SGG 4653
Advance Database System

Data Mining (Clustering)

CLUSTERING

- Objective of this topic:
  - To understand the differences between Classification and Clustering
  - To understand the important of measurement in clustering process
- Contents of this topic:
  - What is Cluster Analysis
  - Clustering Problem
  - What is Similarity?

Clustering

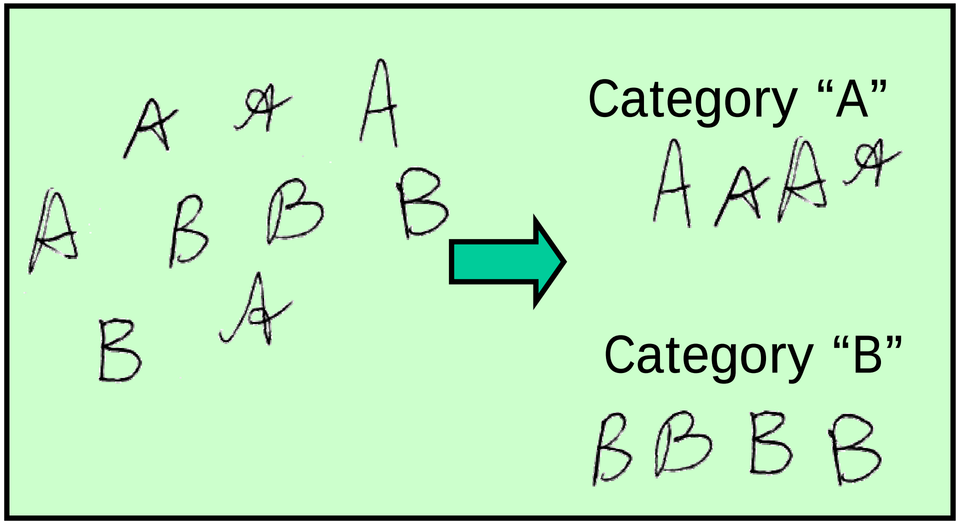
- Partitioning a set of data (or objects) into a set of classes, called clusters, such that members of each class sharing some interesting common properties.

- No prior knowledge
  - Number of clusters
  - Meaning of clusters
- Unsupervised learning

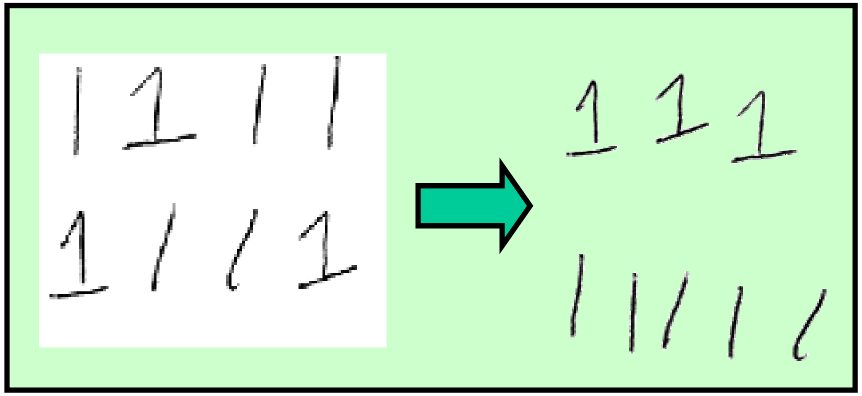
Clustering vs. Classification

- Identification of a pattern as a member of a category (pattern class) we already know, or we are familiar with
  - Supervised Classification (known categories)
  - Unsupervised Classification, or “Clustering” (creation of new categories)

- Classification: The goal is to predict the class variable based on the feature values of samples



Classification



Clustering

What is Similarity?

“The quality or state of being similar; likeness; resemblance; as, a similarity of features”.

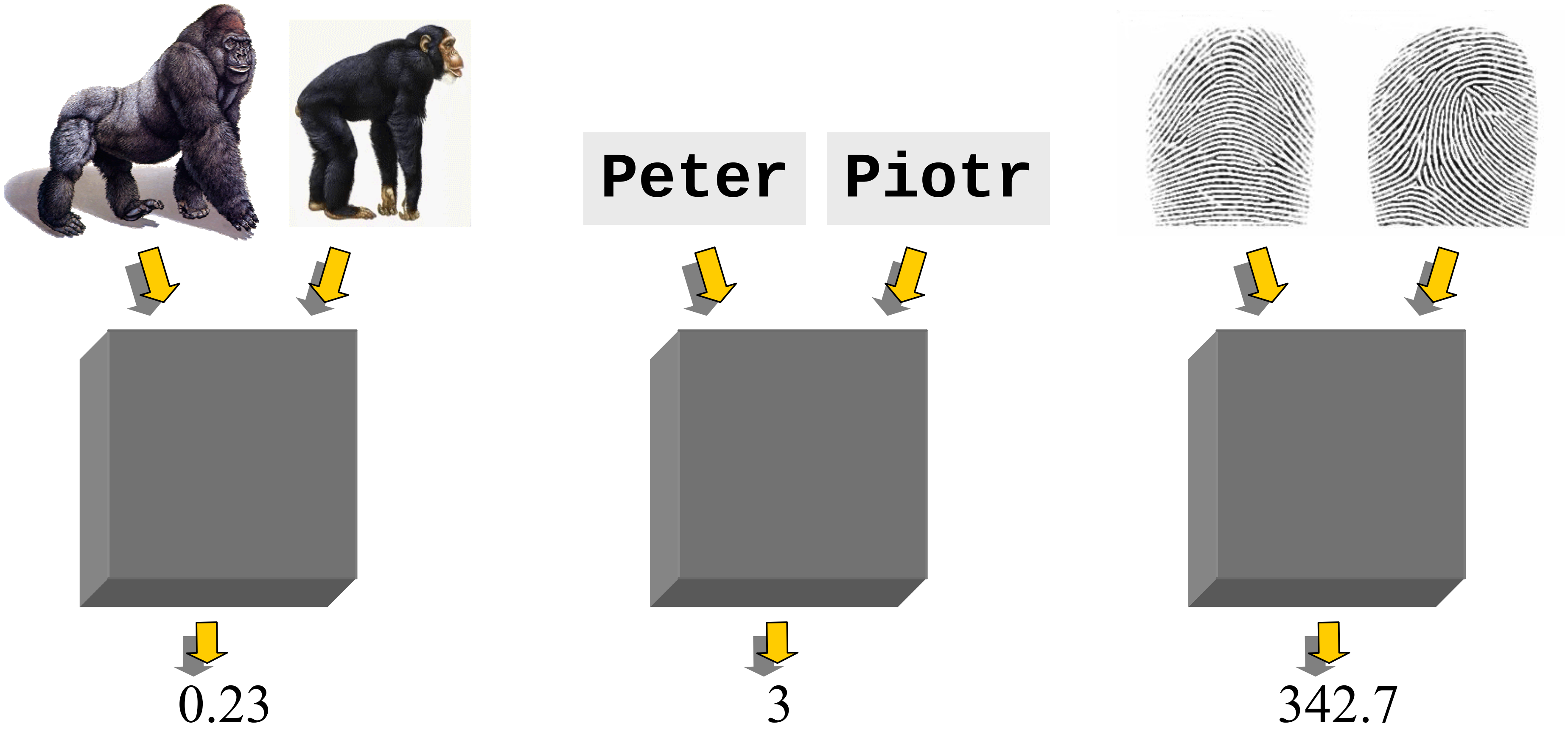
Webster's Dictionary

Similarity is hard to define, but... “*We know it when we see it*”

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Defining Distance Measures

Definition: Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) between O_1 and O_2 is a real number denoted by $D(O_1,O_2)$



Peter Piotr

0.23

3

342.7

What Is Good Clustering?

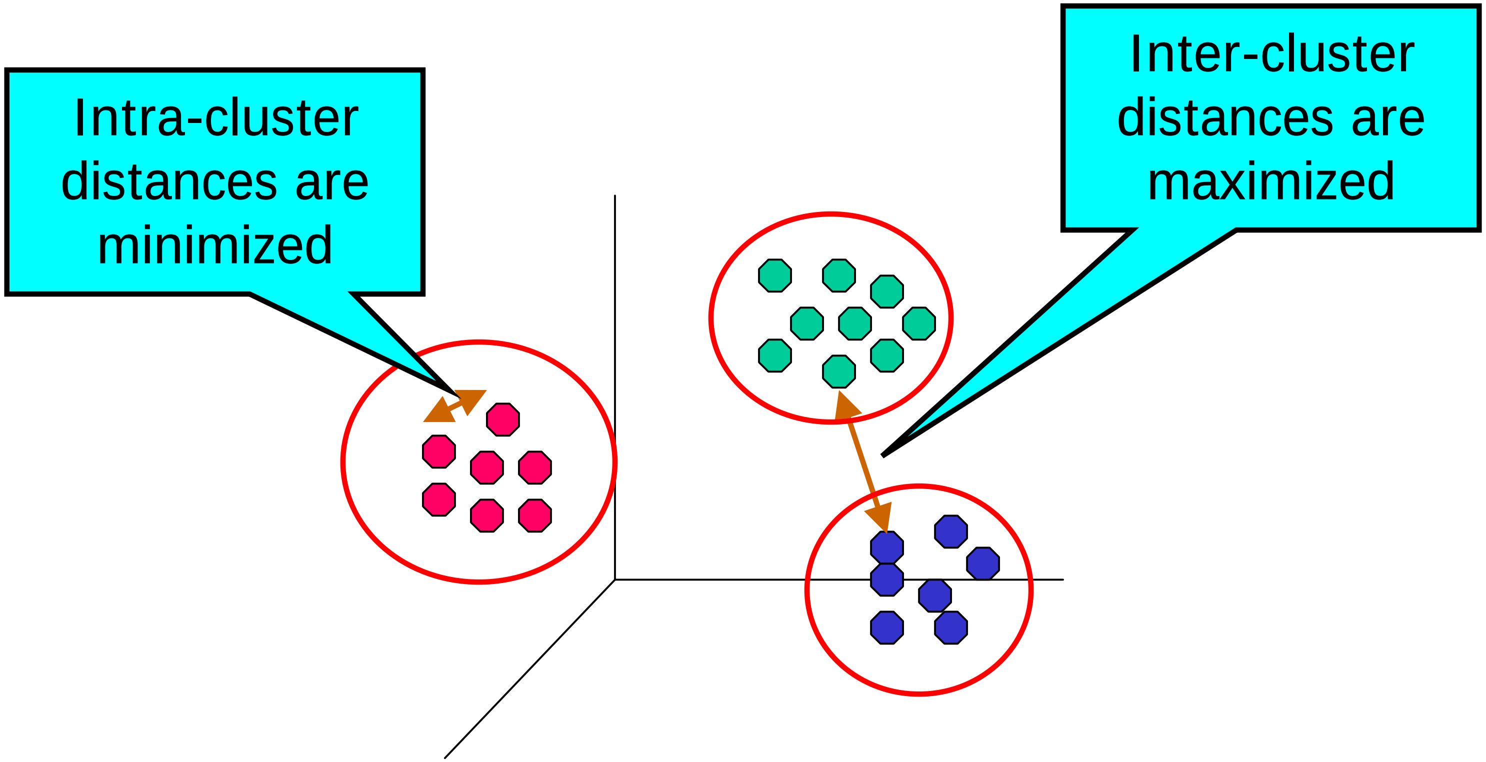
- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity

- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.

- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

What is Cluster Analysis?

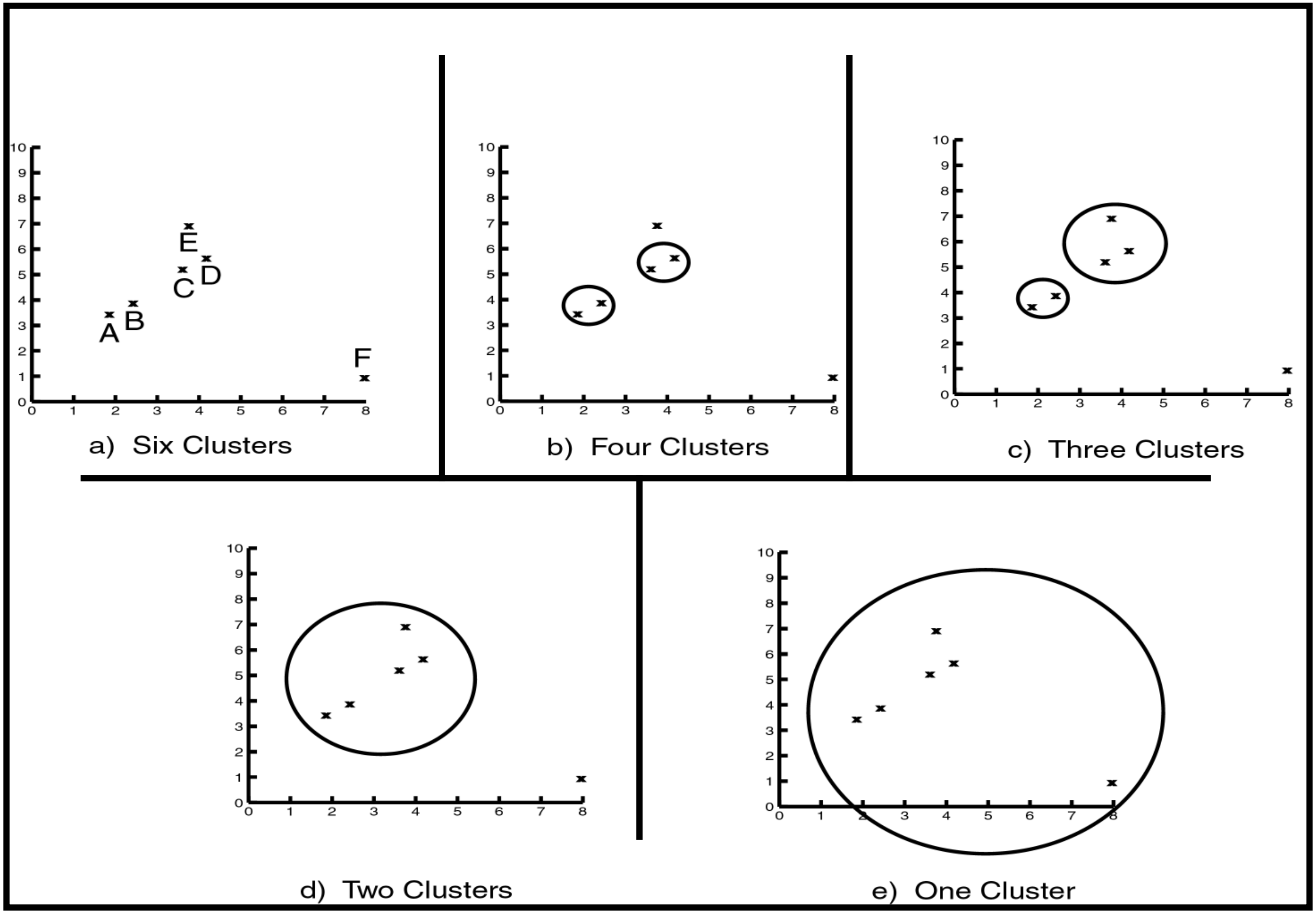
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Intra-cluster distances are minimized

Inter-cluster distances are maximized

Levels of Clustering



a) Six Clusters

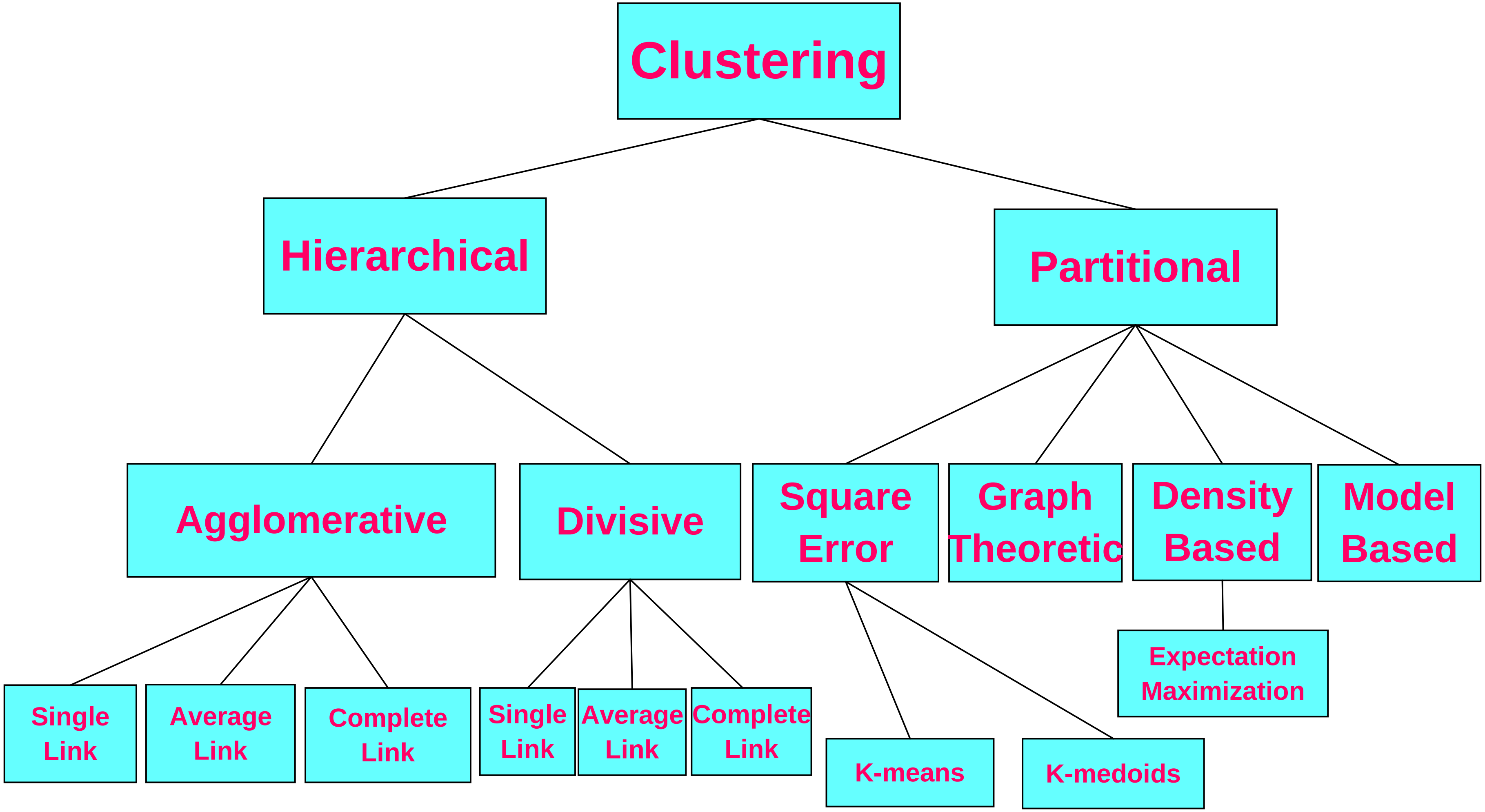
b) Four Clusters

c) Three Clusters

d) Two Clusters

e) One Cluster

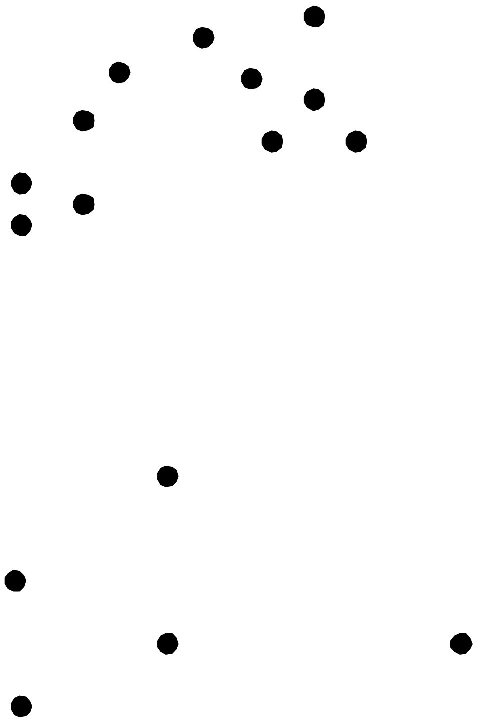
Taxonomy of clustering methods



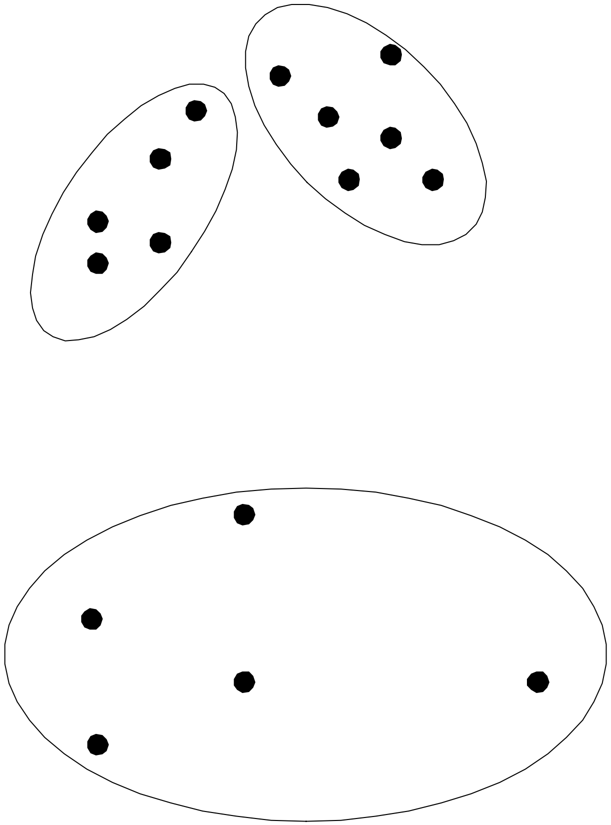
Partitional algorithms and Hierarchical algorithms

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion.
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion.

Partitional Clustering

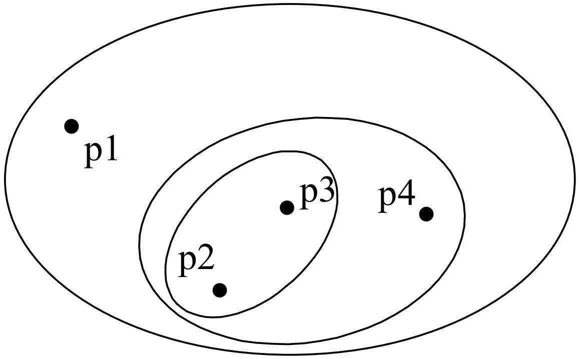


Original Points

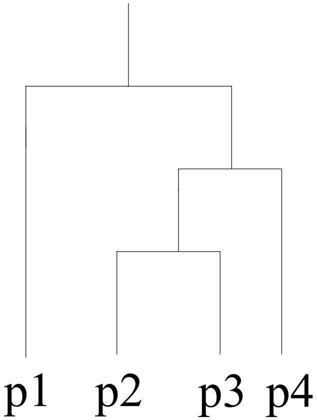


A Partitional Clustering

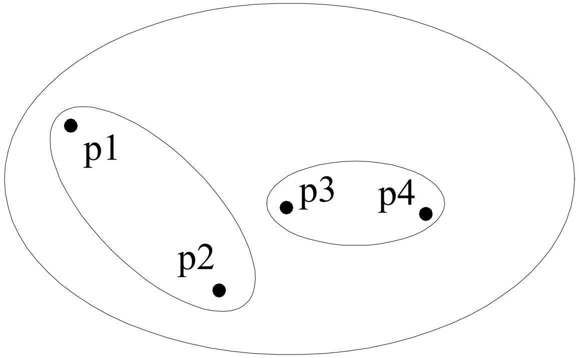
Hierarchical Clustering



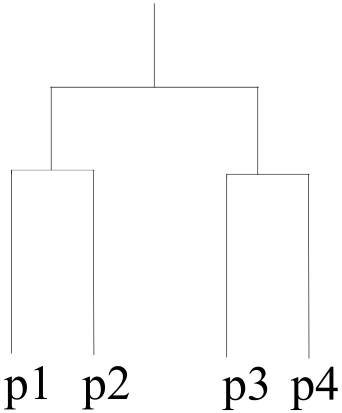
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Partitional Clustering

- Nonhierarchical
- Creates clusters in one step as opposed to several steps.
- Since only one set of clusters is output, the user normally has to input the desired number of clusters, k.
- Usually deals with static sets.

The K-means method

- a most commonly used method

- input variables must be in numerical form

- use the concept of "distance" in assigning each record to the nearest cluster center

- in real application, there could be very high number of variables or "dimentions"

Examples of distance functions

- **Numerical data**

  - Euclidean distance

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

- very popular in many applications

K-Means : Example 1

- Given: {2,4,10,12,3,20,30,11,25}, k=2

- Randomly assign means: $m_1=3, m_2=4$

- $K_1=\{2,3\}$, $K_2=\{4,10,12,20,30,11,25\}$, $m_1=2.5, m_2=16$
- $K_1=\{2,3,4\}, K_2=\{10,12,20,30,11,25\}$, $m_1=3, m_2=18$

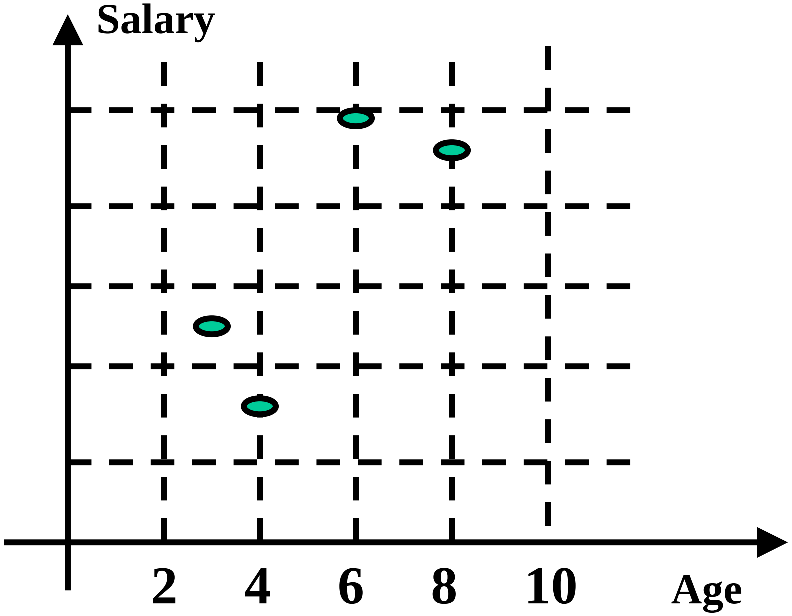
- $K_1=\{2,3,4,10\}, K_2=\{12,20,30,11,25\}$, $m_1=4.75, m_2=19.6$
- $K_1=\{2,3,4,10,11,12\}, K_2=\{20,30,25\}$, $m_1=7, m_2=25$

- Stop as the clusters with these means are the same.

K-Means : Example 2

K- Mean Clustering (K=2)

Phase 3	Normalized Age	Normalized Salary
Case1	**3**	**5**
Case2	**6**	**10**
Case3	**4**	**3**
Case4	**8**	**9**



Ex K- Mean Clustering (K=2) ..Cont.

Initial condition: **choose k cases randomly as initial k cluster centers.**

Stopping condition: cluster number of all cases in the current phase are the same as all cluster numbers in the previous phase.

Phase 1	Normalized Age	Normalized Salary	d(X, Center1)	d(X, Center2)	Assign to Cluster#
Center1	**3**	**5**			
Center2	**6**	**10**			
Case1	**3**	**5**	**0**	**Sqrt(9+25)**	**1**
Case2	**6**	**10**	**Sqrt(9+25)**	**0**	**2**
Case3	**4**	**3**	**Sqrt(1+4)**	**Sqrt(4+49)**	**1**
Case4	**8**	**9**	**Sqrt(25+16)**	**Sqrt(4+1)**	**2**

Ex K- Mean Clustering (K=2) … Cont’.

Before continue : recalculate all cluster centers by computing the average the values of each attribute from all cases in the same cluster.

Phase 2	Normalized Age	Normalized Salary	d(X, Center1)	d(X, Center2)	Assign to Cluster#
Center1	**3.5**	**4**			
Center2	**7**	**9.5**			
Case1	**3**	**5**	**Sqrt(0.25+1)**	**Sqrt(16+4.5*4.5)**	**1**
Case2	**6**	**10**	**Sqrt(6.25+36)**	**Sqrt(1+0.25)**	**2**
Case3	**4**	**3**	**Sqrt(0.25+1)**	**Sqrt(9+6.5*6.5)**	**1**
Case4	**8**	**9**	**Sqrt(4.5*4.5+25)**	**Sqrt(1+0.25)**	**2**

Ex K- Mean Clustering (K=2) … Cont.

Before continue : **Stopping condition is true → Finish !!!**

Phase 3	Normalized Age	Normalized Salary	Assign to Cluster#
Center1	**3.5**	**4**	
Center2	**7**	**9.5**	
Case1	**3**	**5**	**1**
Case2	**6**	**10**	**2**
Case3	**4**	**3**	**1**
Case4	**8**	**9**	**2**

ANSWER

Conclusions

- Cluster analysis groups objects based on their similarity

- Cluster analysis is the process of finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters

- K-mean method:
  - input variables must be in numerical form
  - use the concept of “distance” in assigning each record to the nearest cluster center