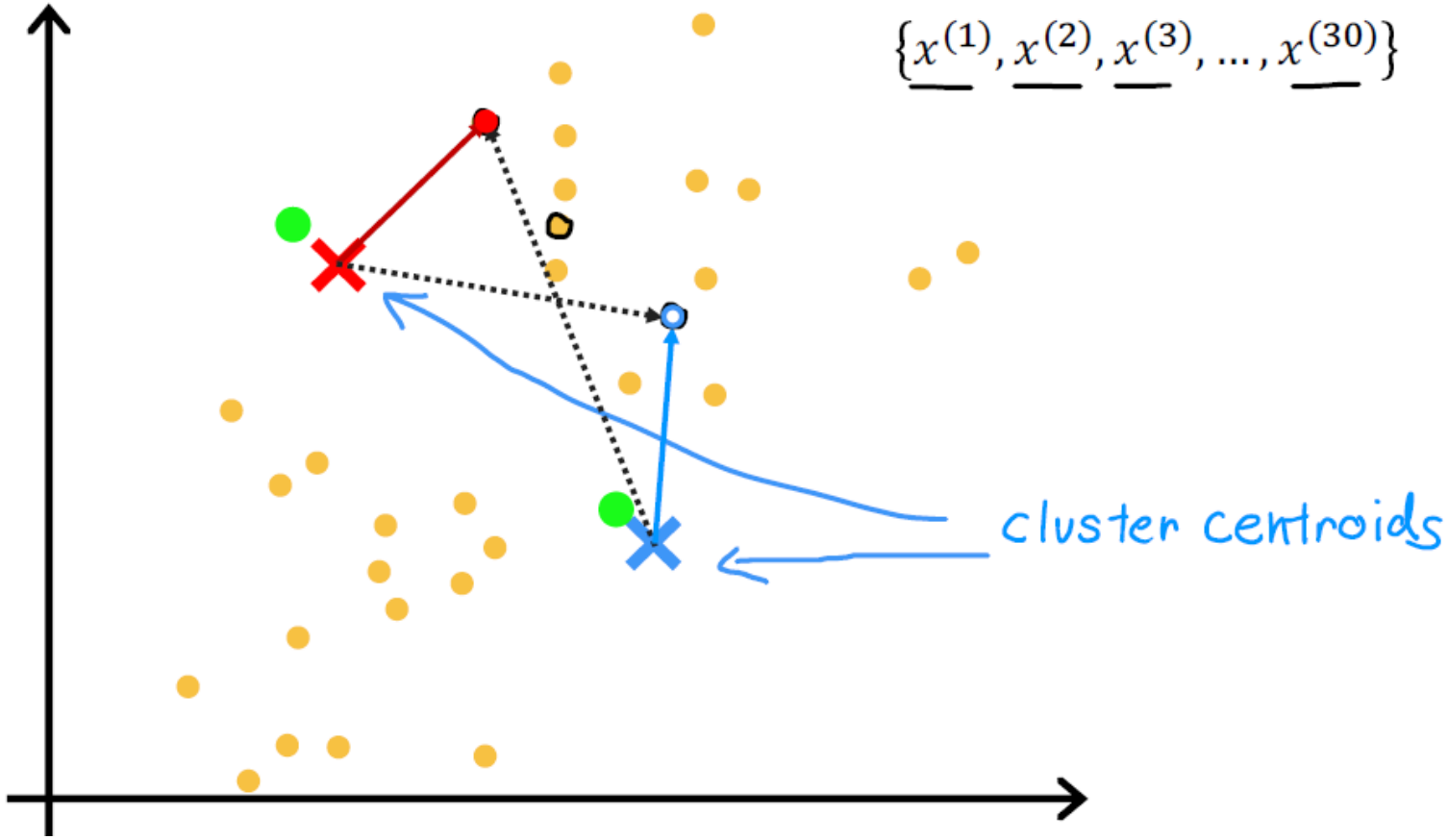
DNA analysis

Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Madison)

Astronomical data analysis

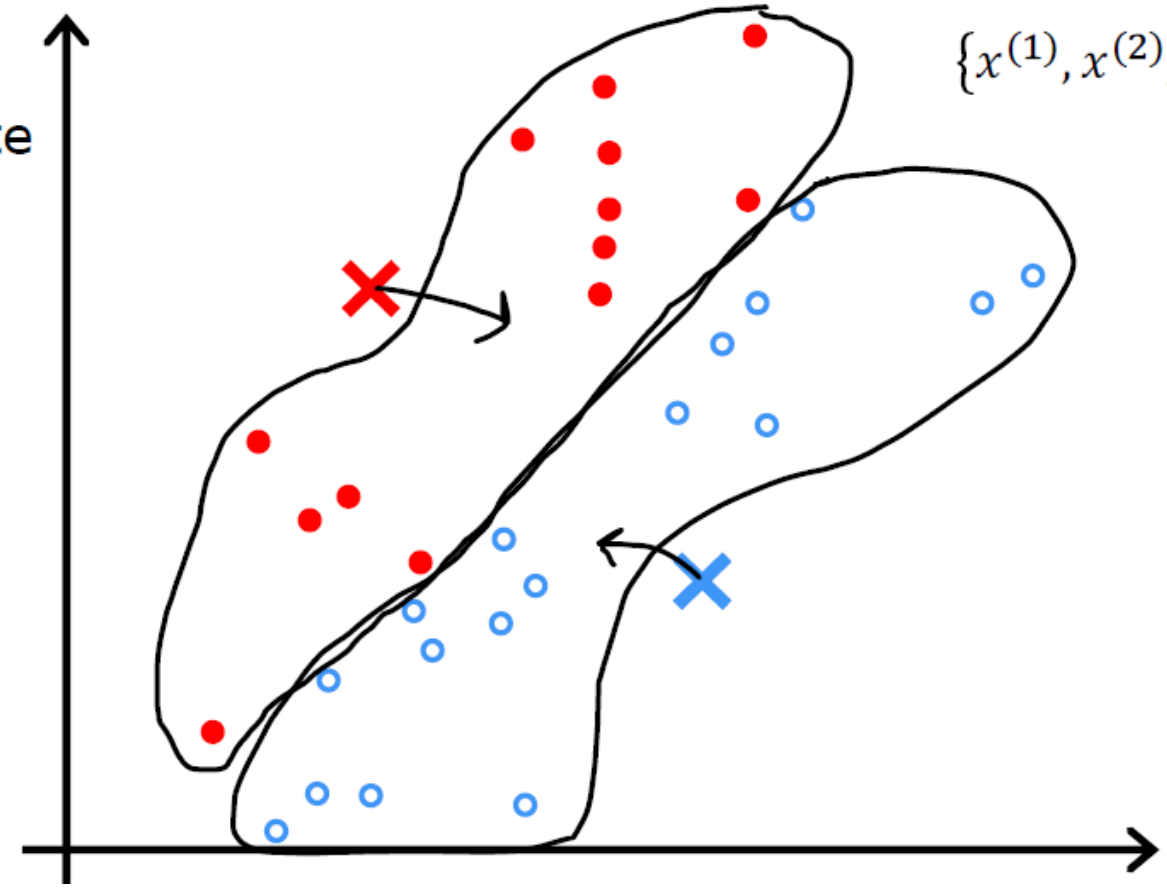
K-mean Clustering intuition

Step 1: Assign each point to its closest centroid



K-mean Clustering intuition

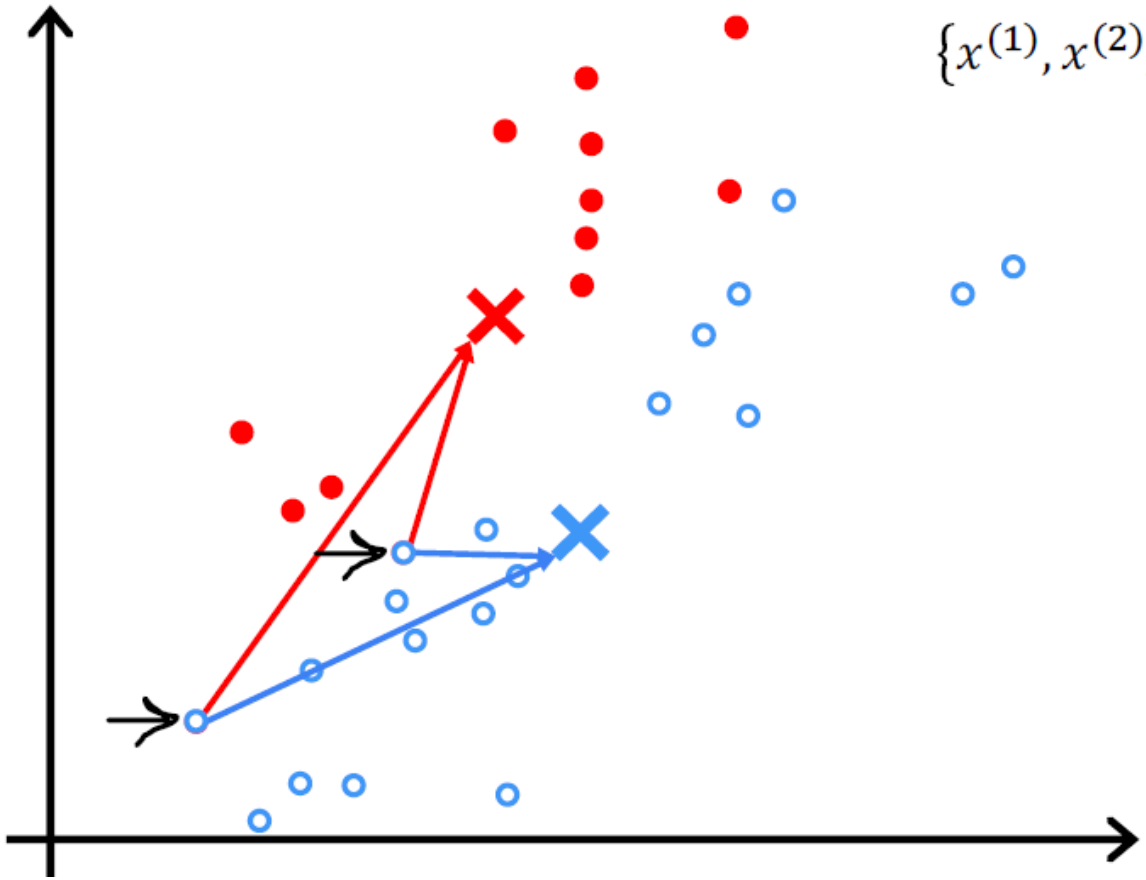
Step 2: Recompute the centroids



$\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(30)}\}$

K-mean Clustering intuition

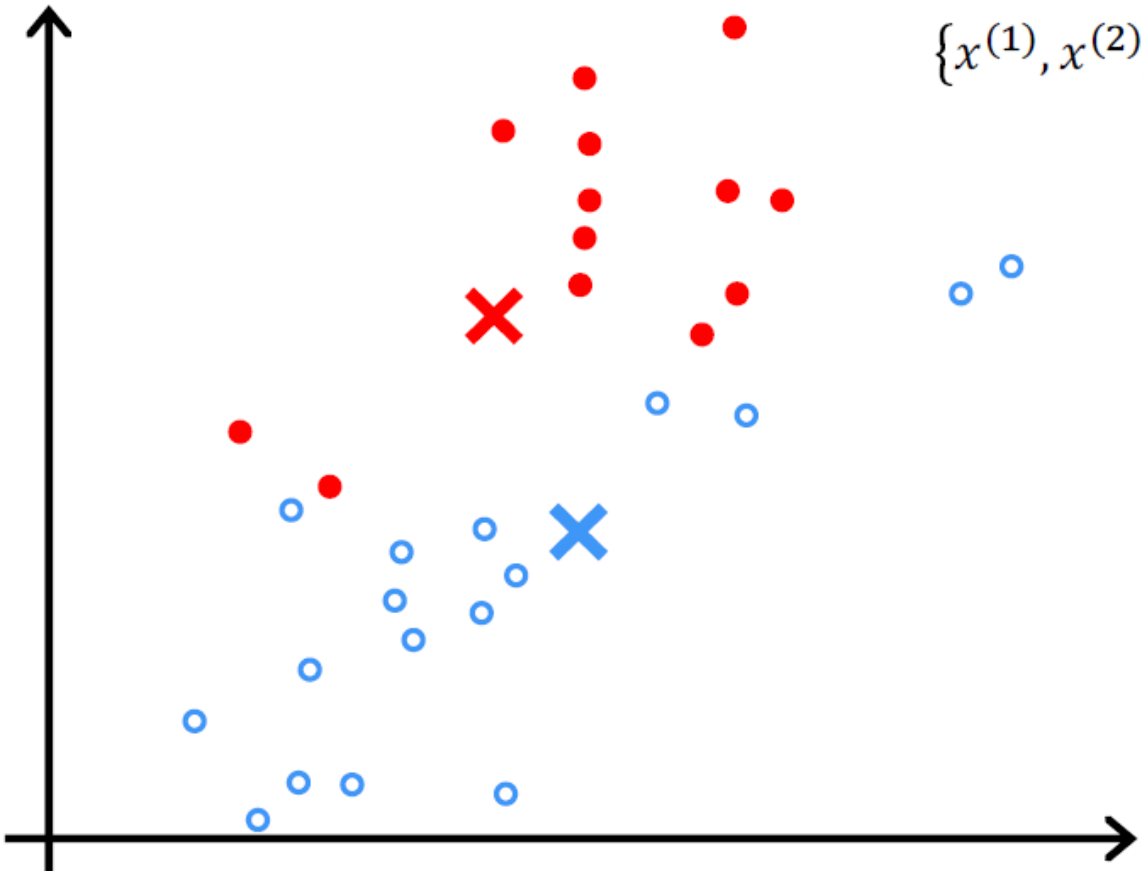
Step 1: Assign each point to its closest centroid



$\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(30)}\}$

K-mean Clustering intuition

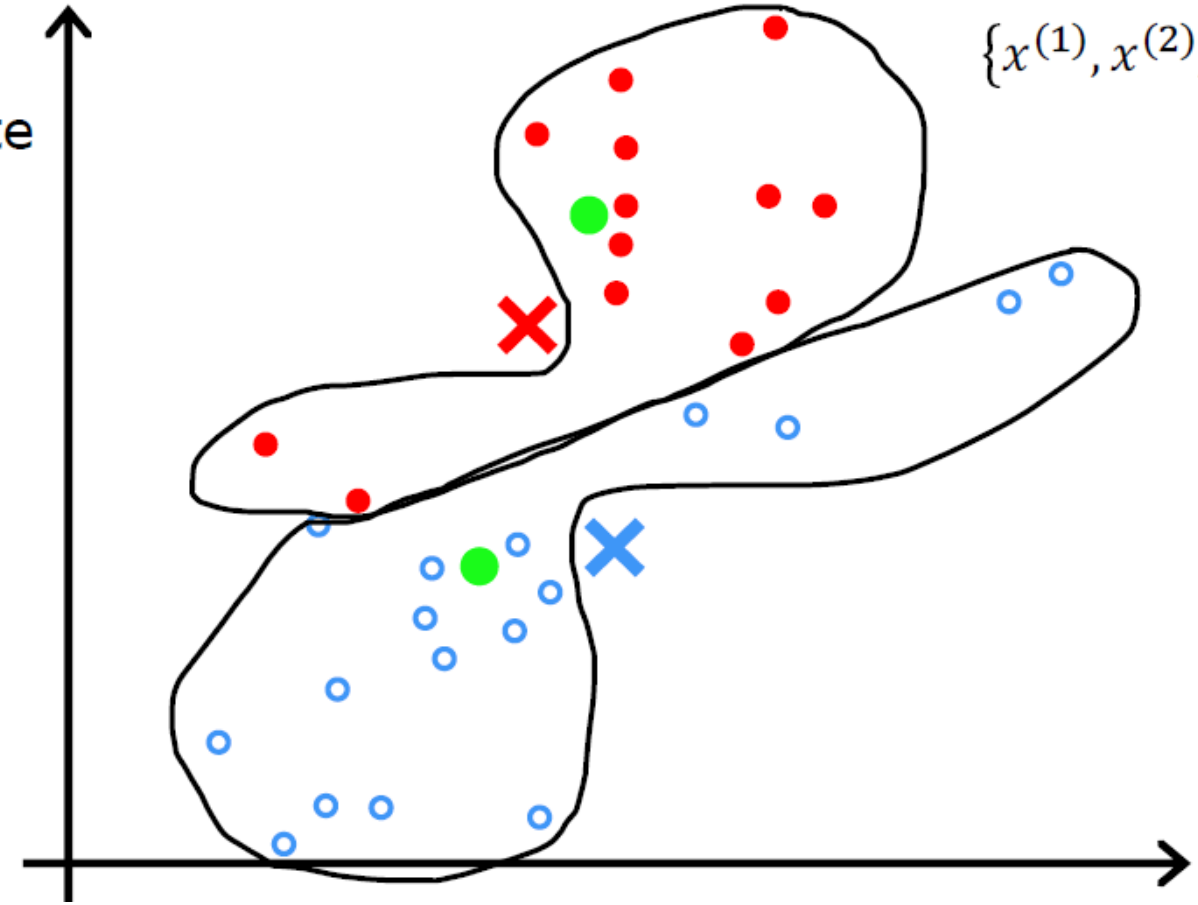
Step 1: Assign each point to its closest centroid



$\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(30)}\}$

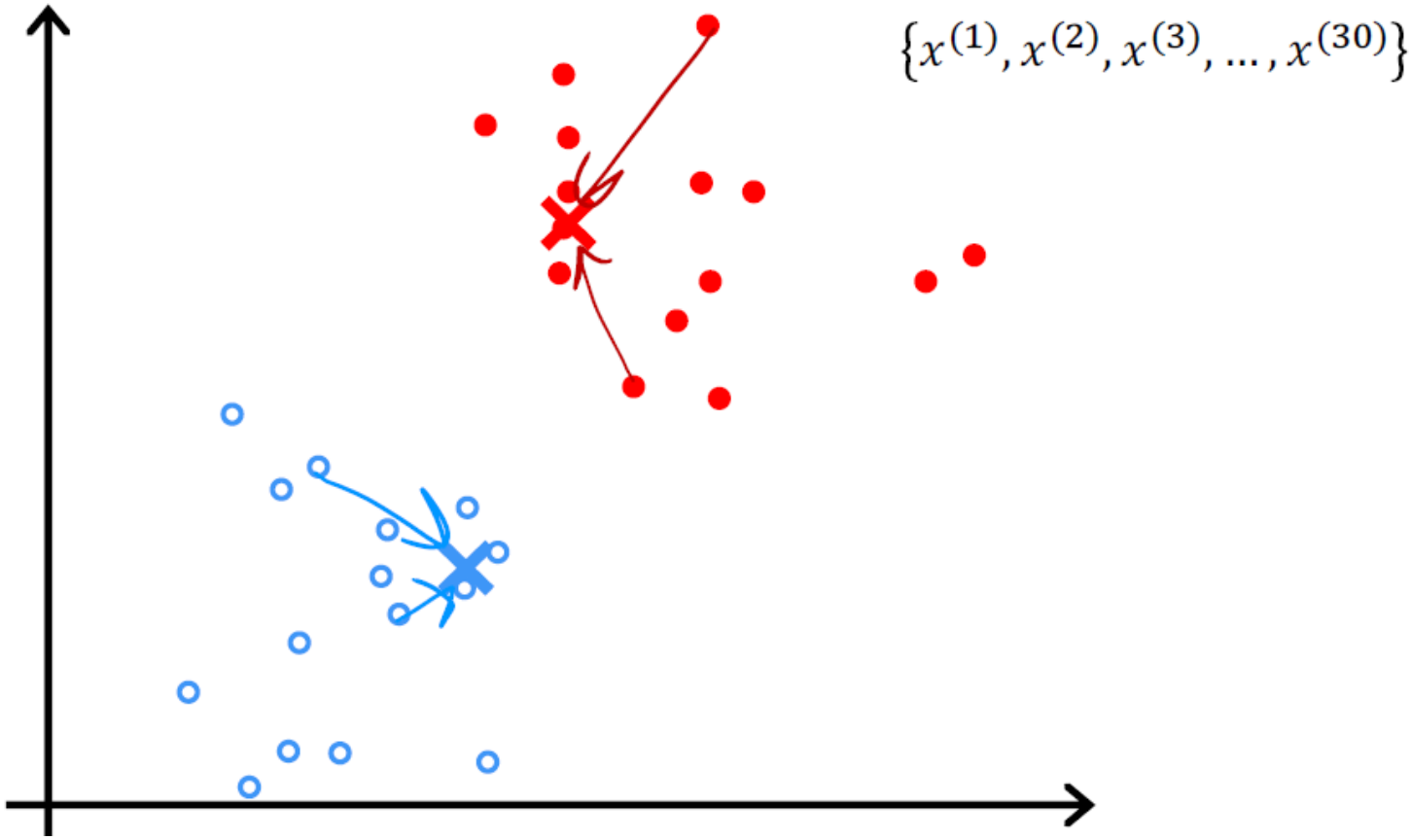
K-mean Clustering intuition

Step 2:
Recompute the centroids

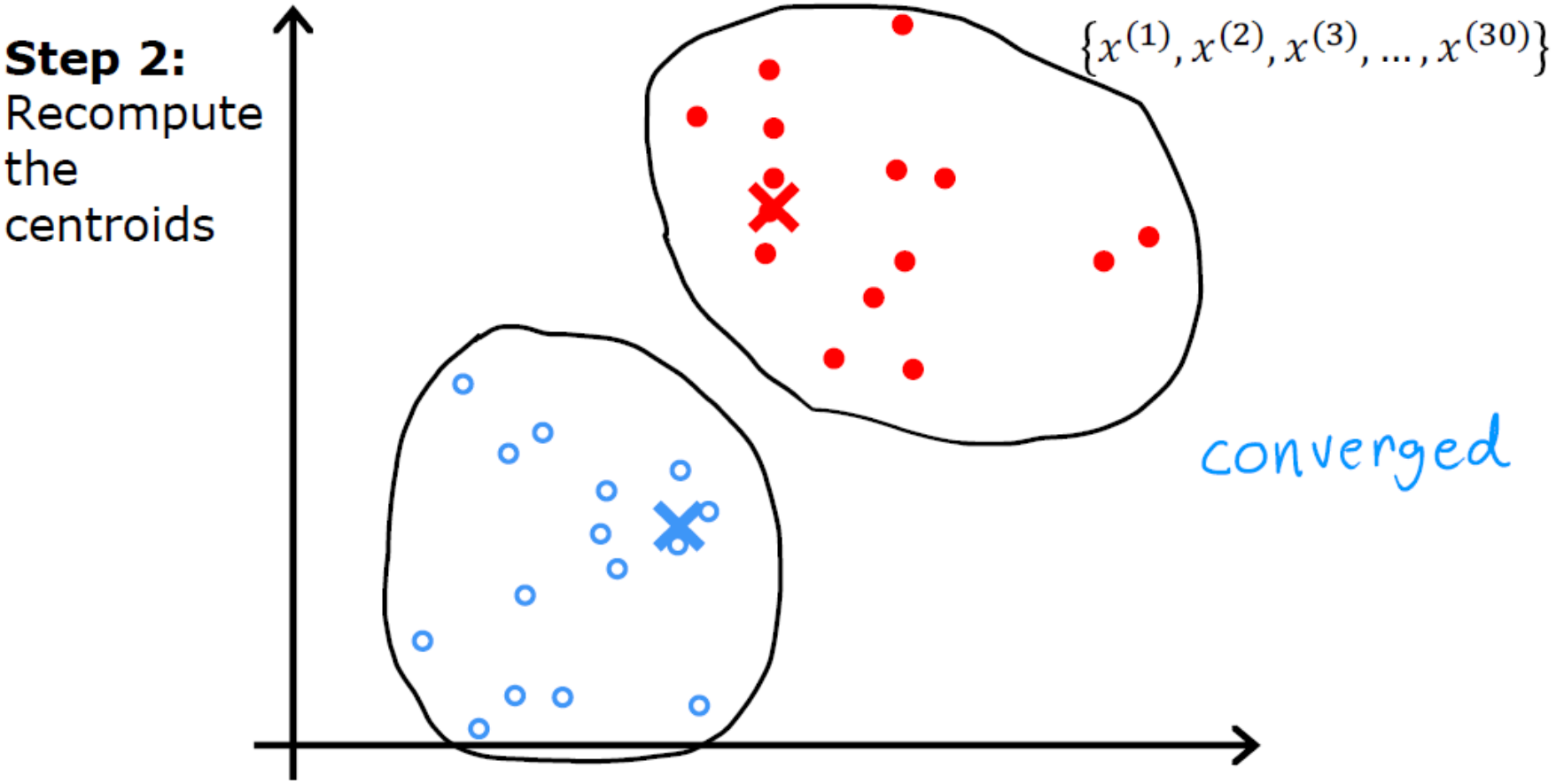


K-mean Clustering intuition

Step 1: Assign each point to its closest centroid

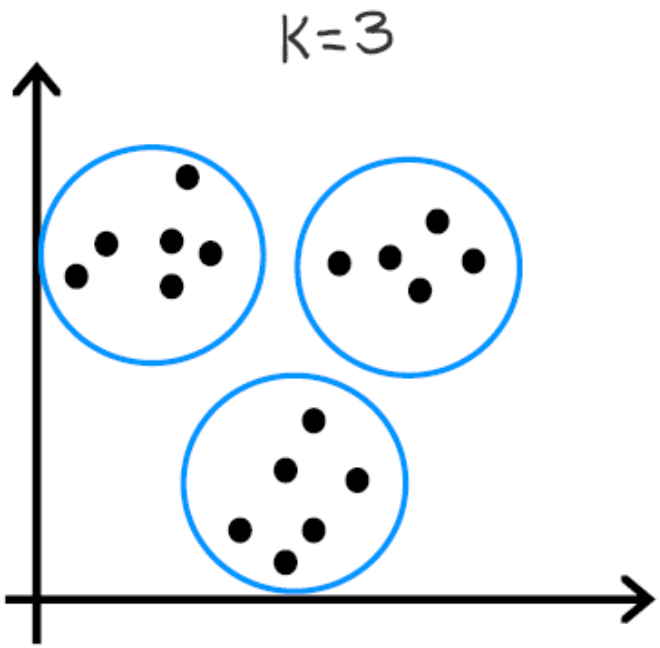


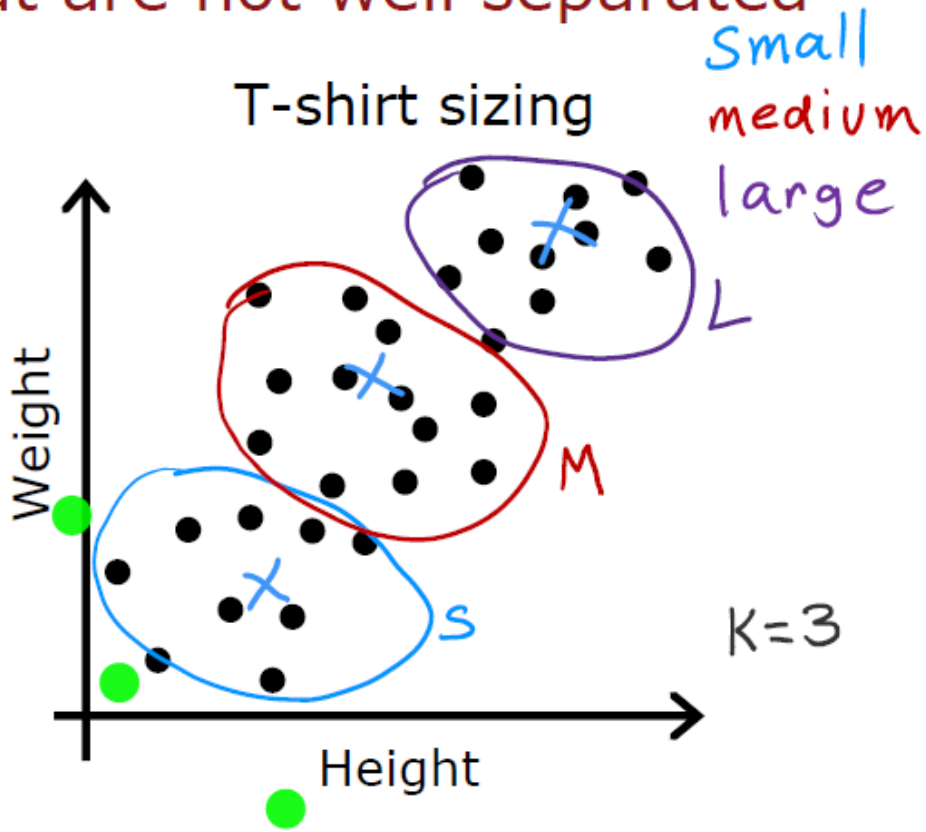
K-mean Clustering intuition



K-mean Algorithm

K-means for clusters that are not well separated





K-mean optimization objectives

$c^{(i)}$ = index of cluster $(1, 2, \ldots, K)$ to which example $x^{(i)}$ is currently assigned

μ_k = cluster centroid k

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

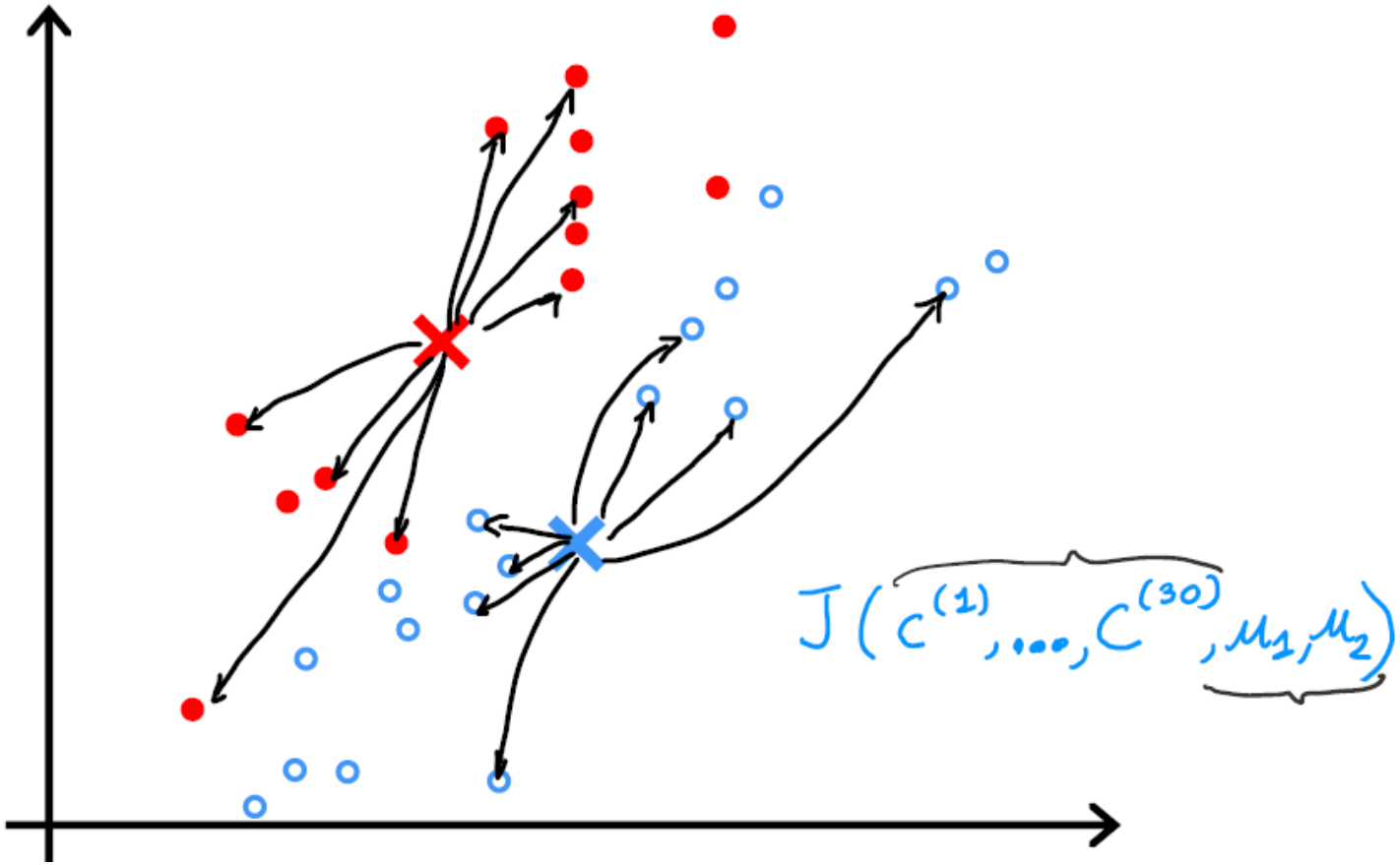
$x^{(10)}$ $c^{(10)}$ $\mu_c^{(10)}$

Cost function

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, .., \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

$$\min_{\substack{c^{(1)}, \ldots, c^{(m)} \\ \mu_1, \ldots, \mu_K}} J(c^{(1)}, \ldots, c^{(m)}, \mu_1, .., \mu_K)$$ distortion

K-mean optimization objectives



Initializing K-means

Step 0: Randomly initialize K cluster centroids $\mu_1, \mu_1, \ldots, \mu_k$

Repeat {

Step 1: Assign points to cluster centroids

Step 2: Move cluster centroids

}

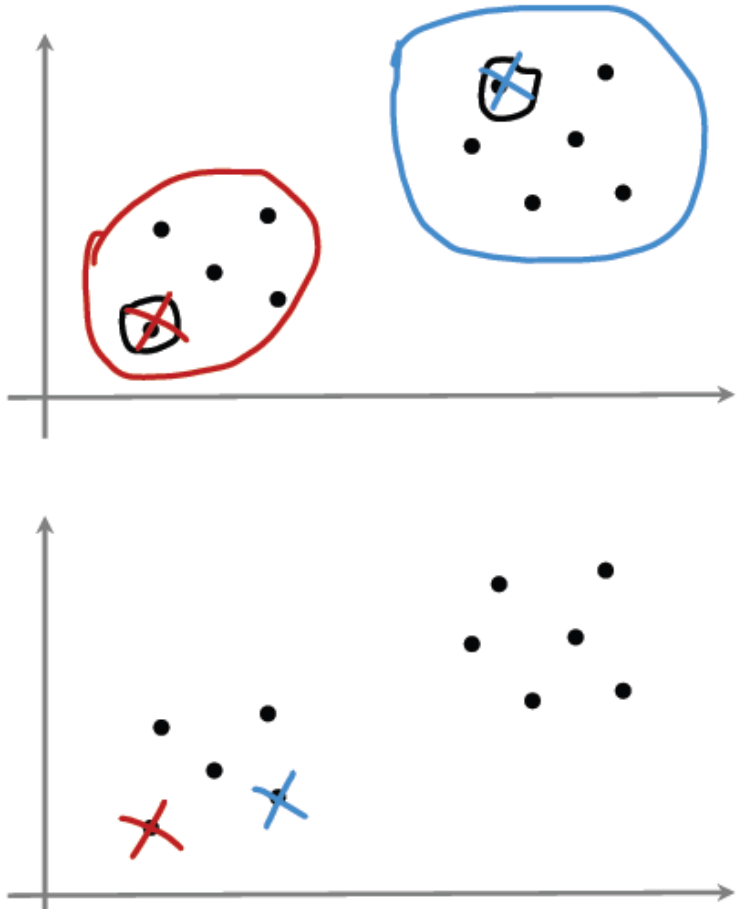
Initializing K-means

Random initialization

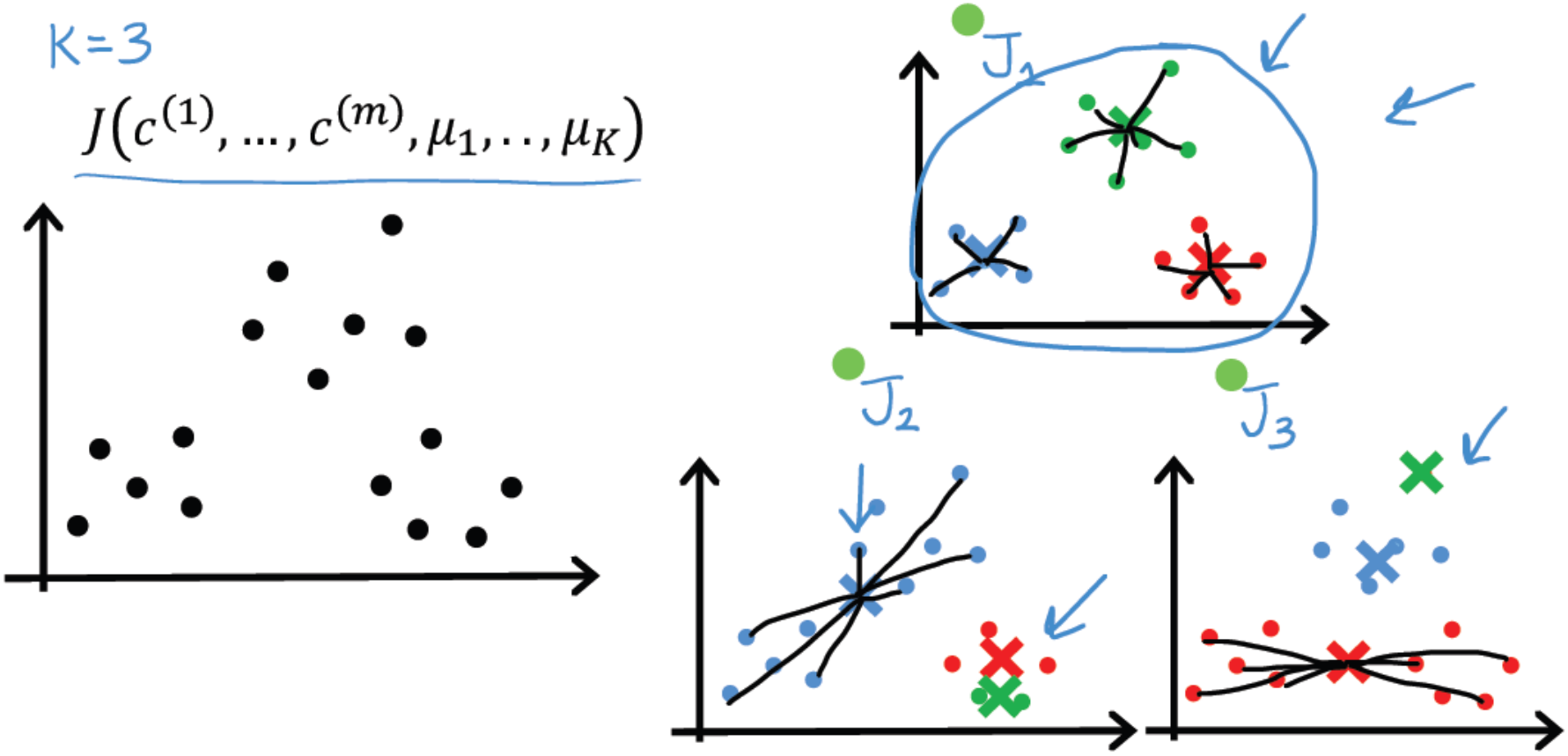
Choose $K < m$

Randomly pick K training examples.

Set μ_1, μ_1,..., μ_k equal to these K examples.



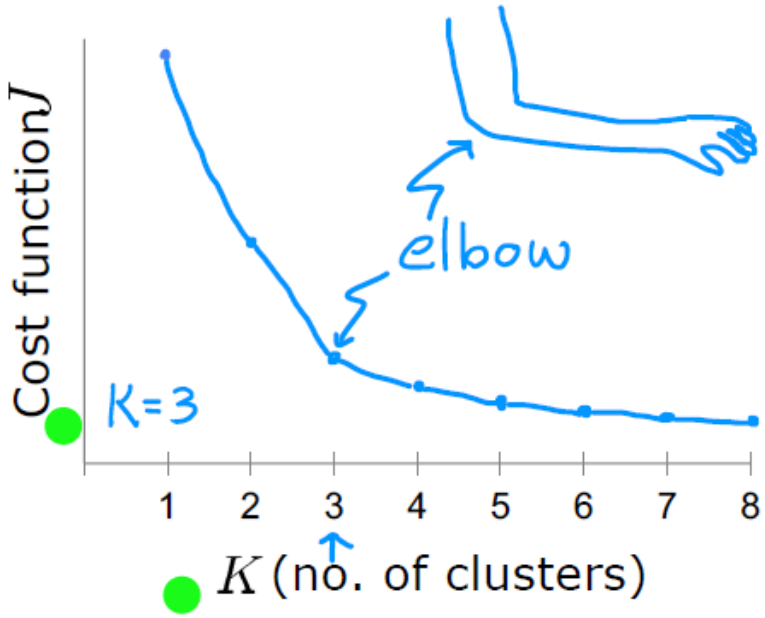
Initializing K-means



Choosing Number of clusters

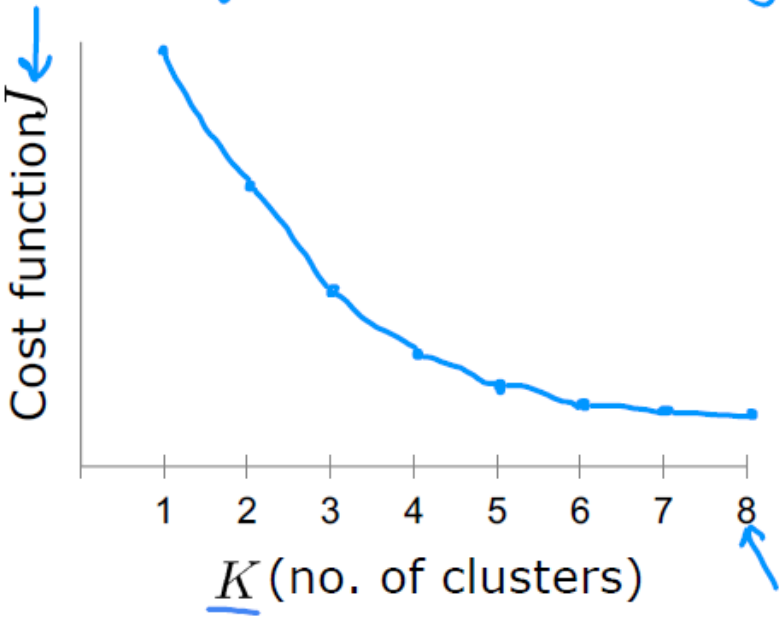
Choosing the value of K

Elbow method



the right "K" is often ambiguous

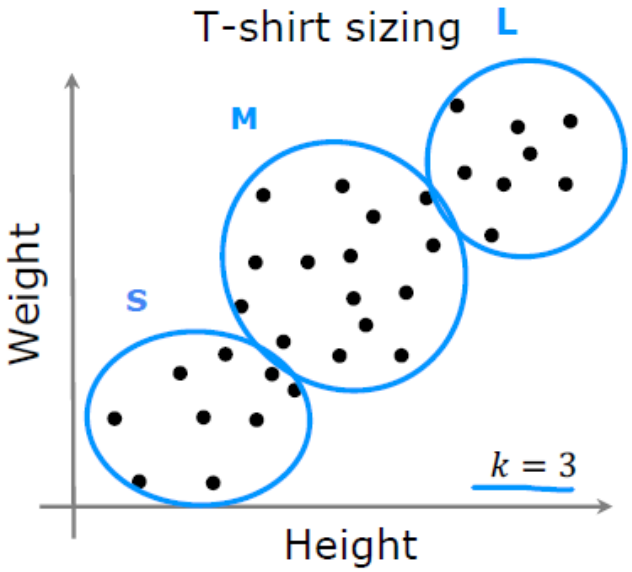
Don't choose K just to minimize cost J

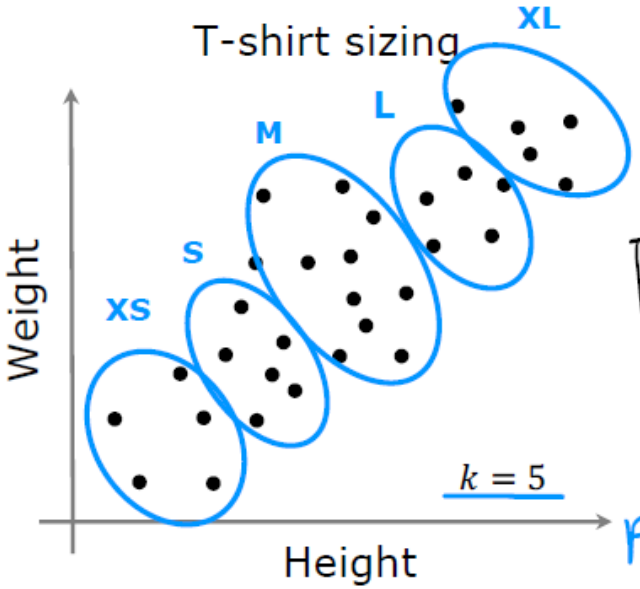


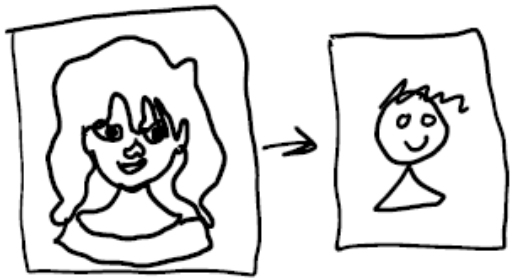
Choosing Number of clusters

Choosing the value of K

Often, you want to get clusters for some later (downstream) purpose.
Evaluate K-means based on how well it performs on that later purpose.







practice lab: image compression!