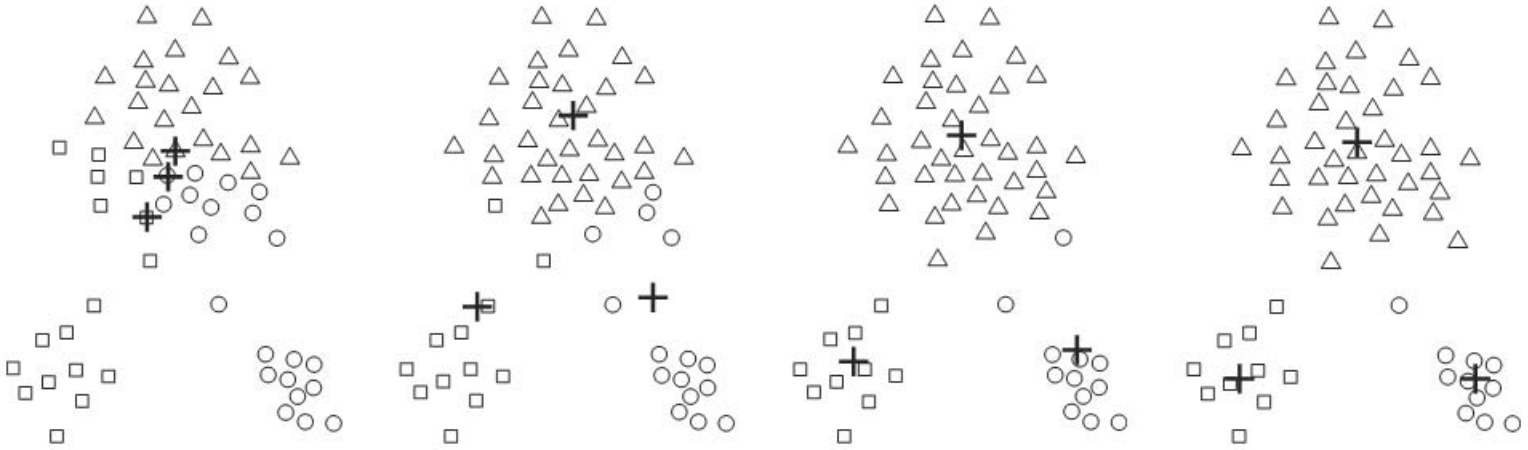
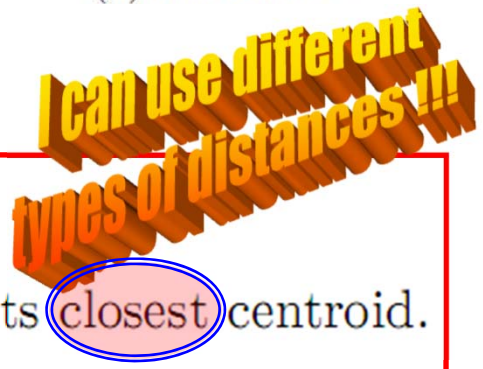
K-Means clustering

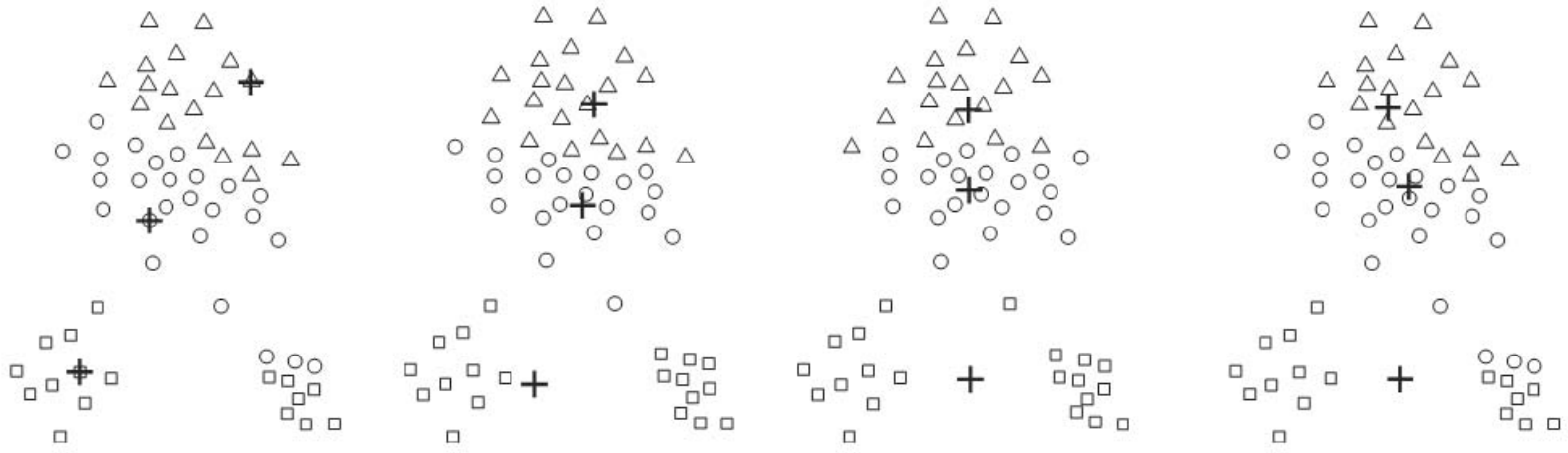


(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.



How many centers? How to initialize them?



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

- Often the cluster centers are initialized to some of the points, picked at random
- Which "optimality criterion" is the k-means algorithm optimizing?

The algorithm attempts to minimize a sum of square criterion:

$$\sum_{j=1}^{K} \sum_{n \in S_j} d(\boldsymbol{x}_n, \boldsymbol{\mu}_j)^2$$

K number of clusters
$\boldsymbol{x}_n$ is a vector representing the n-th data point
$n \in S_j$ indicates the set of points belonging to cluster S_j
$\boldsymbol{\mu}_j$ is the mean of the data points in S_j
$d(\boldsymbol{x}_n, \boldsymbol{\mu}_j)$ indicates the a **distance** between $\boldsymbol{x}_n$ and $\boldsymbol{\mu}_j$

Example of distances: Euclidean distance, corrected Pearson correlation coefficient