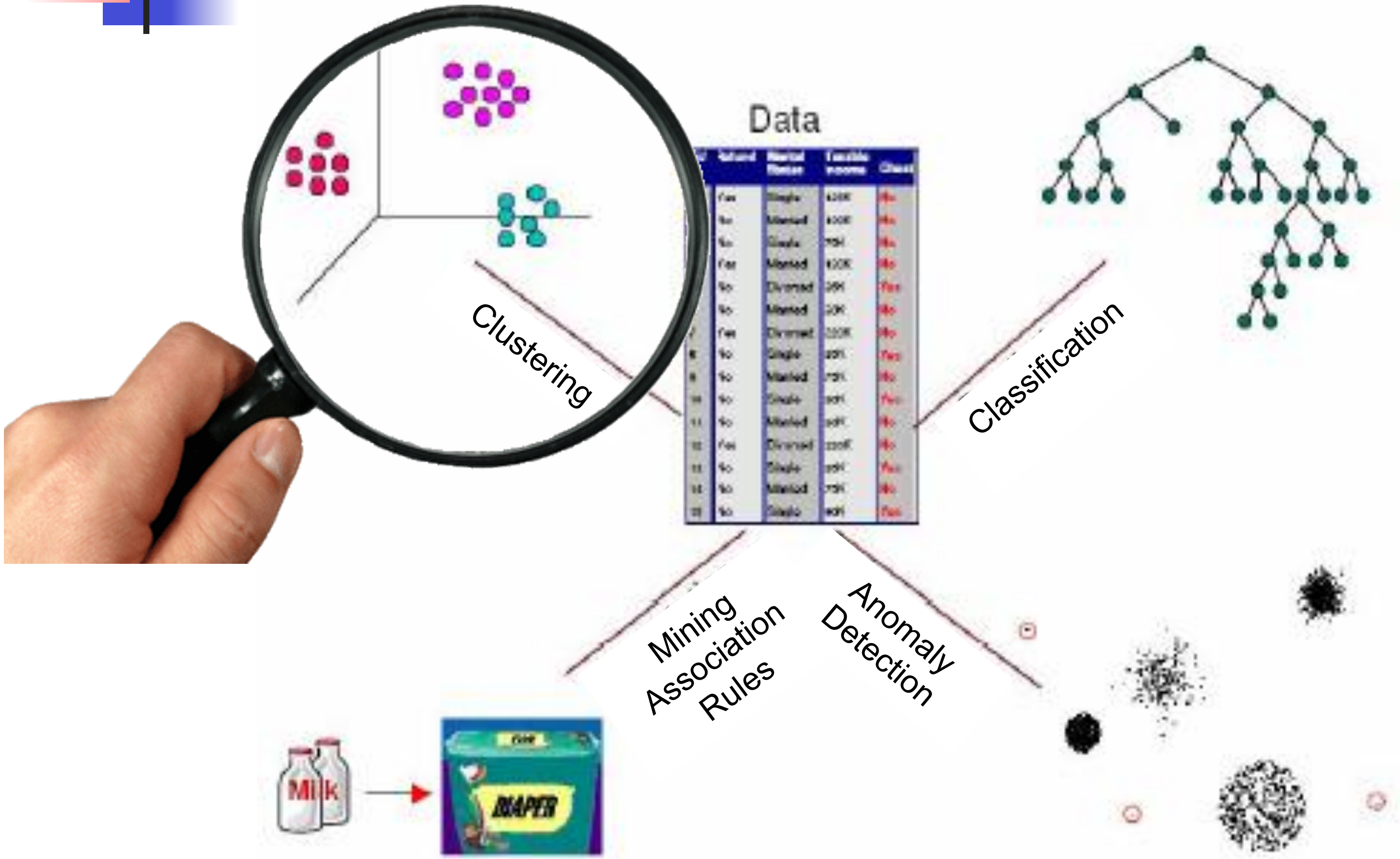
Data Mining Tasks



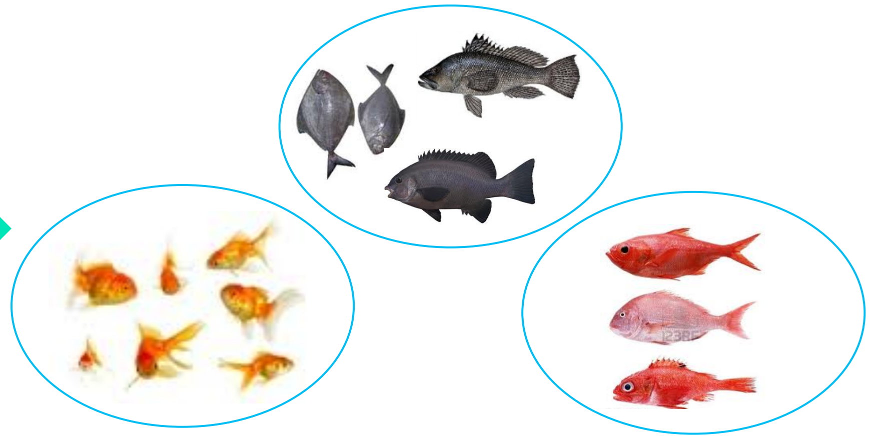
Data Mining

- Clustering (I)

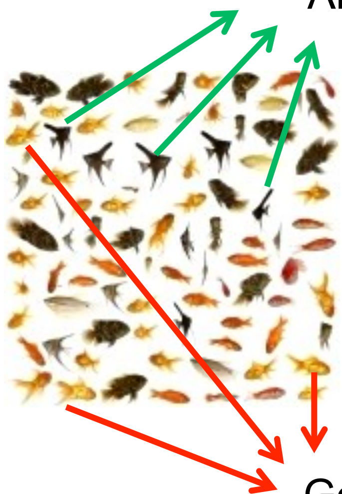
Clustering vs. Classification

clustering

colour feature
no labels



Angelfish



Goldfish

classification

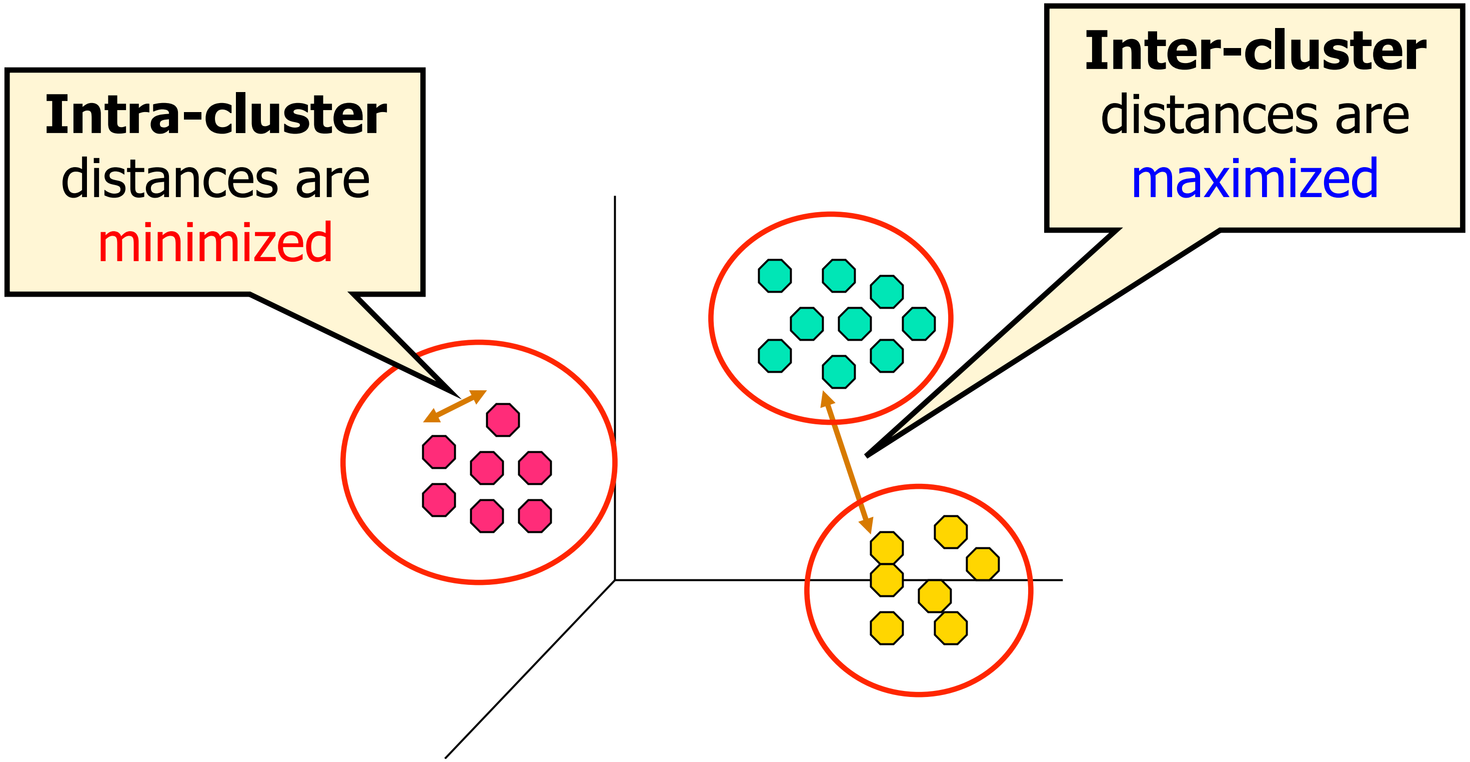
Angelfish:
Up to 6inches or 15cm. Their bodies are very thin, yet tall, their profile rounded, almost disc-shaped.

Salmon:
pinkish in color and have spots on their fins and back

Tuna:
The tuna is a streamlined fish, stout in the middle

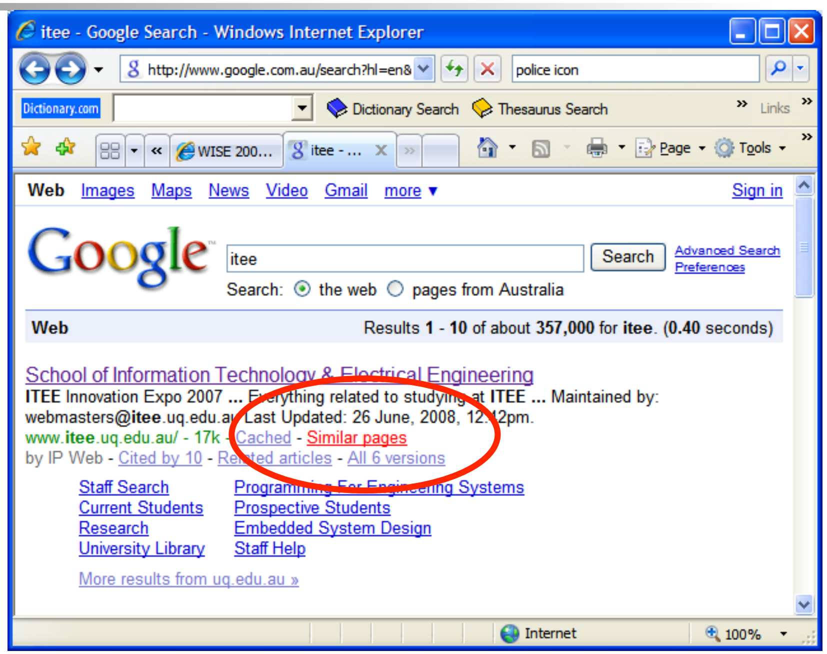
What is Cluster Analysis?

- Finding **groups** of objects such that the objects in a group will be
  1. **similar** (or related) to one another; and
  2. **different** from (or unrelated to) the objects in other groups

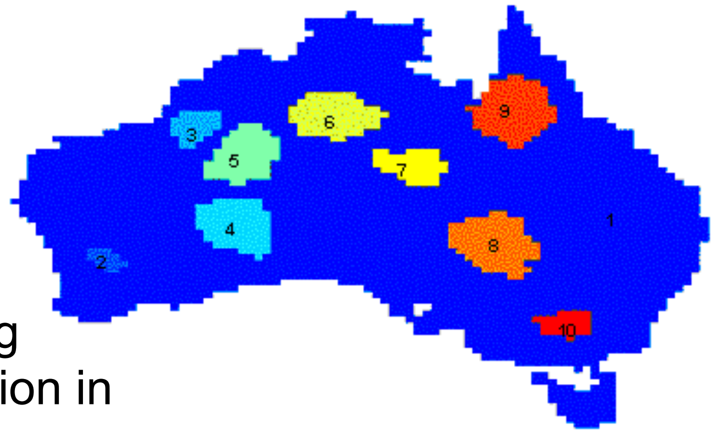


Applications of Cluster Analysis

- **Understanding**
  - Group related documents for browsing,
  - group genes and proteins that have similar functionality, or
  - group stocks with similar price fluctuations



- **Summarization**
  - Reduce the size of large data sets



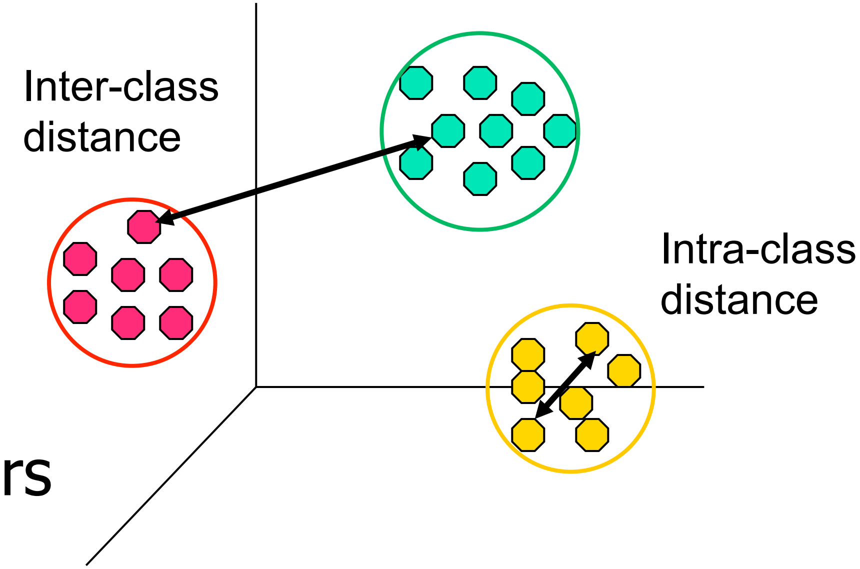
Clustering precipitation in Australia

Clustering as a Preprocessing Tool (Utility)

- **Summarization:**
  - Preprocessing for: easy for expert to clean the data
    - Classification
    - Recommendation
    - ..
- **Outlier detection:**
  - Outliers are often viewed as "far away" from any cluster
- …

What is a Good Clustering?

- A "*good*" clustering method will produce high quality clusters
  - **high** intra-class similarity:
    - cohesive within clusters
  - **low** inter-class similarity:
    - distinctive between clusters



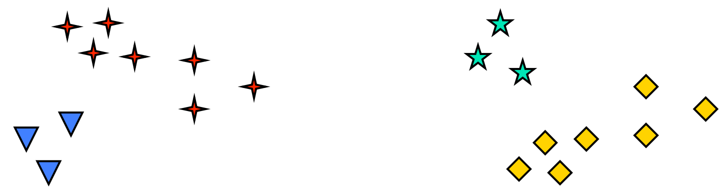
- The "*quality*" of a clustering method depends on
  - the similarity measure used by the method

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters



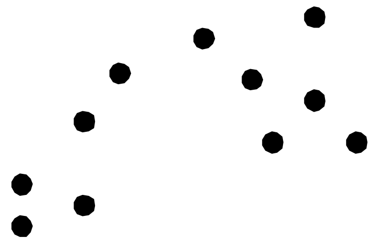
Four Clusters

Types of Clusterings

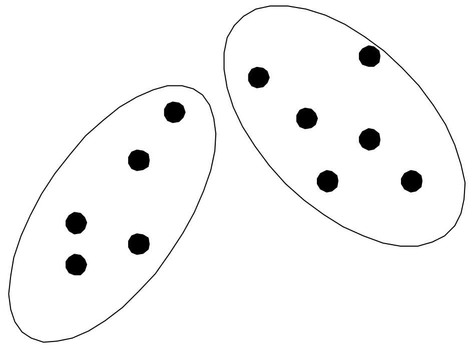
- A clustering is a set of clusters
- A Cluster: a collection of data objects
  - Similar to one another within the same group
  - Dissimilar to the objects in other groups

- Partitional Clustering
  - A division of data objects into **non-overlapping** subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of **nested** clusters organized as a hierarchical tree

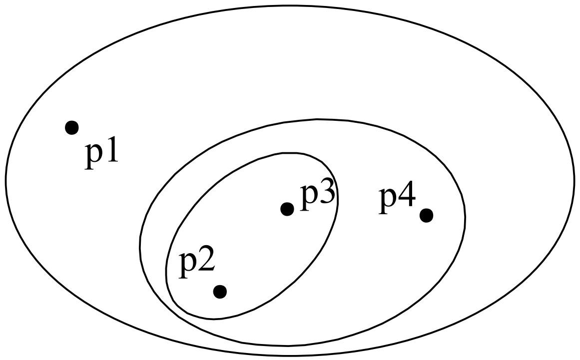
Partitional vs. Hierarchical Clustering



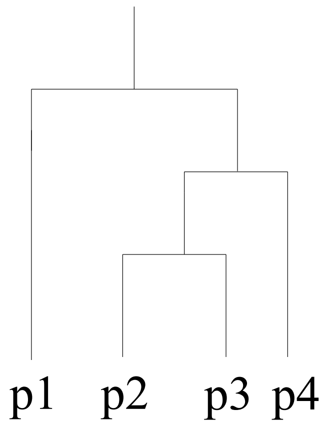
Original Points



A Partitional Clustering



Hierarchical Clustering



Dendrogram

website navigation

Clustering Algorithms

- K-means
- Hierarchical clustering
- Density-based clustering

K-means Clustering

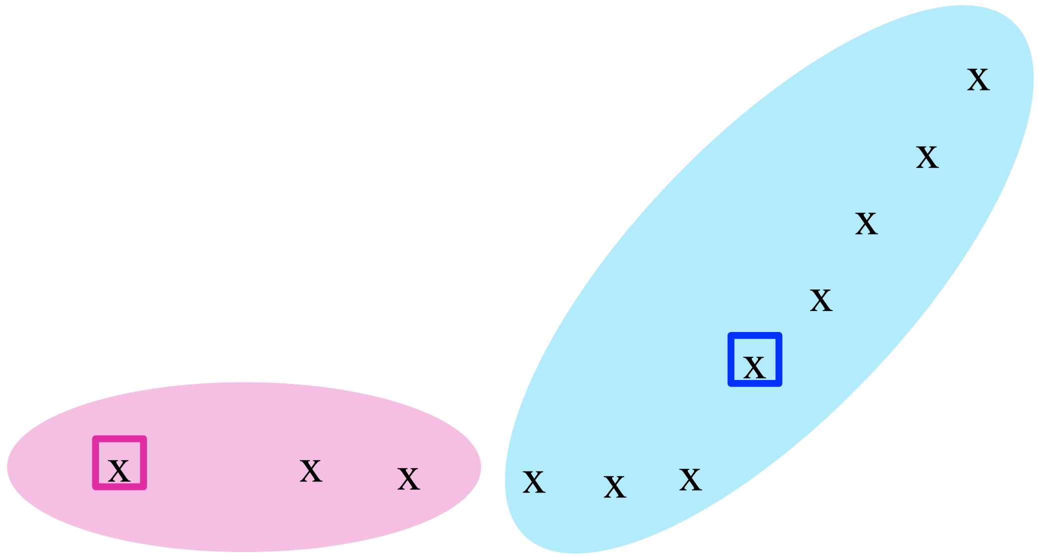
- Partitional clustering approach
- Each cluster is associated with a **centroid** (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters ***K*** must be specified

K-Means Algorithm

- Steps:

```
Select K points as the initial centroids

Repeat

   Form K clusters by Assigning all
   points to the nearest centroid

   Re-compute the centroid of each
   cluster

Until all the centroids do not change
```

Example: Assigning Clusters

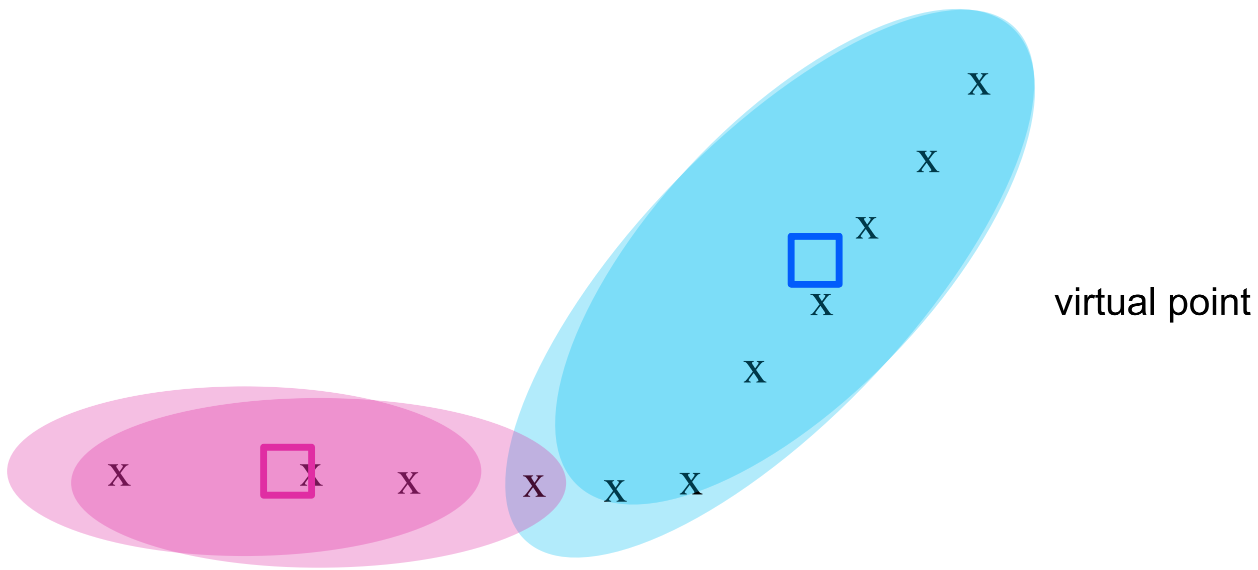


x … data point

□ … centroid

Clusters after round 1

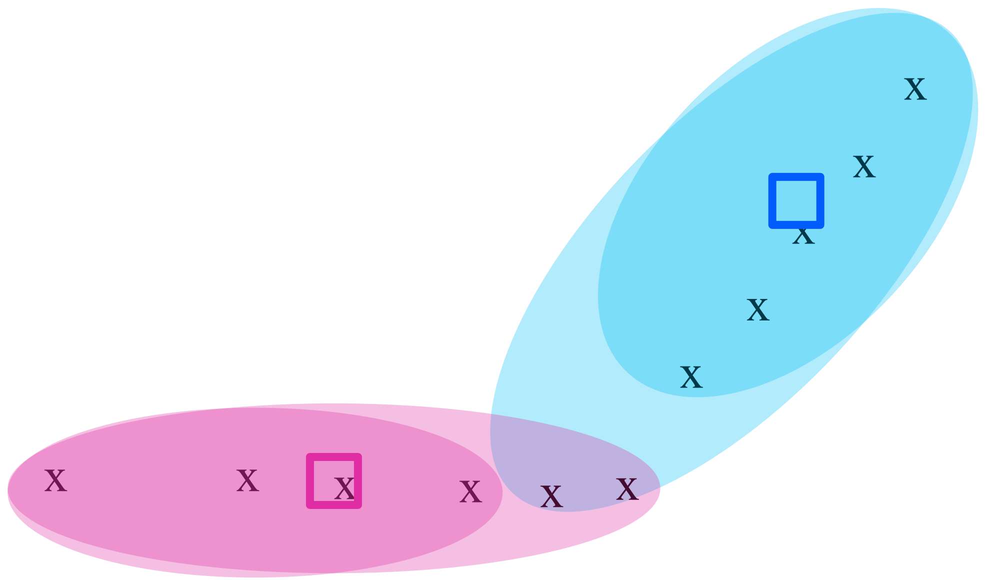
Example: Assigning Clusters



x … data point
□ … centroid

Clusters after round 2

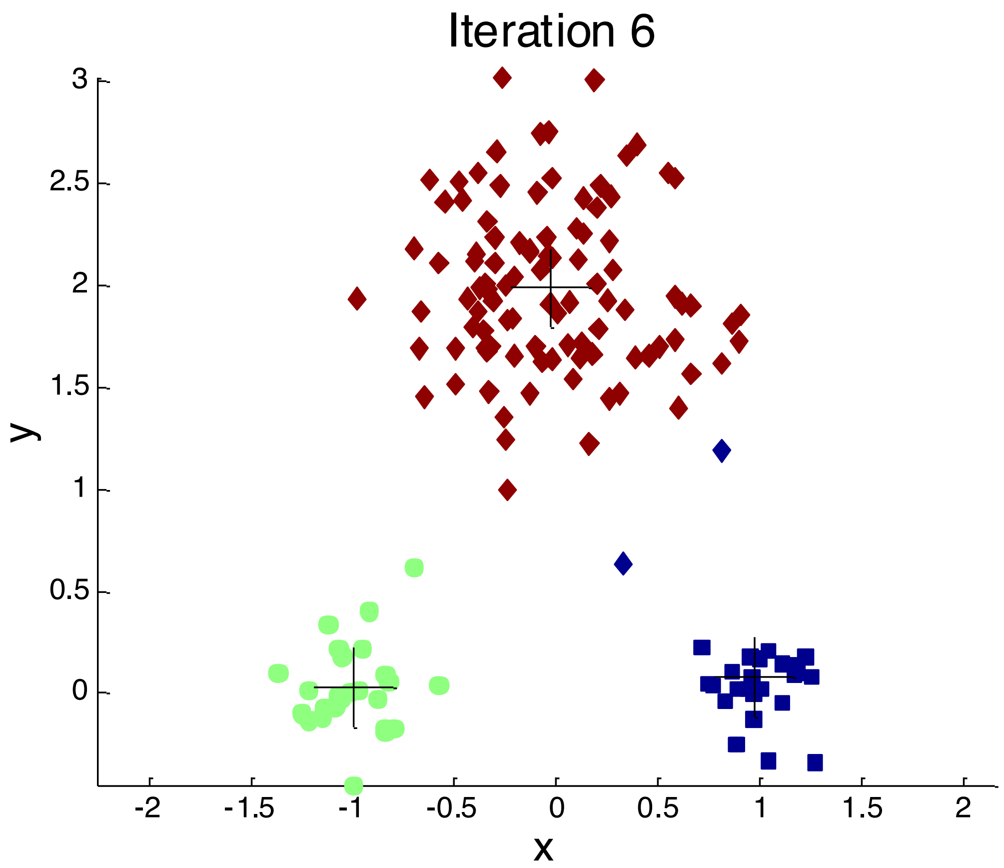
Example: Assigning Clusters



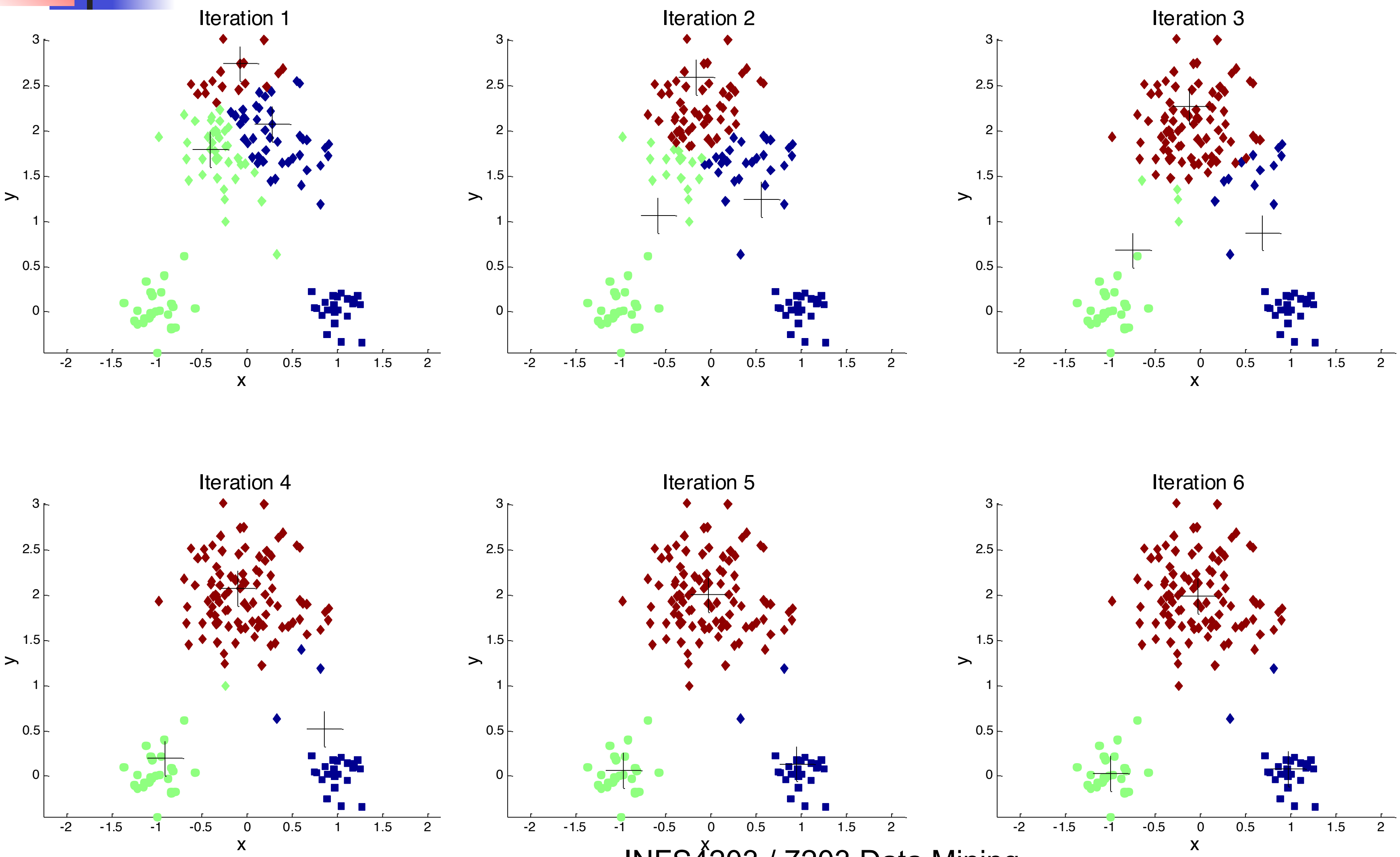
x … data point
□ … centroid

Clusters at the end

K-Means: Example



K-Means: Example



K-means Clustering – Details

- **Select**
  - Initial centroids are often chosen randomly
    - Clusters produced vary from one run to another
- **Nearest**
  - Closeness is measured by Euclidean distance, cosine similarity, etc.
- **Re-compute**
  - A centroid is typically the mean of the points in a cluster

Example

Suppose the data mining task is to cluster the following measurements of *age* into **three** groups:

18, 22, 25, 42, 27, 43, 33, 35, 56, 28,

Use *k-means* algorithm to show the clustering procedure

Suppose the initial centroids are 22, 35 and 43, show the final three clusters.

Example

Cluster#	Old Centroid	Cluster Elements	new Centroid
1	22	18, 22, 25, 27, 28	24
2	35	33, 35	34
3	43	42, 43, 56	47

Cluster#	Old Centroid	Cluster Elements	new Centroid
1	24	18, 22, 25, 27, 28	24
2	34	33,35	34
3	47	42,43,56	47

K-means Clustering – Details

- K-means will **converge** for the common similarity measures mentioned above

- Most of the convergence happens in the first few iterations

  - Often the stopping condition is changed to:

`'Until relatively few points change clusters'`

  - Convergence does not necessarily mean **optimal** clustering!

    - How to evaluate clustering?

Evaluating K-means Clusters

- Most common measure is Sum of Squared Error **(SSE)**
  - For each point, the **error** is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(c_i, x)$$

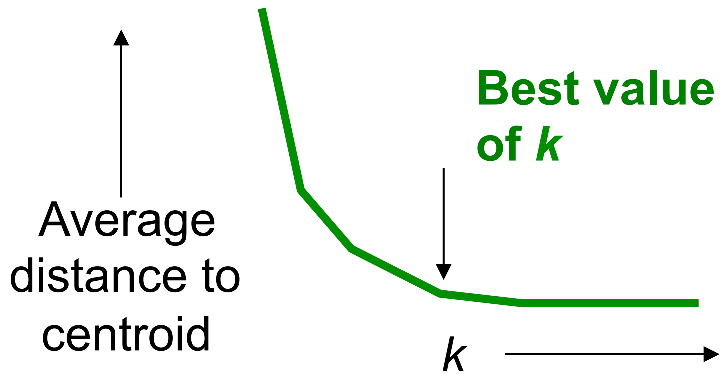
value is high, this is bad

  - K is the number of clusers
  - x is a data point in cluster C_i
  - c_i is the centroid point for cluster C_i
- SSE is basically the sum of SSE of each cluster
- Given two clusters, we choose the one with smaller error

Getting the *k* right

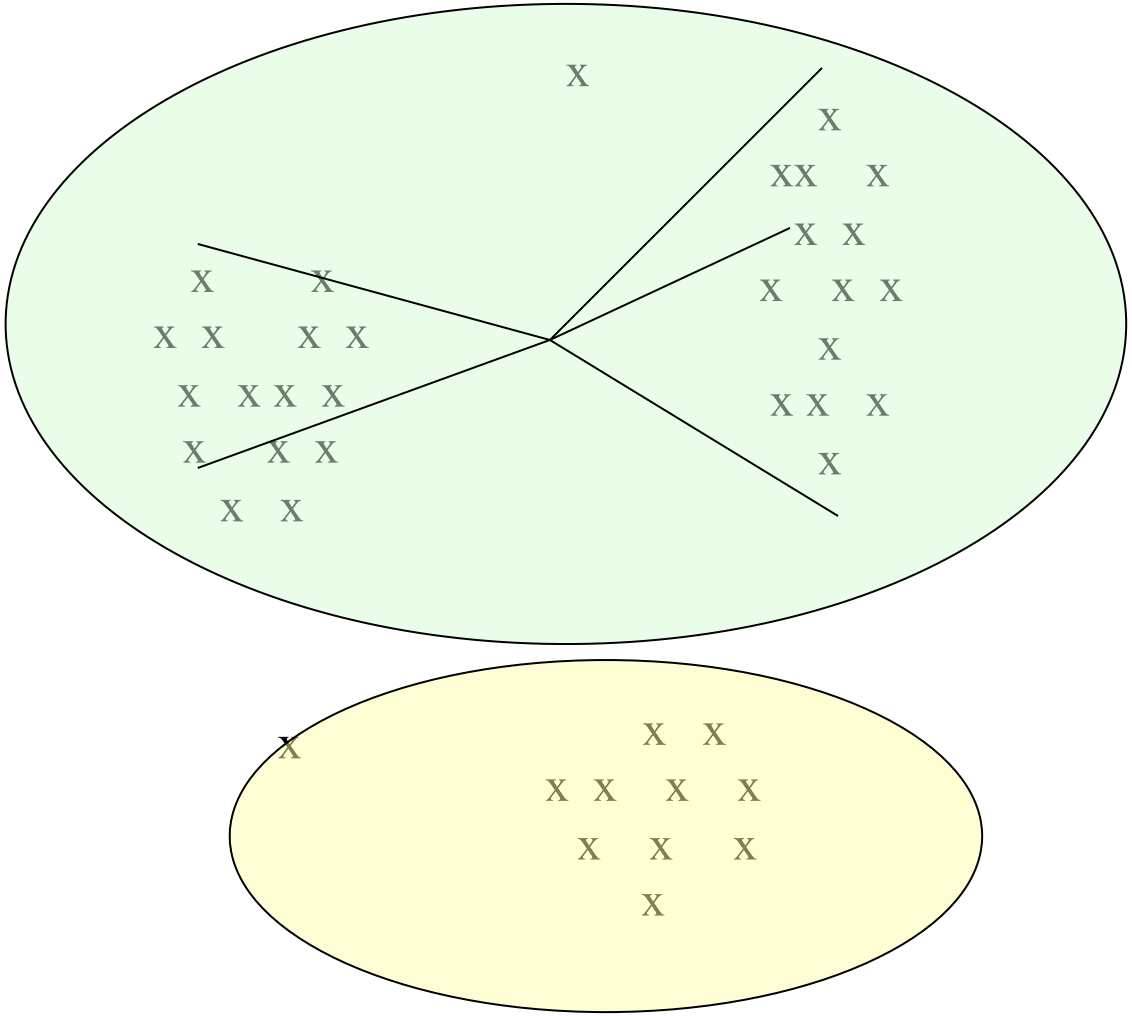
How to select *k*?

- 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



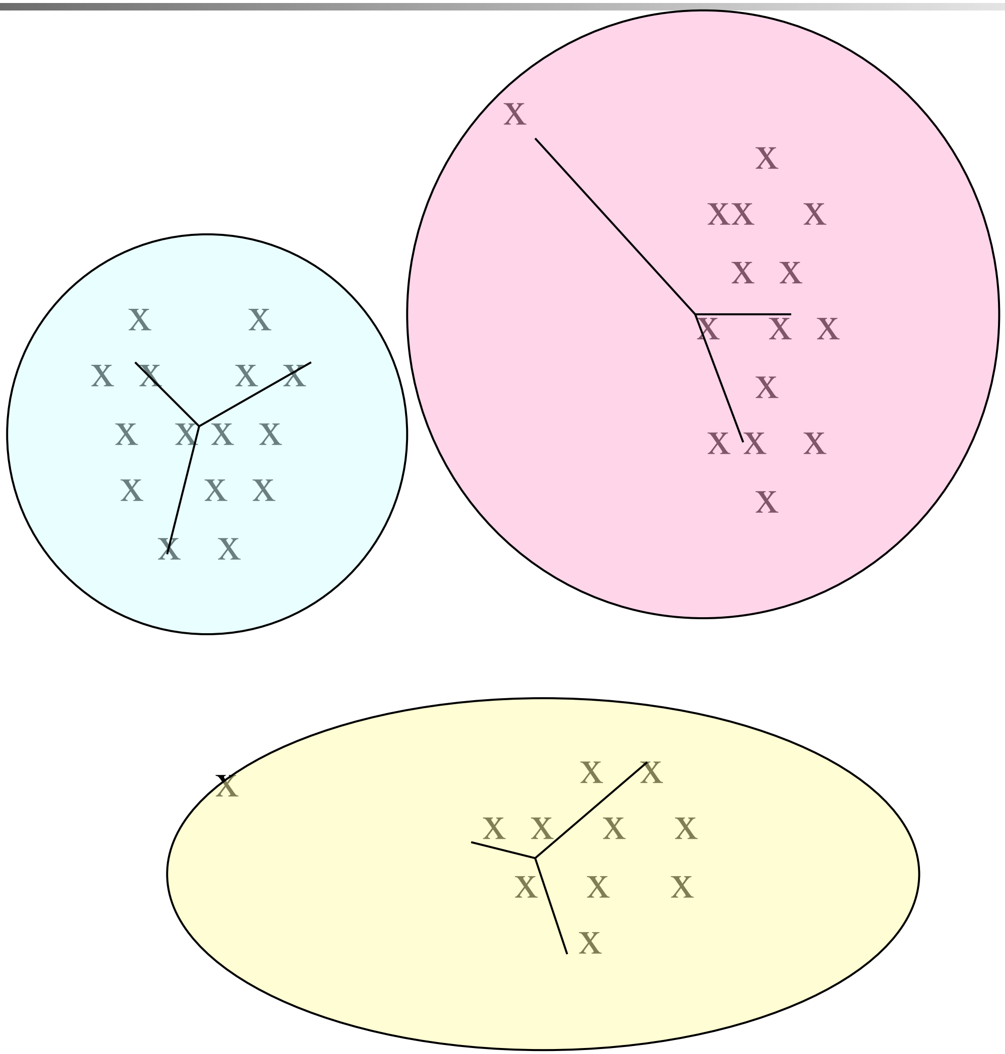
Example: Picking *k*

Too few;
many long
distances
to centroid.



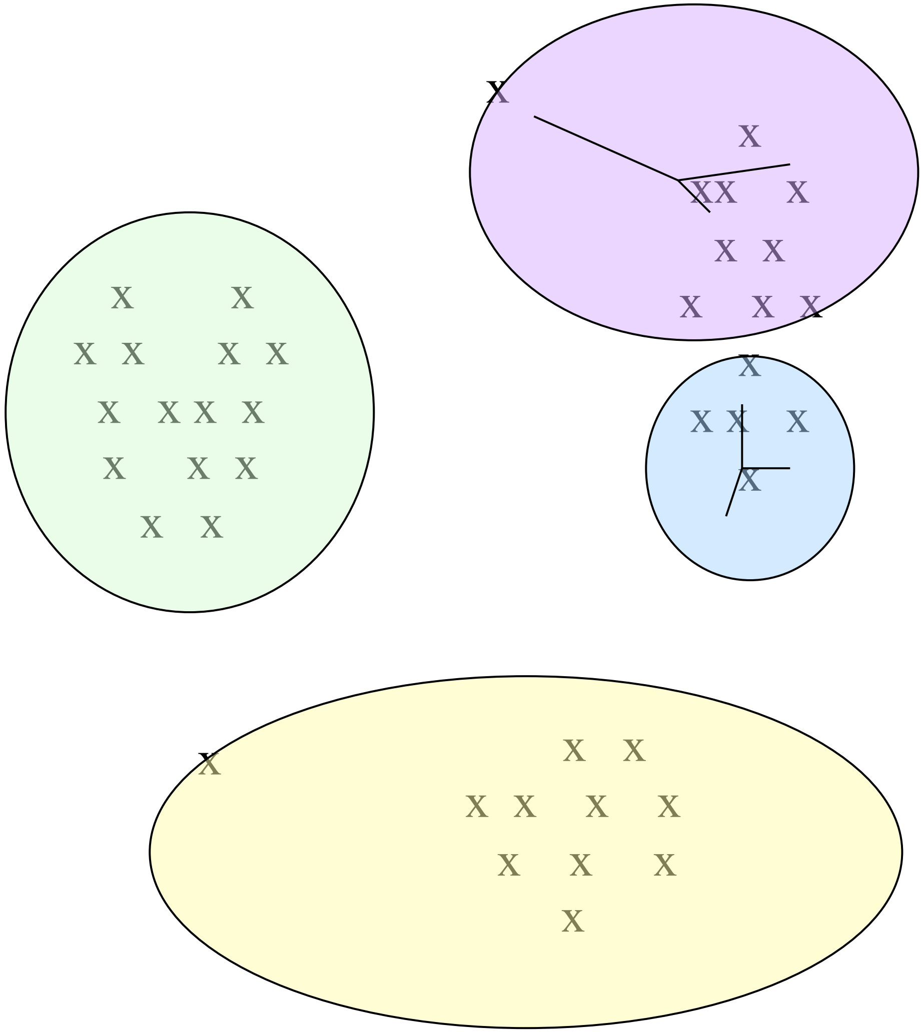
Example: Picking *k*

Just right;
distances
rather short.

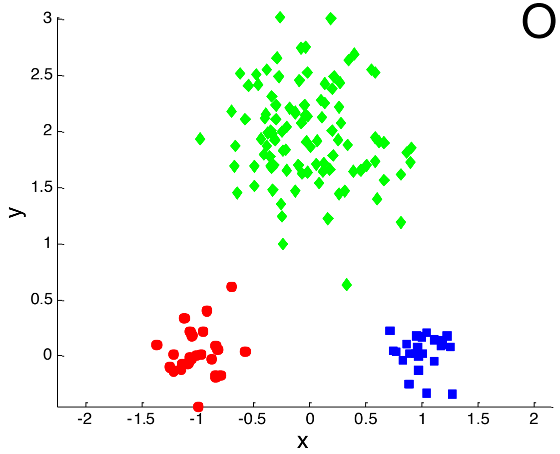


Example: Picking *k*

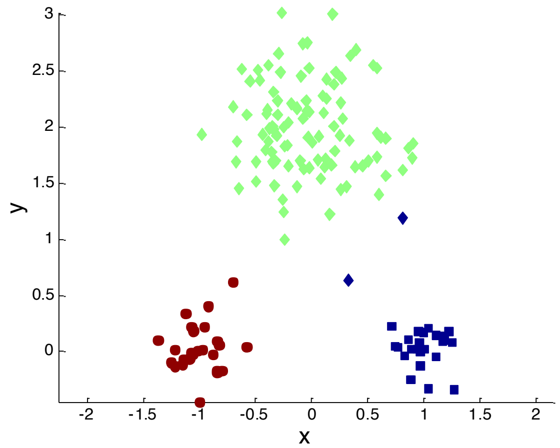
Too many;
little improvement
in average
distance.



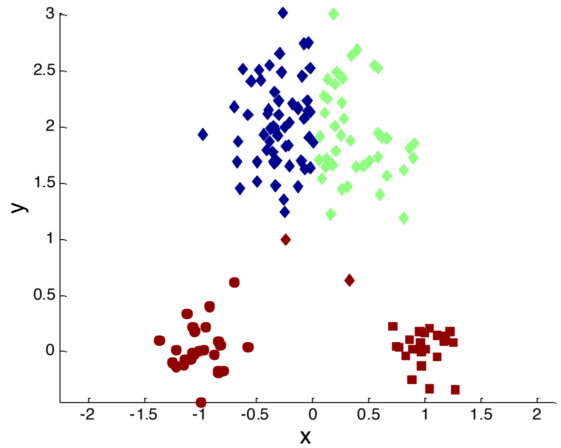
Two different K-means Clusterings



Original Points

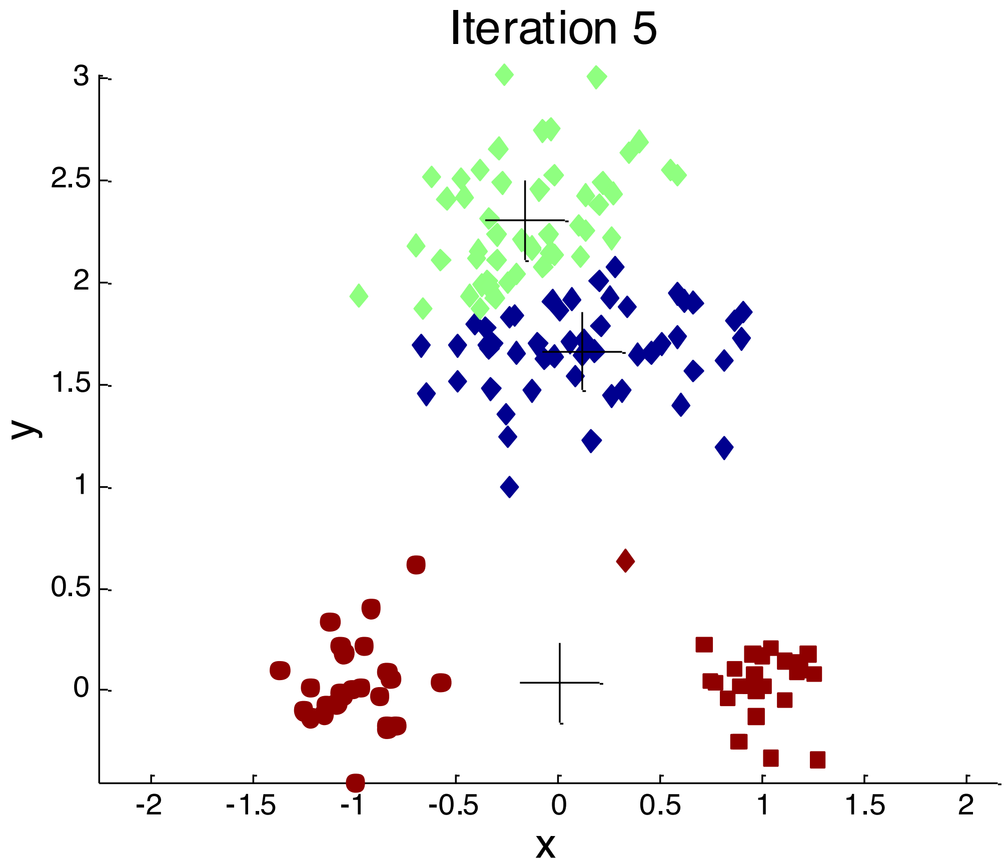


Optimal Clustering

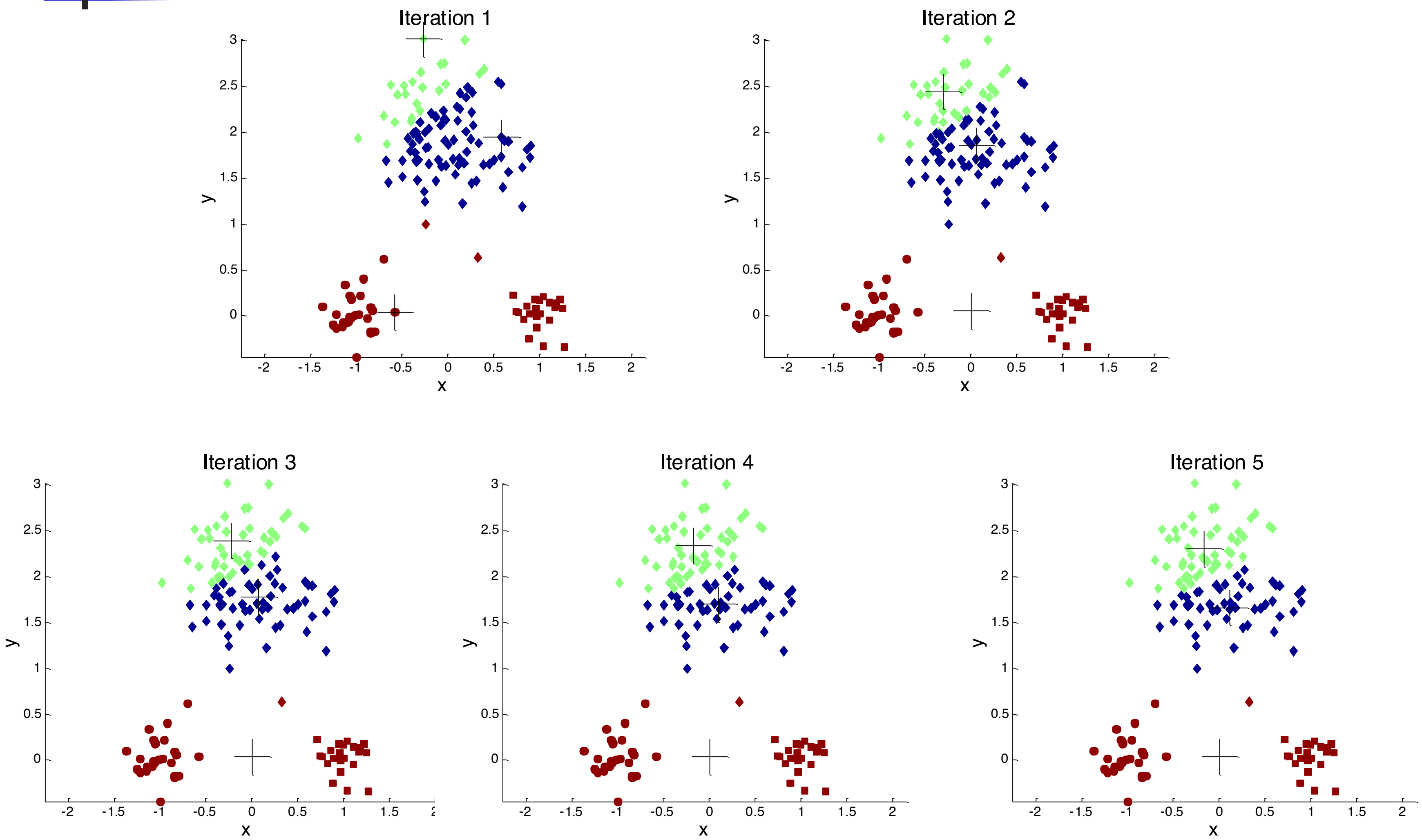


Sub-optimal Clustering

Importance of Choosing Initial Centroids …



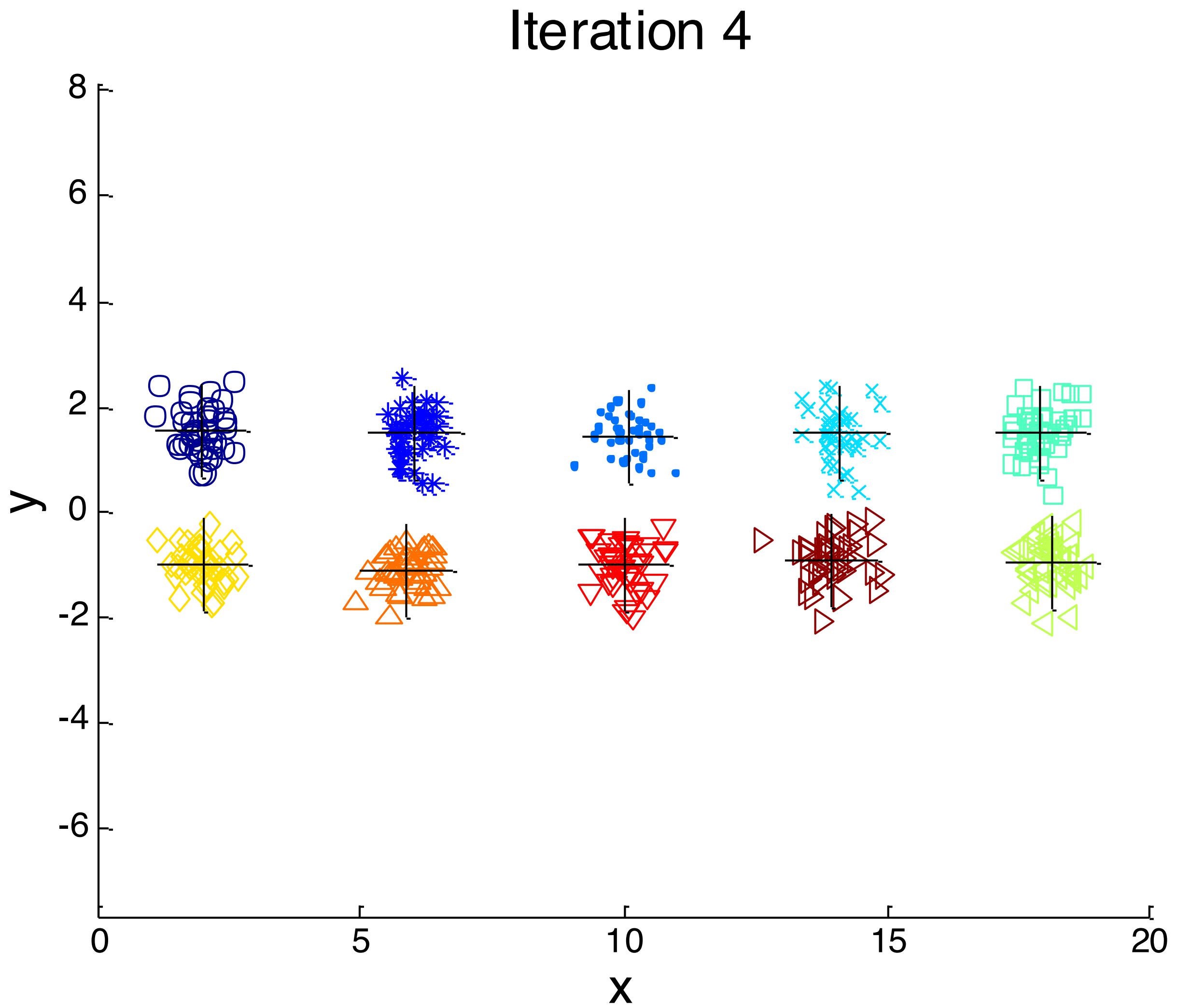
Importance of Choosing Initial Centroids …



Problems with Selecting Initial Points

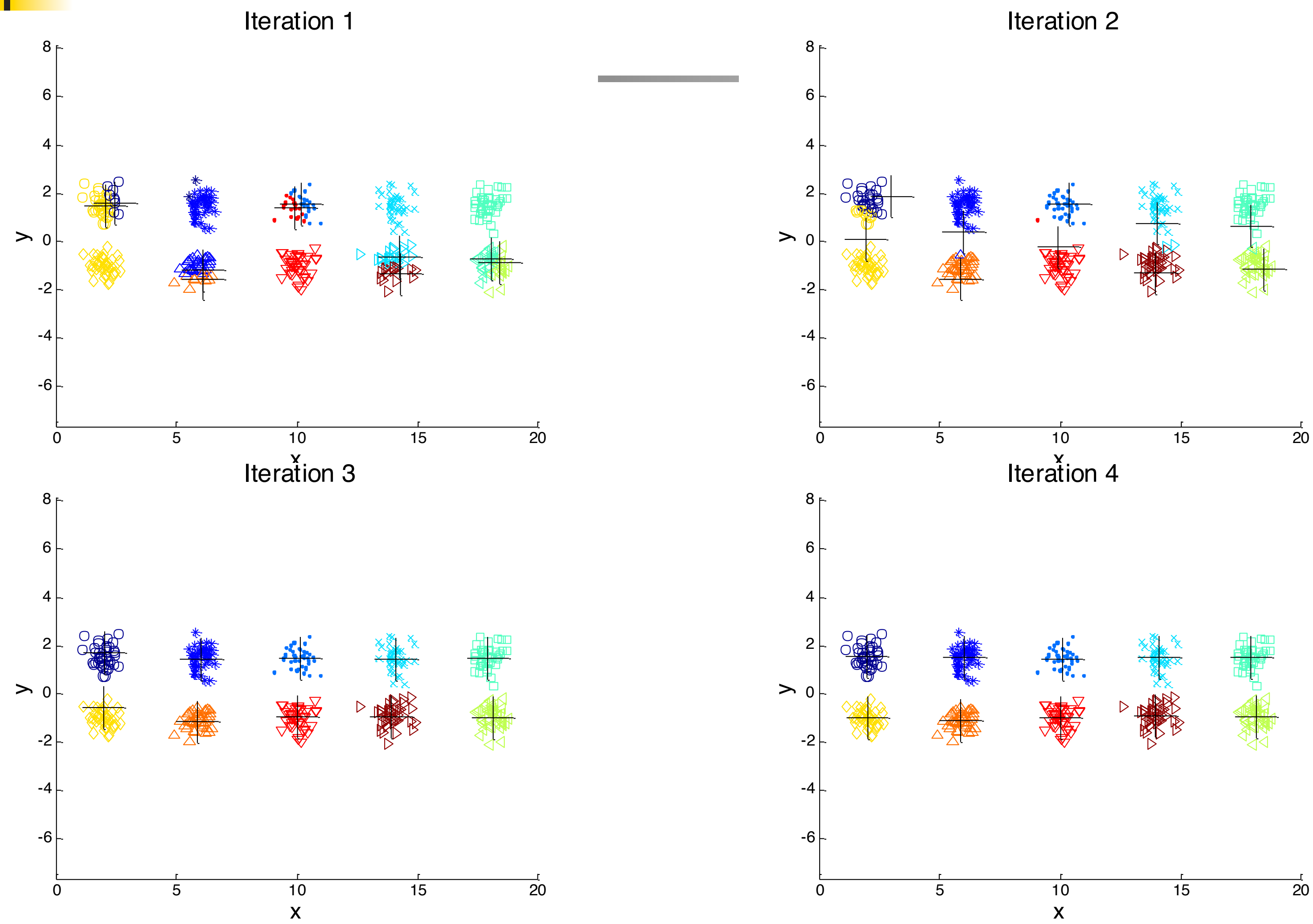
- Sometimes the initial centroids will readjust themselves in the 'right' way,
  - and sometimes they don't!
- Consider the following example of five pairs of clusters..

10 Clusters Example



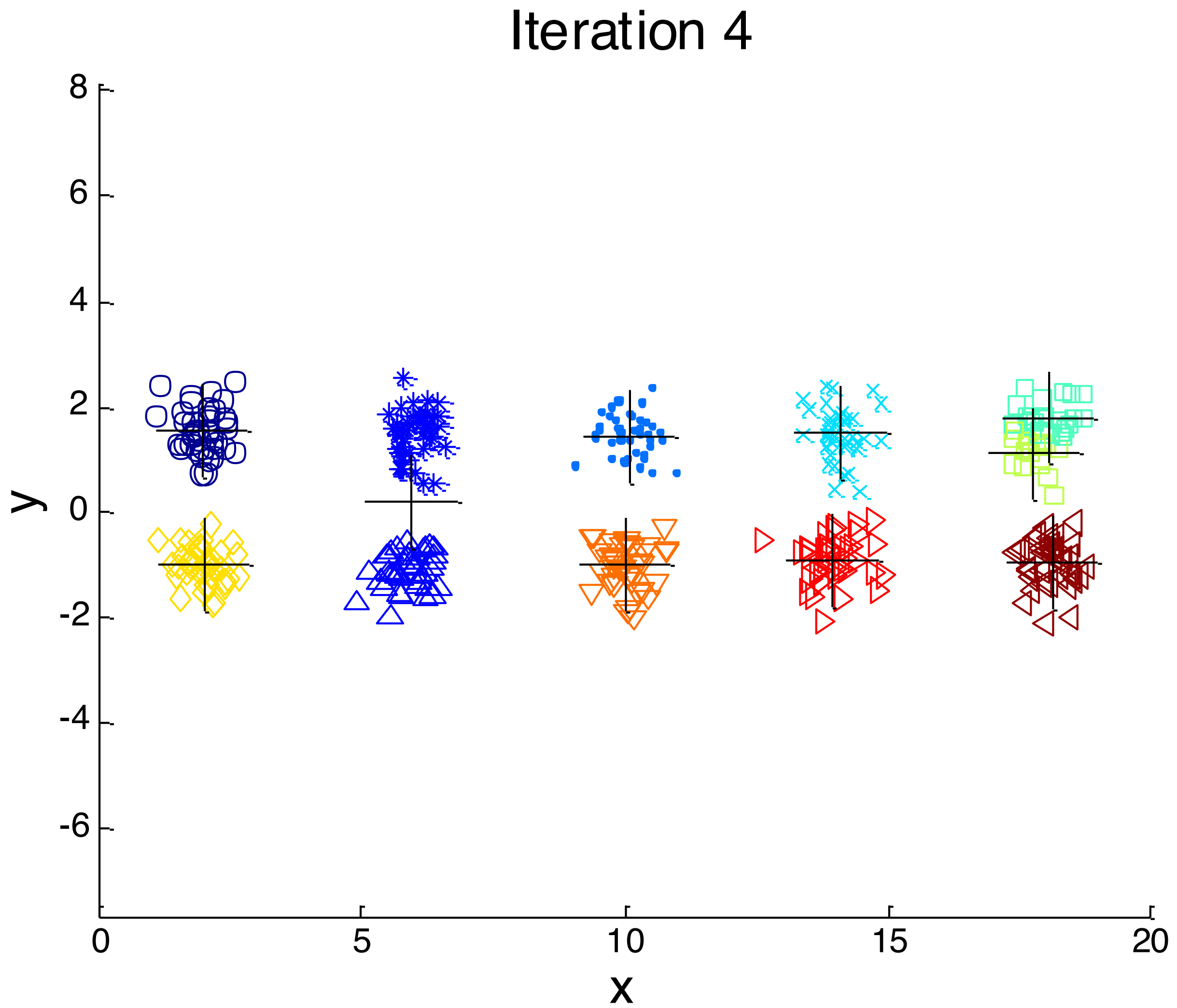
Starting with **two** initial centroids in one cluster of each pair of clusters

10 Clusters Example



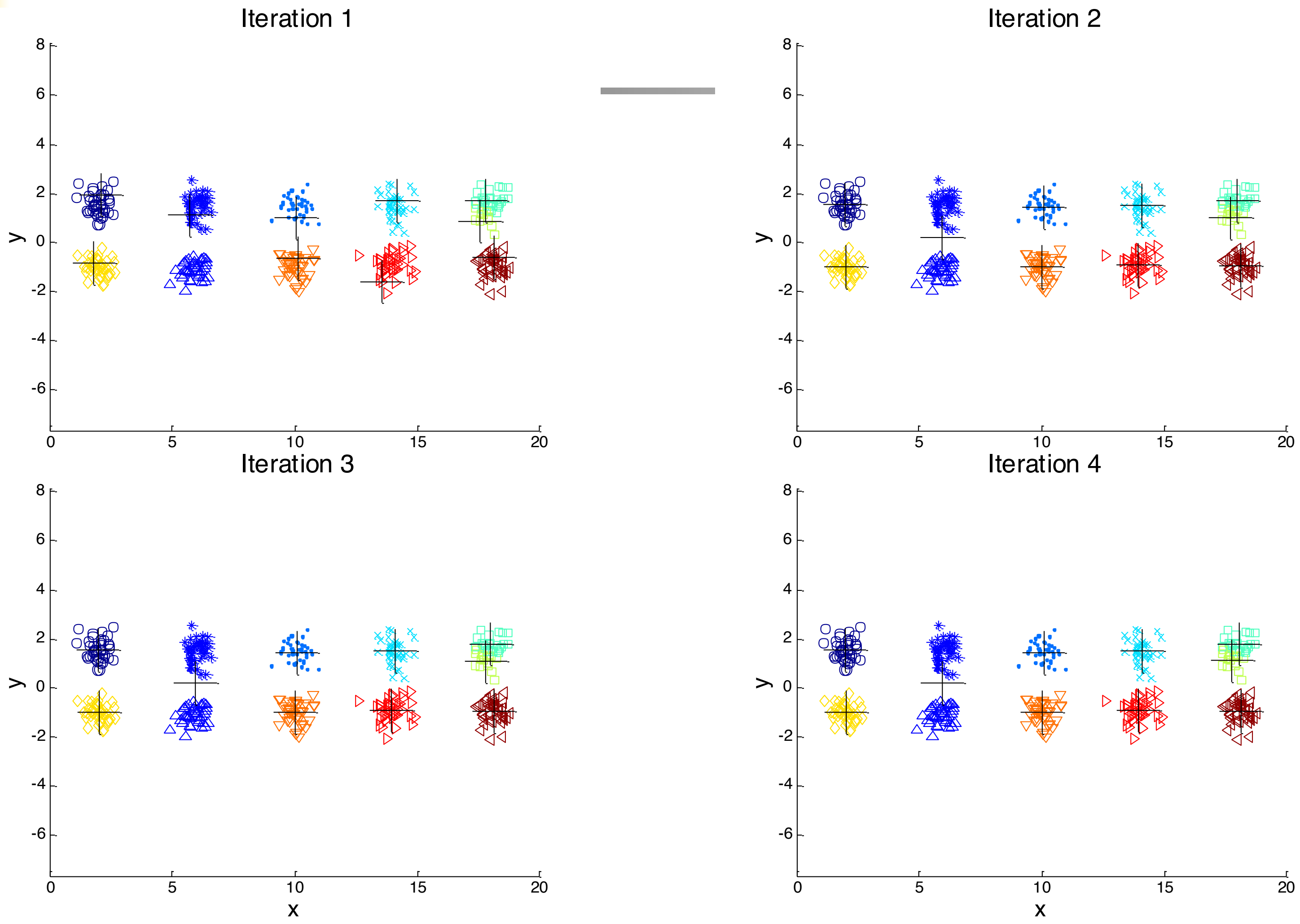
Starting with **two** initial centroids in one cluster of each pair of clusters

10 Clusters Example



Starting with some pairs of clusters having **three** initial centroids, while other have only **one**.

10 Clusters Example



Starting with some pairs of clusters having **three** initial centroids, while other have only **one**.

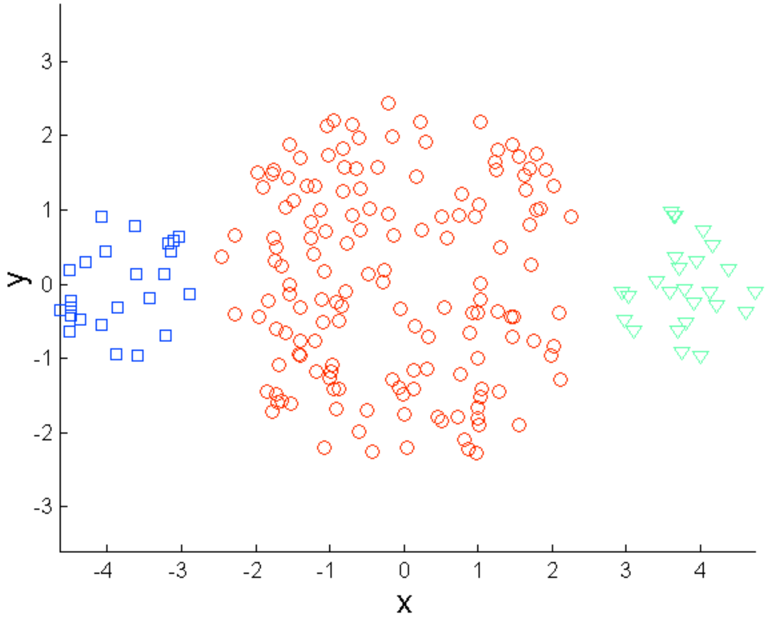
Solutions to Initial Centroids Problem

- Multiple runs
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing
  - **Eliminate** 'small' clusters that may represent outliers
  - **Split** 'loose' clusters (clusters with relatively high SSE)
  - **Merge** 'close' clusters (clusters with relatively low SSE)

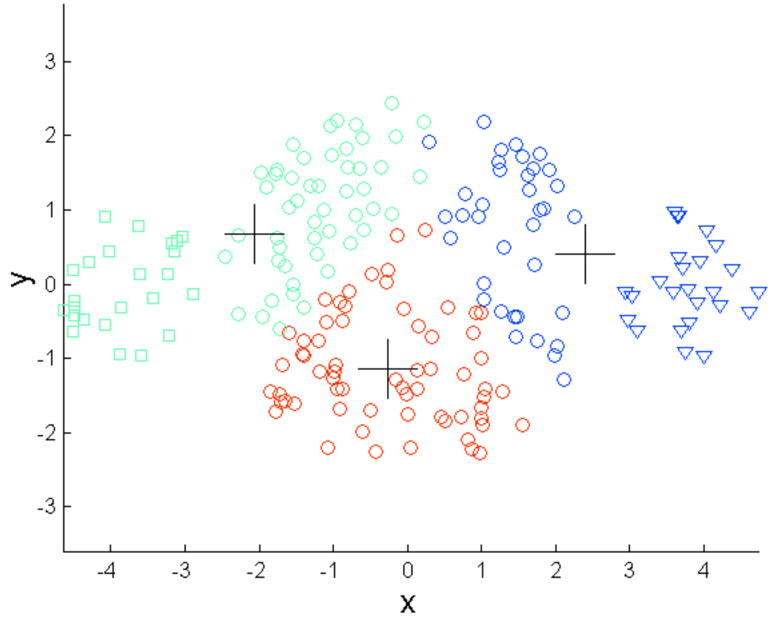
Limitations of K-means

- K-means is simple and suitable for many types of data
- K-means has problems when clusters are of different:
  - Sizes
  - Densities
  - Non-spherical shapes

Limitations of K-means: Different Sizes

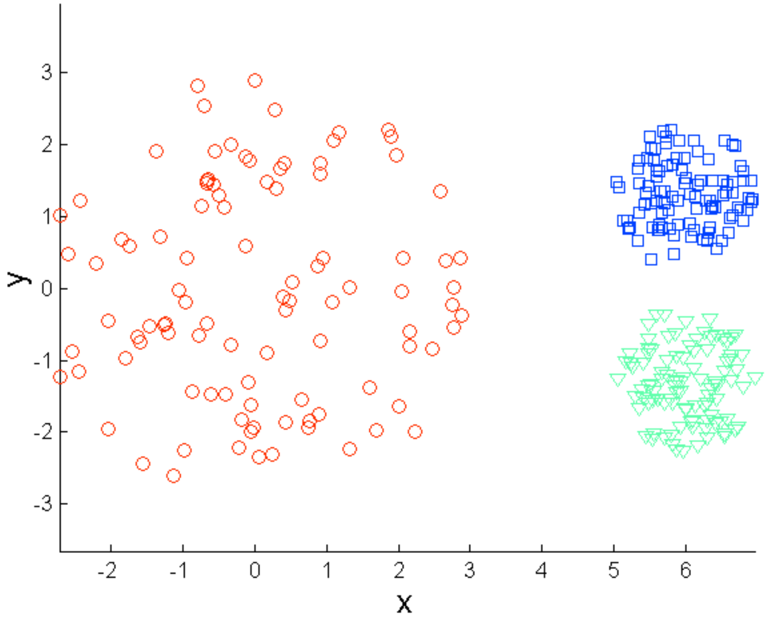


Original Points

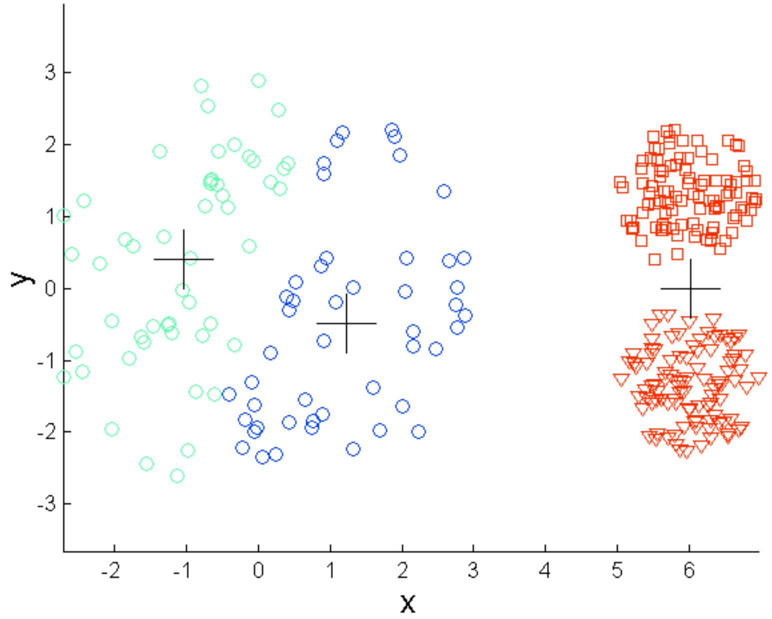


K-means (3 Clusters)

Limitations of K-means: Different Density

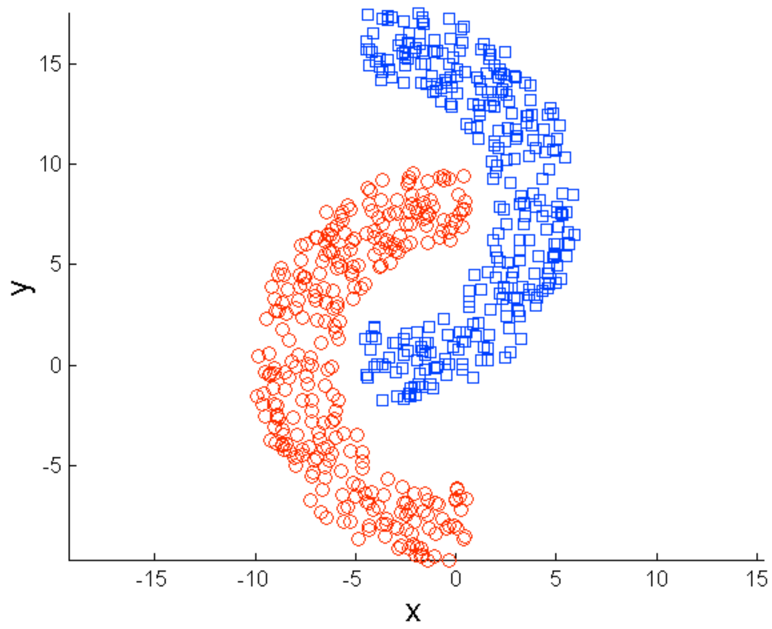


Original Points

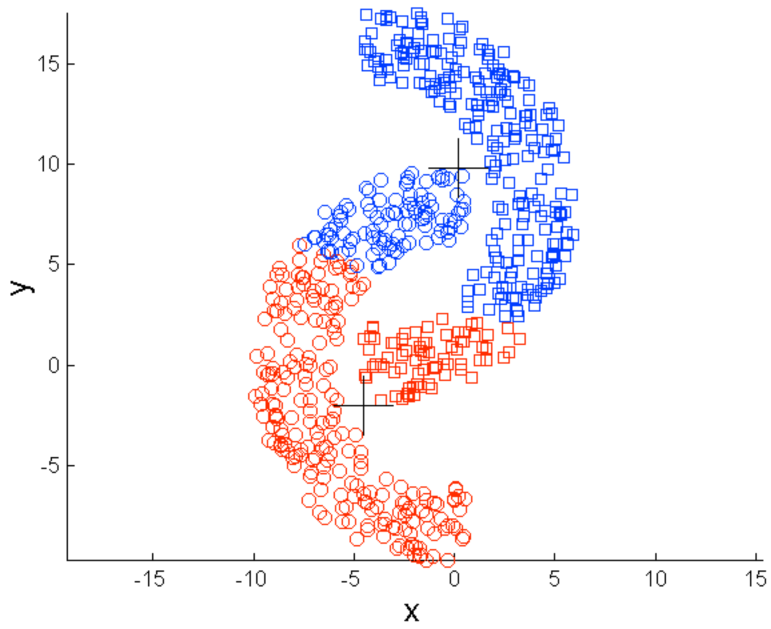


K-means (3 Clusters)

Limitations of K-means: Non-spherical Shapes



Original Points



K-means (2 Clusters)

Clustering Algorithms

- K-means

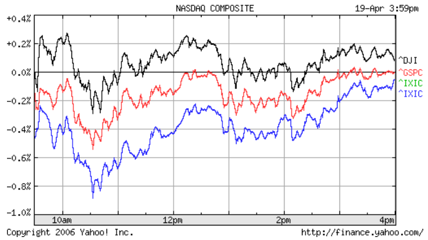
- Hierarchical clustering

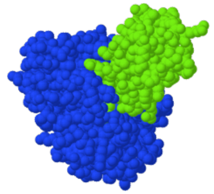
- Density-based clustering

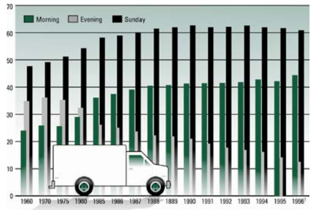
- *But, first…*

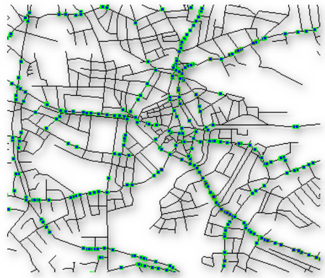
Complex Data Types

- Complex data
  - Text Data
  - Temporal data
  - Spatial data
  - Spatial-temporal data
  - Multimedia data









- How to measure "distance"?