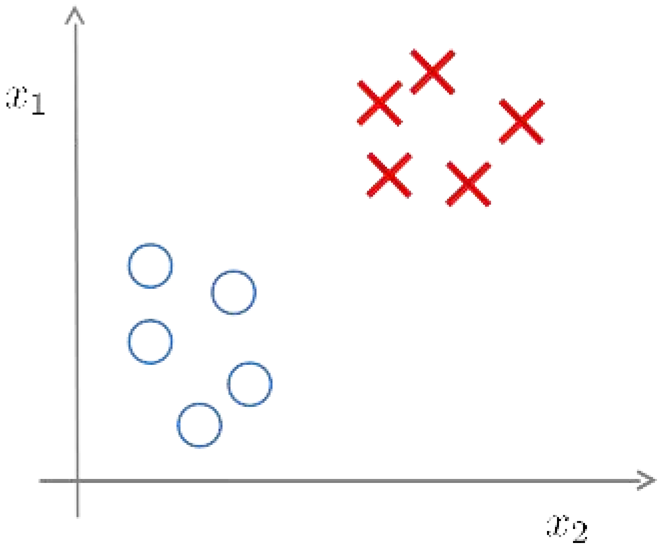
CENG 463
Machine Learning

Lecture 09 - Clustering with K-Means

Slides were prepared using the material provided in Stanford's Machine Learning Course by Andrew Ng.

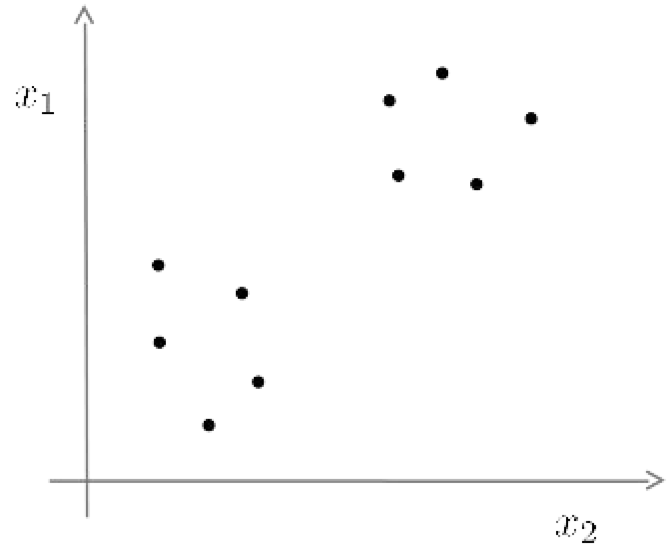
Unsupervised Learning

- Supervised Learning



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

- Unsupervised Learning

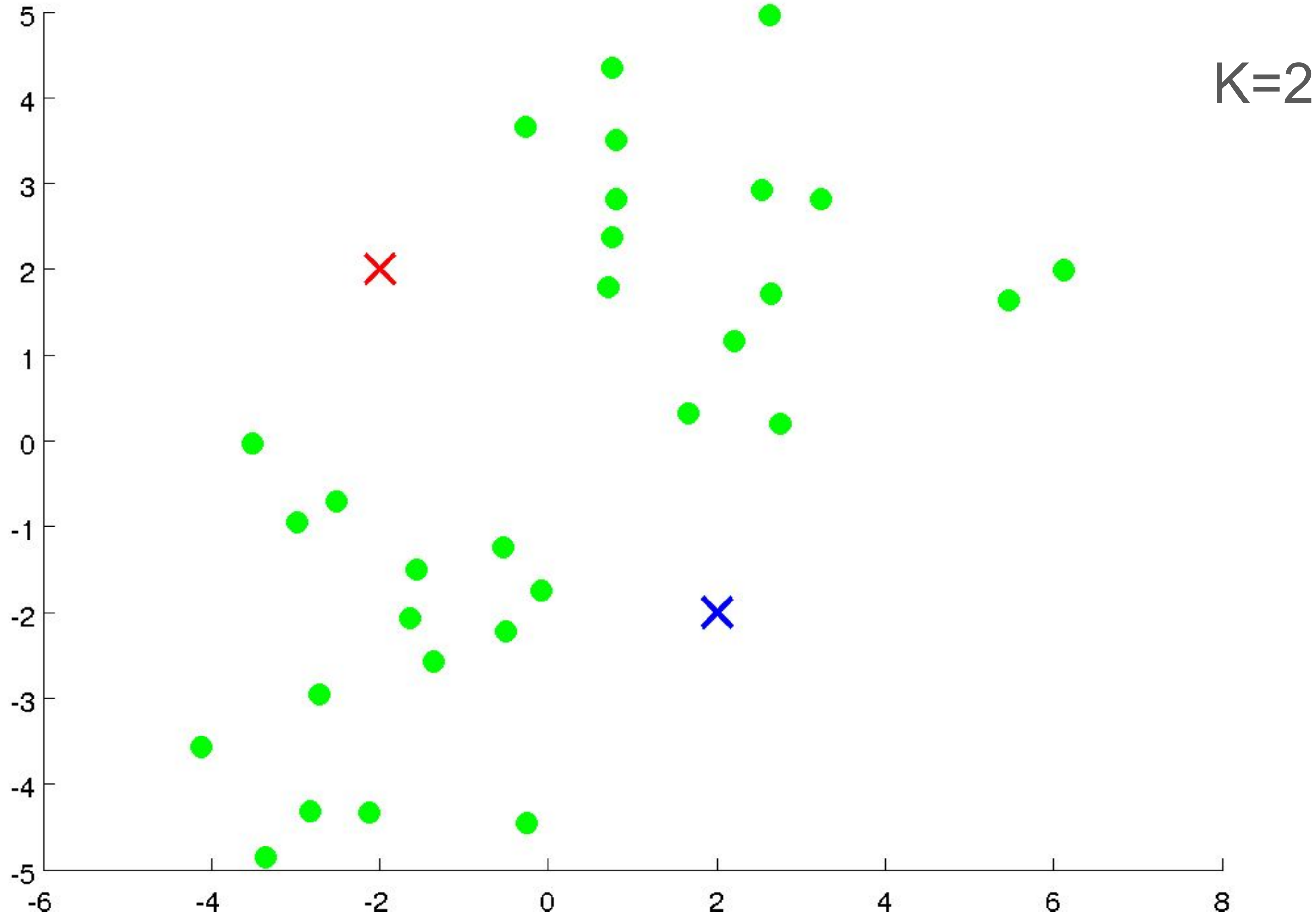


Training set:
$\{(x^{(1)},y^{(1)}), \ldots, (x^{(m)},y^{(m)})\}$

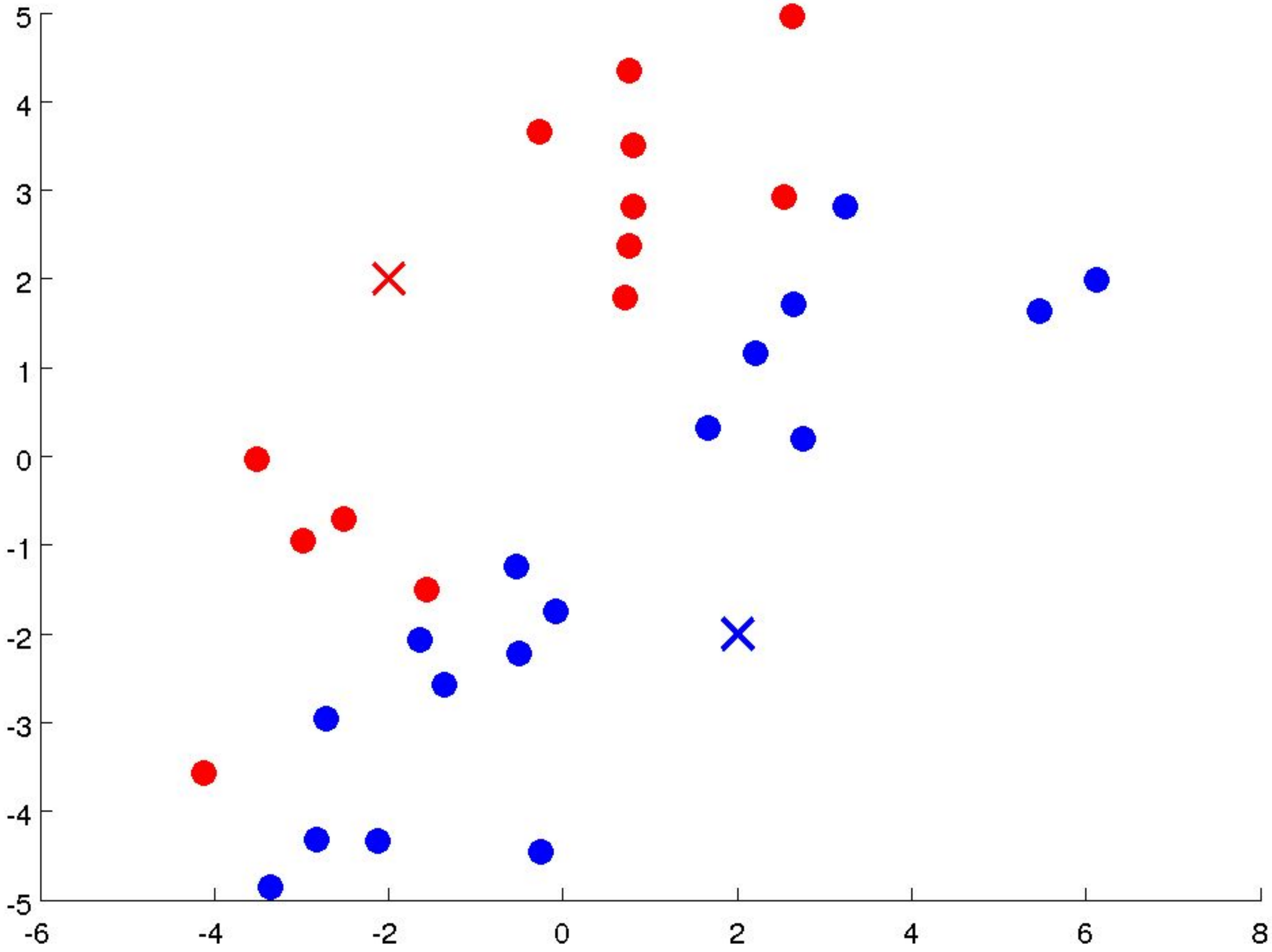
K-Means Algorithm

- K-means is an iterative clustering algorithm.
- It has two steps in each iteration:
  - Cluster assignment step:
    - Assign each sample to the closest cluster centroid
  - Move centroids step:
    - Recompute cluster centroids using assigned samples

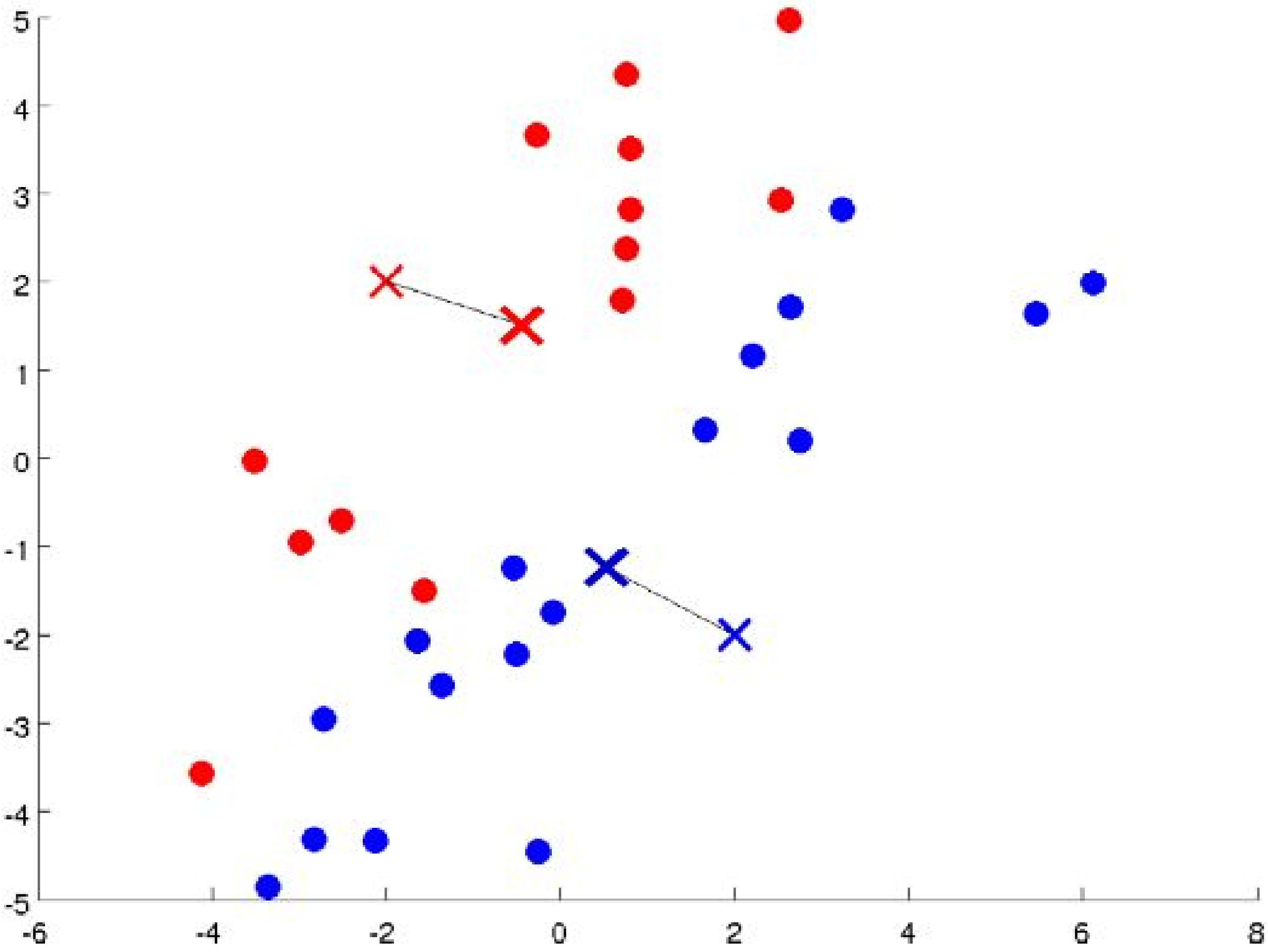
K-Means Algorithm



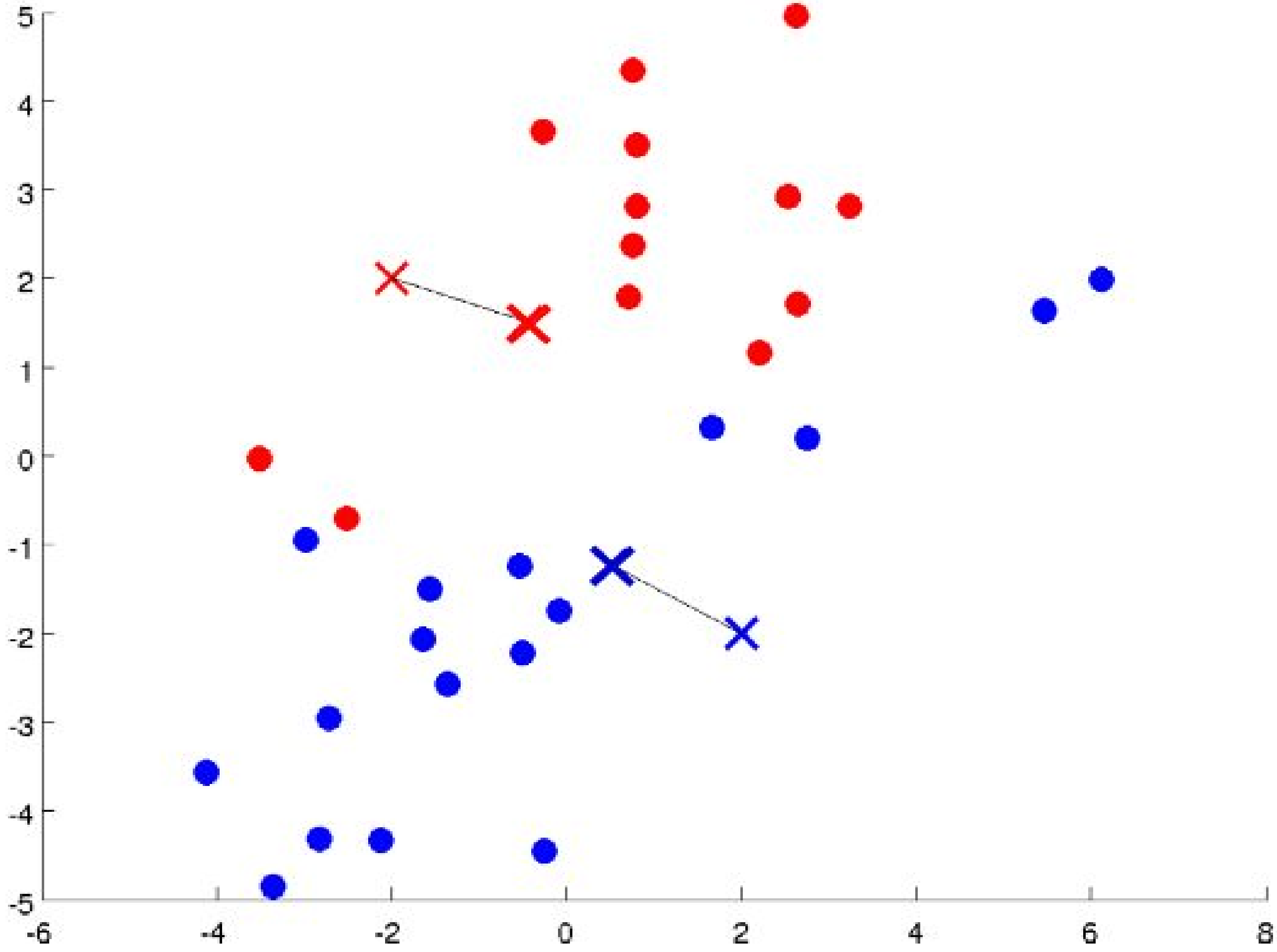
K-Means Algorithm



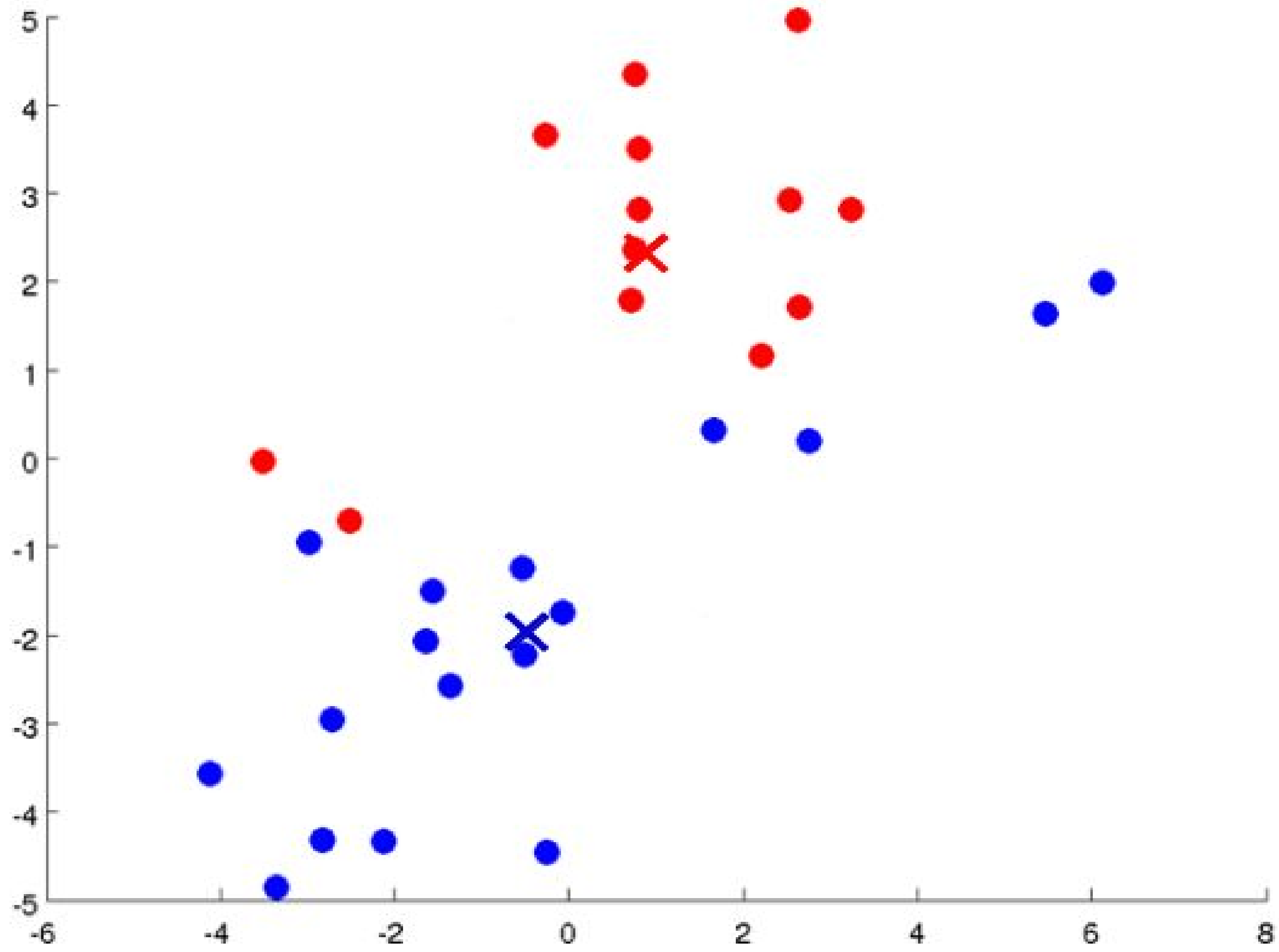
K-Means Algorithm



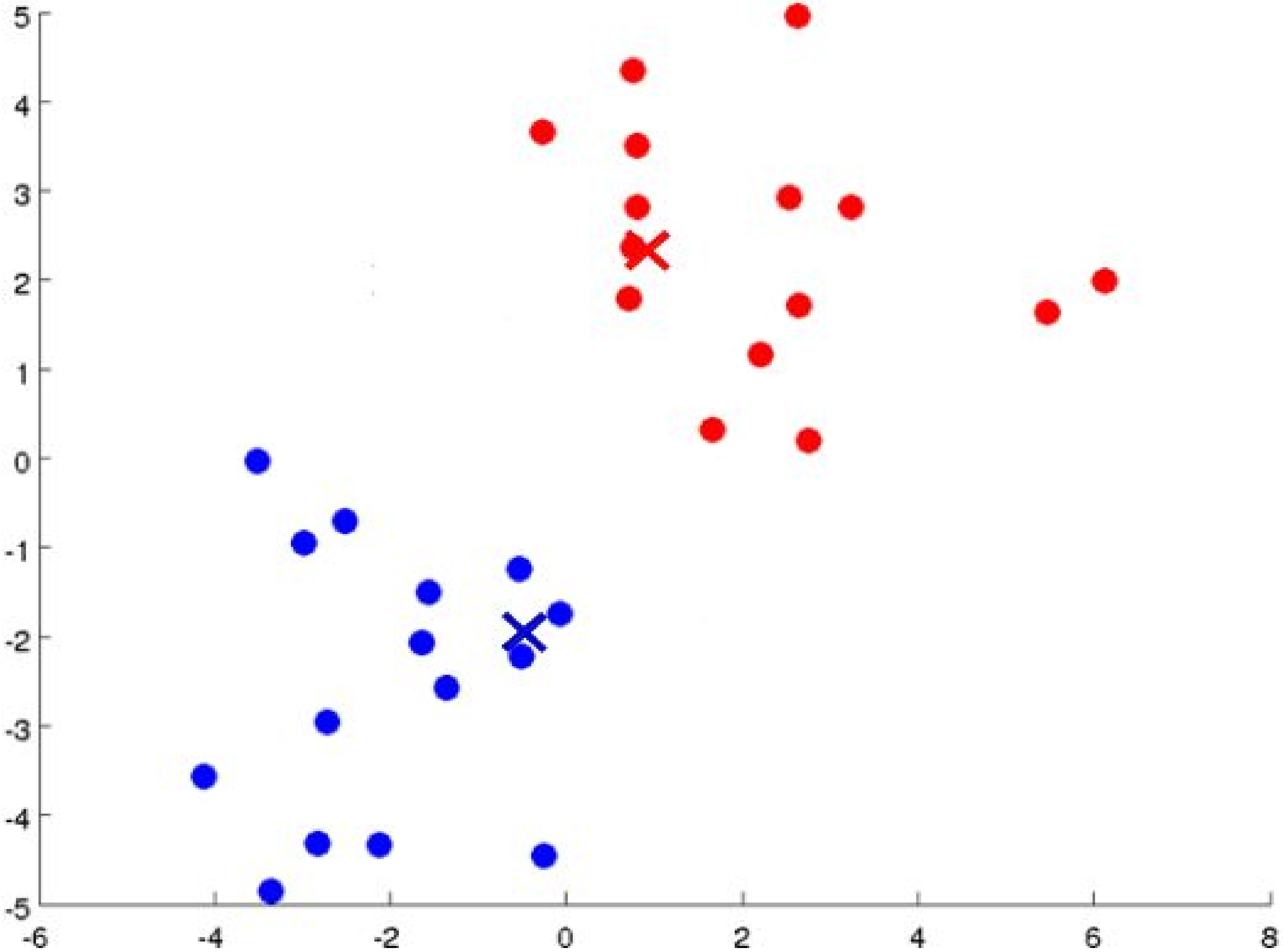
K-Means Algorithm



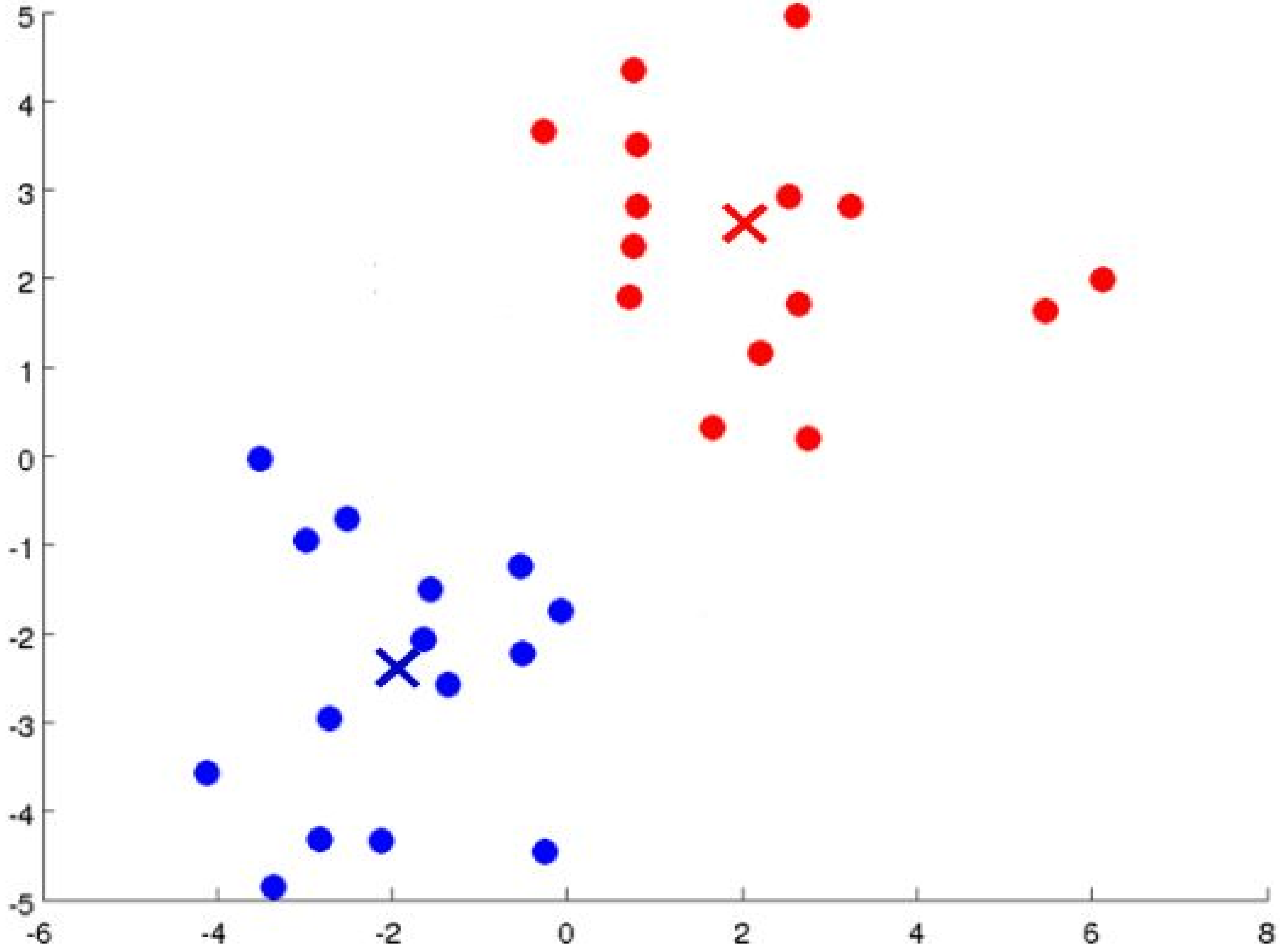
K-Means Algorithm



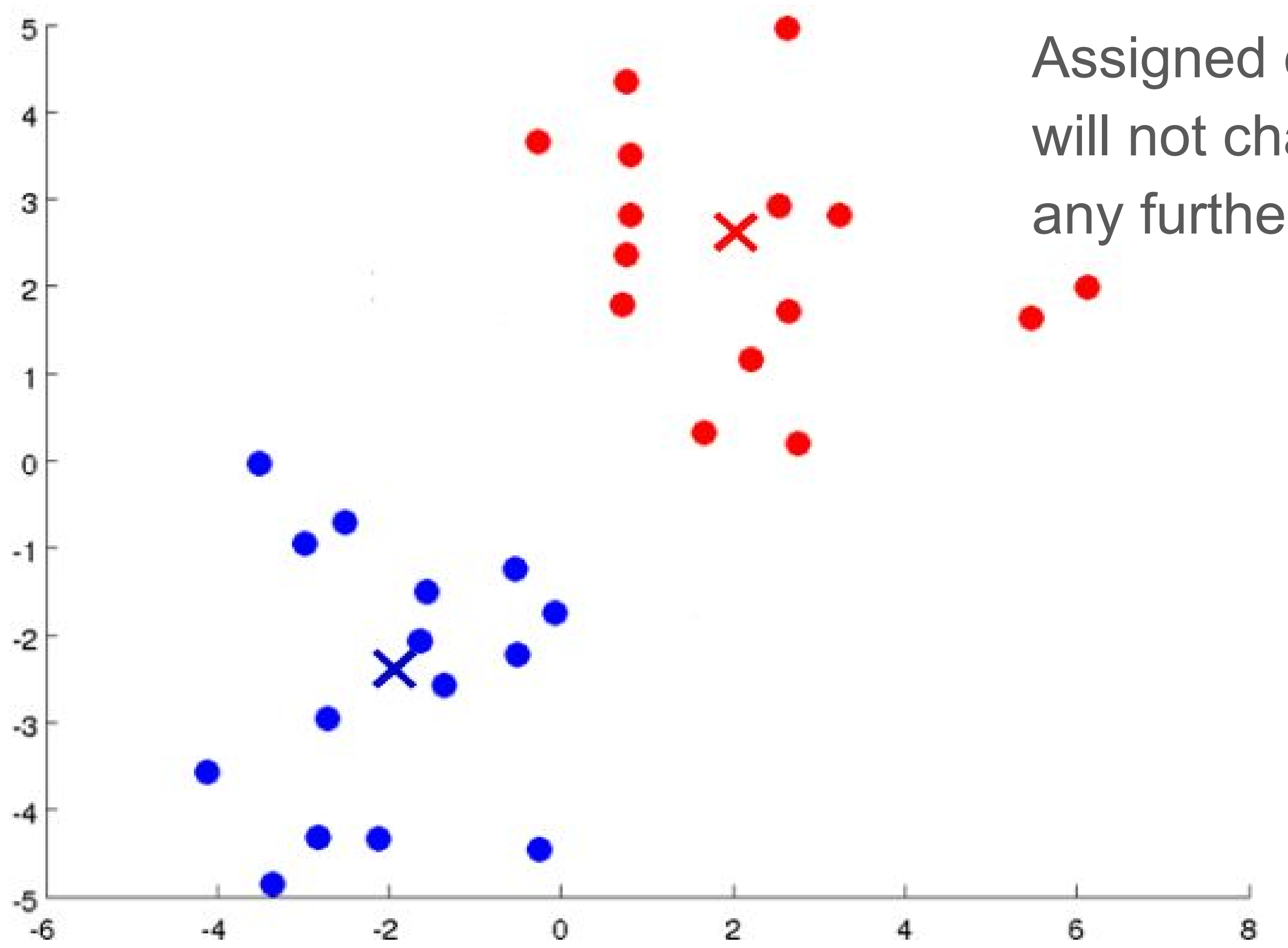
K-Means Algorithm



K-Means Algorithm



K-Means Algorithm



Assigned clusters will not change any further

K-Means Algorithm

- Input:
  - K (number of clusters)
  - Training set: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$
  - n is the number of features
    - e.g. $x_4^{(2)}$: 4th feature of 2^{nd} sample.
    - note: we do not use x_0=1 for K-means

K-Means Algorithm

- Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K$.
- Repeat {

  for i = 1 to m

  $c^{(i)} \coloneqq$ index (from 1 to K) of cluster centroids closest to $x^{(i)}$

  for k = 1 to K

  $\mu_k \coloneqq$ mean (centroid) of points assigned to cluster k

  }

  until centroids stop moving.

Cluster Assignment

Centroid Recalculation

K-Means Algorithm

- If an iteration of the algorithm results in the situation of ‘no sample is assigned to one of the clusters’, i.e. ‘empty cluster’, then you can eliminate that cluster and continue with K-1 clusters.
- If you are sure that there are K clusters, then you need to randomly initialize centroids and run K-means again.

K-Means Optimization Objective

- $c^{(i)}$ = index of cluster (1,2,…, K) to which example $x^{(i)}$ is currently assigned
- μ_k = centroid of cluster k
- $\mu_{c^{(i)}}$ = centroid of cluster to which example $x^{(i)}$ has been assigned
- <u>Optimization objective:</u>

$$J(c^{(1)},\dots,c^{(m)},\mu_{1,\dots},\mu_K)=\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-\mu_{c^{(i)}}\right\|^2$$

$$\min_{\substack{c^{(1)},\dots,c^{(m)}\\ \mu_{1,\dots},\mu_K}} J(c^{(1)},\dots,c^{(m)},\mu_{1,\dots},\mu_K)$$

K-Means Optimization Objective

- One can see that the cost is minimized in the
  - Cluster assignment step by changing $c^{(i)}$
  - Centroid recalculation step by changing μ_k

- Repeat {

  for i = 1 to m

    $c^{(i)}$ ≔ index (from 1 to K) of cluster centroids closest to $x^{(i)}$

  for k = 1 to K

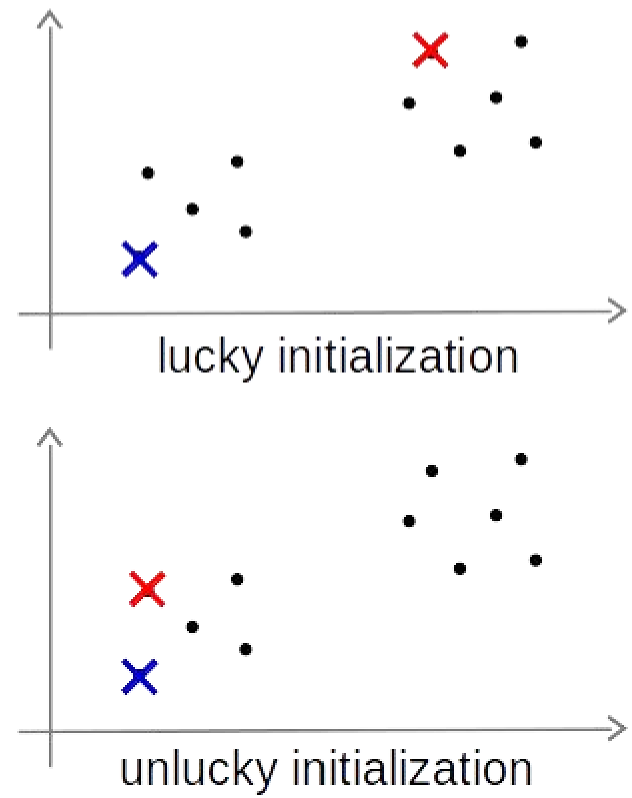
    μ_k ≔ mean (centroid) of points assigned to cluster k

}

until centroids stop moving.

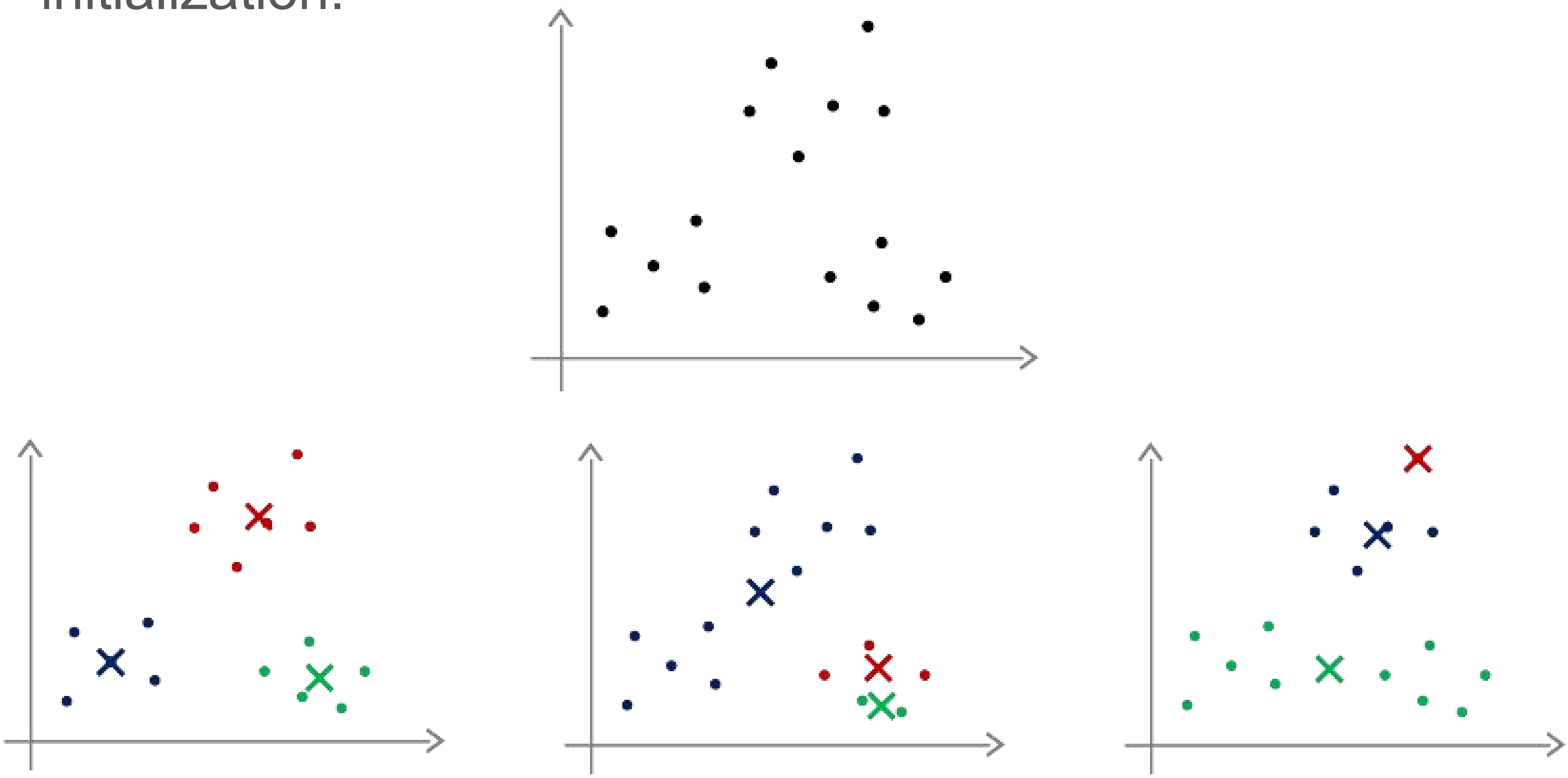
Random Initialization

- The initial centroid locations are randomly picked.
- One way to initialize cluster centroids is randomly picking K training samples and setting μ_1, μ_2, …, μ_K equal to these K samples.
- K-means **can get stuck in local optimum** point depending on the initialization.



Random Initialization

- K-means can get stuck in local optimum point depending on the initialization.



Multiple Random Initialization

For i = 1 to 100 {

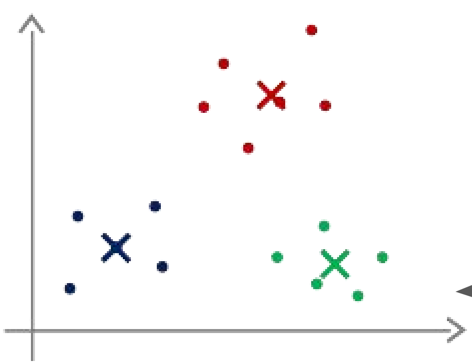
Randomly initialize K-means.

Run K-means. Get $c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_K$

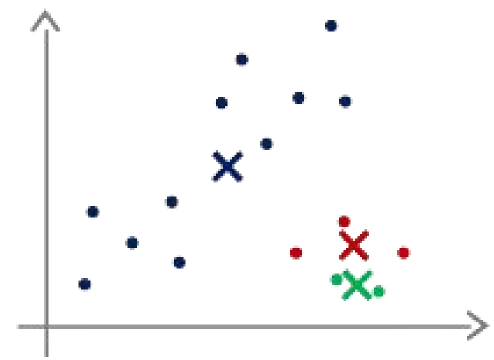
Compute cost function $J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_K)$

}

Pick clustering that gave lowest cost J

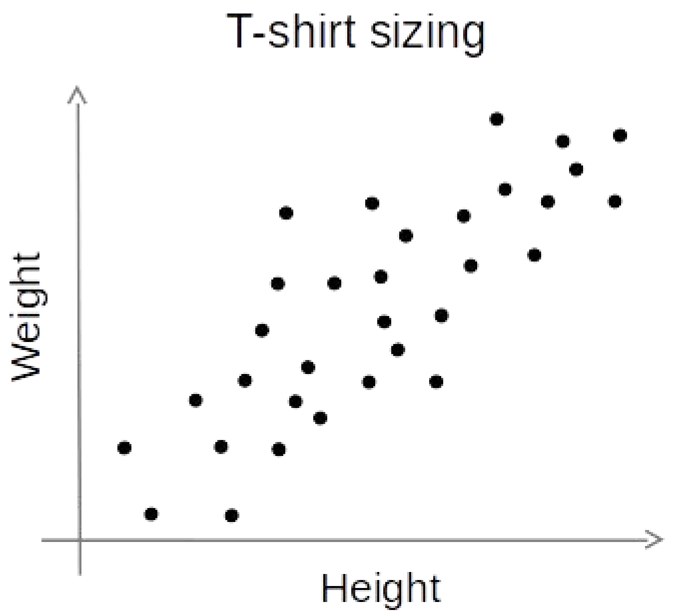


Better clustering (left one) will end up with lower cost



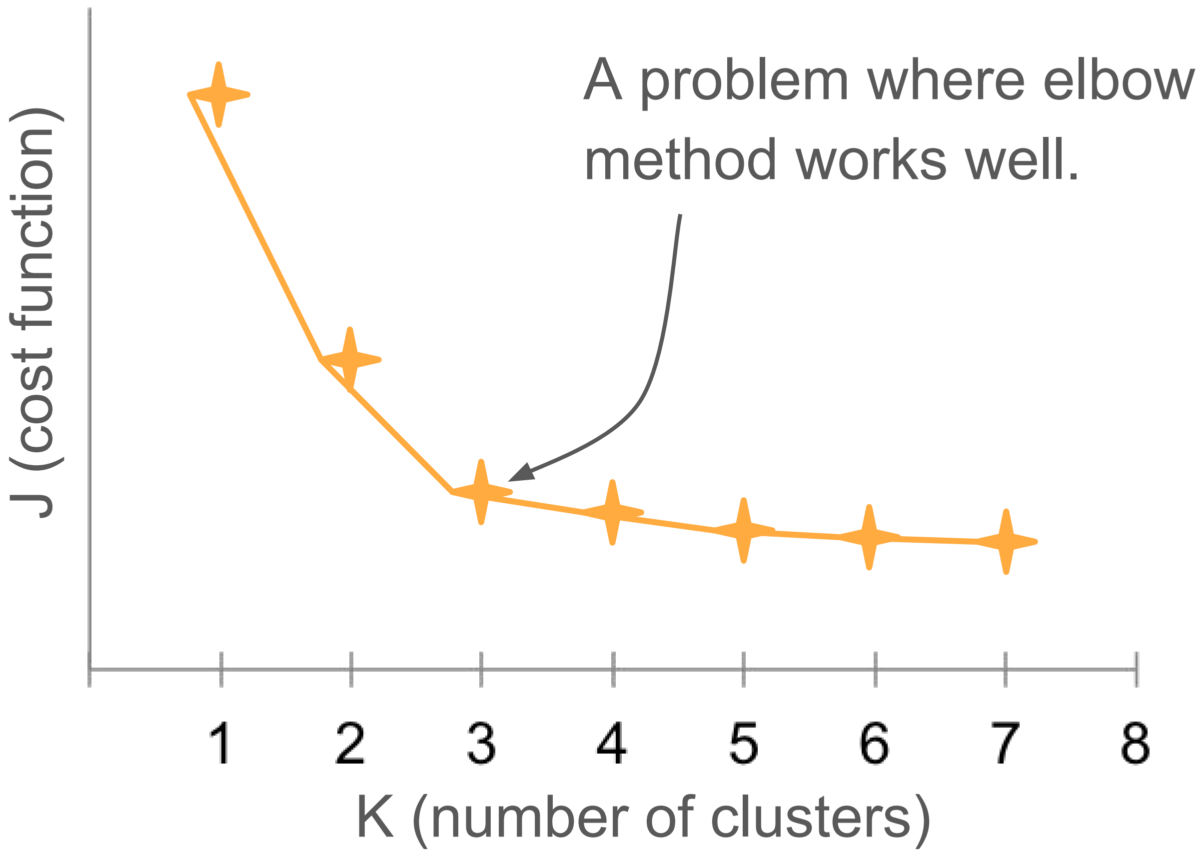
Choosing K

- For non-well-separated clusters, what is the right value of K ?



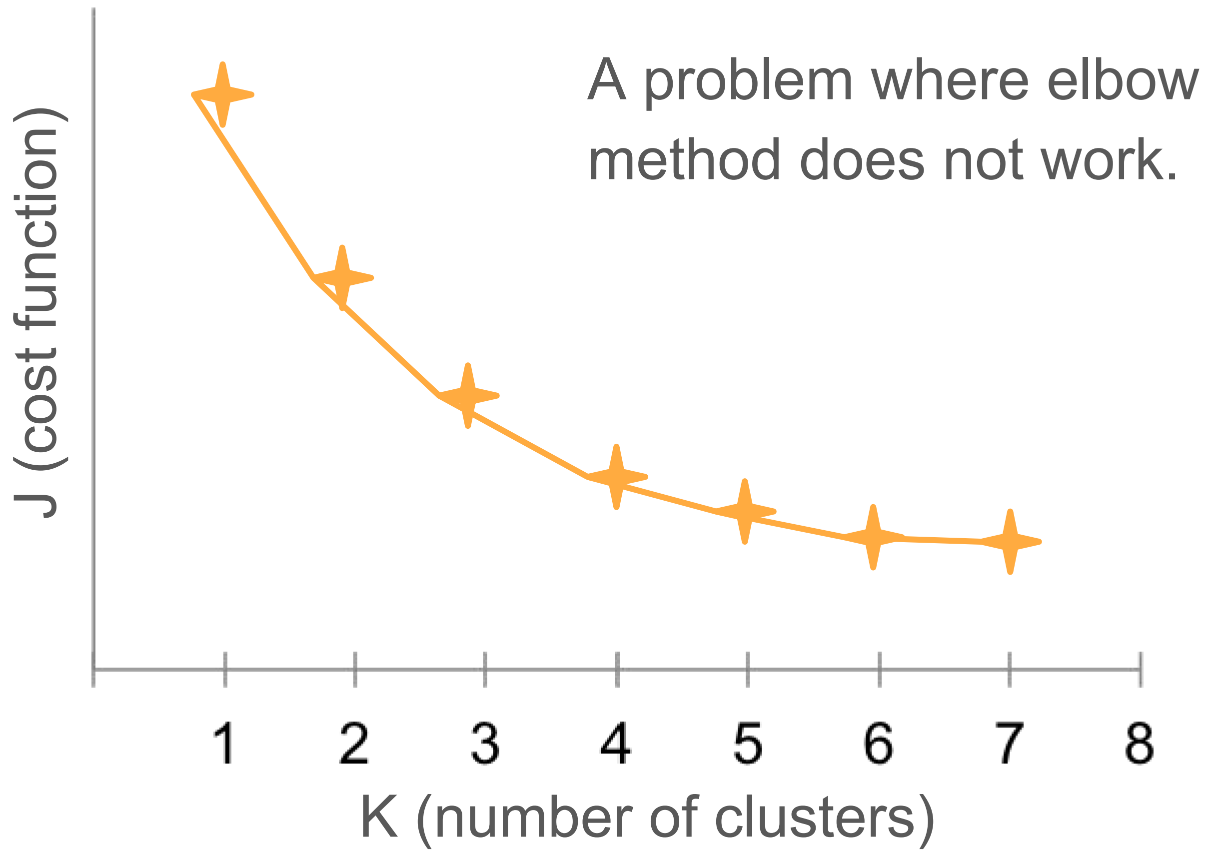
Choosing K

- Elbow method:



Choosing K

- Elbow method:



Choosing K

- Usually K is selected manually considering the clustering purpose.
- If you can find a metric to evaluate the needs of your problem (production cost, customer satisfaction etc.), use it to choose K.