K-means Clustering

- What is clustering?
- Why would we want to cluster?
- How would you determine clusters?
- How can you do this efficiently?

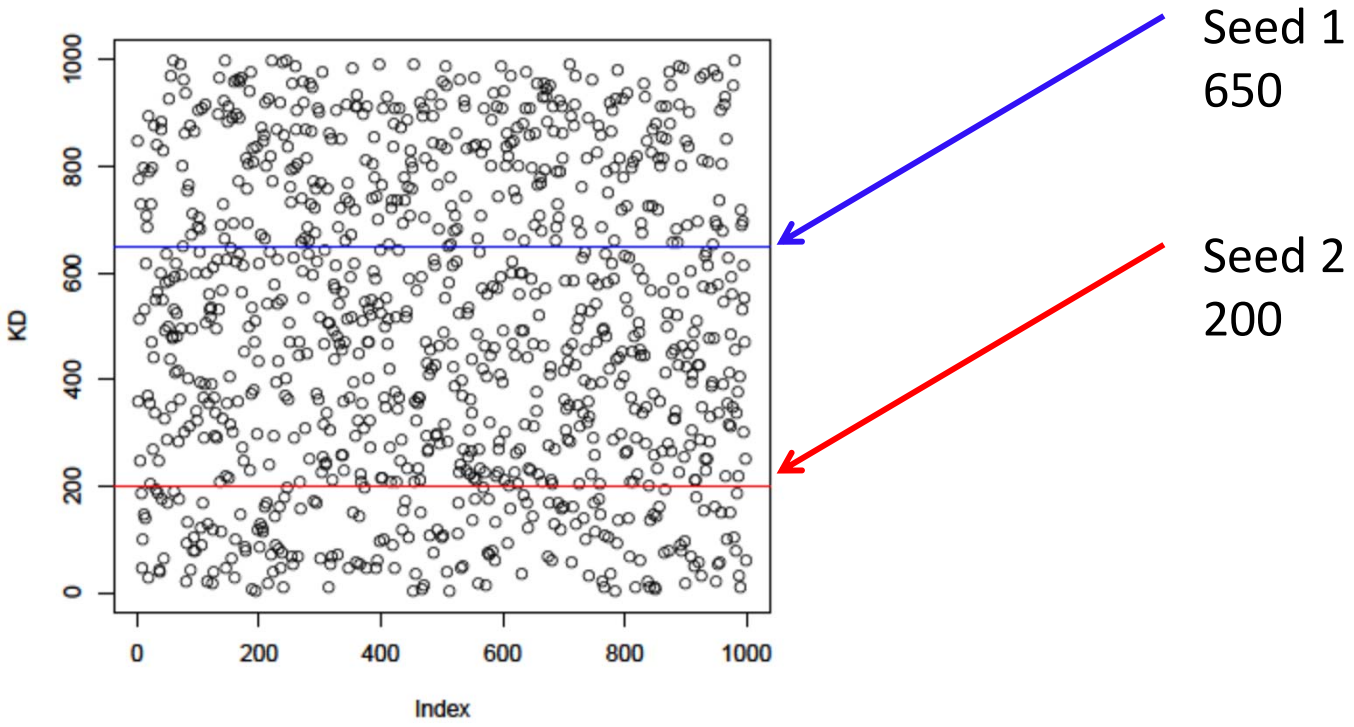
K-means Clustering

- Strengths
  - Simple iterative method
  - User provides “K”
- Weaknesses
  - Often too simple → bad results
  - Difficult to guess the correct “K”

K-means Clustering

Basic Algorithm:

- Step 0: select K
- Step 1: randomly select initial cluster seeds



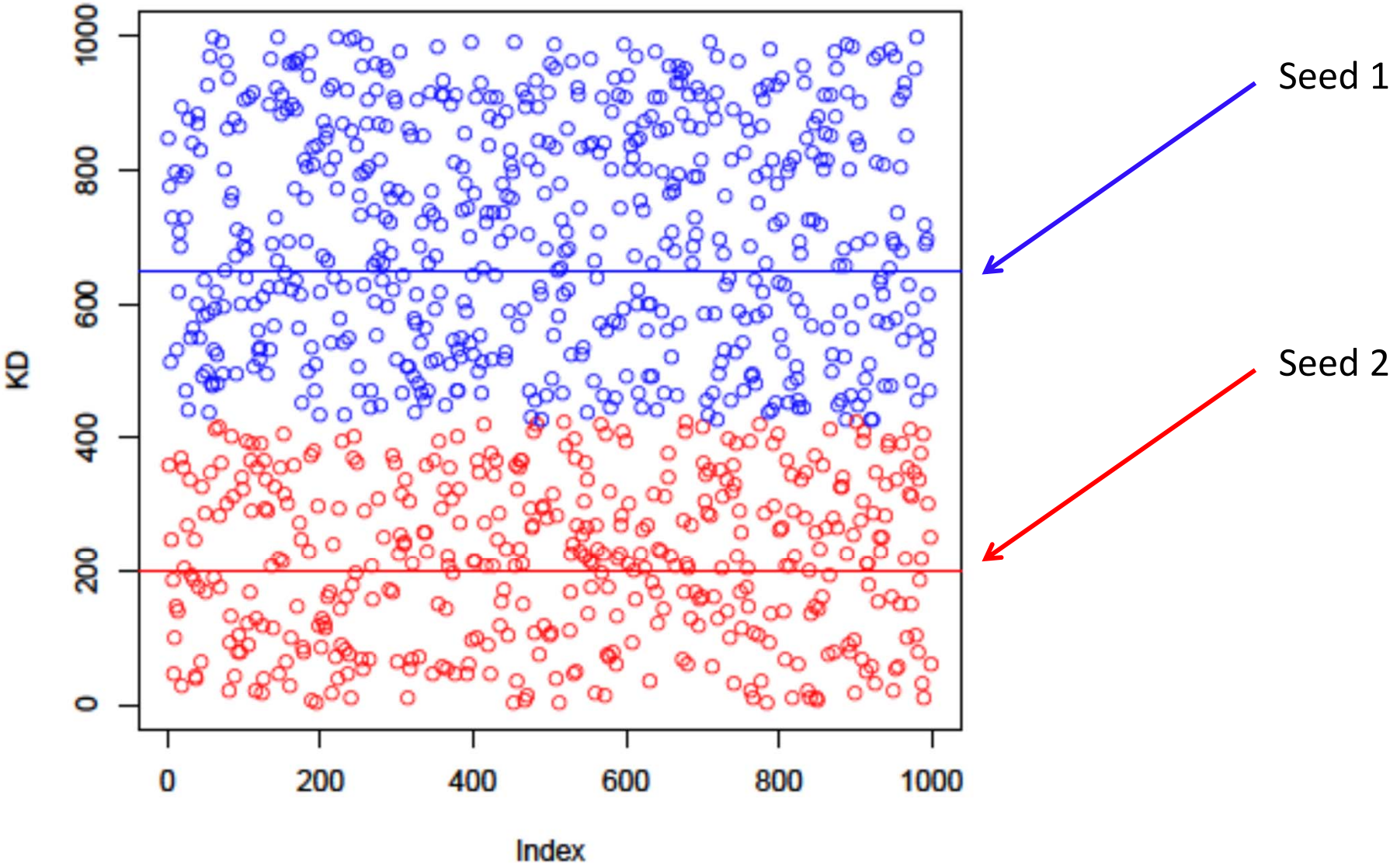
K-means Clustering

- An initial cluster seed represents the “mean value” of its cluster.
- In the preceding figure:
  - Cluster seed 1 = 650
  - Cluster seed 2 = 200

K-means Clustering

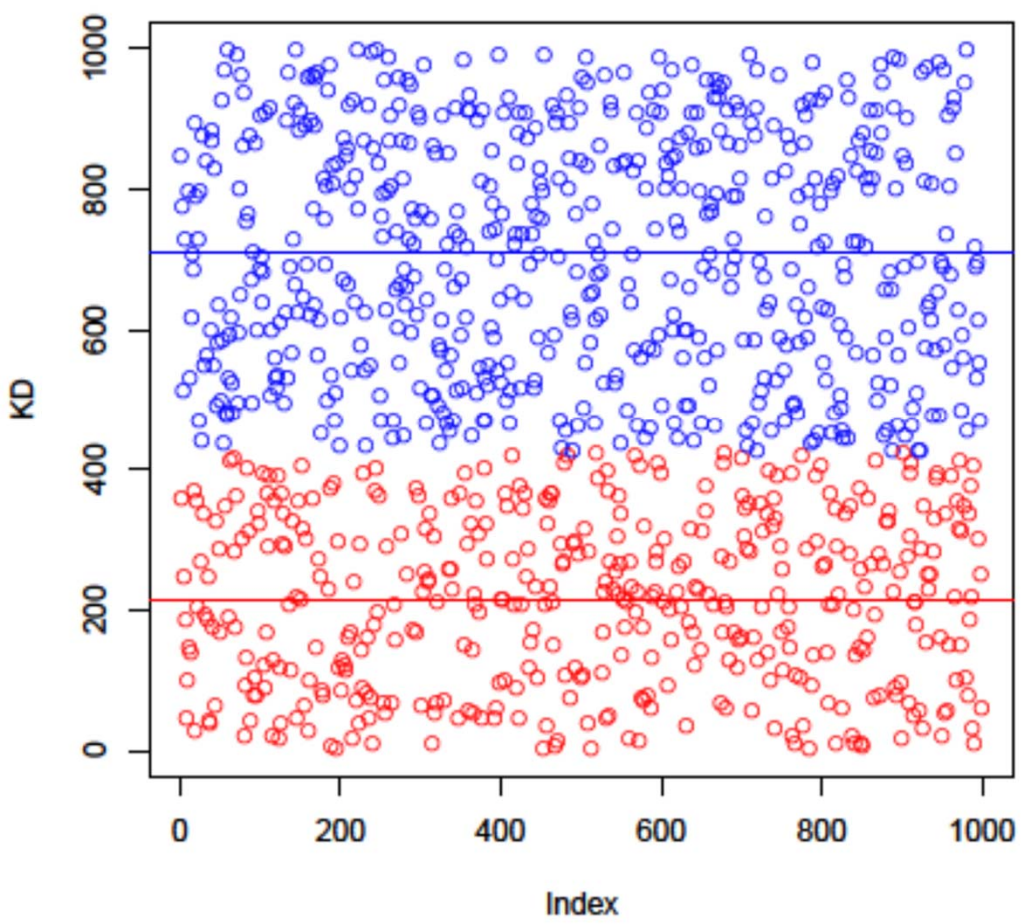
- Step 2: calculate distance from each object to each cluster seed.
- What type of distance should we use?
  - Squared Euclidean distance
- Step 3: Assign each object to the closest cluster

K-means Clustering



K-means Clustering

- Step 4: Compute the new centroid for each cluster



K-means Clustering

- Iterate:
  - Calculate distance from objects to cluster centroids.
  - Assign objects to closest cluster
  - Recalculate new centroids
- Stop based on convergence criteria
  - No change in clusters
  - Max iterations

K-means Issues

- Distance measure is squared Euclidean
  - Scale should be similar in all dimensions
    - Rescale data?
  - Not good for nominal data. Why?
- Approach tries to minimize the within-cluster sum of squares error (WCSS)
  - Implicit assumption that SSE is similar for each group

WCSS

- The over all WCSS is given by:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

- The goal is to find the smallest WCSS
- Does this depend on the initial seed values?
- Possibly.

Bottom Line

- K-means
  - Easy to use
  - Need to know K
  - May need to scale data
  - Good initial method
- Local optima
  - No guarantee of optimal solution
  - Repeat with different starting values

K-Means Lab

Pause this set of slides and switch to lab slides