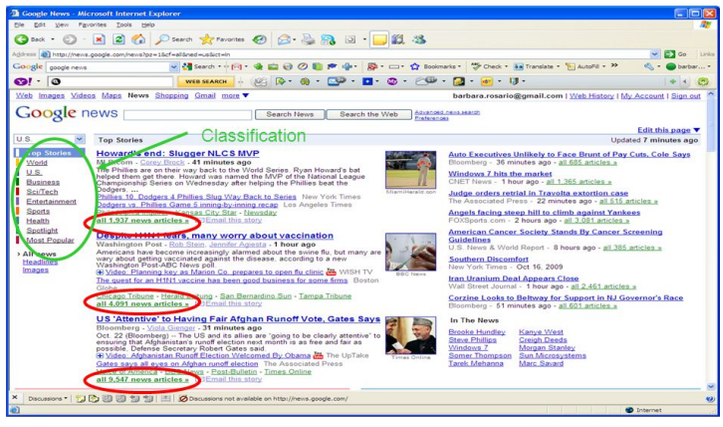
Clustering and K-means

Qiang Liu
UT Austin

Clustering

- Clustering: Partition the dataset into groups based on their similarity.

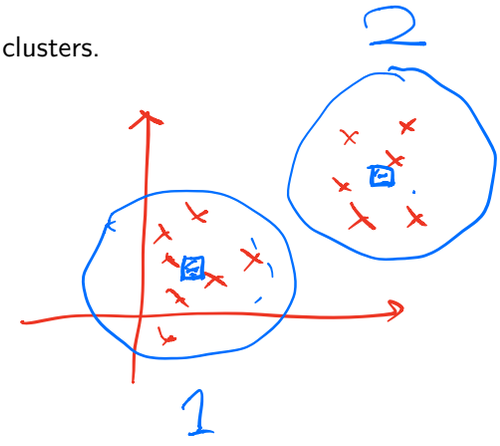
- Google News: automatic clustering gives an effective news presentation metaphor



- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs

K-means Algorithm

- **Inputs**: n objects (data points) $\{x_i\}_{i=1}^{n}$ and a number K of clusters.



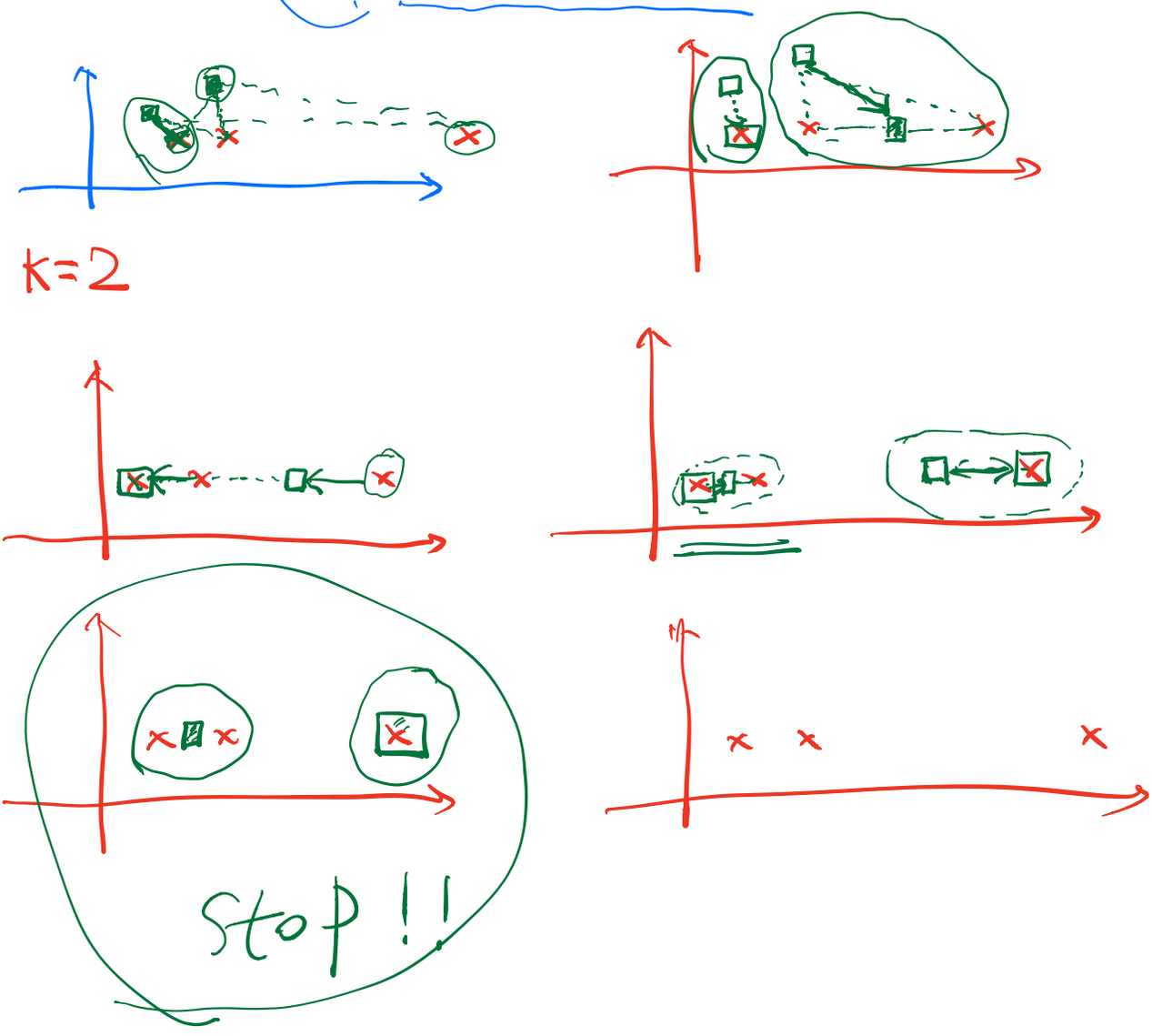
- **K-means**:
  - Idea:

    Alternatively optimize
    - i) the **assignment** of the data points different clusters.
    - ii) the **centroids** of the clusters.

  - Algorithm:

    Initialize the centroids or assignments randomly;

    Iterate until convergence:
    - i) **Assign each data point** (x_i) to the cluster that has the closest centroid.
    - ii) **Update the centroids** of all the clusters.



K-means Algorithm

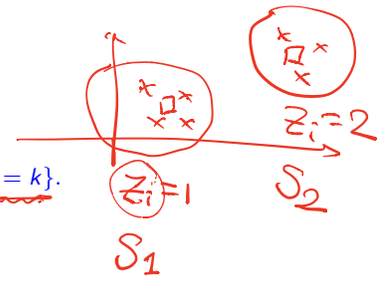
- **Inputs**: n objects (data points) $\{x_i\}_{i=1}^n$ and a number K of clusters.

- **K-means Algorithm**:
  - Initialization: randomly place K points (as the centroids)
  - Iterate until convergence:
    - i) Assign each object to the group that has the closest centroid

$$z_i = \arg\min_{k=1,\ldots,K} ||x_i - \mu_k||$$

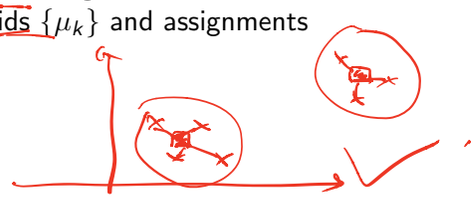
    - ii) Recalculate the position of the K centroids

$$\mu_k = \frac{1}{|S_k|} \sum_{i \in S_k} x_i, \qquad S_k = \{i : z_i = k\}.$$



K-means as Optimization

- K-means can be viewed as a "coordinate descent" algorithm for optimizing an objective function of the centroids $\{\mu_k\}$ and assignments $\{z_i\}$.

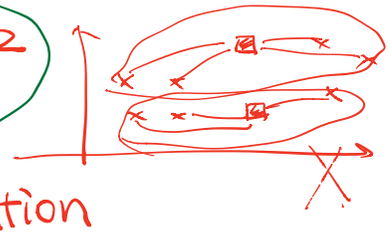


Given $\{x_i\}_{i=1}^n$

Want $\{\mu_k\}_{k=1}^K$ $\{z_i\}_{i=1}^n$

μ z

Define: $L(\mu, z) = \sum_{i=1}^n \|x_i - \mu_{z_i}\|^2$



$\min_{\mu, z} L(\mu, z)$

mixed optimization

$\mu \in \mathbb{R}^{d \times K}$ $z \in \{1, \cdots K\}^n$

Coordinate descent

Initialize μ_0

Repeat (t-iteration)

① update z with fixed μ

$z_t = \operatorname{argmin}_z L(\mu_t, z)$

② update μ with z fixed

$\mu_{t+1} = \operatorname{argmin}_\mu L(\mu, z_t)$

$L(\mu, z) = \sum_{i=1}^n \|x_i - \mu_{z_i}\|^2$

Find optimal $z_i \leftarrow \operatorname{argmin}_{z_i} \|x_i - \mu_{z_i}\|^2$ (Assigment step)

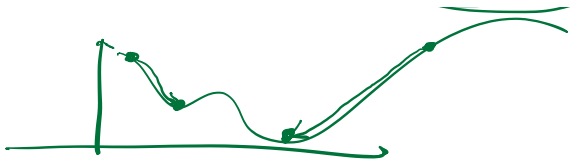
Find optimal $\mu_k \leftarrow \operatorname{argmin}_{\mu_k} \sum_{i \in S_k} \|x_i - \mu_k\|^2$

$S_k = \{i : z_i = k\}$

$= \sum_{i \in S_k} \|x_i\|^2 - 2x_i^T \mu_k + \|\mu_k\|^2$

$= \text{const} - 2\left(\sum_{i \in S_k} x_i\right)^T \mu_k + |S_k| \|\mu_k\|^2$

$\Rightarrow \mu_k \leftarrow \frac{\sum_{i \in S_k} x_i}{|S_k|}$ (centroid update step)



(μ_t, z_t)

$\Rightarrow L(\mu_{t+1}, z_{t+1}) \leq L(\mu_t, z_t)$

$L(\mu_t, z_t) \geq L(\mu_{t+1}, z_t)$

$\geq L(\mu_{t+1}, z_{t+1})$