بسم الله الرحمن الرحيم

BISMILLAH ARRAHMAN ARRAHEEM

Artificial Intelligence (CS-401)

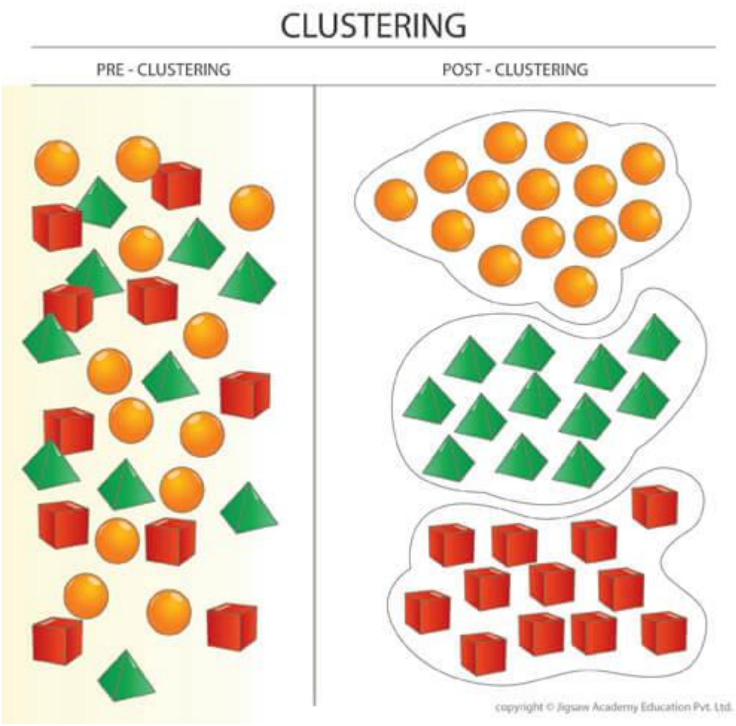
CLUSTERING-K MEANS

Dr. Fahad Sherwani (Assistant Professor – FAST NUCES)
PhD in Artificial Intelligence
Universiti Tun Hussein Onn Malaysia
fahad.sherwani@nu.edu.pk

Clustering

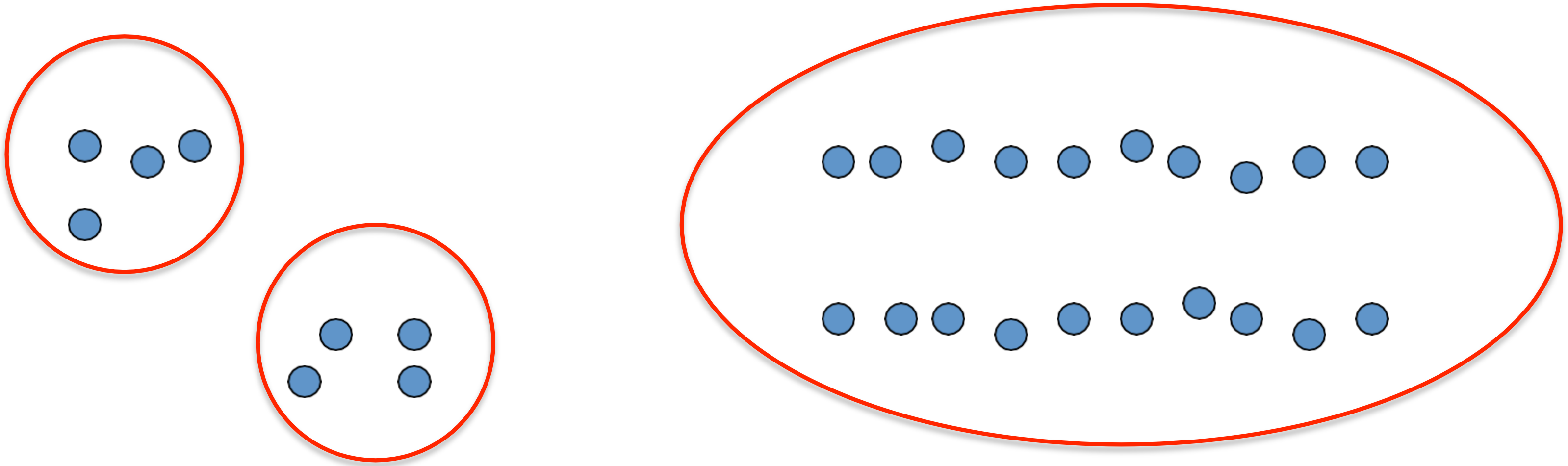
Clustering:

- – Unsupervised learning
- – Requires data, but no labels
- – Detect patterns e.g. in
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- – Useful when don’t know what you’re looking for
- – But: can get gibberish



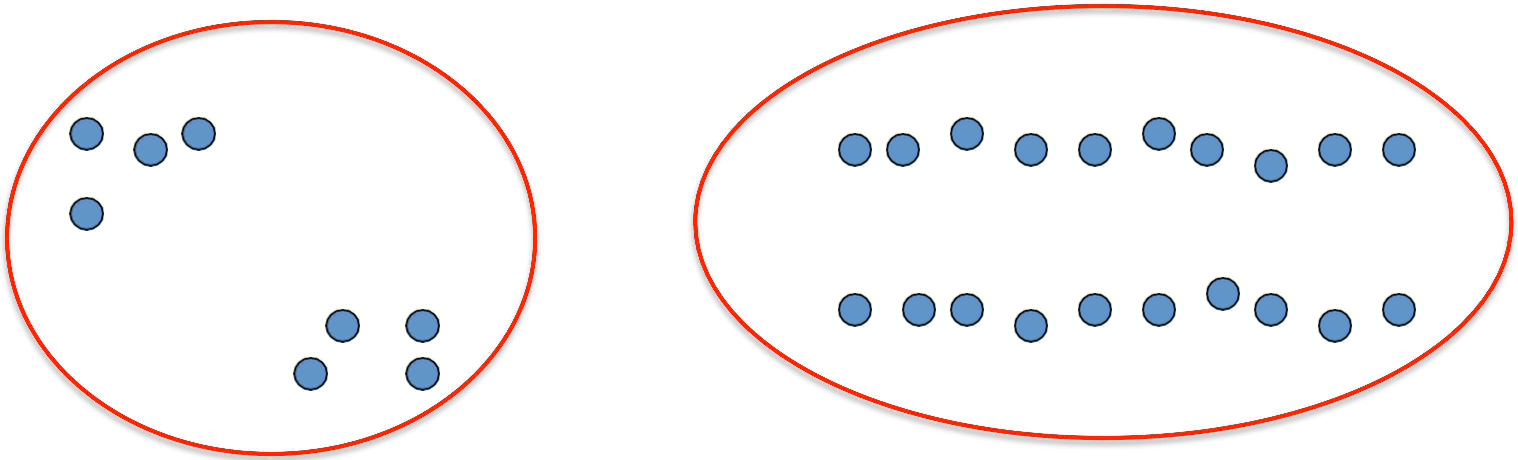
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



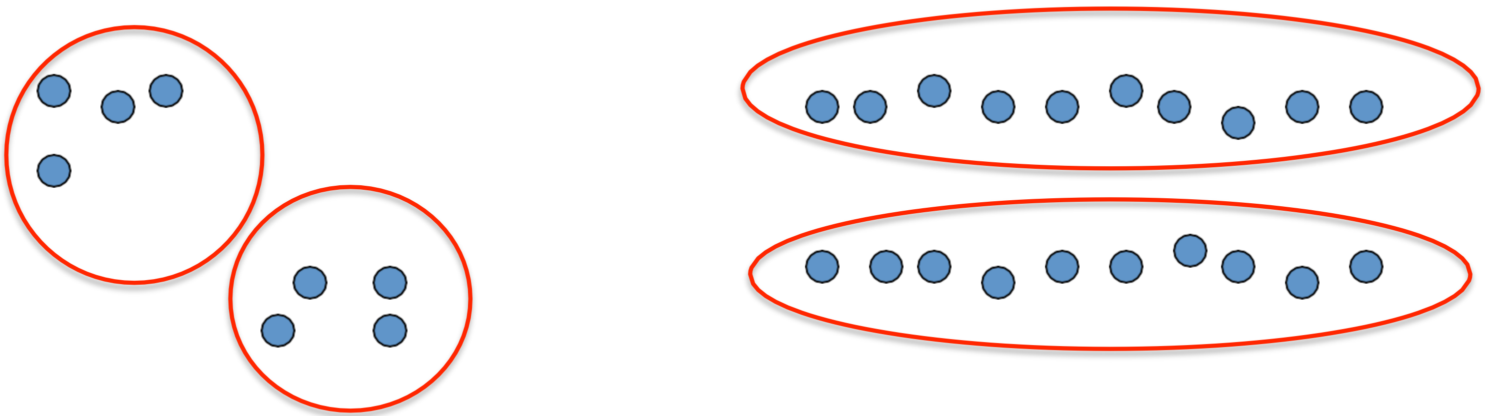
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



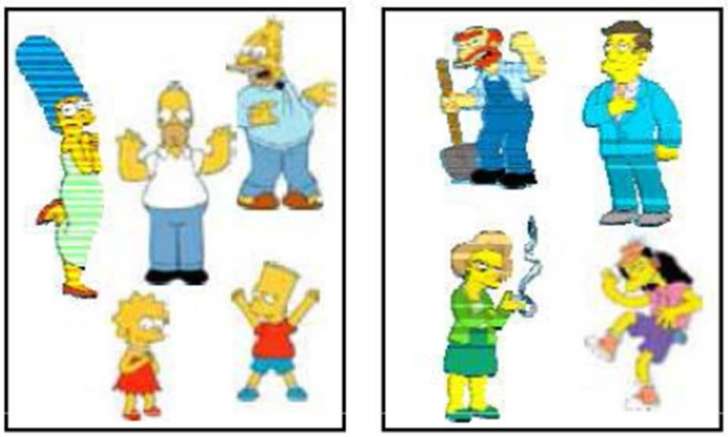
- What could “similar” mean?
  – One option: small Euclidean distance (squared)

$$\mathrm{dist}(x, y) = \|x - y\|_2^2$$

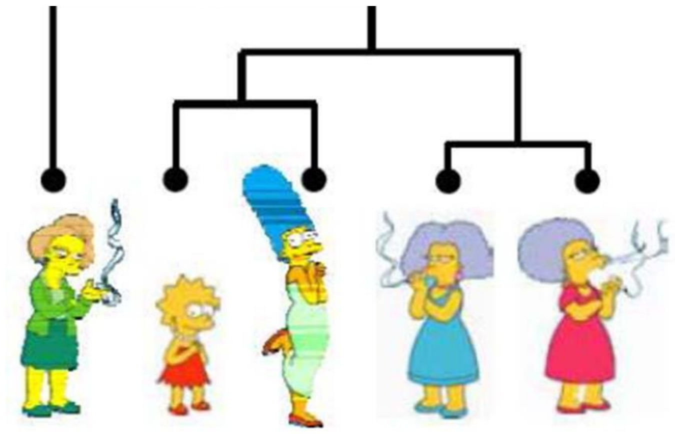
  – Clustering results are crucially dependent on the measure of similarity (or distance) between “points” to be clustered

Clustering algorithms

- Partition algorithms (Flat)
  - K-means
  - Mixture of Gaussian
  - Spectral Clustering



- Hierarchical algorithms
  - Bottom up – agglomerative
  - Top down – divisive



Clustering examples

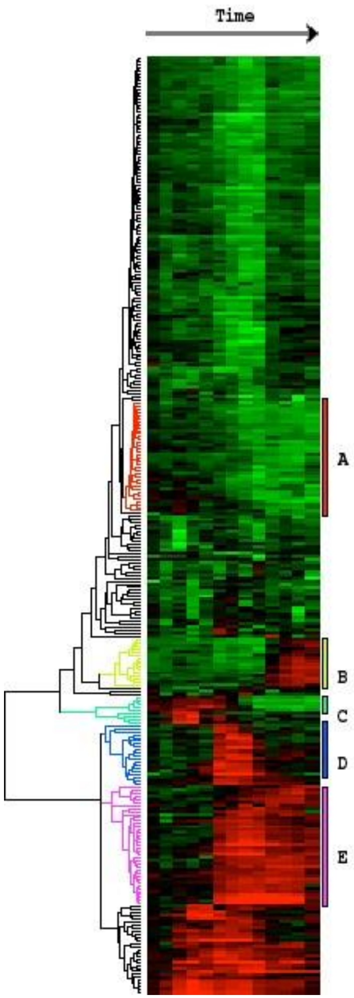
Image segmentation

Goal: Break up the image into meaningful or perceptually similar regions

[Slide from James Hayes]

Clustering examples

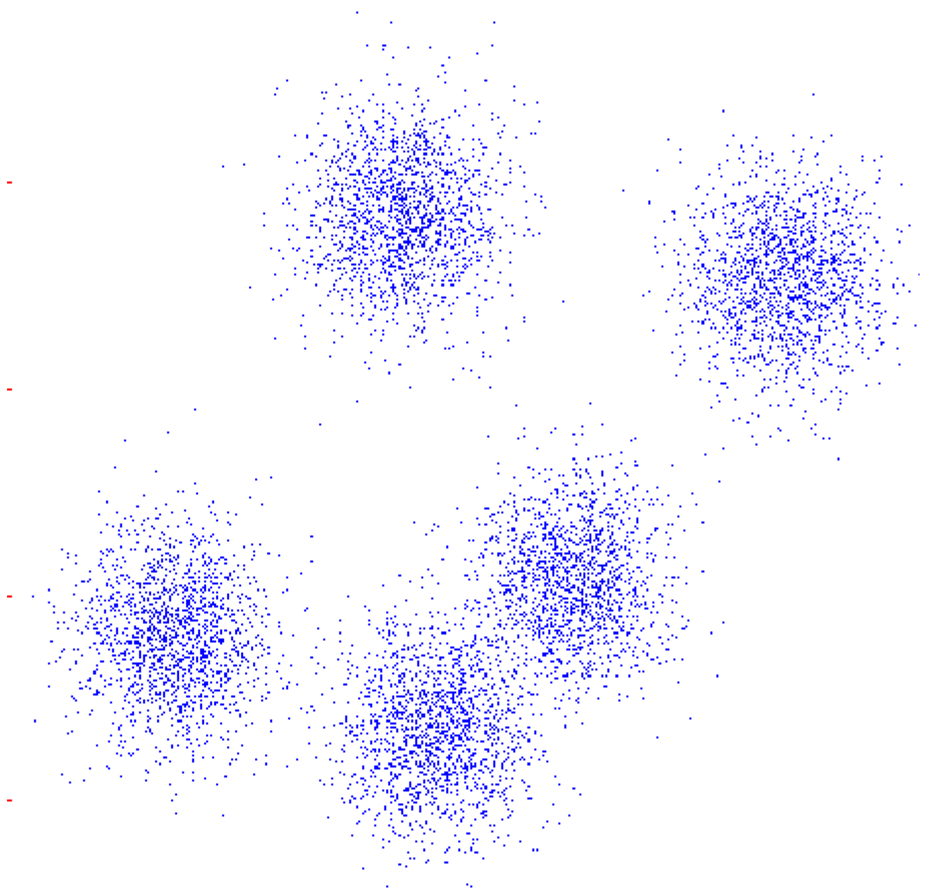
Clustering gene expression data



Eisen et al, PNAS 1998

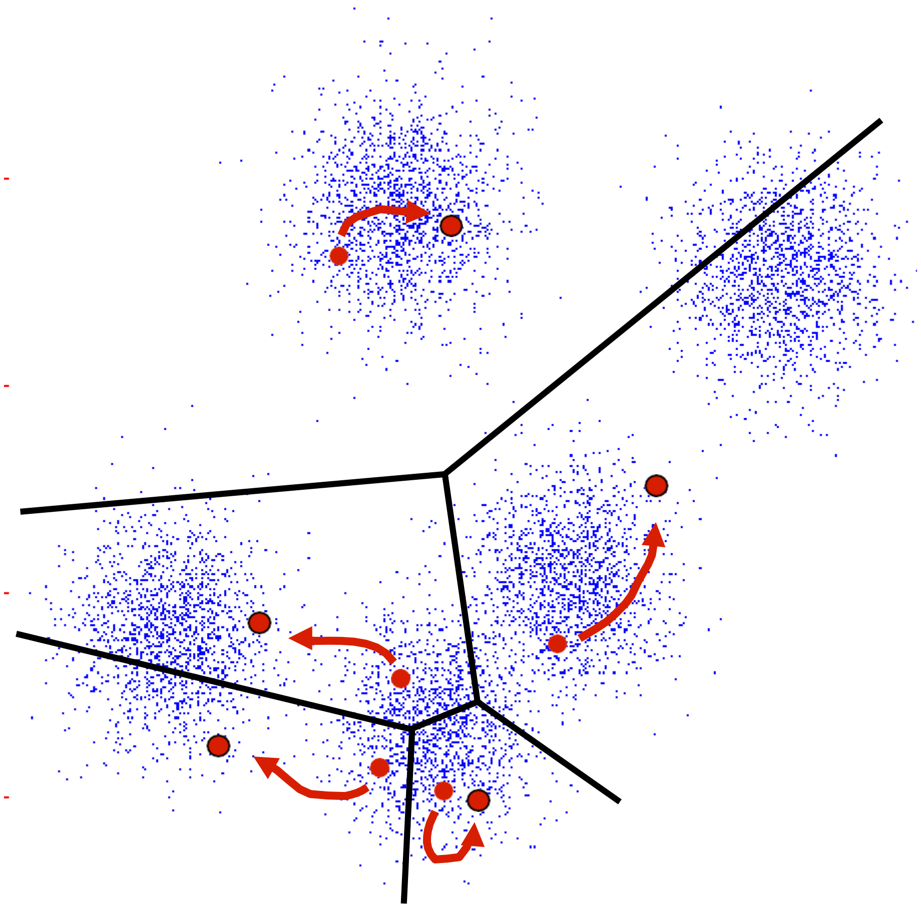
K-Means

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change

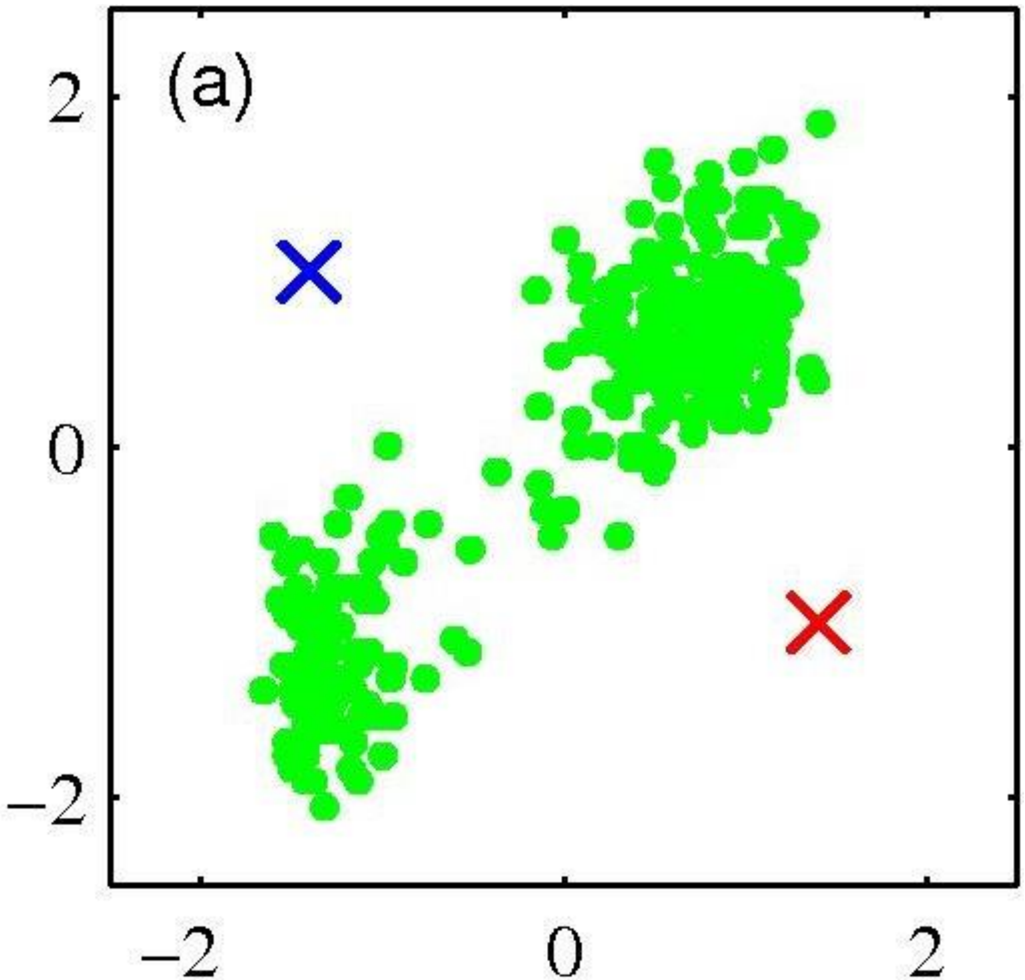


K-Means

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



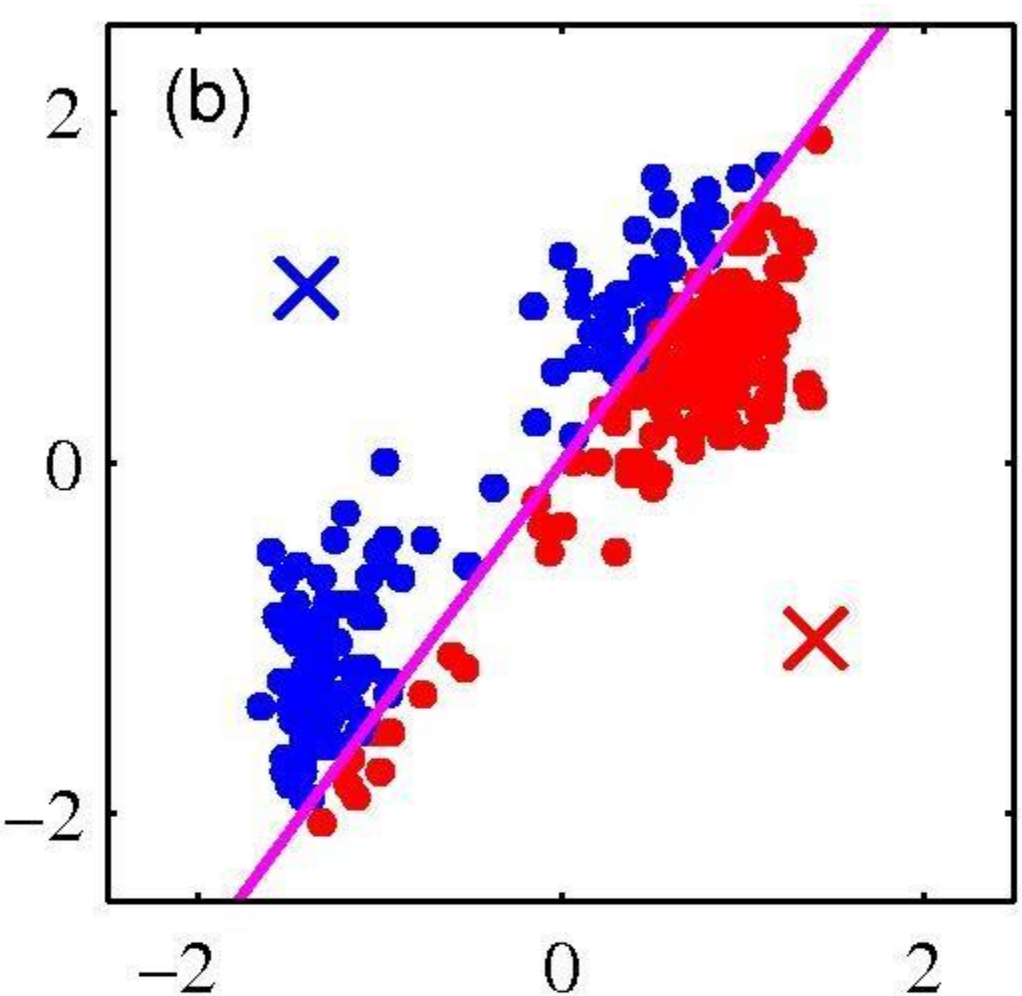
K-means clustering: Example



- Pick K random points as cluster centers (means)

Shown here for K=2

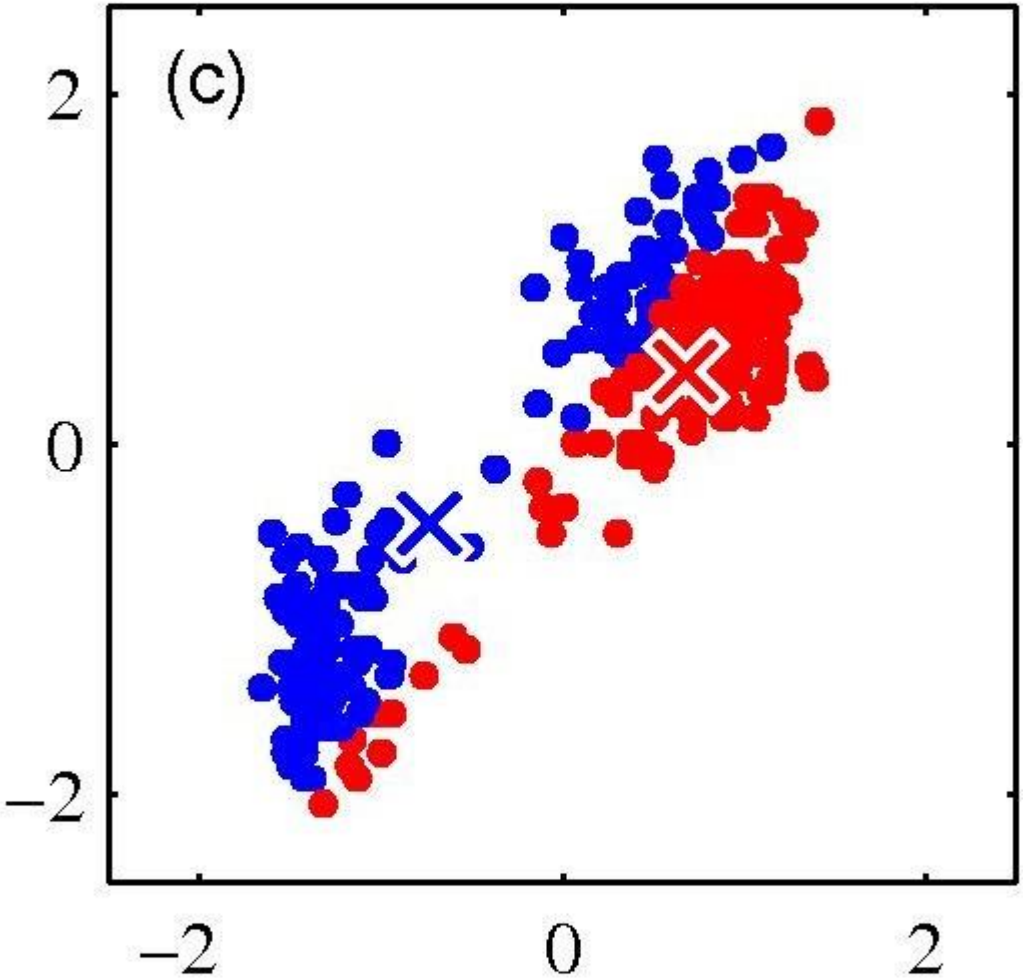
K--means clustering: Example



Iterative Step 1

- Assign data points to closest cluster center

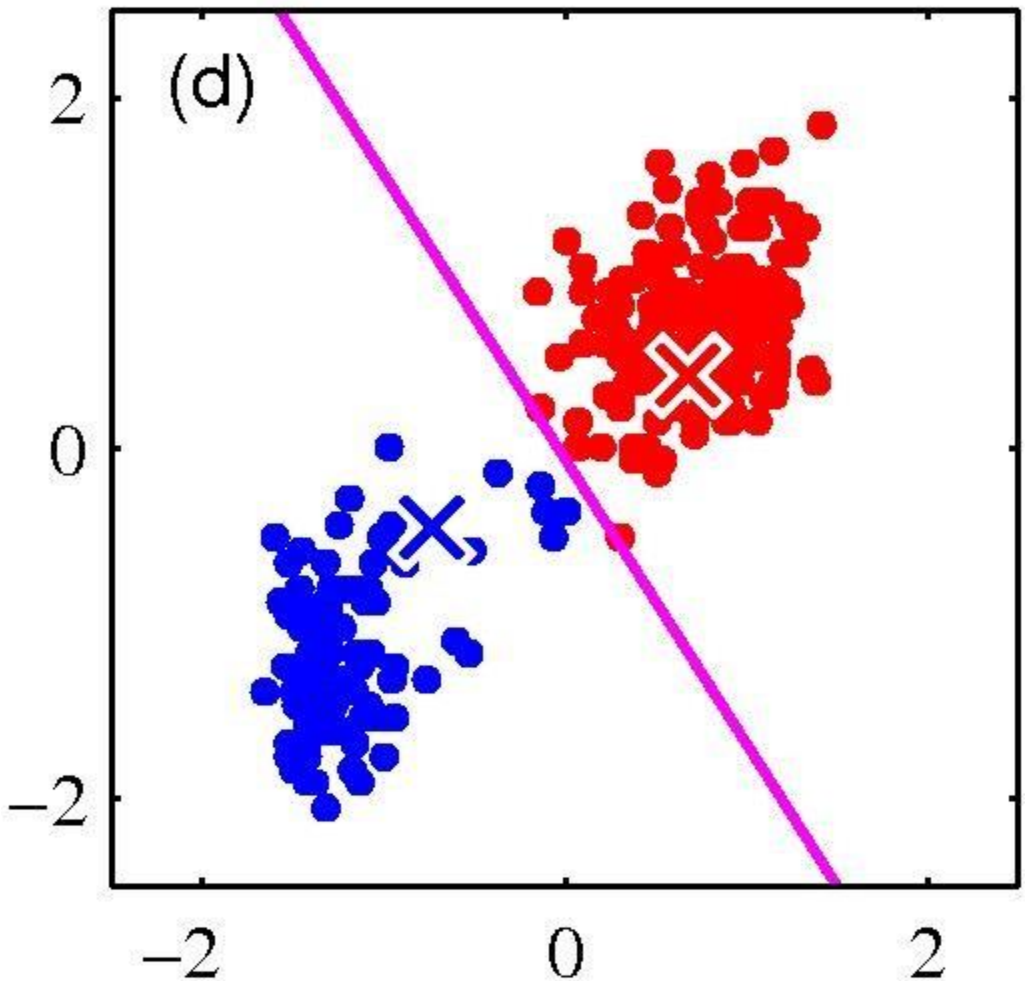
K--means clustering: Example



Iterative Step 2

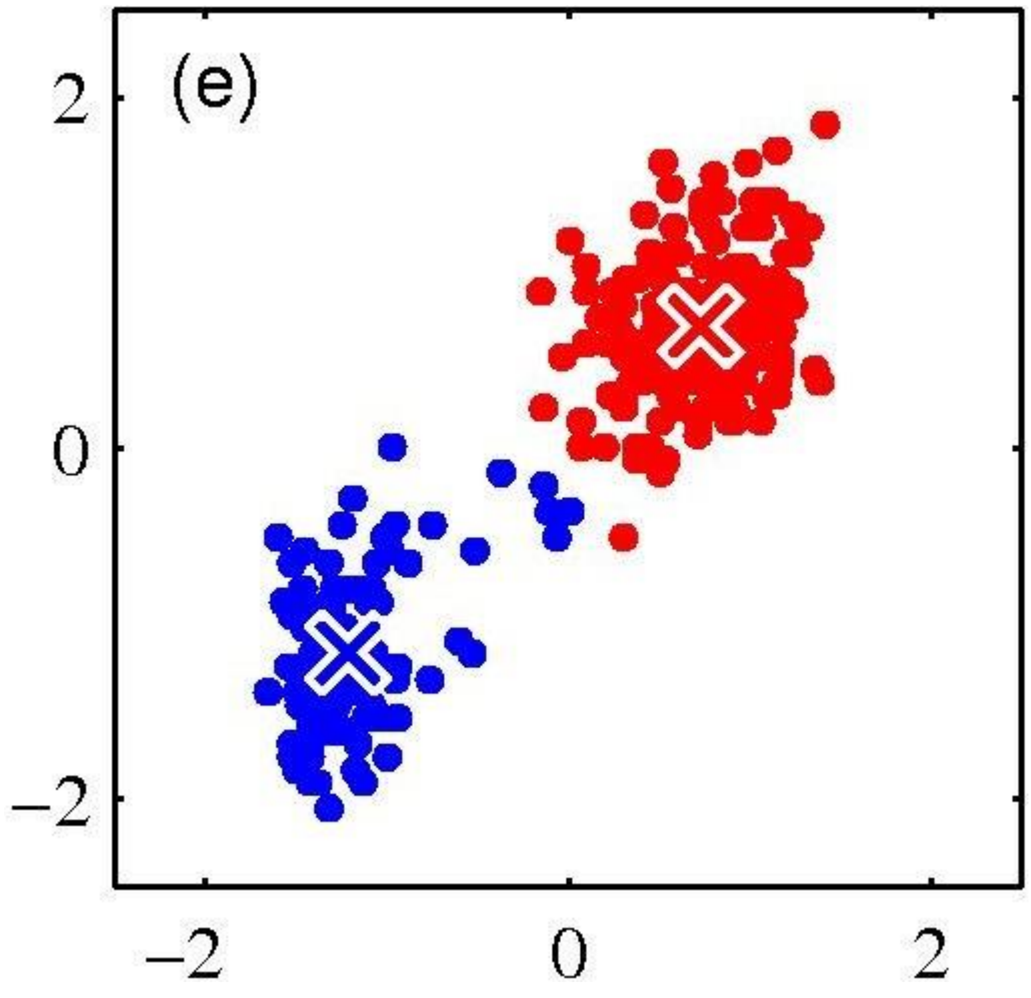
- Change the cluster center to the average of the assigned points

K--means clustering: Example



- Repeat until convergence

K--means clustering: Example



Properties of K--means **algorithm**

- Guaranteed to converge in a finite number of iterations

- Running time per iteration:
    1. Assign data points to closest cluster center

       O(KN) time

    2. Change the cluster center to the average of its assigned points

       O(N)

Example: K-Means for Segmentation

K=2

Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.

Original

Example: K-Means for Segmentation

K=2 K=3 Original





Example: K-Means for Segmentation

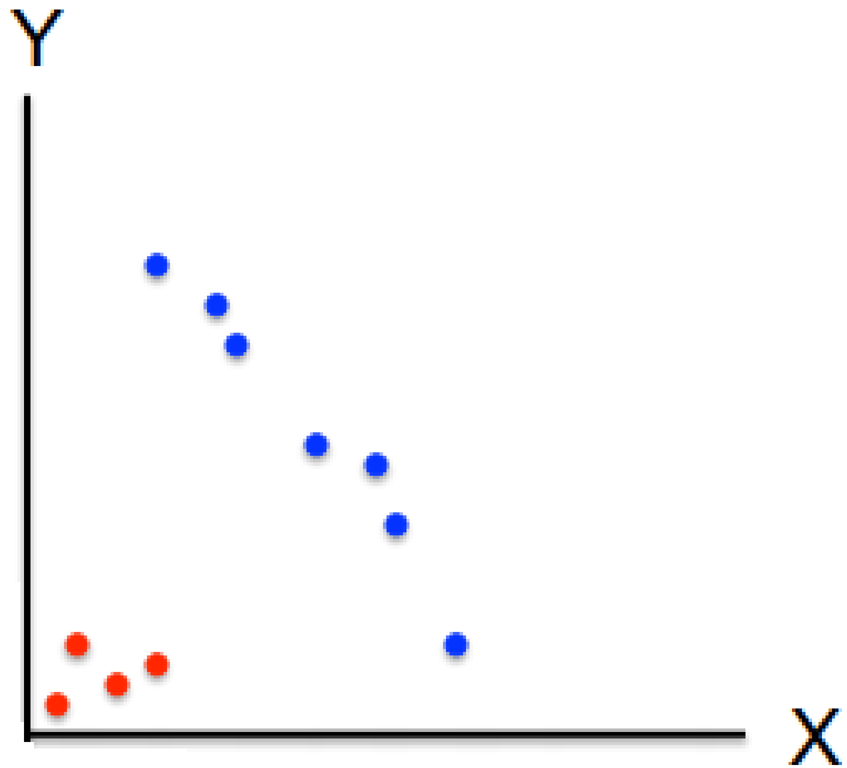
K=2 K=3 K=10 Original

4% 8% 17%

Example 1: K-Means

$k = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

$k = 2$



Example 1: K-Means

$K = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

$\underline{K = 2}$

$m_1 = 4 \qquad m_2 = 12$

$K_1 = \{2, 3, 4\} \qquad K_2 = \{10, 11, 12, 20, 25, 30\}$

Example 1: K-Means

$$k = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$$

$$k = 2$$

$$m_2 = \frac{108}{6} = 18$$

$$m_1 = 3$$

$$k_1 = \{2, 3, 4, 10\} \quad k_2 = \{11, 12, 20, 25, 30\}$$

Example 1: K-Means

$k = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

$k = 2$

$m_1 = 4.75$ $\qquad$ $m_2 = 19.6$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $\qquad$ $K_2 = \{20, 25, 30\}$

Example 1: K-Means

$k = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

$k = 2$

$m_1 = 7$ $\quad$ $m_2 = 25$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $\quad$ $K_2 = \{20, 25, 30\}$

$m_1 = 7$ $\quad$ $m_2 = 25$

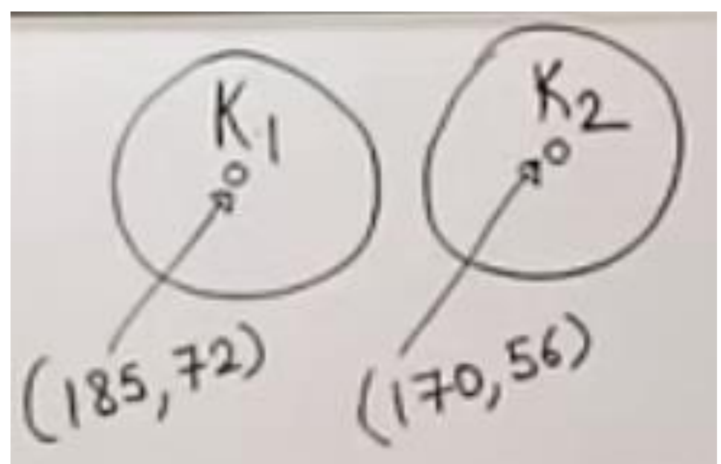
Example 2: K-Means

	Height	weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

Euclidean Distance

$$\sqrt{(X_0 - X_c)^2 + (Y_0 - Y_c)^2}$$

O = Observed
C = Centeroid



Example 2: K-Means

	Height	weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

ED for 3 $\rightarrow K_1 \rightarrow \sqrt{(168-185)^2+(60-72)^2}$

20.80

$\rightarrow K_2 \rightarrow \sqrt{(168-170)^2+(60-56)^2}$

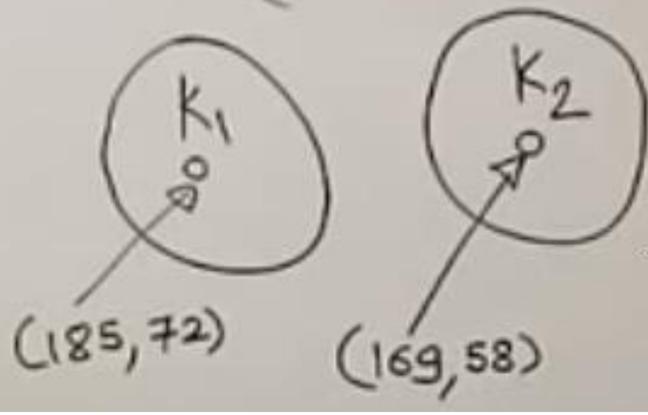
$= 4.48$

Example 2: K-Means

	Height	weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

New centroid calculation :-

for $K_2 = \left(\frac{170+168}{2}, \frac{60+56}{2}\right) = (169, 58)$



Example 2: K-Means

	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

E.D for 4 $\rightarrow k_1 = \sqrt{(179-185)^2+(68-72)^2}$

$= (6.32)$

$\rightarrow k_2 = \sqrt{(179-169)^2+(68-58)^2}$

$= 14.14$

$K_1 \rightarrow \{1,4,5,6,7,8,9,10,11,12\}$

$K_2 \rightarrow \{2,3\}$