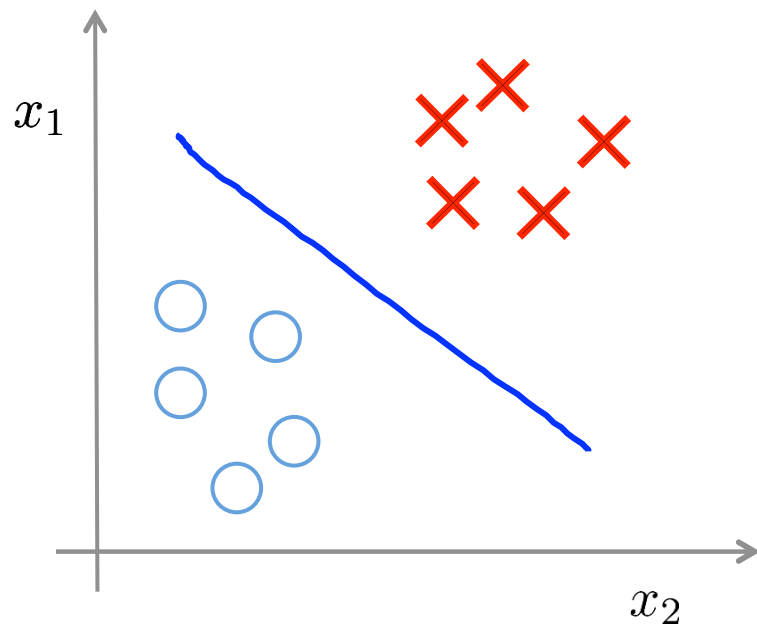


Machine Learning

Clustering

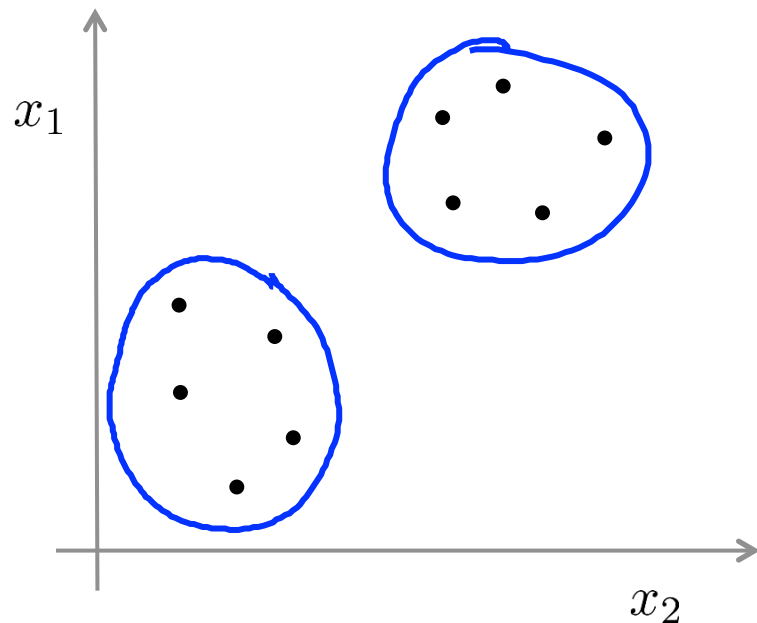
Unsupervised learning
introduction

Supervised learning



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

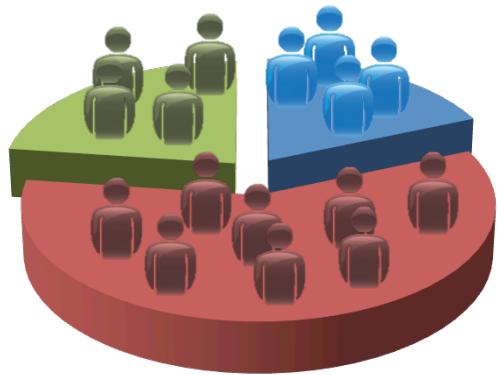
Unsupervised learning



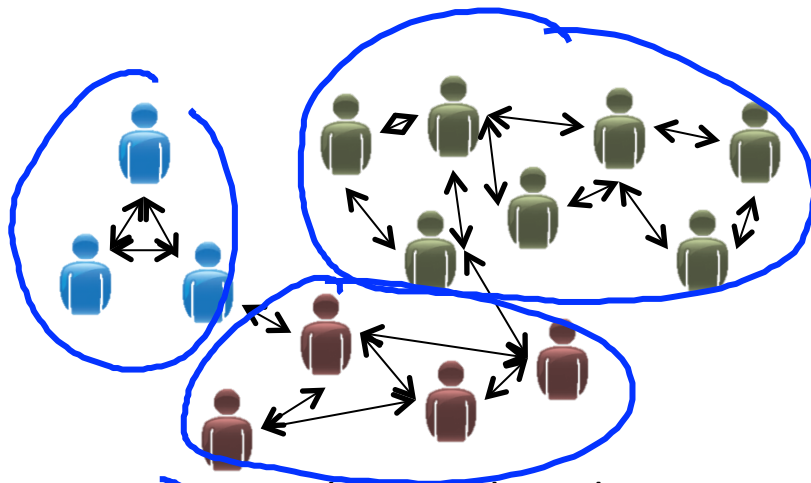
Clustering algorithm

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

Applications of clustering



→ Market segmentation



→ Social network analysis



→ Organize computing clusters

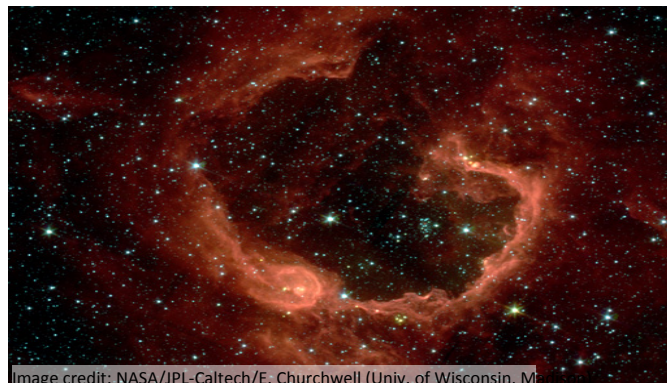
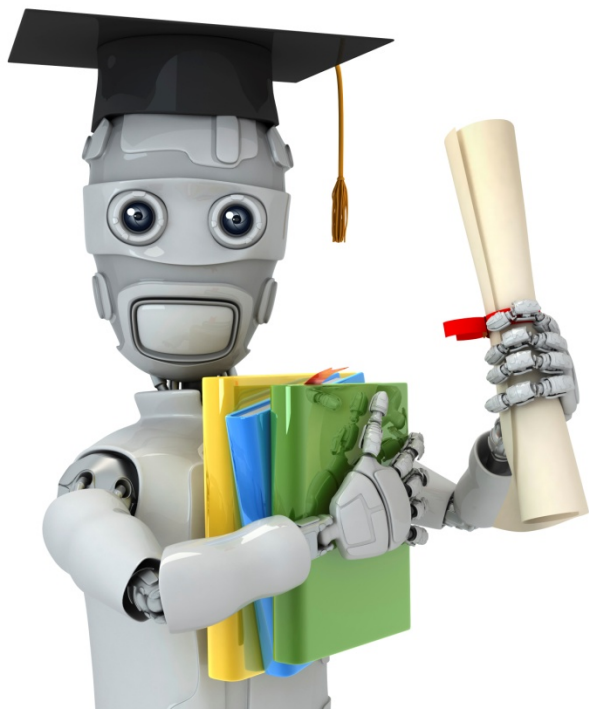


Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, M

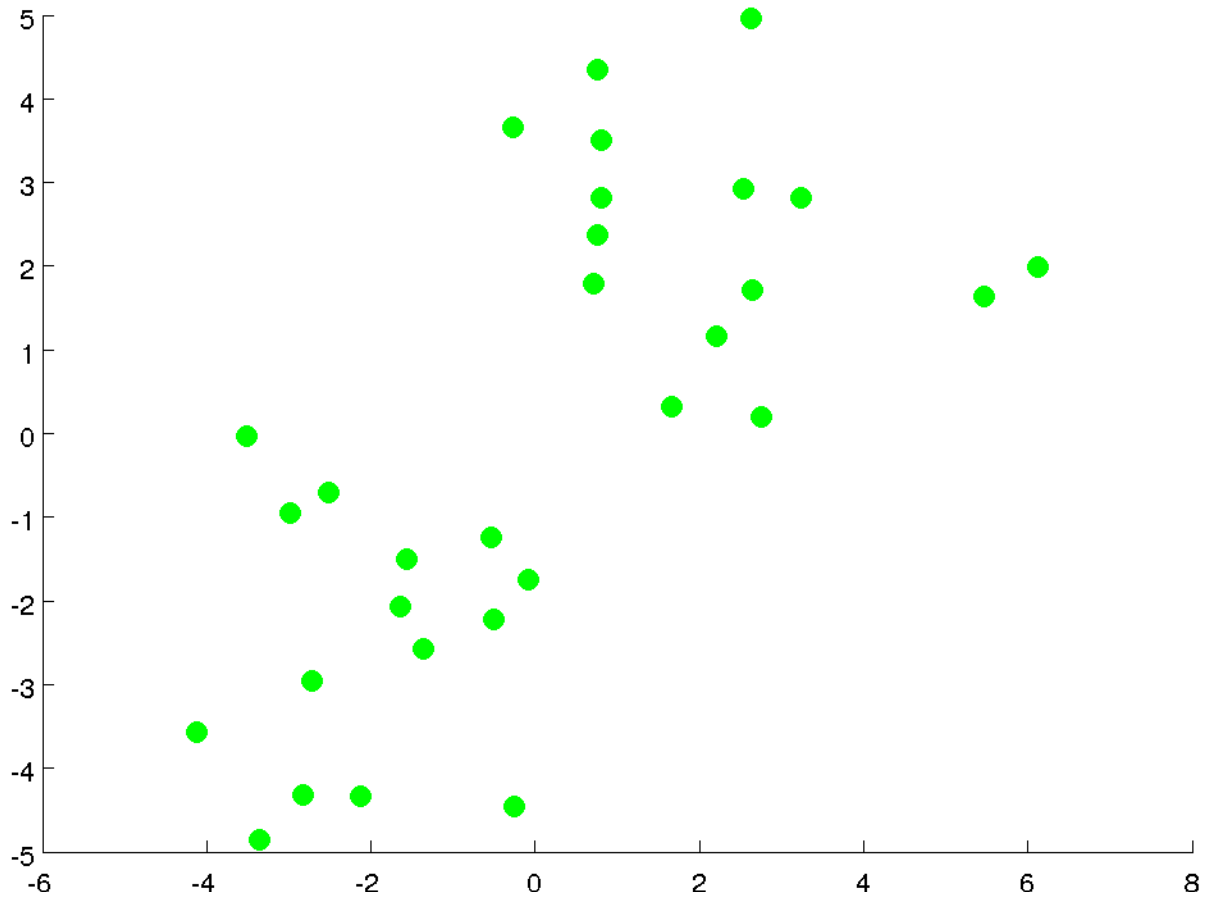
→ Astronomical data analysis

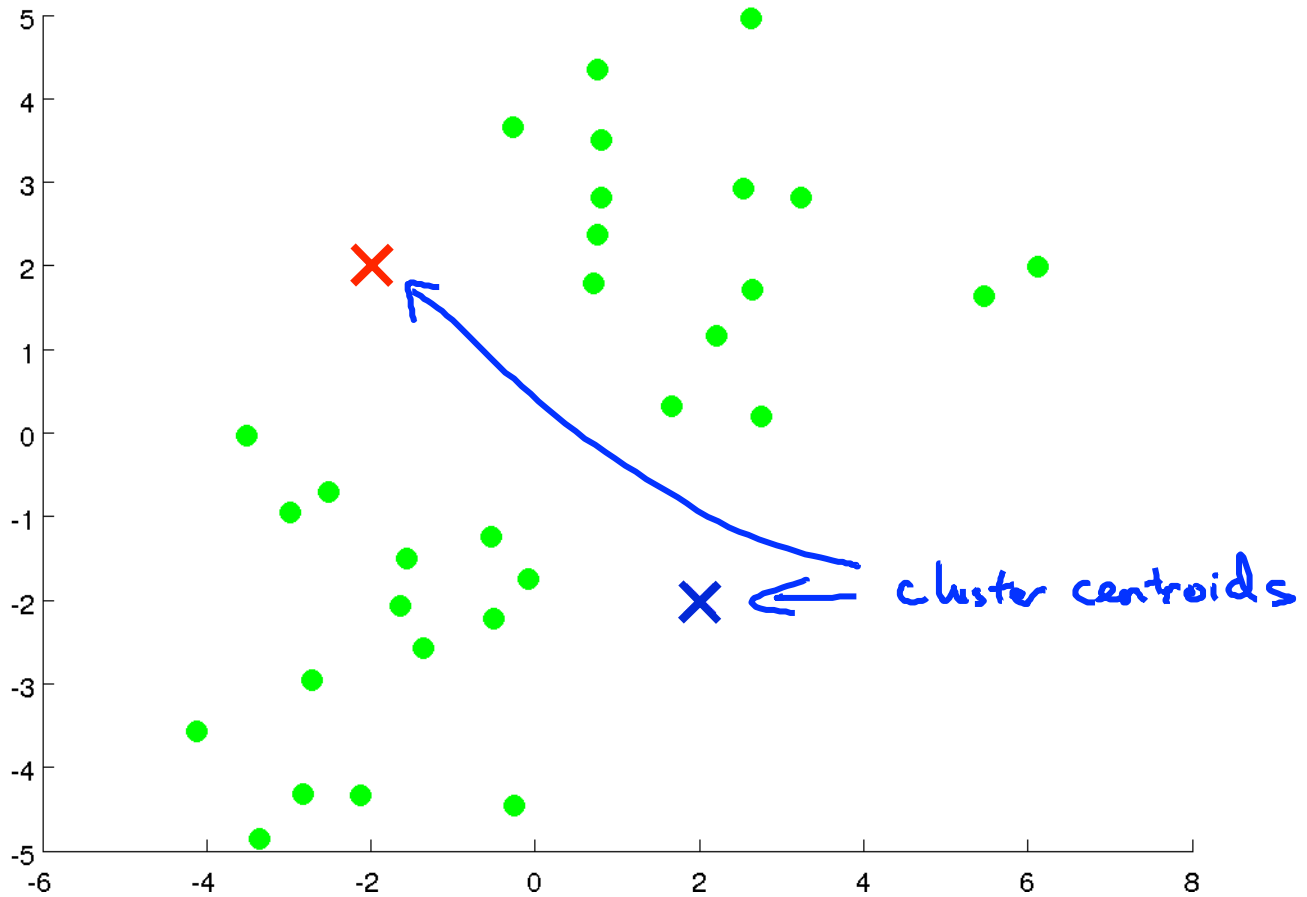


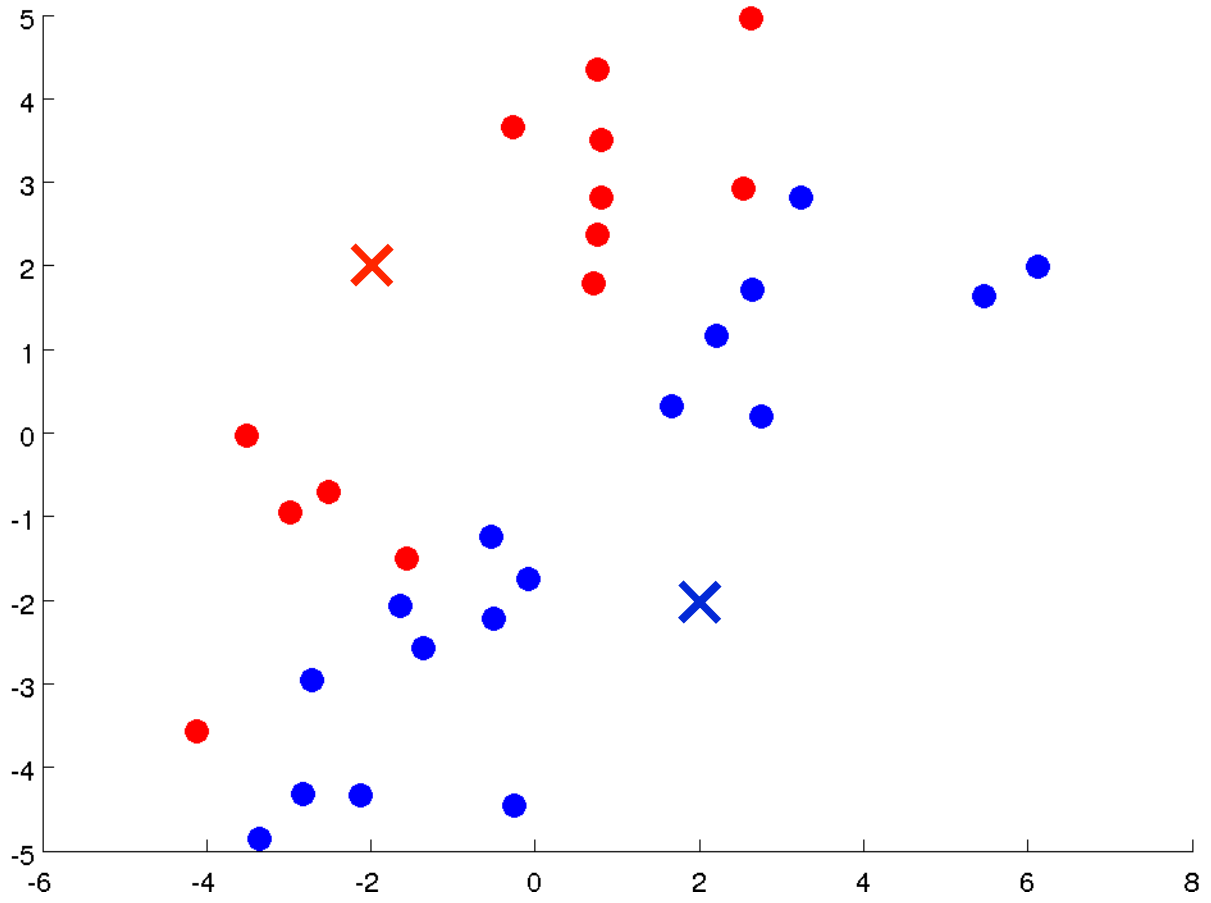
Machine Learning

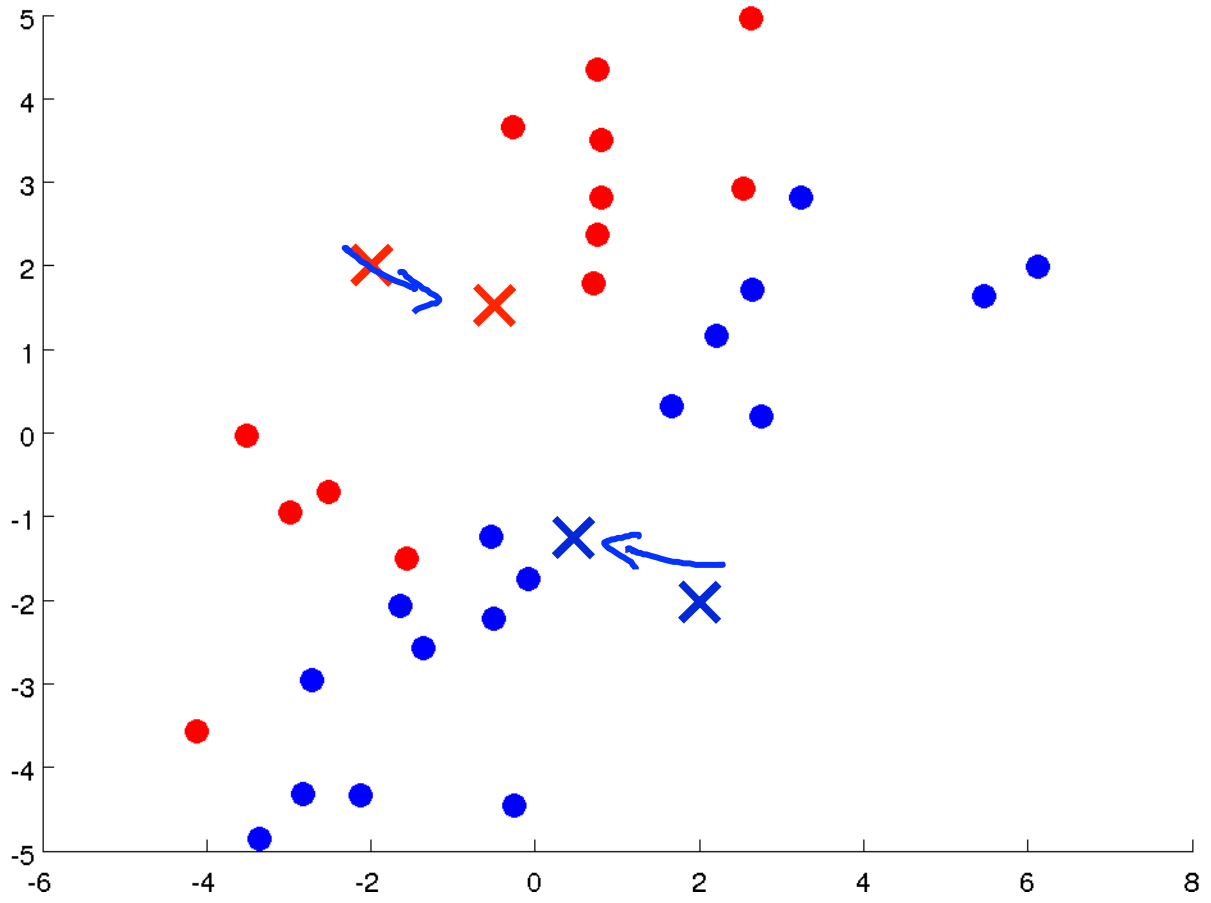
Clustering

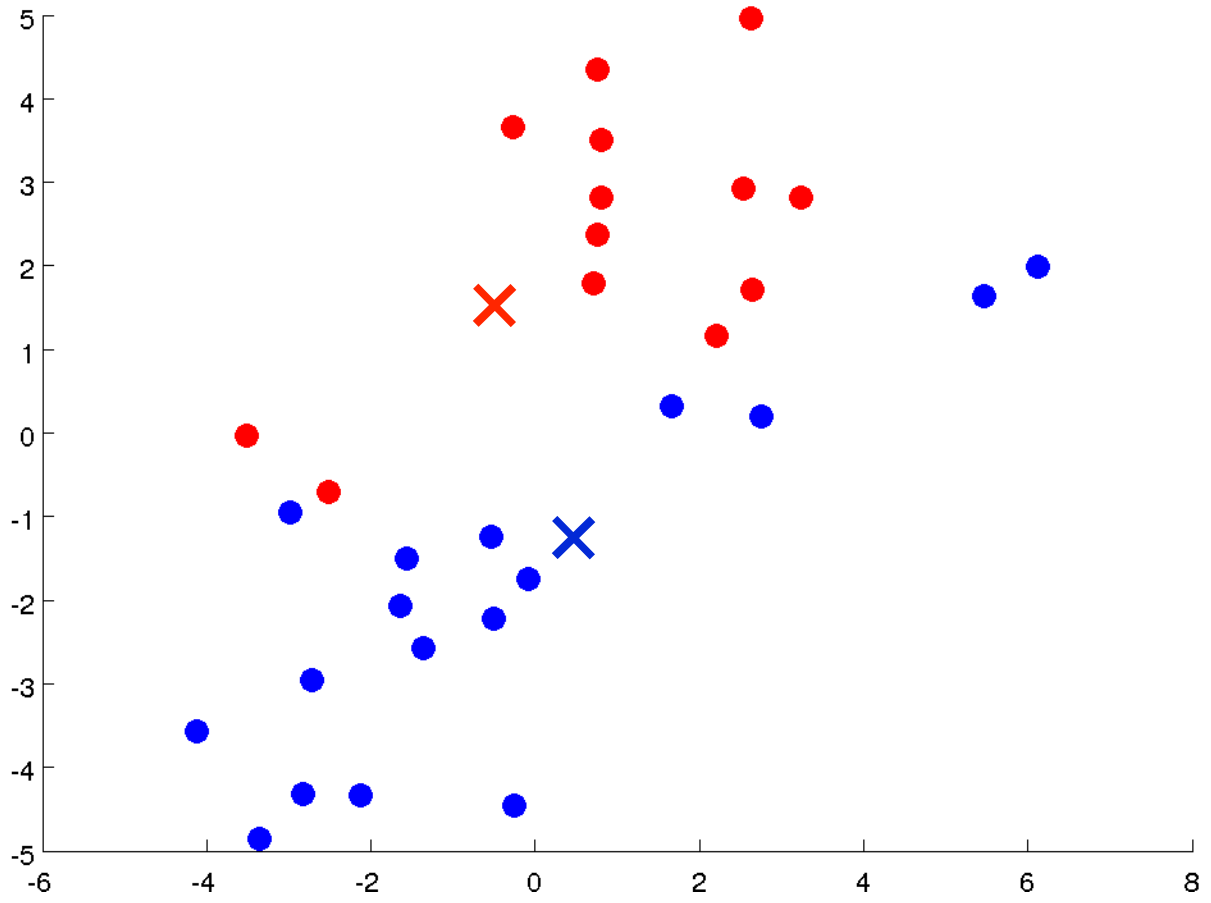
K-means
algorithm

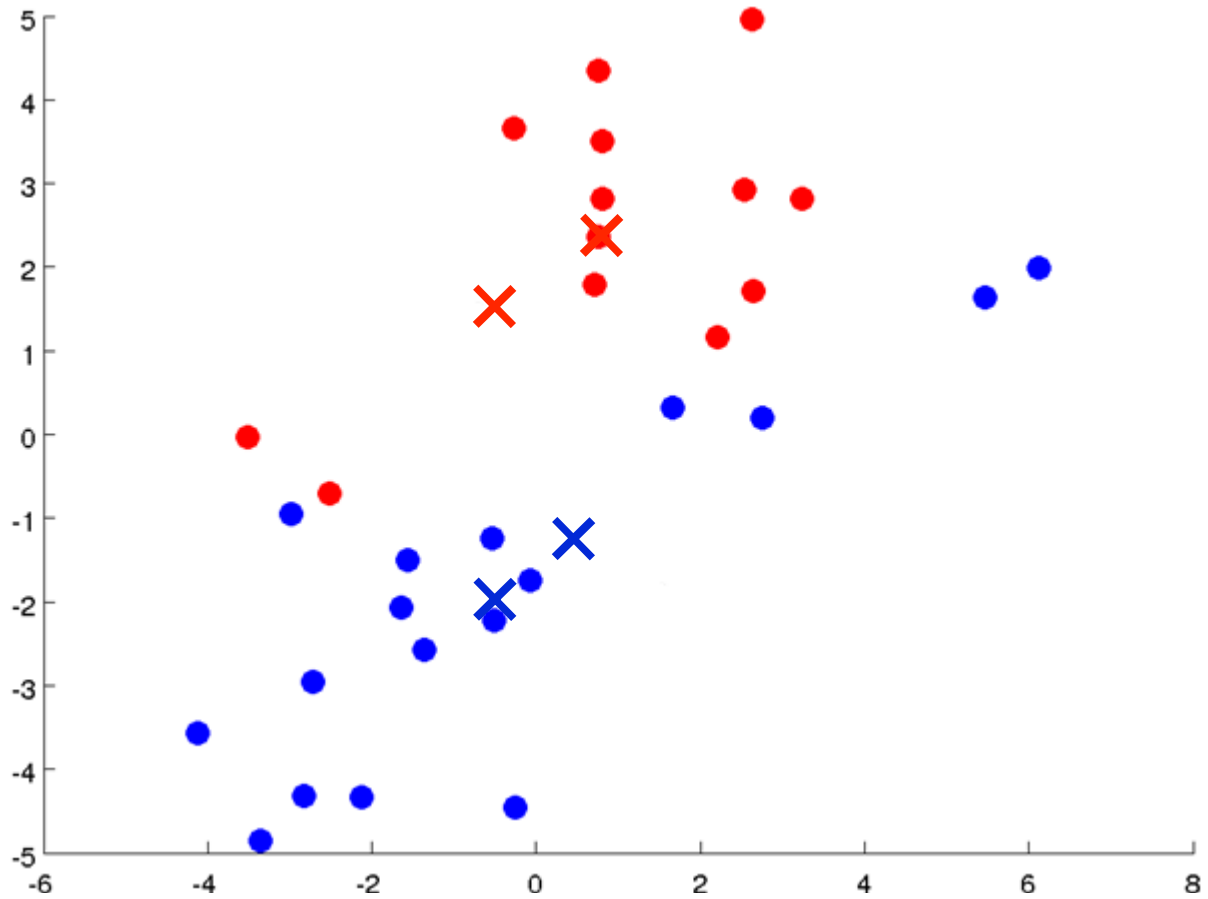


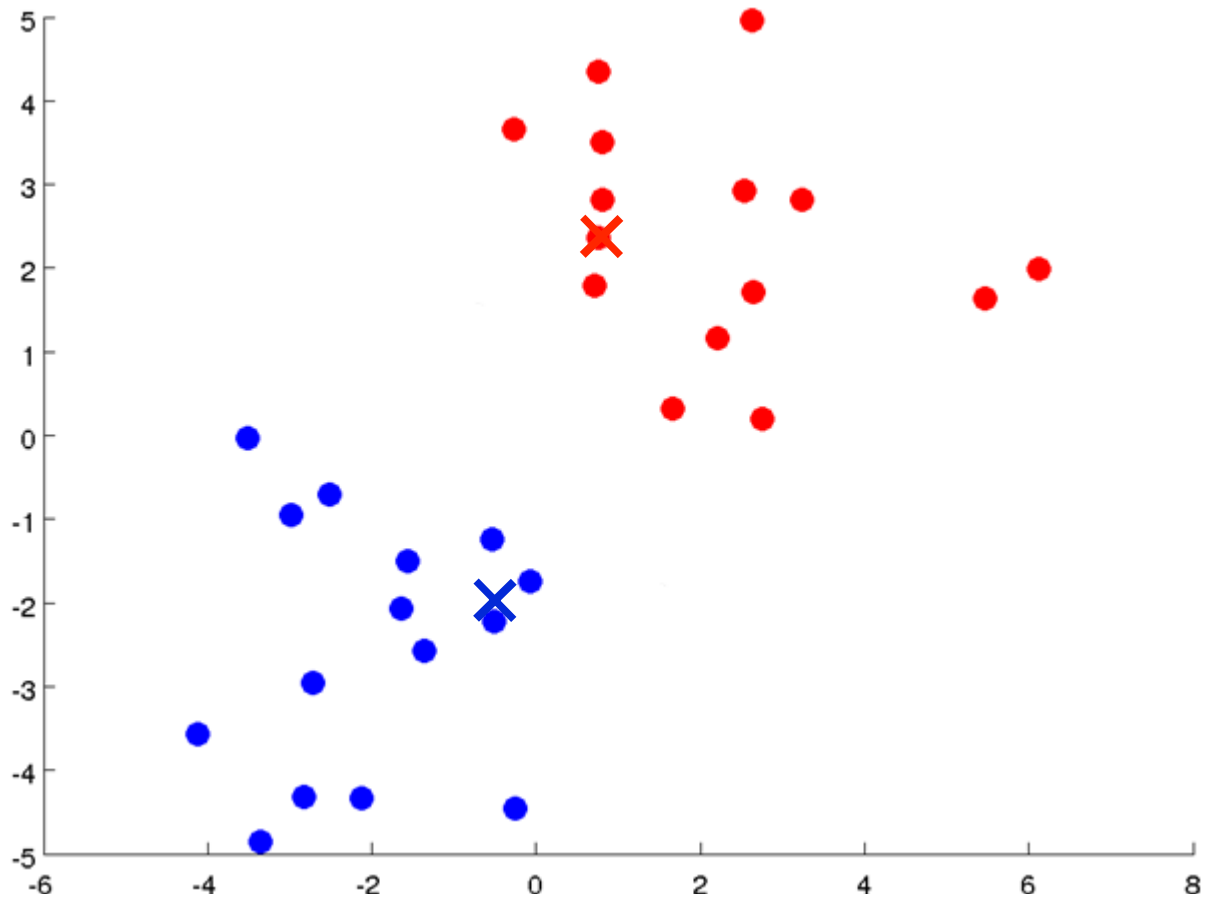


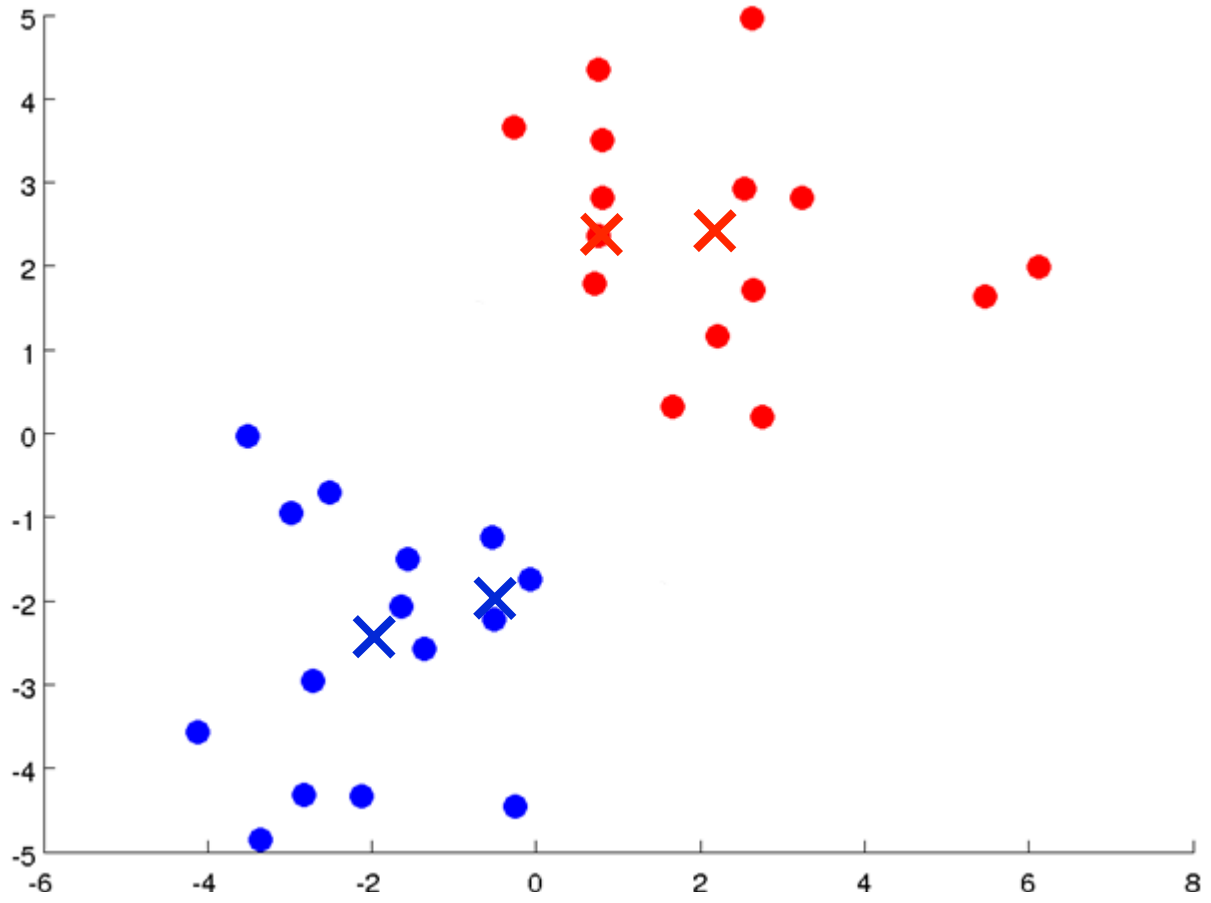


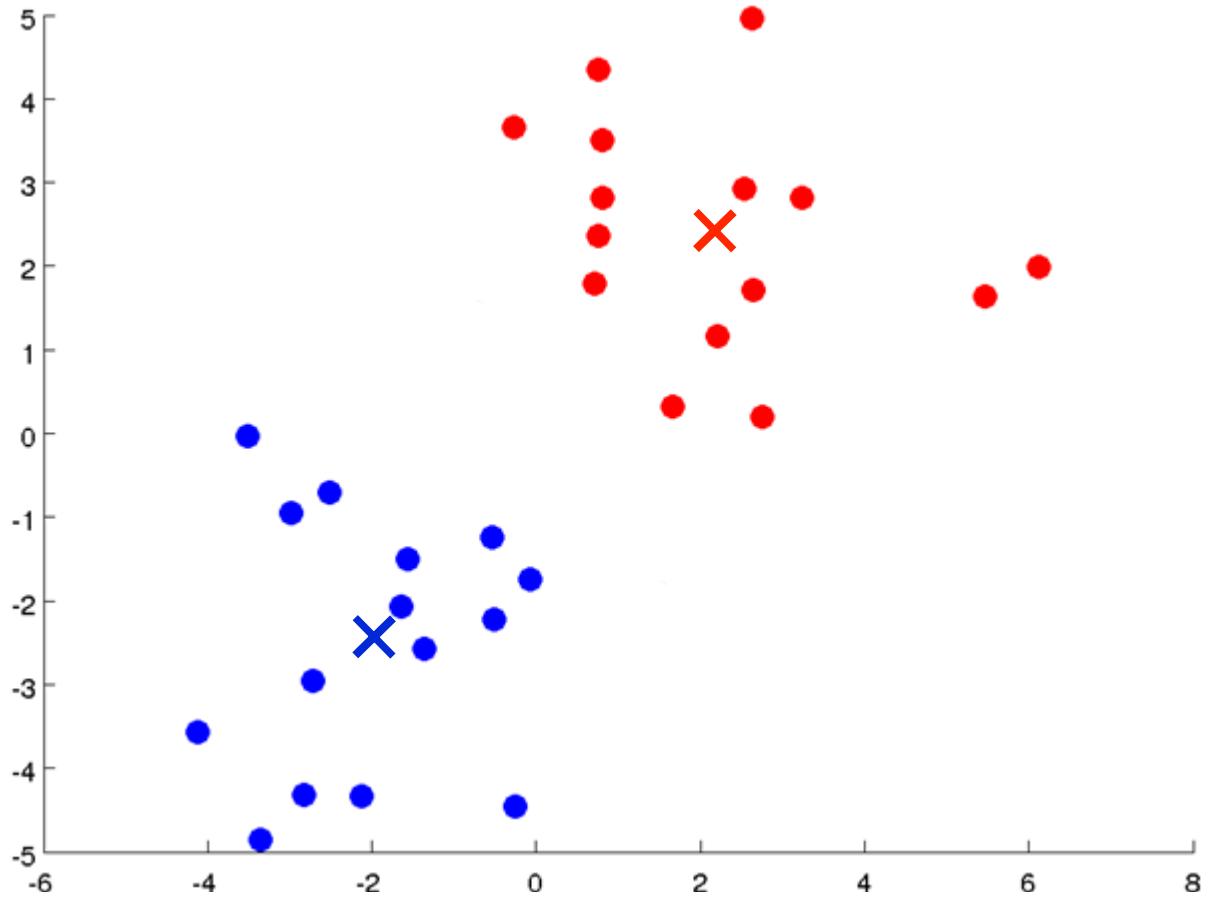














K-means algorithm

Input:

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

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

K-means algorithm

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

Repeat {

 for $i = 1$ to m

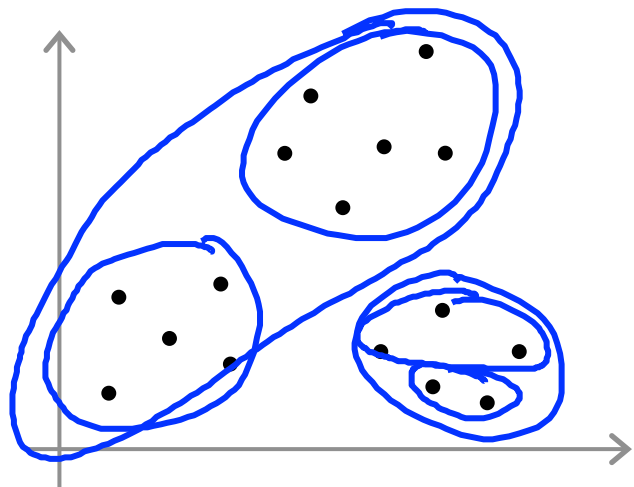
$c^{(i)} :=$ index (from 1 to K) of cluster centroid
 closest to $x^{(i)}$

 for $k = 1$ to K

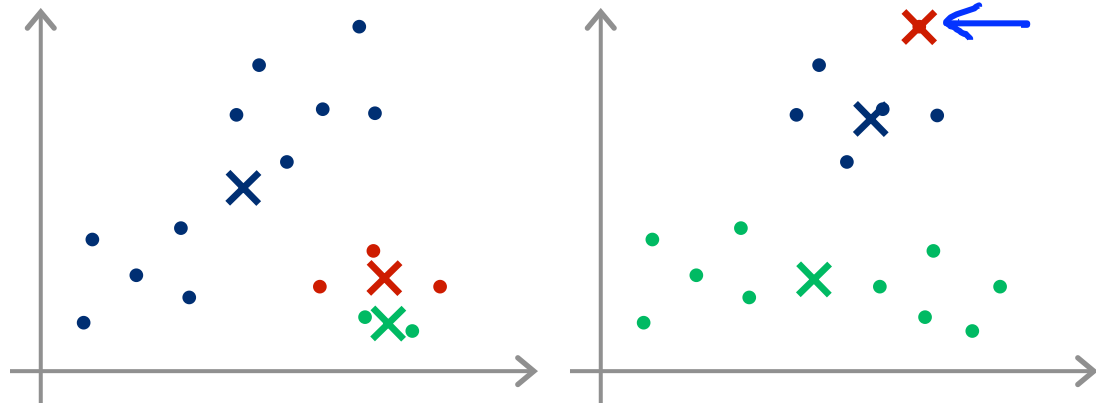
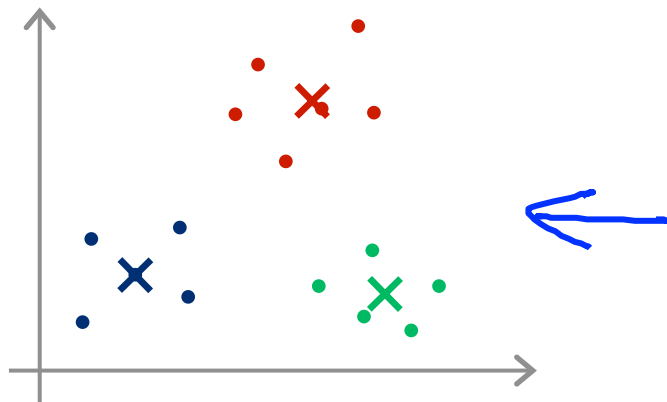
$\mu_k :=$ average (mean) of points assigned to cluster k

}

Local optima



$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_k)$$



Random initialization

For $i = 1$ to 100 {

Randomly initialize K-means.

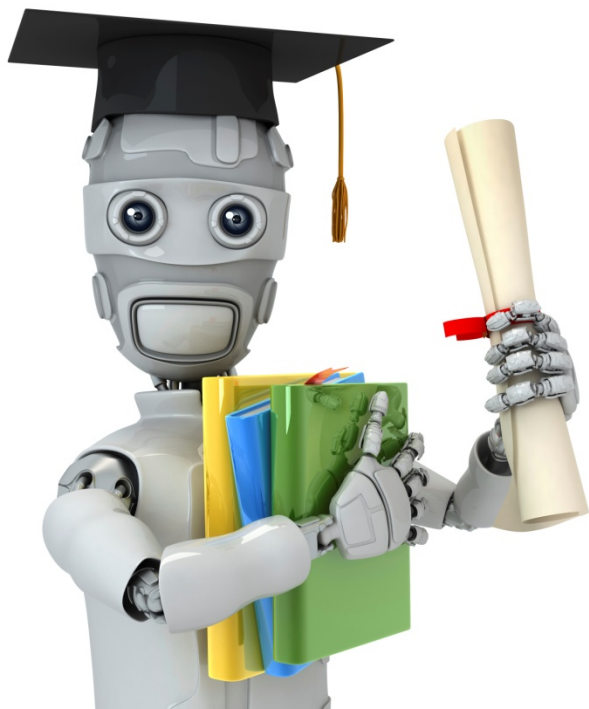
Run K-means. Get $c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K$.

Compute cost function (distortion)

$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$$

}

Pick clustering that gave lowest cost $J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$

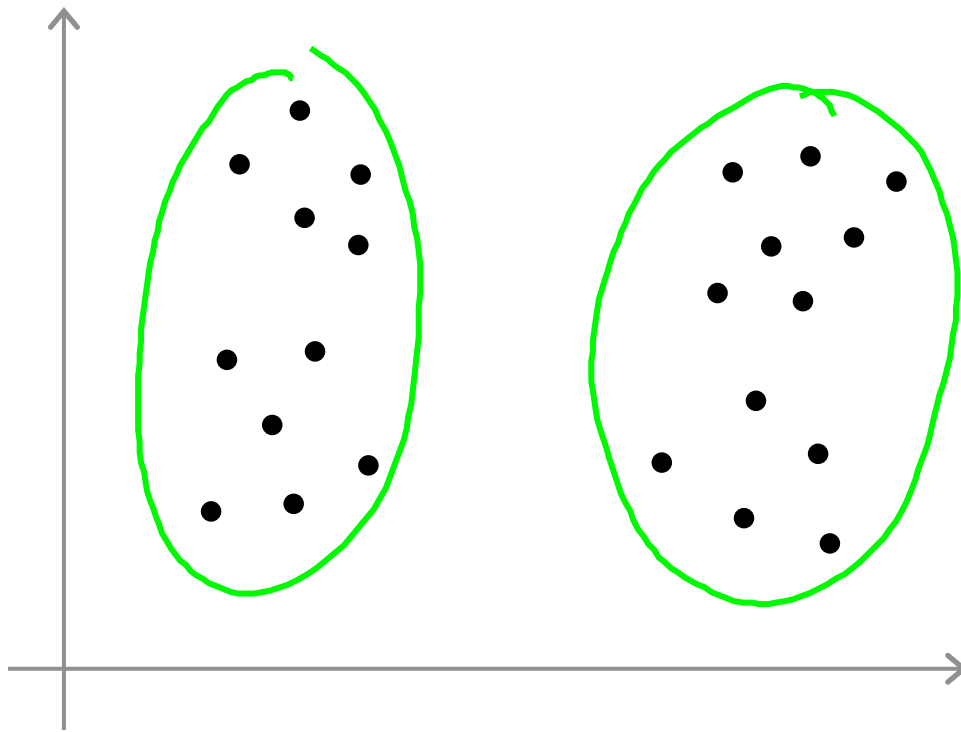


Machine Learning

Clustering

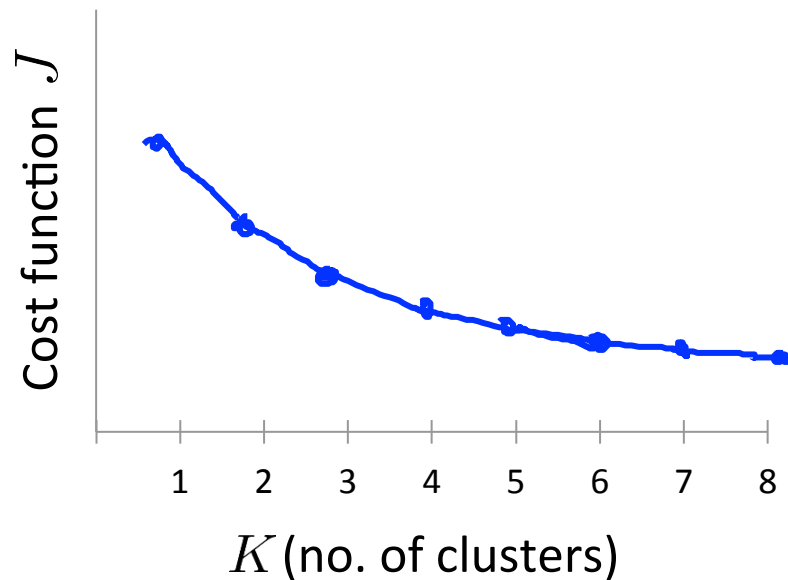
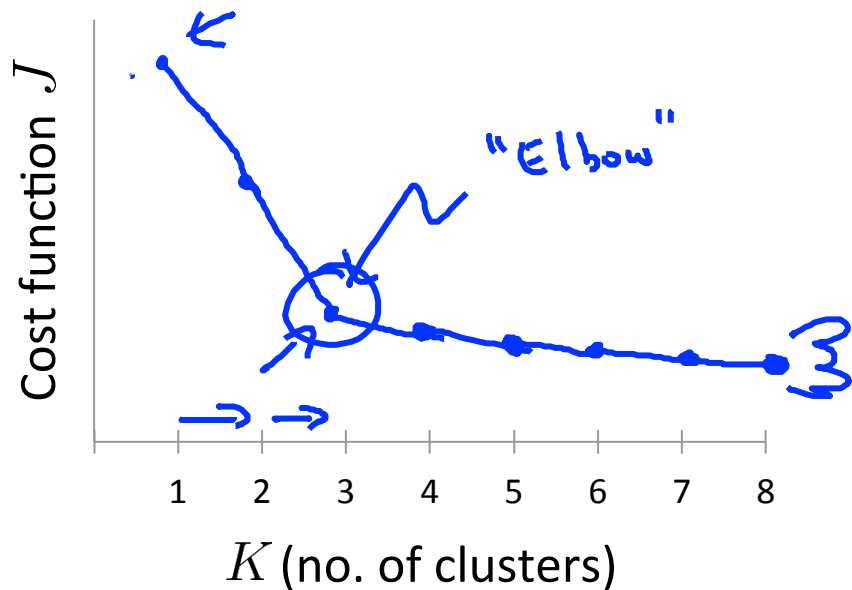
Choosing the
number of clusters

What is the right value of K?



Choosing the value of K

Elbow method:

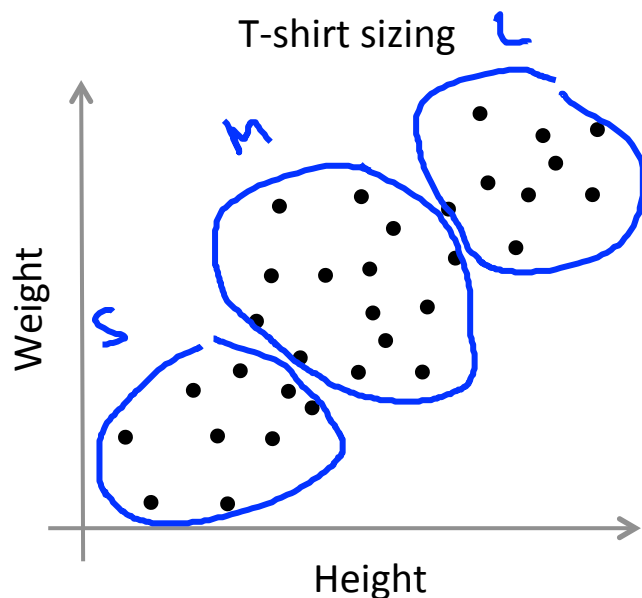


Choosing the value of K

Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

$K=3$ S, M, L

E.g.



$K=5$ XS, S, M, L, XL

