Clustering analysis

József Mezei

Business Analytics I

What is Cluster Analysis?

- Cluster: a collection of data objects

  - Similar to one another within the same cluster

  - Dissimilar to the objects in other clusters

- Cluster analysis

  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters

- Unsupervised learning: no predefined classes

- Typical applications

  - As a stand-alone tool to get insight into data distribution
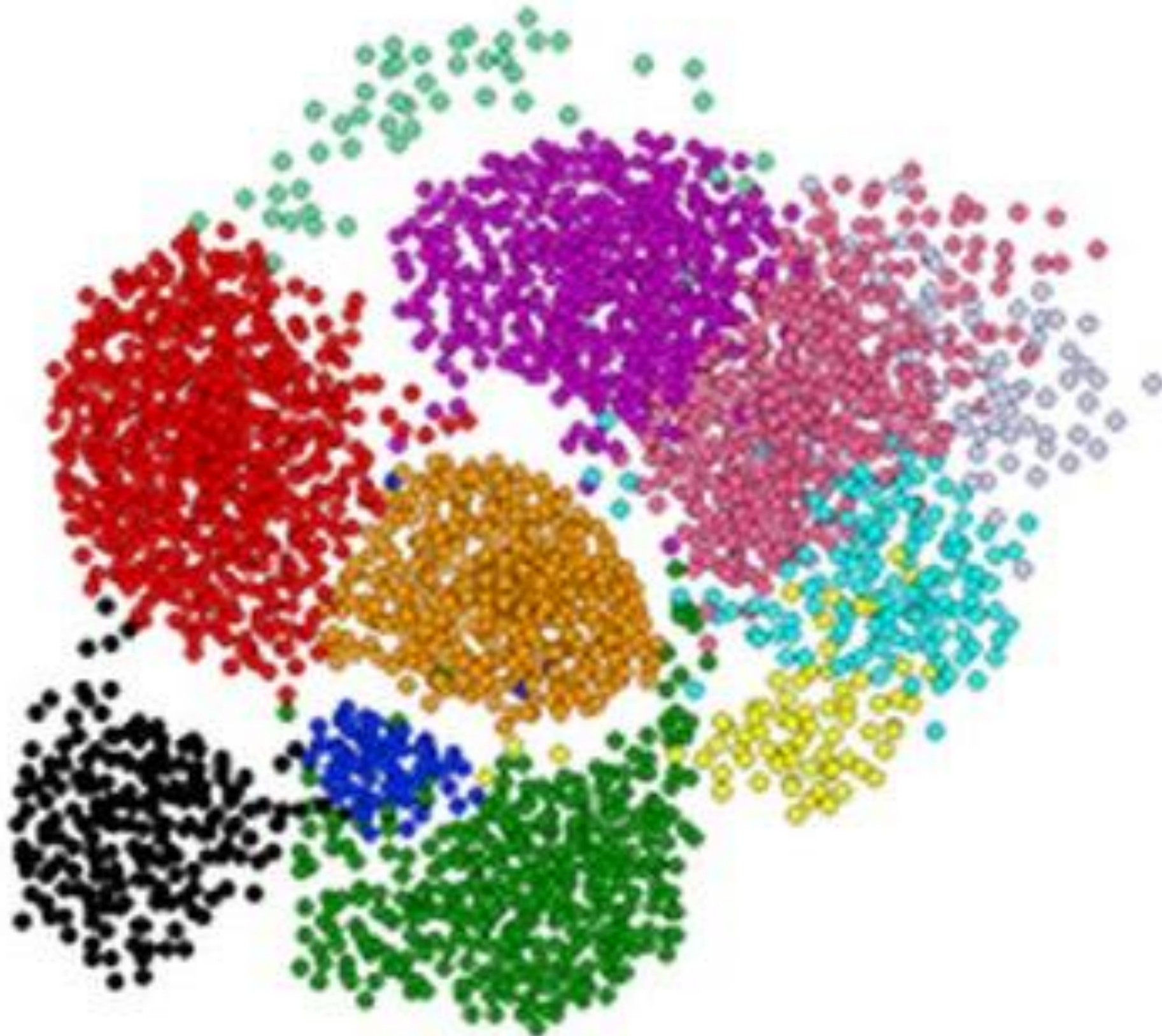
  - As a preprocessing step for other algorithms

High Dimensional Data

Example: Clusters & Outliers
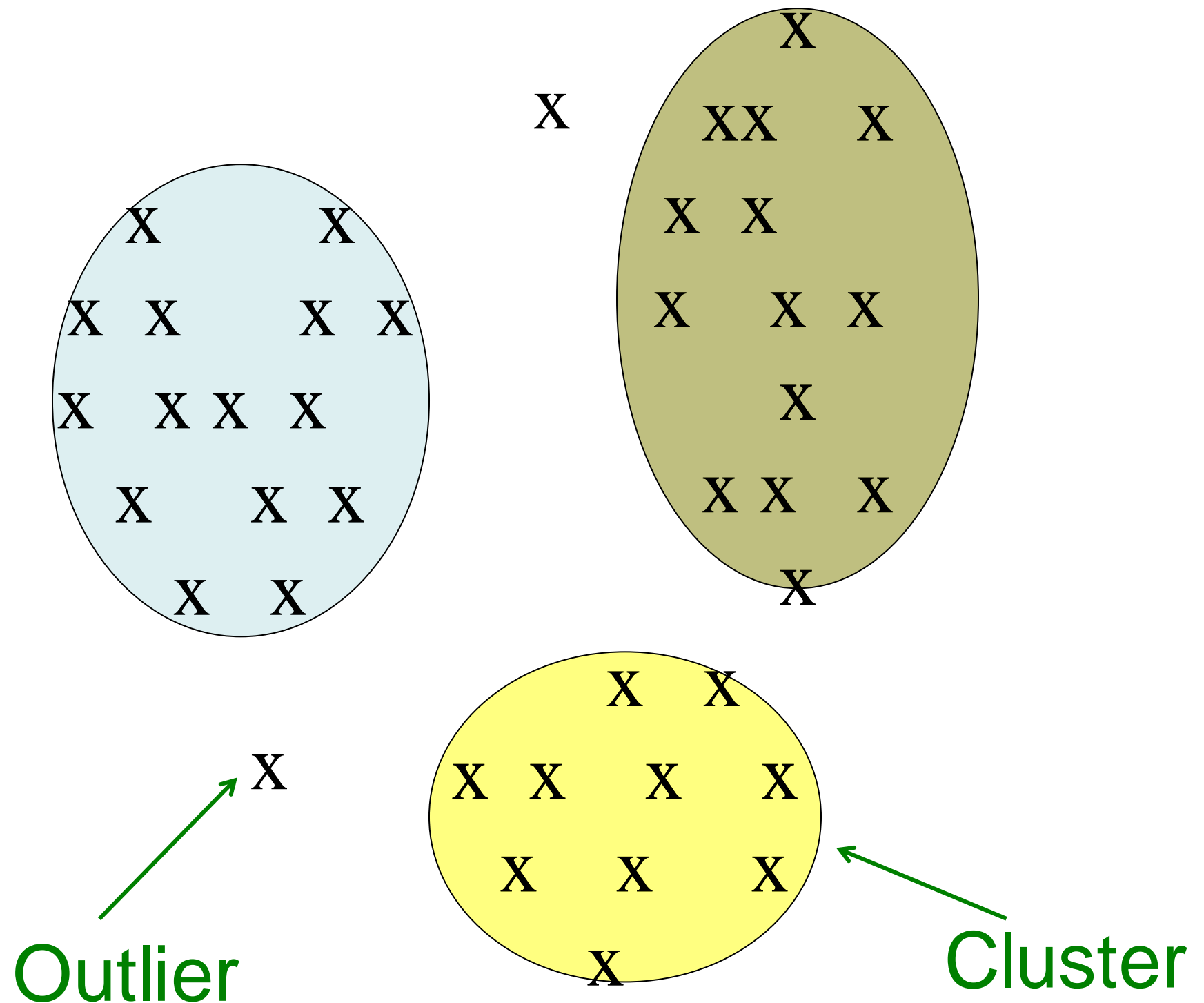
Clustering is a hard problem!

Why is it hard?

- Clustering in two dimensions looks easy

- Clustering small amounts of data looks easy

- And in most cases, looks are not deceiving

- Many applications involve not 2, but 10 or 10,000 dimensions

- **High-dimensional spaces look different:** Almost all pairs of points are at about the same distance

Clustering Problem: Music album

- **Intuitively: Music divides into categories, and customers prefer a few categories**

  - But what are categories?

- Represent an album by a set of customers who bought it

- Similar albums have similar sets of customers, and vice-versa

Clustering Problem: Music CDs

Space of all albums:

- Think of a space with one dim. for each customer

    - Values in a dimension may be 0 or 1 only

    - An album is a point in this space $(x_1, x_2,..., x_k)$, where $x_i = 1$ if the i th customer bought it

- For Amazon, the dimension is tens of millions

- **Task:** Find clusters of similar albums

Clustering

- Pattern Recognition

- Spatial Data Analysis

- Image Processing

- Economic Science (especially market research)

- WWW: Document classification

Examples of Clustering Applications

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

- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost

- City-planning: Identifying groups of houses according to their house type, value, and geographical location

Quality: What Is Good Clustering?

- A <u>good clustering</u> method will produce high quality clusters with

  - high <u>intra-class</u> similarity

  - low <u>inter-class</u> similarity

- The <u>quality</u> of a clustering result depends on both the similarity measure used by the method and its implementation

Measure the Quality of Clustering

- **Dissimilarity/Similarity metric**: Similarity is expressed in terms of a distance function

- There is a separate "quality" function that measures the "goodness" of a cluster.

- The definitions of distance functions are usually very different for interval-scaled, logical, categorical, ordinal variables.

- Weights should be associated with different variables based on applications and data semantics.

- It is hard to define "similar enough" or "good enough"

  - the answer is typically highly subjective.

Data Structures

- ## Data matrix
  - (two modes)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- ## Dissimilarity matrix
  - (one mode)

$$\begin{bmatrix} 0 \\ d(2,1) & 0 \\ d(3,1) & d(3,2) & 0 \\ \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Type of data in clustering analysis

- <u>Numerical variables</u>

- <u>Binary variables</u>

- <u>Nominal and ordinal variables</u>

Similarity and Dissimilarity Between Objects

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

- Also, one can use weighted distance, parametric Pearson product moment correlation, or other disimilarity measures

Major Clustering Approaches (I)

- <u>Partitioning approach</u>:

  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors

  - Typical methods: k-means, k-medoids, CLARANS

- <u>Hierarchical approach</u>:

  - Create a hierarchical decomposition of the set of data (or objects) using some criterion

  - Typical methods: Diana, Agnes, BIRCH, ROCK, CAMELEON

- <u>Density-based approach</u>:

  - Based on connectivity and density functions

  - Typical methods: DBSACN, OPTICS, DenClue

Major Clustering Approaches (II)

- Grid-based approach:

  - based on a multiple-level granularity structure

  - Typical methods: STING, WaveCluster, CLIQUE

- Model-based:

  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other

  - Typical methods: EM, SOM, COBWEB

- Frequent pattern-based:

  - Based on the analysis of frequent patterns

  - Typical methods: pCluster

- User-guided or constraint-based:

  - Clustering by considering user-specified or application-specific constraints

  - Typical methods: COD (obstacles), constrained clustering

Typical Alternatives to Calculate the Distance between Clusters

- **Single link:** smallest distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = min(t_{ip}, t_{jq})$

- **Complete link:** largest distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = max(t_{ip}, t_{jq})$

- **Average:** avg distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = avg(t_{ip}, t_{jq})$

- **Centroid:** distance between the centroids of two clusters, i.e., $dis(K_i, K_j) = dis(C_i, C_j)$

- **Medoid:** distance between the medoids of two clusters, i.e., $dis(K_i, K_j) = dis(M_i, M_j)$

  - Medoid: one chosen, centrally located object in the cluster

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:

    - Partition objects into *k* nonempty subsets

    - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)

    - Assign each object to the cluster with the nearest seed point

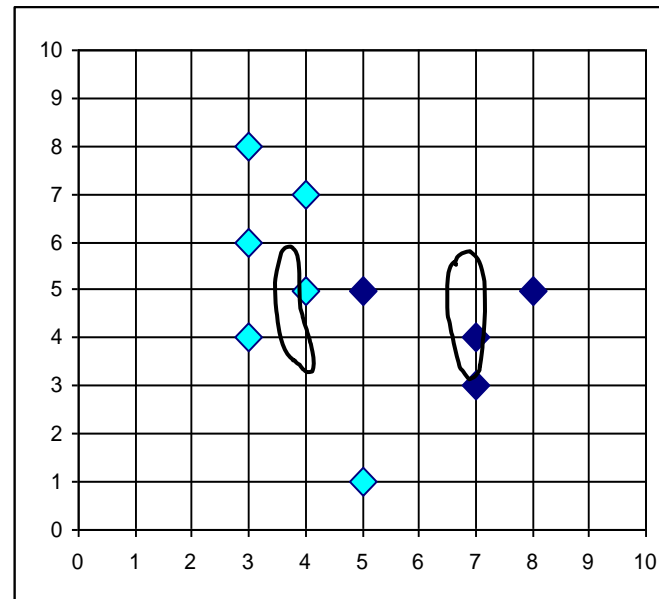   - Go back to Step 2, stop when no more new assignment
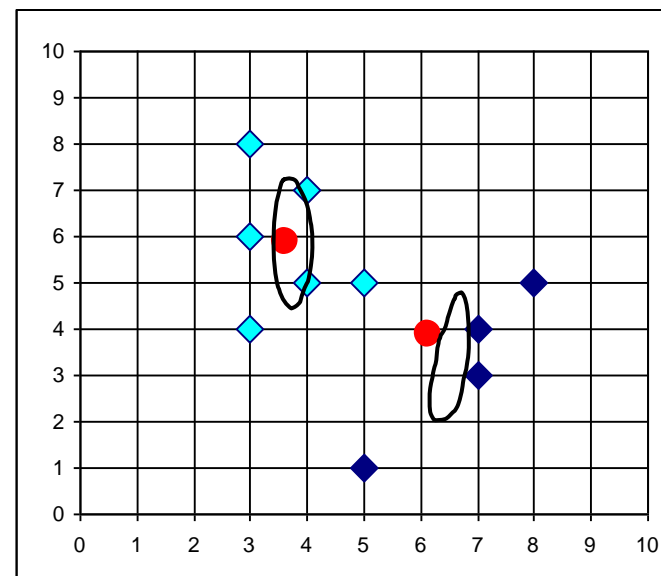
The *K-Means* Clustering Method

K=2

Arbitrarily choose K object as initial cluster center
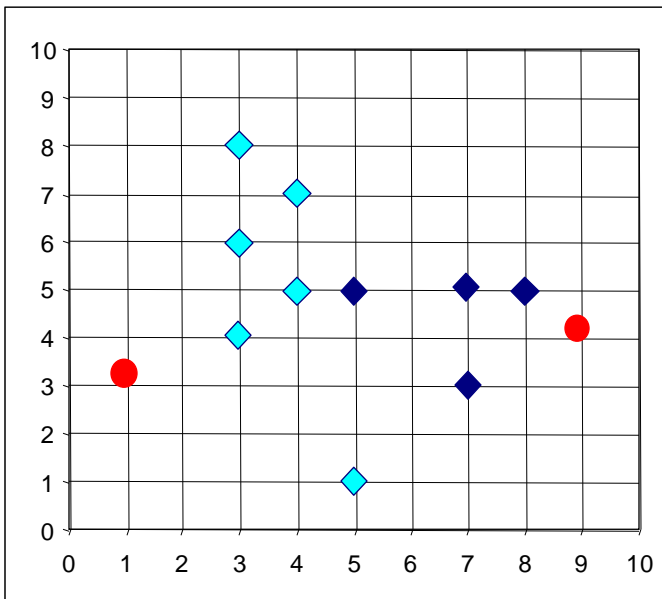
Assign each objects to most similar center

Update the cluster means

reassign

Update the cluster means
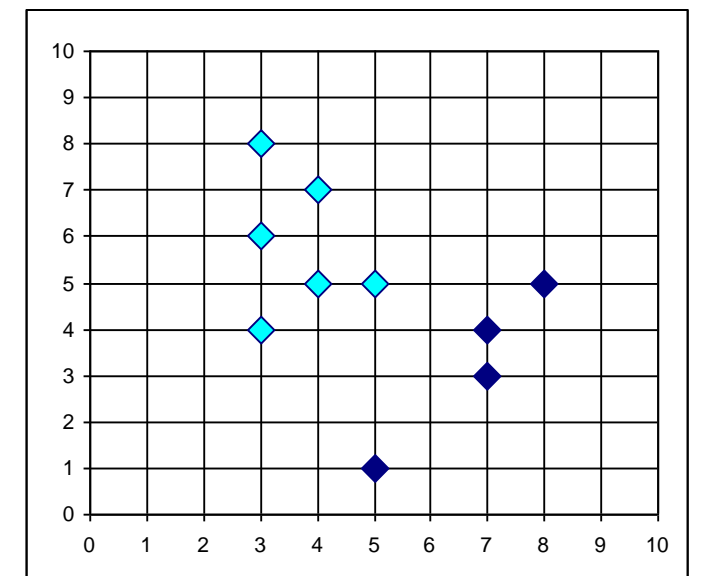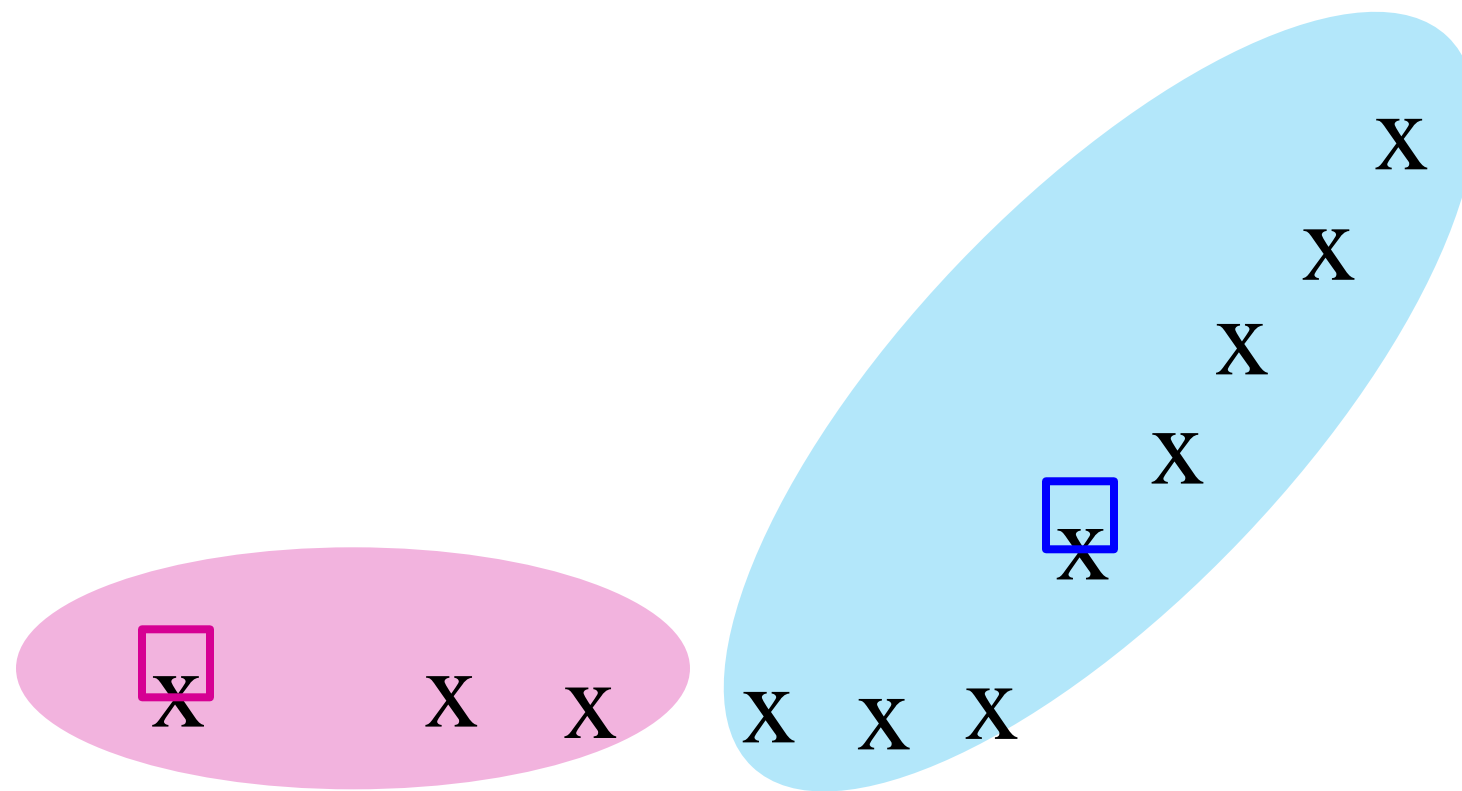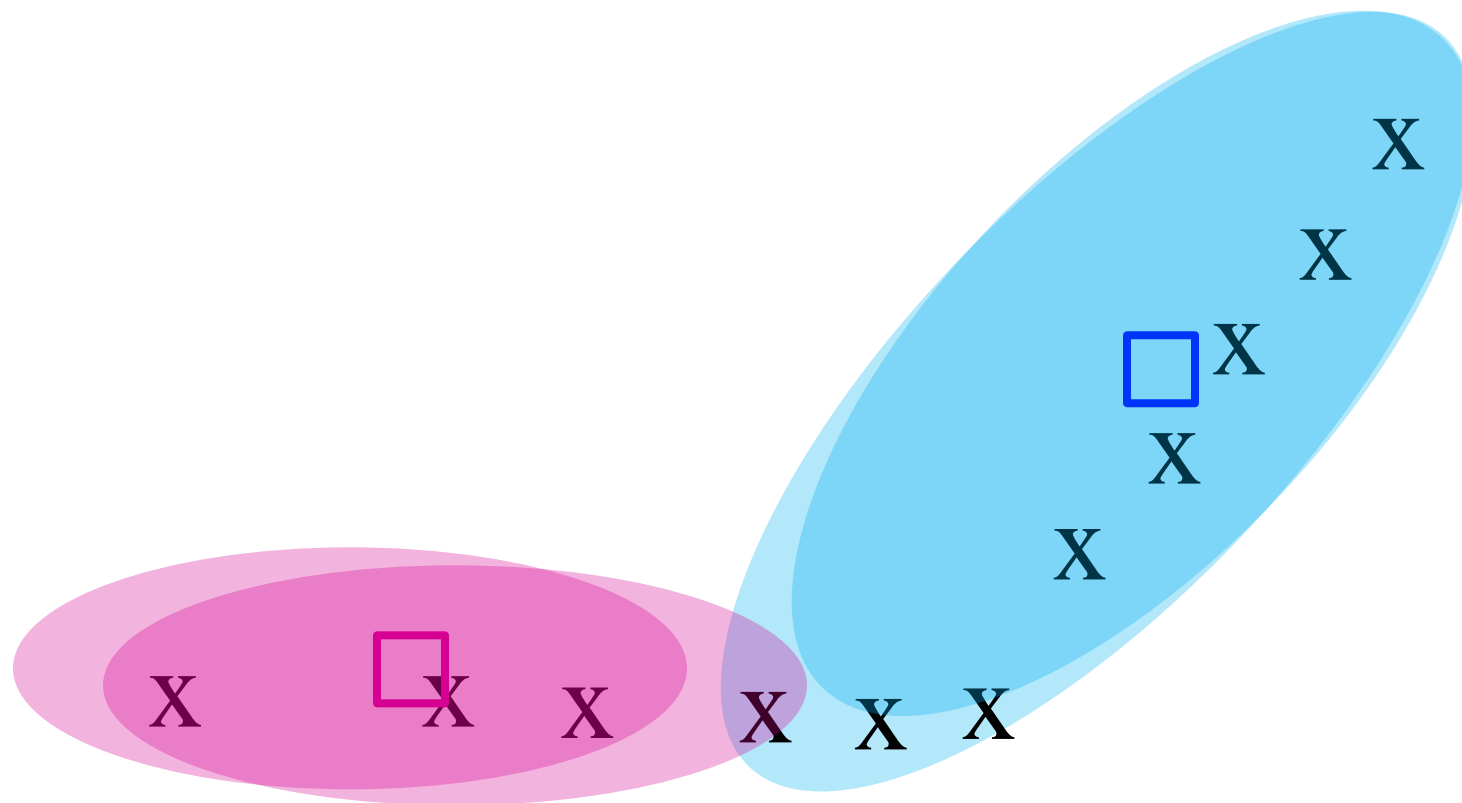
reassign

Example: Assigning Clusters

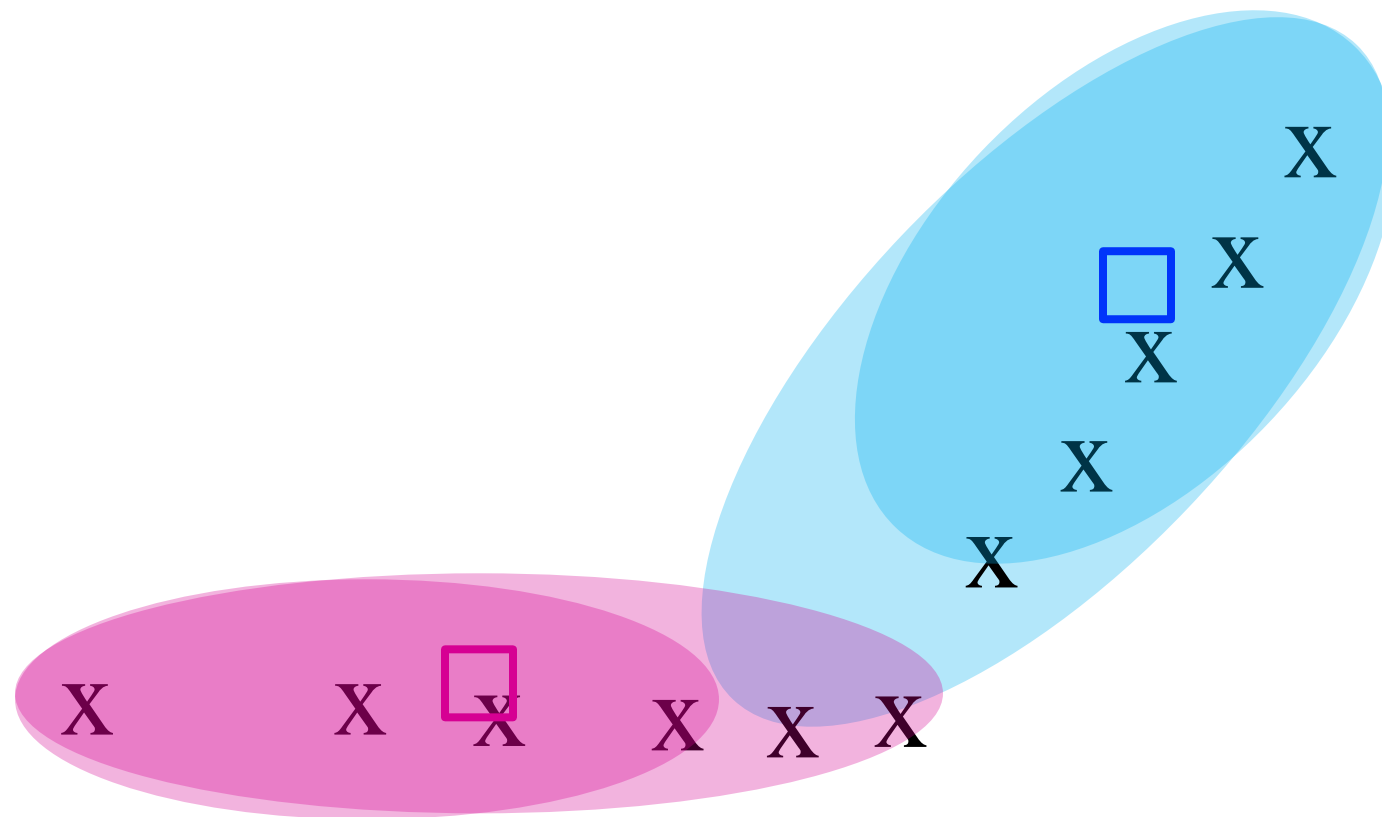X … data point

☐ … centroid

Clusters after round 1

Example: Assigning Clusters

Clusters after round 2
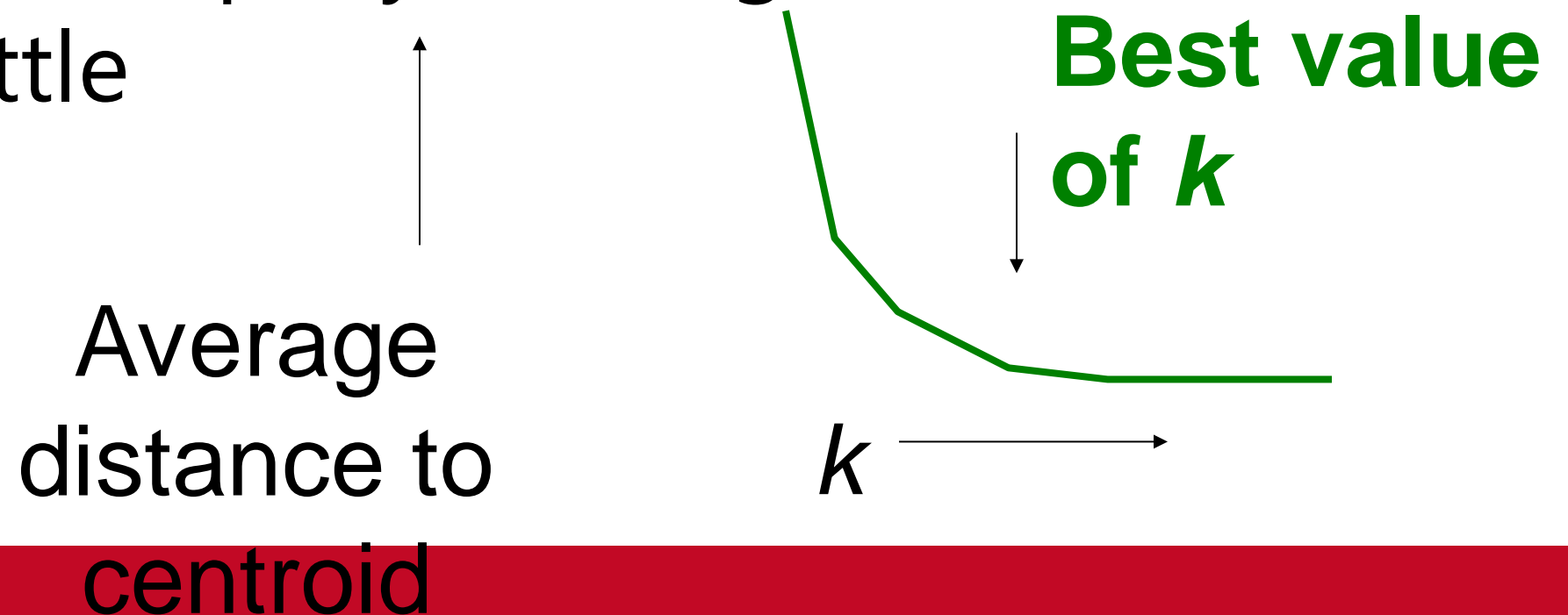
Example: Assigning Clusters

Clusters at the end
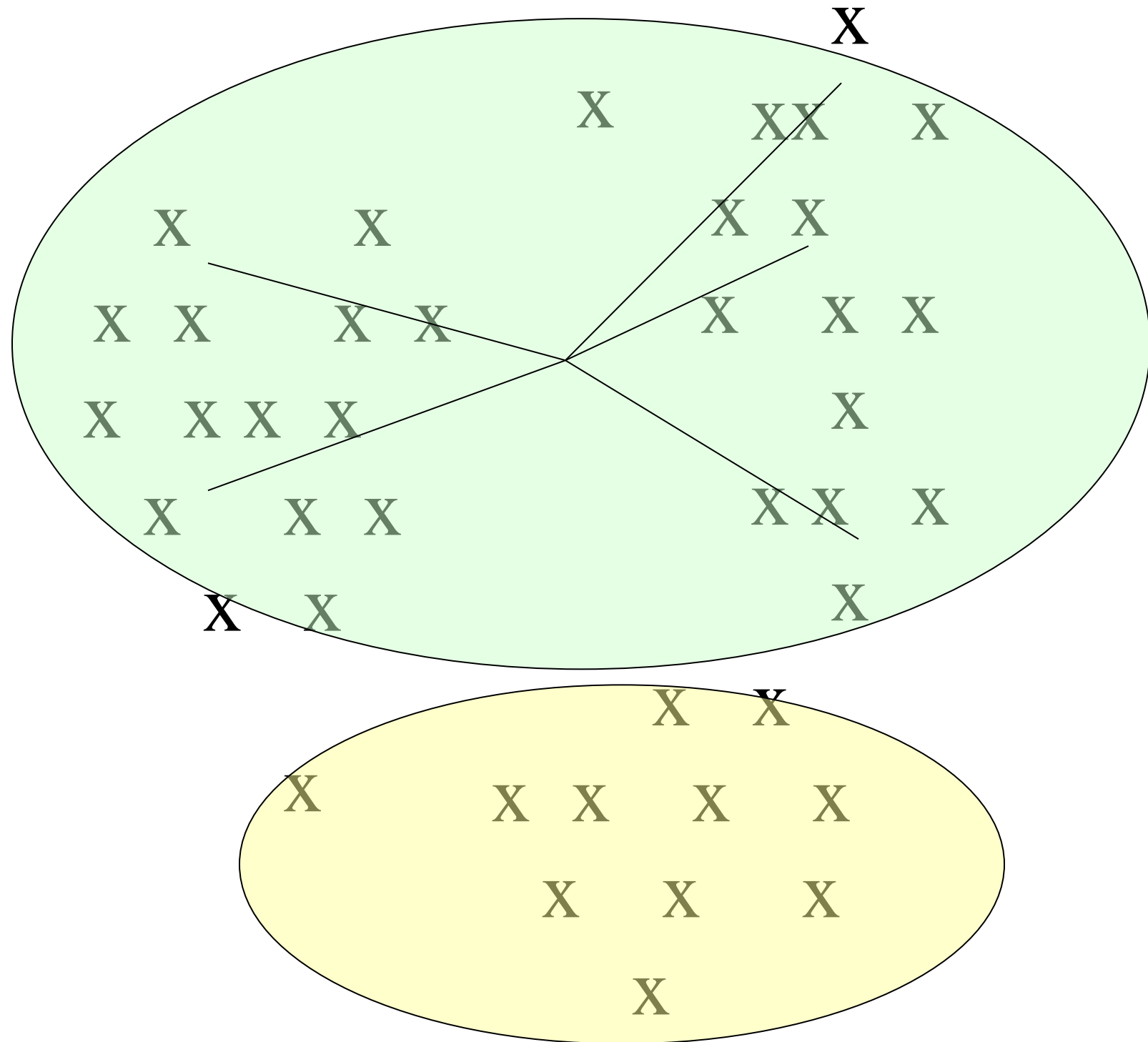
Getting the *k* right

How to select *k*?

- Try different **k**, looking at the change in the average distance to centroid as **k** increases

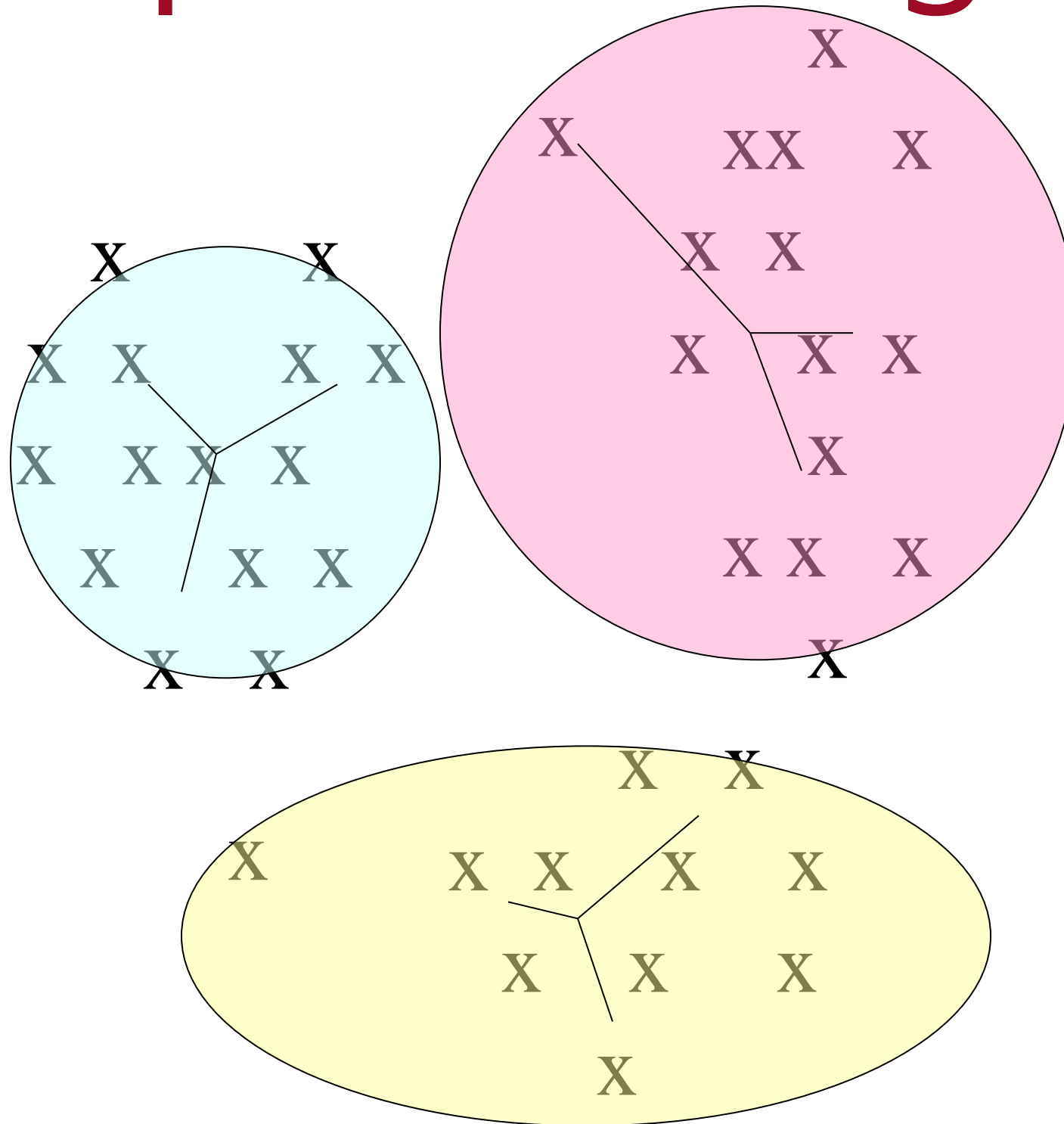- Average falls rapidly until right **k**, then changes little

Best value of *k*

Average distance to centroid

k

Example: Picking *k*

Too few;
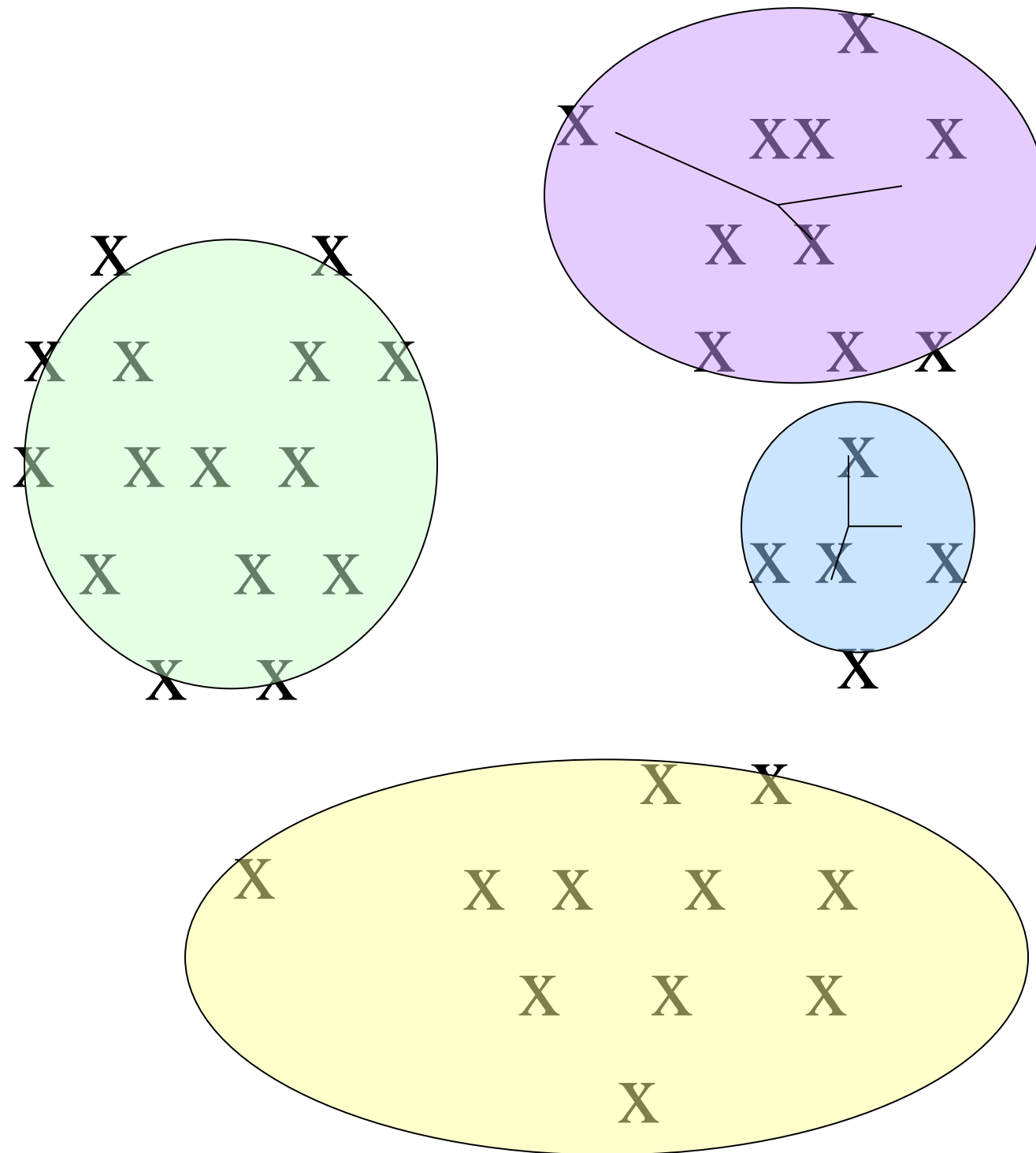many long
distances
to centroid.

Example: Picking *k*

Just right; distances rather short.

Example: Picking *k*

**Too many;** little improvement in average distance.

Variations of the *K-Means* Method

- A few variants of the *k-means* which differ in

  - Selection of the initial *k* means

  - Dissimilarity calculations

  - Strategies to calculate cluster means

- Handling categorical data: *k-modes* (Huang'98)

  - Replacing means of clusters with <u>modes</u>

  - Using new dissimilarity measures to deal with categorical objects

  - Using a <u>frequency</u>-based method to update modes of clusters

  - A mixture of categorical and numerical data: *k-prototype* method

What Is the Problem of the K-Means Method?

- The k-means algorithm is sensitive to outliers !

  - Since an object with an extremely large value may
    substantially distort the distribution of the data.
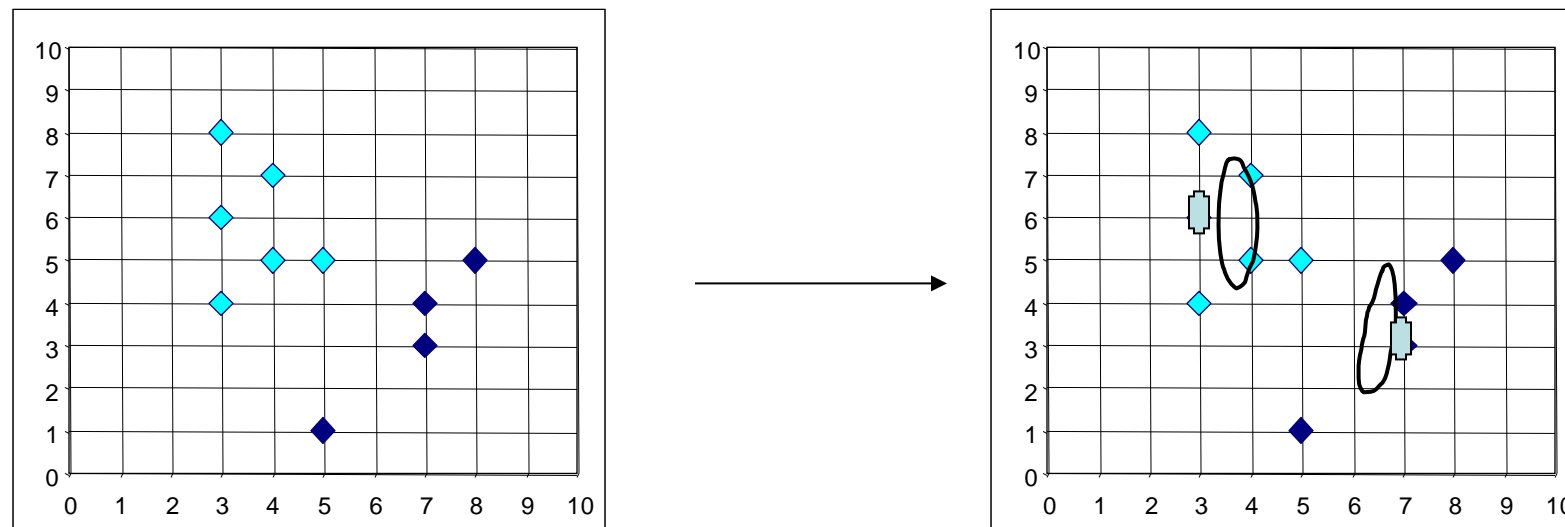
Dendrogram:
Shows How the clusters are Merged

Decompose data objects into a several levels of nested partitioning (<u>tree</u> of clusters), called a <u>dendrogram</u>.

A <u>clustering</u> of the data objects is obtained by <u>cutting</u> the dendrogram at the desired level, then each <u>connected component</u> forms a cluster.