

# The *K-Means* Clustering Method

---

- ❑ *K-Means* (MacQueen'67, Lloyd'57/'82)
  - ❑ Each cluster is represented by the center of the cluster
- ❑ Given  $K$ , the number of clusters, the *K-Means* clustering algorithm is outlined as follows
  - ❑ Select  $K$  points as initial centroids
  - ❑ **Repeat**
    - ❑ Form  $K$  clusters by assigning each point to its closest centroid
    - ❑ Re-compute the centroids (i.e., *mean point*) of each cluster
  - ❑ **Until** convergence criterion is satisfied
- ❑ Different kinds of measures can be used
  - ❑ Manhattan distance ( $L_1$  norm), **Euclidean distance ( $L_2$  norm)**, Cosine similarity

## Initialization of K-Means

---

- ❑ Different initializations may generate rather different clustering results (some could be far from optimal)
- ❑ Original proposal (MacQueen'67): Select  $K$  seeds randomly
  - ❑ Need to run the algorithm multiple times using different seeds
- ❑ There are many methods proposed for better initialization of  $k$  seeds
  - ❑ ***K-Means++*** (Arthur & Vassilvitskii'07):
    - ❑ The first centroid is selected at random
    - ❑ The next centroid selected is the one that is farthest from the currently selected (selection is based on a weighted probability score)
    - ❑ The selection continues until  $K$  centroids are obtained

