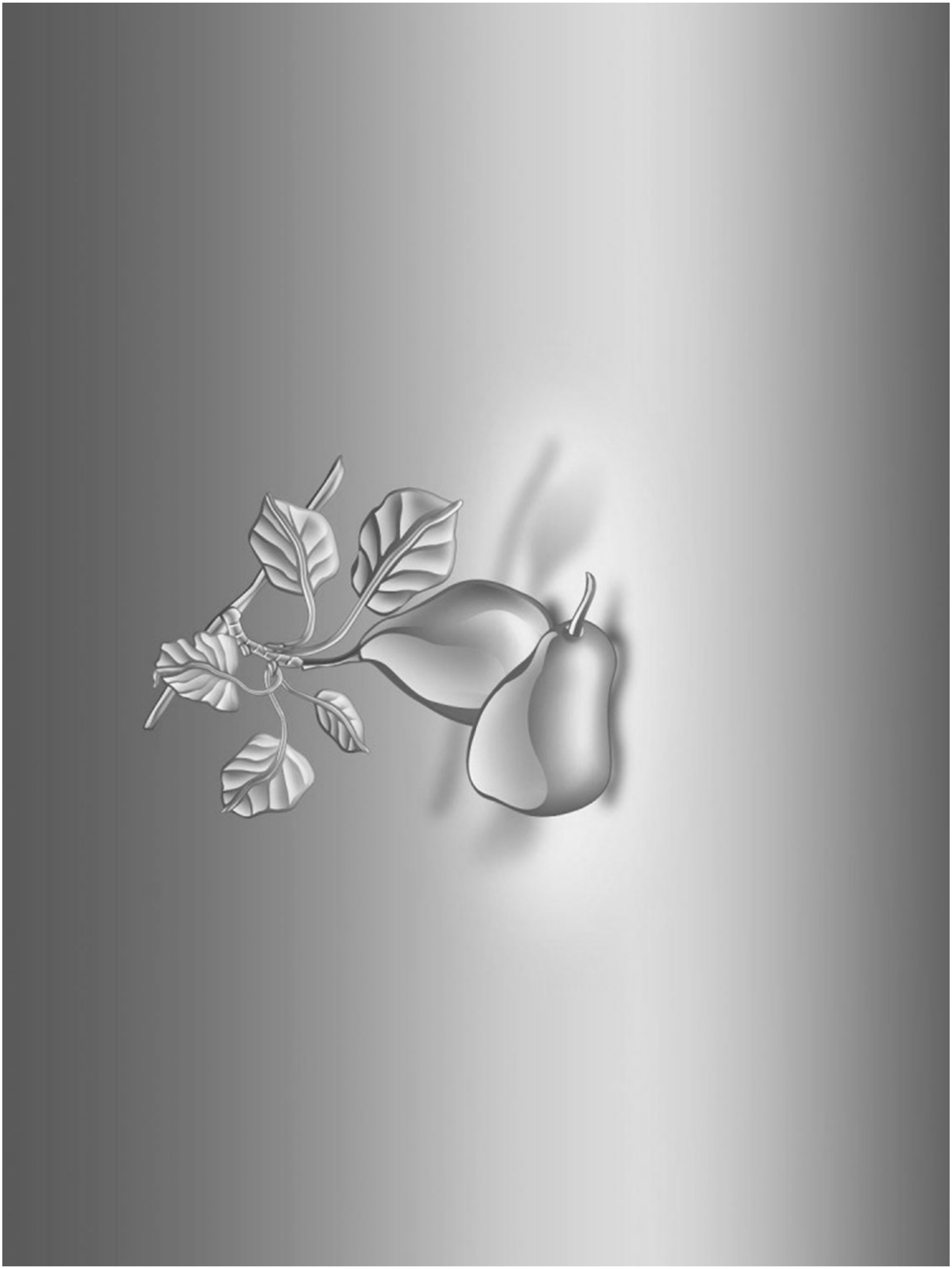


What is Cluster Analysis?

- 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

Clustering: Rich Applications and Multidisciplinary Efforts

- Pattern Recognition
- Image Processing
- Document classification in WWW
- Marketing
- Land use
- Insurance
- City-planning
- Earth-quake studies etc..

Quality: What Is Good Clustering?

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

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with random shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

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)

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

Type of data in clustering analysis

- Interval-scaled variables
- Binary variables
- Nominal
- Ordinal
- ratio variables
- Variables of mixed types

Interval-valued variables

- Interval scale is a scale which represents quantity.
- Examples of interval data:
  -Temperature (Degrees F)
  -Dates
  -Dollars
  -Years
  -Sea Level etc....

Interval-valued variables

- Standardize data
  - Calculate the mean absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}).$$

  - Calculate the standardized measurement (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is more robust than using standard deviation

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

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

Similarity and Dissimilarity Between Objects

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

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

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures

Binary Variables

- A contingency table for binary data

		Object j		
		1	0	*sum*
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Distance measure for symmetric binary variables:

$$d(i, j) = \frac{b+c}{a+b+c+d}$$

- Distance measure for asymmetric binary variables:

$$d(i, j) = \frac{b+c}{a+b+c}$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i, j) = \frac{a}{a+b+c}$$

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- gender is a symmetric attribute
- the remaining attributes are asymmetric binary
- let the values Y and P be set to 1, and the value N be set to 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Nominal (Categorical) Variables

- For example, gender is a categorical variable having two categories (male and female

- Hair color is also a categorical variable having a number of categories (blonde, brown, brunette, red, etc.)

Nominal (Categorical) Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: use a large number of binary variables
  - creating a new binary variable for each of the *M* nominal states

Ordinal Variables

- An ordinal variable is similar to a categorical variable. The difference between the two is that there is a clear ordering of the variables. For example, suppose you have a variable, economic status, with three categories (low, medium and high).

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, ..., M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Ratio-Scaled Variables

- These are continuous positive measurements on a nonlinear scale
- Ratio-scaled variable: a positive measurement on a nonlinear scale, approximately at exponential scale - Ae^{Bt} or Ae^{-Bt}
- A typical example is the growth of bacterial population (say, with a growth function Ae^{Bt}.). In this model, equal time intervals multiply the population by the same ratio.

Ratio-Scaled Variables

- Ratio-scaled variable: a positive measurement on a nonlinear scale, approximately at exponential scale - Ae^{Bt} or Ae^{-Bt}
- Methods:
  - treat them like interval-scaled variables—the scale can be distorted
  - apply logarithmic transformation and treat as interval-scaled

$$y_{if} = log(x_{if})$$

  - treat them as continuous ordinal data treat their rank as interval-scaled

Variables of Mixed Types

- A database may contain all the six types of variables
  - Interval-scaled, symmetric binary, asymmetric binary, nominal, ordinal and ratio
- Bring all variables onto a common scale

Variables of Mixed Types

- Bring all variables onto a common scale [0.0,1.0]

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$$

-

- if x_{if} or x_{jf} is missing or $x_{if}=x_{jf}$ and variable f is asymmetric

otherwise 1.

$$\delta_{ij}^{(f)}=0$$

Variables of Mixed Types

- $d_{ij}^{(f)}$ is computed based on its type
  - f is interval-based: $d_{ij}^{(f)}=|x_{if} - x_{jf}|/(\max_h x_{hf} - \min_h x_{hf})$
  - f is binary or nominal:

    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$
  - f is ordinal: compute ranks r_{if} and

    $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
  - f is ratio-scaled : perform logarithmic transformation and treat as interval-scaled or ordinal data

Major Clustering Approaches/Methods (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**, **WaveCluste**r, 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

Partitioning Algorithms: Basic Concept

- Partitioning method: Given D, a data set of n objects, and k, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions(k<=n), where each partition represents a cluster.
- Given a *D*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - *k-means* : Each cluster is represented by the center of the cluster.
  - *k-medoids* or PAM (Partition Around Medoids) : Each cluster is represented by one of the objects in the cluster .

The *K-Means* Clustering Algorithm

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

  Step 1: Partition objects into *k* nonempty subsets

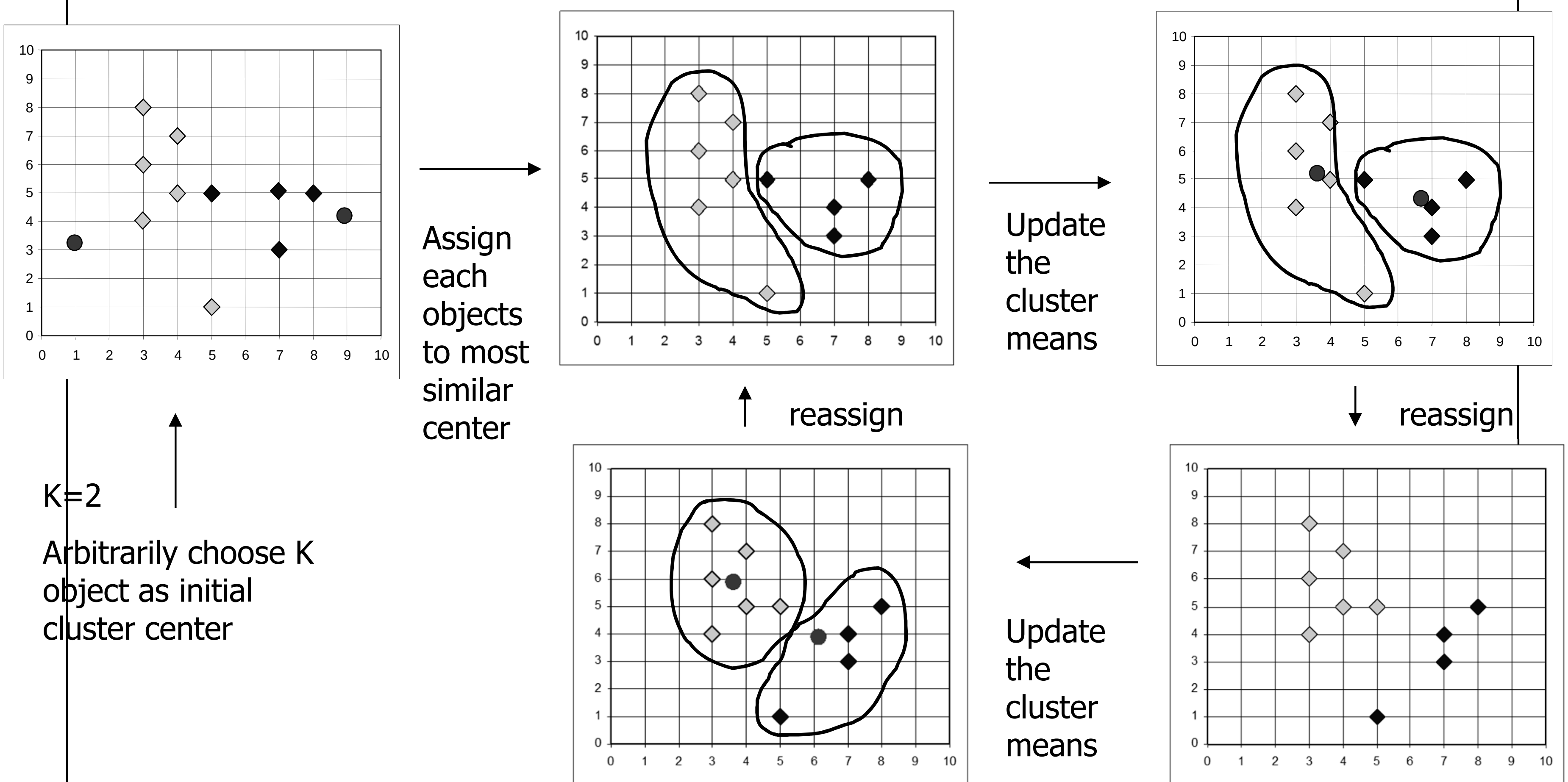
  Step 2: Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)

  Step 3: Assign each object to the cluster with the nearest seed point

  Step4: Go back to Step 2**, stop** when no more new assignment.

The *K-Means* Clustering Method

- Example



Example

- Problem: Cluster the following 8 points into 3 clusters
- A1(2,10), A2(2,5), A3(8,4) , A4(5,8) , A5(7,5), A6(6,4) , A7(1,2) , A8(4,9).

Three cluster centers **A1(2,10)** **A4(5,8)** **A7(1,2**)

The distance function between two points a=(x1,y1) and b=(x2,y2) is defined as

P(a,b)=|x2-x1| + |y2-y1|. Use k-means algorithm to find the three cluster centers after the second iteration.

Solution:

- Iteration

		(2,10)	(5,8)	(1,2)	
	Point	Distance Mean 1	Distance Mean 2	Distance Mean 3	Cluster
A1	(2,10)				
A2	(2,5)				
A3	(8,4)				
A4	(5,8)				
A5	(7,5)				
A6	(6,4)				
A7	(1,2)				
A8	(4,9)				

Point mean1

x1, y1 x2, y2 p(a, b)=**)=|x2-x1| + |y2-y1|**

(2, 10) (2, 10) =|2 – 2| + | 10 – 10|

P(a,b)=|x2-x1| + |y2-y1| = 0+0 = 0

Point mean1

x1, y1 x2, y2 p(a,b)=**)=|x2-x1| + |y2-y1|**

(2, 10) (5, 8) =|5 – 2| + | 8 – 10|

P(a,b)=|x2-x1| + |y2-y1| = 3+2 = 5

Point mean1

x1, y1 x2, y2 p(a,b)=**)=|x2-x1| + |y2-y1|**

(2, 10) (1, 1) =|1 – 2| + | 2 – 10|

P(a,b)=|x2-x1| + |y2-y1| = 1+8 = 9

		(2,10)	(5,8)	(1,2)	
	Point	Distance Mean 1	Distance Mean 2	Distance Mean 3	Cluster
A1	(2,10)	0	5	9	1
A2	(2,5)	5	6	4	3
A3	(8,4)	12	7	9	2
A4	(5,8)	5	0	10	2
A5	(7,5)	10	5	9	2
A6	(6,4)	10	5	7	2
A7	(1,2)	9	10	0	3
A8	(4,9)	3	2	10	2

- Next we recomputed the new cluster centers(means).
- For Cluster 2 , (8+5+7+6+4)/5,(4+8+5+4+9)/5=(6,6)
- For Cluster 3 , (2+1)/2,(5+2)/2=(1.5, 3.5)
- Repeat same iteration with new cluster centers until clusters remain unchanged

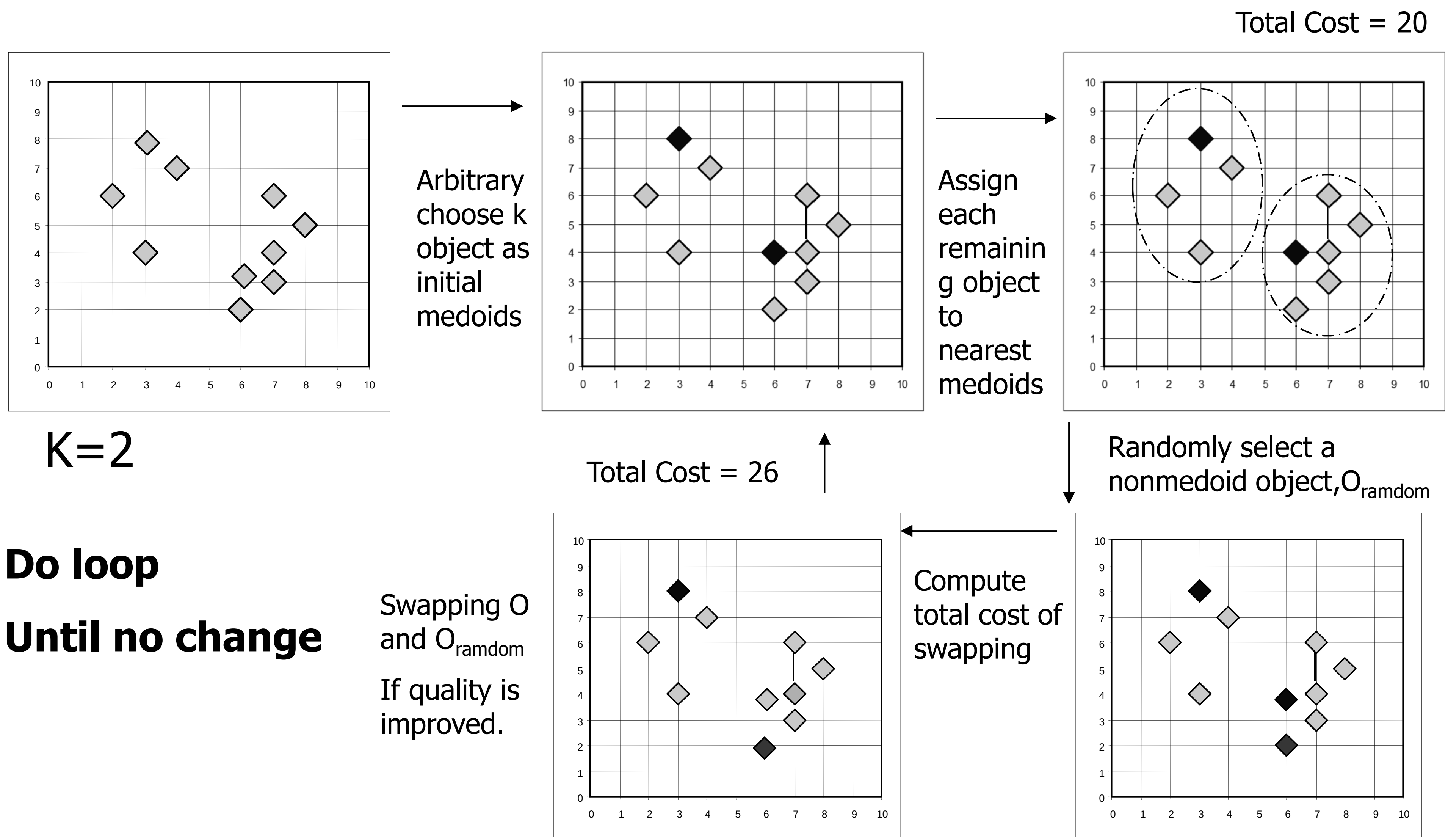
Comments on the *K-Means* Method

- Strength: *Relatively efficient*:
- Comment: Often terminates at a *local optimum*.
- Weakness
  - Applicable only when *mean* is defined.
  - Need to specify *k*, the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with different size.

The *K-Medoids* Clustering Method

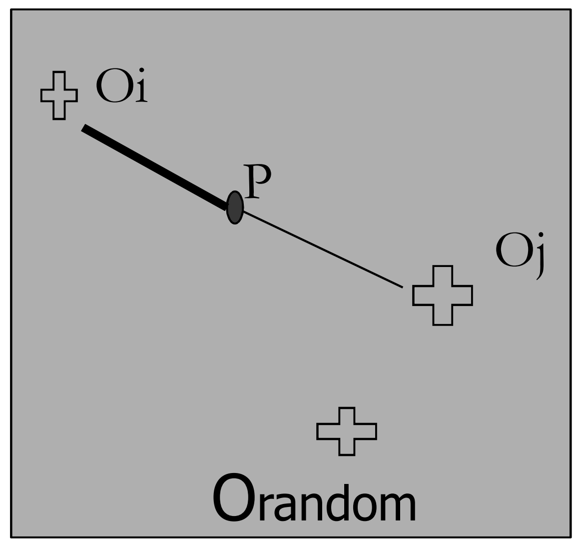
- Find *representative* objects, called medoids, in clusters which is the **most centrally located** object in a cluster.
- *PAM* (Partitioning Around Medoids)
  - Starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - *PAM* works effectively for small data sets, but does not scale well for large data sets

A Typical K-Medoids Algorithm (PAM)



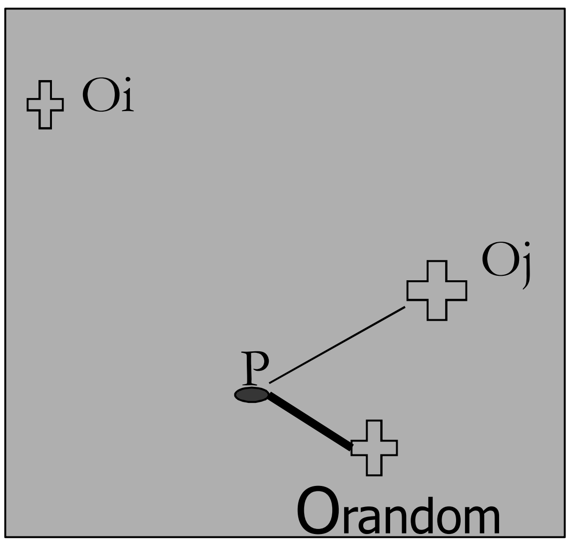
- The initial representative objects(or seeds) are chosen arbitrarily.
- The iterative process of replacing representative objects by non-representative objects continues as long as the quality of the resulting clustering is improved.
- This quality is estimated using a cost function that measures the average dissimilarity between an object and rep. object of its cluster.

- Case 1: p currently belongs to representative object, oj. If oj is replaced by o random as a representative object and p is closest to one of the other representative objects oi, i!= j, then p is reassigned to Oi.



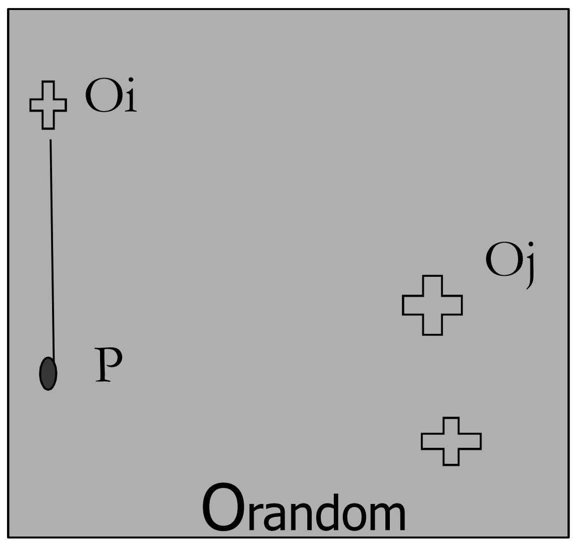
1. Reassigned to Oi

- Case 2: p currently belongs to representative object, oj. If oj is replaced by orandom as a representative object and p is closest to orandom then p is reassigned to orandom.



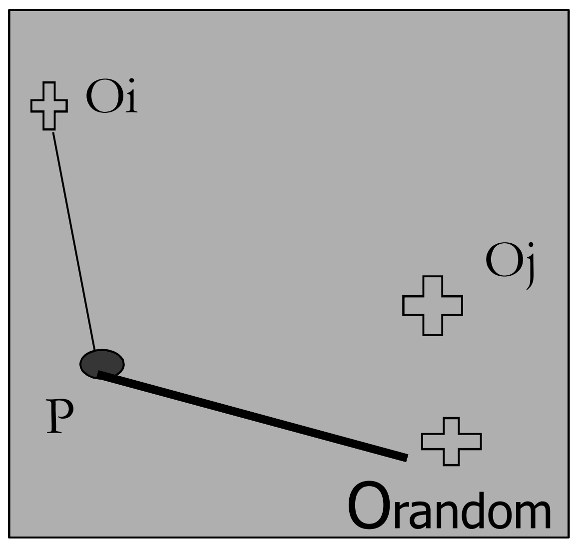
2. Reassigned to Orandom

- Case 3: p currently belongs to representative object, oi i!=j. If oj is replaced by orandom as a representative object and p is still closest to oi then the assignment does not change.



3. No change

- Case 4: p currently belongs to representative object, oi i!=j. If oj is replaced by Orandom as a representative object and p is closest to Orandom then p is assigned to Orandom.



3. No change

CLARA (Clustering LARge Applications)

- It draws ***multiple samples*** of the data set, applies *PAM* on each sample, and gives the best clustering as the output
- Strength: deals with larger data sets than *PAM* (Partitioning Around Medoids)
- Weakness:
  - Efficiency depends on the sample size
  - A good clustering based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased

Hierarchical Clustering

- Works by grouping data objects into a tree of clusters
- Classified as:
  - Agglomerative(bottom-up)
  - Divisive(top-down)
- No backtracking is possible

Agglomerative hierarchical clustering

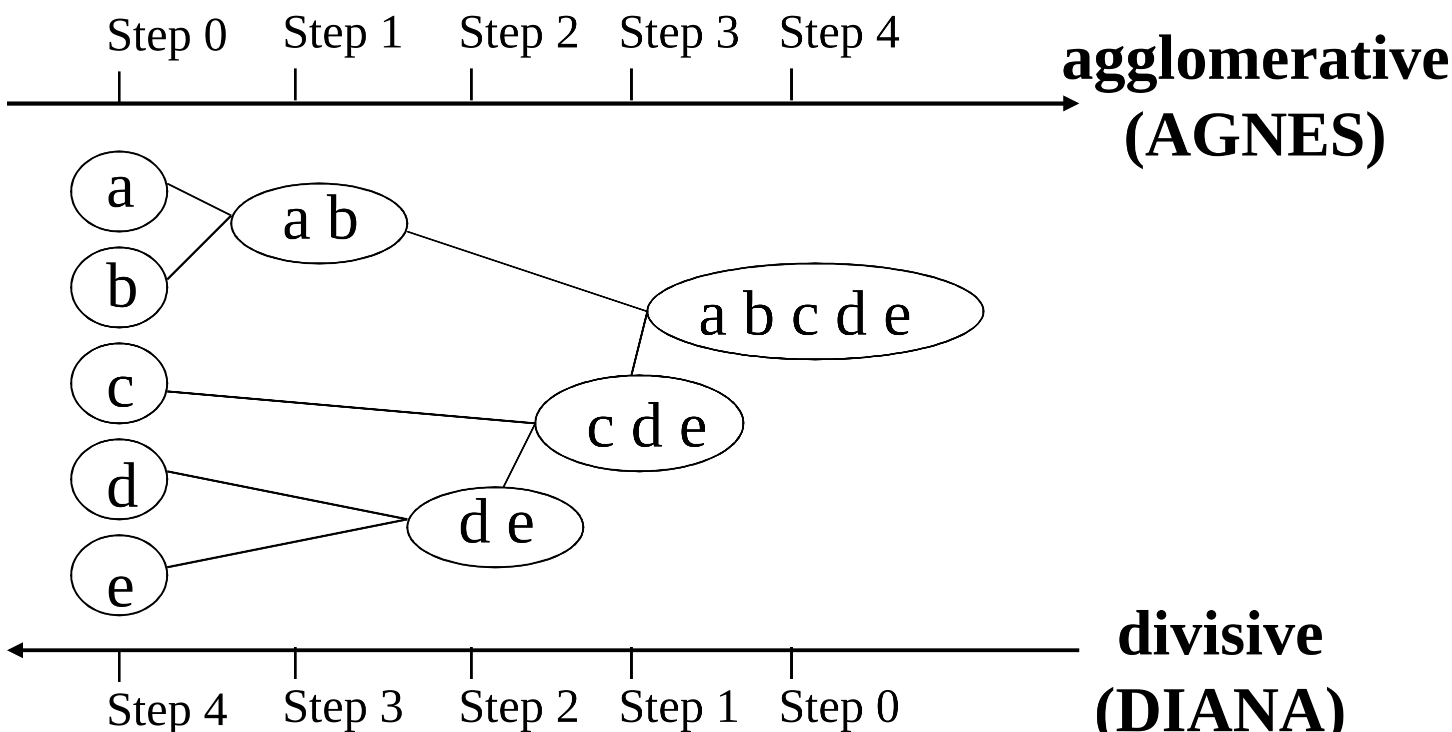
- Follows bottom up strategy
- Place each objects in its own cluster and then merges this atomic clusters into larger clusters.
- AGNES is an example for agglomerative hierarchical clustering.

Divisive hierarchical clustering

- Follows top down strategy.
- Reverse of agglomerative strategy
- Subdivides the cluster into smaller and smaller pieces.
- DIANA is an example for divisive hierarchical clustering

Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition



AGNES (AGglomerative NESting)

Employee	Skill X	Skill Y
1	2	8
2	8	15
3	3	6
4	6	9
5	8	7
6	10	10

Iteration number(i)	No of clusters	Nearest Clusters	Centroid of nearest clusters	Distance b/w nearest clusters	Set of clusters after meging nearest clusters
1	6	C1,C3	(2.5,7)	2.236	C13,C2,C4,C5,C6
2	5	C4,C5	(7,8)	2.828	C13,C2,C45,C6)
3	4	C45,C6	(8,8.7)	3.600	C13,C2,C456
4	3	C456,C2	(8,10.3)	6.300	C13,C4562
5	2	C4562,C13	(6.2,9.2)	6.414	C134562

AGNES (AGglomerative NESting)

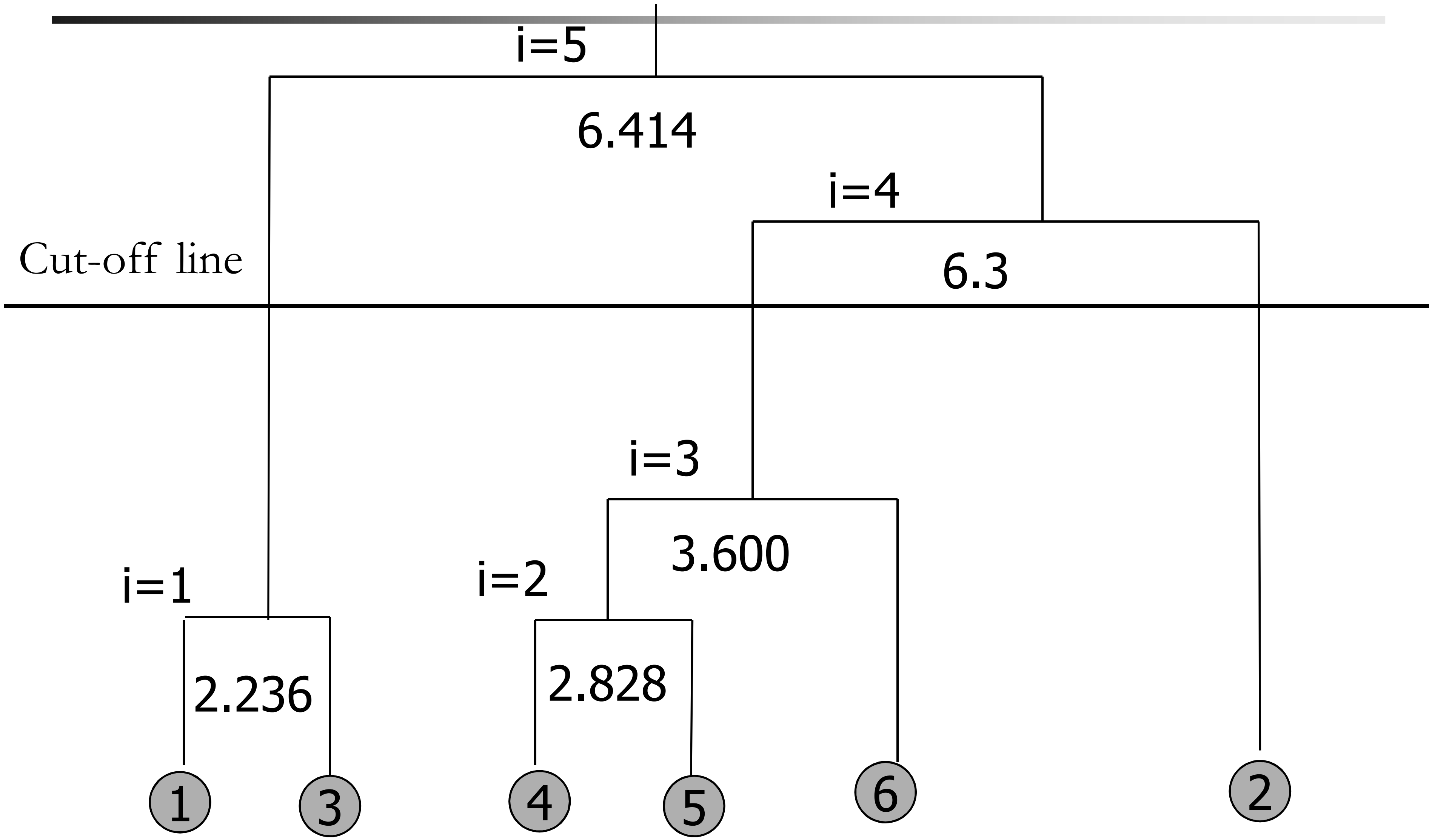
- Use the Single-Link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity.
- Single-Linkage – Clustering process is terminated when the distance between nearest clusters exceeds an arbitrary threshold.

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.

Dendrogram: Shows How the Clusters are Merged



DIANA (DIvisive ANAlysis)

- Inverse order of AGNES
- Eventually each node forms a cluster on its own

Recent Hierarchical Clustering Methods

- Major weakness of previous clustering methods
  - do not scale well
  - Can never undo what was done previously
- Integration of hierarchical with distance-based clustering
- BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies
  - uses CF-tree(**clustering feature tree)** and incrementally adjusts the quality of sub-clusters
- ROCK: RObust Clustering using linKs
  - clustering categorical data by neighbor and link analysis
- CHAMELEON: hierarchical clustering using dynamic modeling

Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points or based on an explicitly constructed density function.
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition

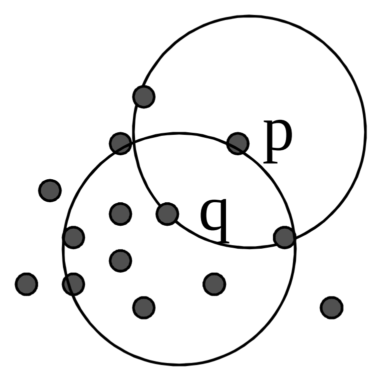
Density-Based Clustering Methods

- Several interesting studies:
  - DBSCAN – algorithm that grows clusters according to a density based connectivity analysis
  - OPTICS – algorithm that extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter settings
  - DENCLUE- algorithm that clusters objects based on a set of density distribution functions

Density-Based Clustering: Background

- Two parameters:
  - ***Eps***: Maximum radius of the neighbourhood
  - ***MinPts***: Minimum number of points in an Eps-neighbourhood of that point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid dist(p,q) \le Eps\}$
- Directly density-reachable**:** A point ***p*** is directly density-reachable from a point ***q*** wrt. ***Eps***, ***MinPts*** if
  - 1) ***p*** belongs to $N_{Eps}(q)$
  - 2) core point condition:

$$|N_{Eps}(q)| \ge MinPts$$

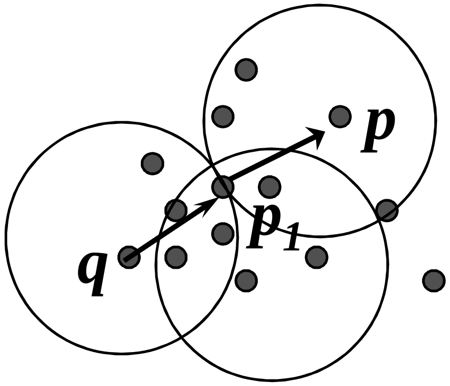


MinPts = 5

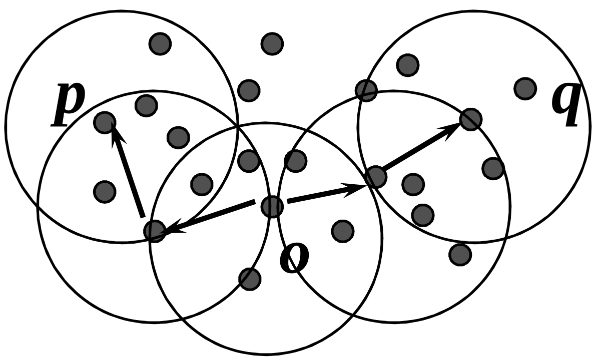
Eps = 1 cm

Density-Based Clustering: Background (II)

- Density-reachable:
  - A point p is density-reachable from a point q wrt. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i

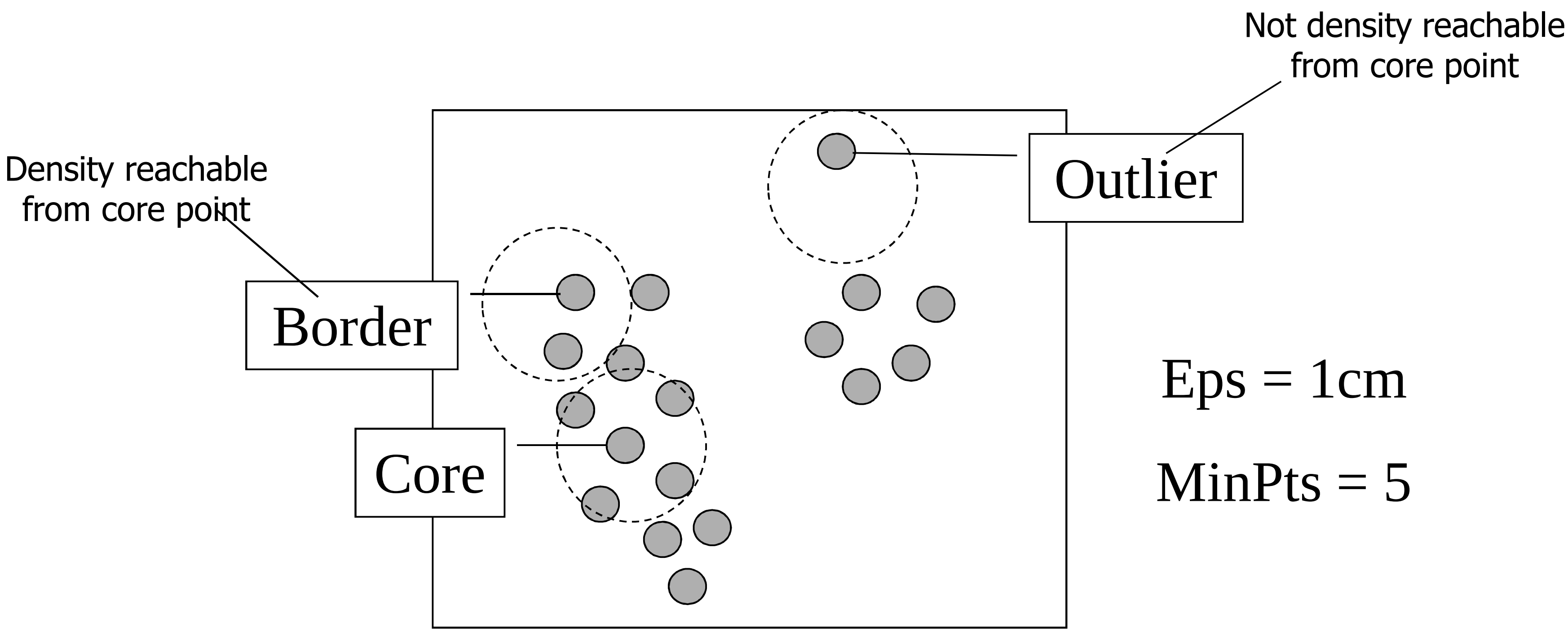


- Density-connected
  - A point p is density-connected to a point q wrt. *Eps*, *MinPts* if there is a point o such that both, p and q are density-reachable from o wrt. *Eps* and *MinPts*.



DBSCAN: Density Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



DBSCAN: The Algorithm

- Arbitrary select a point ***p***
- Retrieve all points density-reachable from ***p*** wrt ***Eps*** and ***MinPts***.
- If ***p*** is a core point, a cluster is formed.
- If ***p*** ia not a core point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.