

# Data Mining

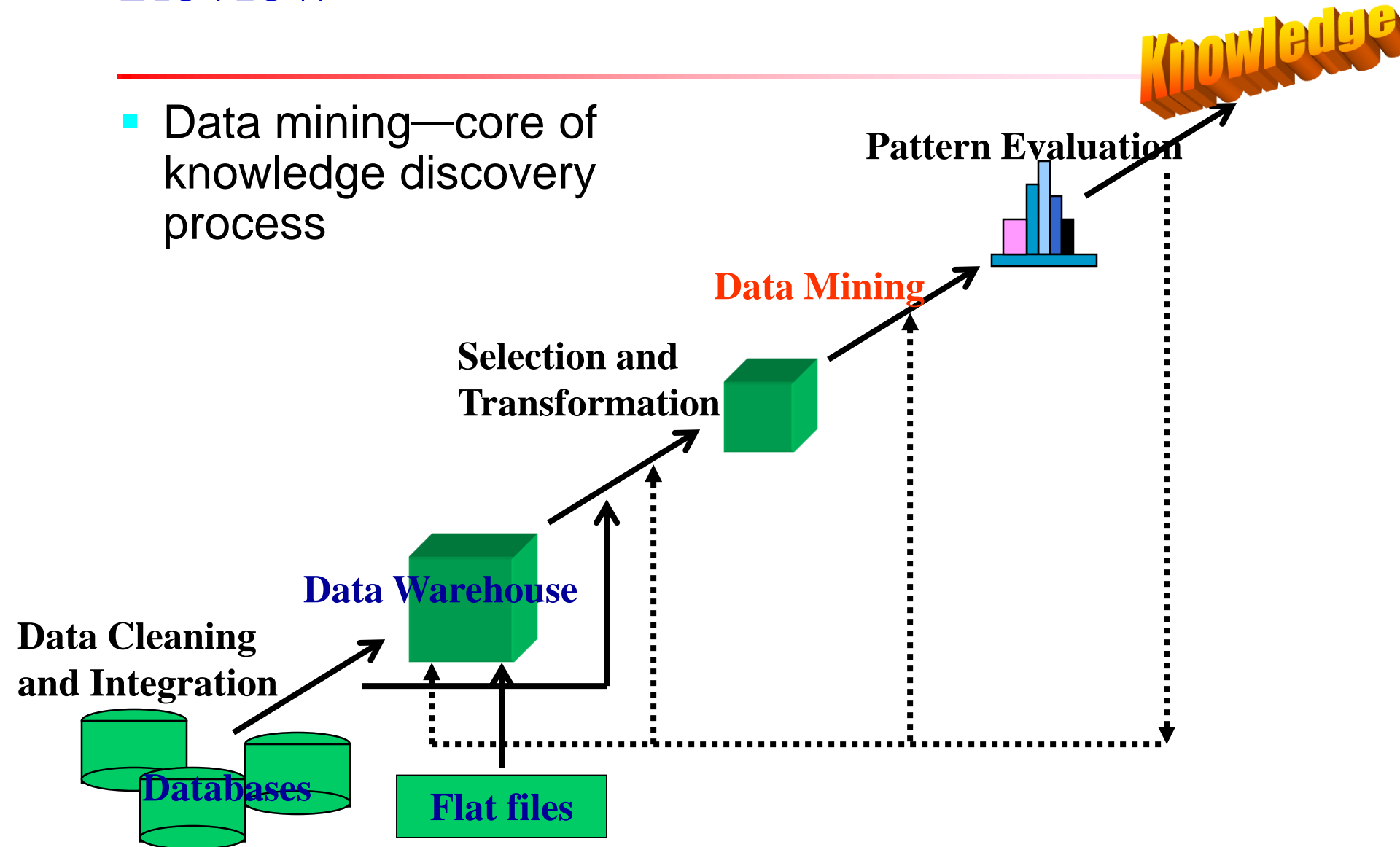
---

**Ying Liu, Prof., Ph.D**

University of Chinese Academy of Sciences

# Review

- Data mining—core of knowledge discovery process



# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Outlier Analysis
- Summary

# 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 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

# Examples of Clustering Applications

---

- Marketing: Help marketers discover distinct groups according to their customer databases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- 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

# Examples of Clustering Applications

---

- Earth-quake studies: Observed earth quake epicenters are clustered along continent faults
- Biology: categorize genes with similar functionality
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar accessing patterns

# Clustering: Rich Applications and Multidisciplinary Efforts

---

- Pattern Recognition
- GIS
  - Create thematic maps in GIS by clustering feature spaces
  - Detect spatial clusters or for other spatial mining tasks
- Image Processing
- Economic Science (especially marketing research)
- Software package
  - S-Plus, SPSS, SAS, R

# 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



# Measure the Quality of Clustering

---

- **Dissimilarity/Similarity metric**: Similarity is expressed in terms of a distance function, typically metric:  $d(i, j)$
- The definitions of **distance functions** are usually very different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables
- Weights may be assigned to different variables based on applications and data semantics
- It is hard to define “similar enough” or “good enough”
  - The answer is typically highly subjective

# Requirements of Clustering

---

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with arbitrary 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

# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Outlier Analysis
- Summary

# Data Structures

---

## ■ Data matrix

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

## ■ Dissimilarity matrix

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

# Type of Data in Clustering Analysis

---

- Interval-scaled variables 区间值
- Binary variables 二进制
- Nominal variables
- Ordinal variables 顺序变量
- Ratio-scaled variables 级数变化
- Variables of mixed types

# Interval-valued Variables

---

## ■ Standardize data

- Calculate the mean absolute deviation:

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

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

- Calculate the standardized measurement (z-score)

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

平均绝对离差

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

标准差

# 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**:

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

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

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

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

# Similarity and Dissimilarity Between Objects

欧几里得度量

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

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

- 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



# Binary Variables

- A contingency table for binary data
- Distance measure for symmetric binary variables: 对称
- Distance measure for asymmetric binary variables: 不对称

		Object <i>j</i>		
		1	0	<i>sum</i>
Object <i>i</i>	1	<i>a</i>	<i>b</i>	<i>a+b</i>
	0	<i>c</i>	<i>d</i>	<i>c+d</i>
<i>sum</i>		<i>a+c</i>	<i>b+d</i>	<i>p</i>

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

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

D是双方都为0的情况，对于聚类的贡献不大，因此选择不对的计算方法。

# 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 Variables

---

- A generalization of 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 nominal 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 can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace  $x_{if}$  by their rank  $r_{if} \in \{1, \dots, M_f\}$
  - 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

---

- Ratio-scaled variable: a positive measurement on a nonlinear scale, approximately at exponential scale, such as  $Ae^{Bt}$  or  $Ae^{-Bt}$
- Methods:
  - 不能像对待区间变量一样去处理级数变量
  - treat them like interval-scaled variables — *not a good choice!*
  - apply logarithmic transformation 使用对数来处理

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

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

# Variables of Mixed Types

- A database may contain all the six types of variables
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio

- One may use a weighted formula to combine their effects

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

- $\delta_{ij}^{(f)} = 0$  if  $x_{if}$  or  $x_{jf}$  is missing, or  $x_{if} = x_{jf} = 0$  and  $f$  is asymmetric attribute; otherwise,  $\delta_{ij}^{(f)} = 1$
- $f$  is binary or nominal:
  - $d_{ij}^{(f)} = 0$  if  $x_{if} = x_{jf}$ , or  $d_{ij}^{(f)} = 1$  otherwise
- $f$  is interval-based: use the normalized distance
- $f$  is ordinal
  - compute ranks  $r_{if}$  and treat  $z_{if}$  as interval-scaled
- $f$  is ratio-scaled
  - Transform  $f$ , and treat  $f$  as interval-scaled.  $y_{if} = \log(x_{if})$

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

# Exercise

---

1. Please compute the dissimilarity matrix for the data set.

ID	Test-1 (categorical)	Test-2 (ordinal)	Test-3 (ratio-scaled)
1	A	excellent	445
2	B	fair	22
3	C	good	164
4	A	excellent	1,210

# Solution

For test-1, use simple matching

0			
d(2,1)	0		
d(3,1)	d(3,2)	0	
d(4,1)	d(4,2)	d(4,3)	0

?

=

0			
1	0		
1	1	0	
0	1	1	0

For test-2

0			
1	0		
0.5	0.5	0	
0	1	0.5	0



# Solution

---

For test-3, use log transformation

- Convert test-3 to 2.65, 1.34, 2.21, 3.08
- Normalize to 0.75, 0, 0.5, 1

0			
0.75	0		
0.25	0.5	0	
0.25	1	0.5	0

Dissimilarity matrix

0			
0.92	0		
0.58	0.67	0	
0.08	1	0.67	0

# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Outlier Analysis
- Summary

# Major Clustering Approaches (I)

---

## ■ Partitioning approach:

- 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

## ■ Hierarchical approach: 分级聚合

- Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Typical methods: Diana, Agnes, BIRCH, ROCK, CHAMELEON

上面这俩 relies on distance

## ■ Density-based approach: 依赖于密度，阈值的确定是一个缺点

- 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
- Probabilistic Model-based approach:
  - Typical methods: EM

# 质心 Centroid, Radius and Diameter of a Cluster (for numerical data sets)

---

- Centroid: the “middle” of a cluster

$$C = \frac{\sum_{i=1}^N (t_i)}{N}$$

- Radius: square root of average distance from any point of the cluster to its centroid

$$R = \sqrt{\frac{\sum_{i=1}^N (t_i - c)^2}{N}}$$

- Diameter: square root of average mean squared distance between all pairs of points in the cluster

$$D = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^N (t_i - t_j)^2}{N(N-1)}}$$

# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- **Partitioning Methods**
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Outlier Analysis
- Summary

# Partitioning Algorithms: Basic Concept

---

- Partitioning method: Construct a partition of a database  $D$  of  $n$  objects into a set of  $k$  clusters, s.t., min sum of squared distance

$$\sum_{m=1}^k \sum_{t_{mi} \in K_m} (C_m - t_{mi})^2$$

- Given a  $k$ , find a partition of  $k$  clusters that optimizes the chosen partitioning criterion
  - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

# The *K-Means* Clustering Method

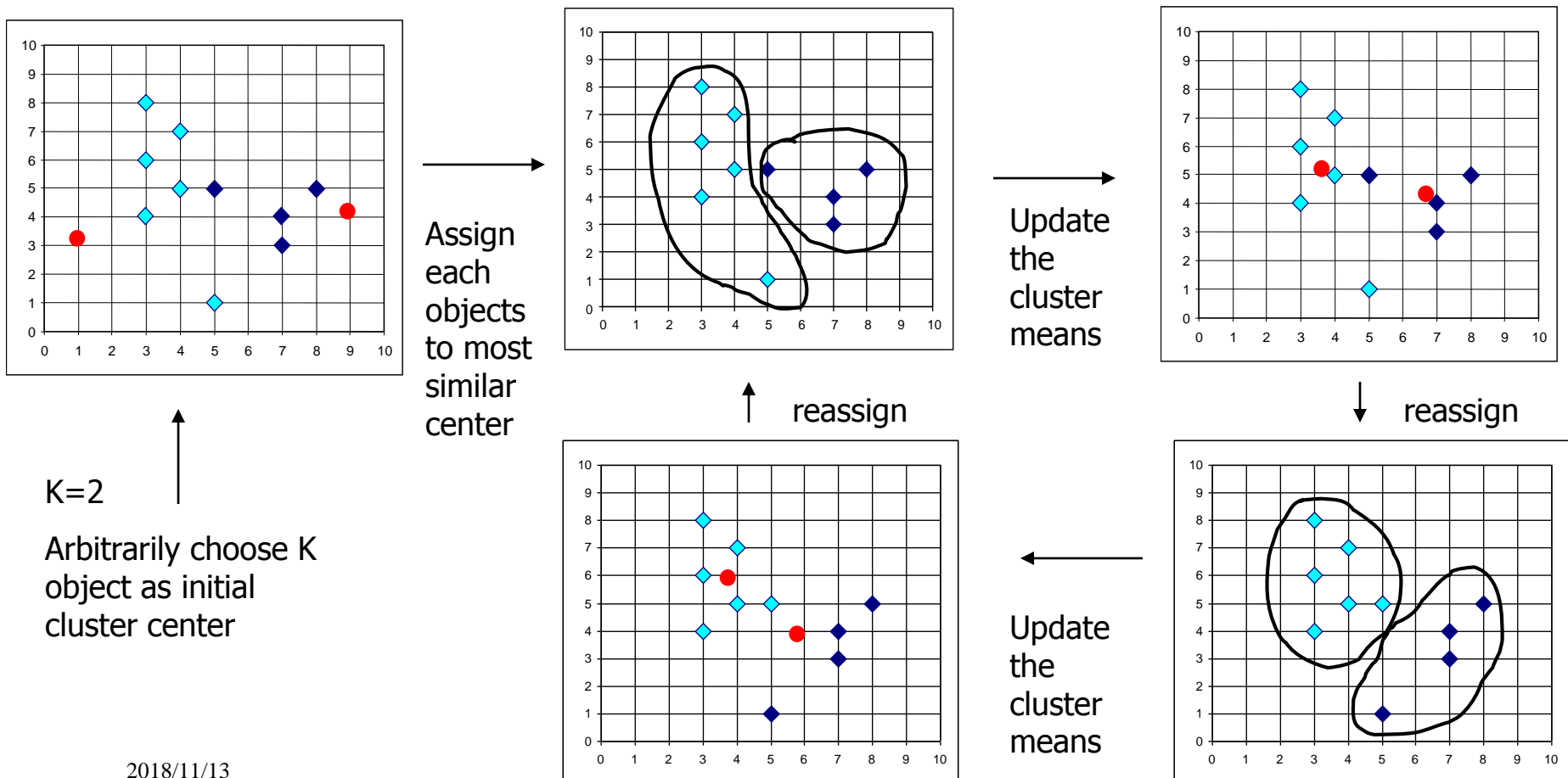
---

- Given a  $k$ , the *k-means* algorithm is implemented in four steps:
  - Give  $k$  random seeds as the initial centroids
  - Compute the centroid of each cluster of the current partition (the centroid is the center, i.e., *mean point*)
  - For each object, compute its distance to the centroids
    - Assign it to the cluster with the nearest centroid
  - Go back to Step 2, stop when no more new assignment



# The *K-Means* Clustering Method

## ■ Example



# Comments on the *K-Means* Method

---

- Strength: Relatively efficient:  $O(tkn)$ , where  $n$  is # objects,  $k$  is # clusters, and  $t$  is # iterations. Normally,  $k, t \ll n$
- Comment: Often terminates at a **local optimum**. The *global optimum* may be found using techniques such as: **deterministic annealing and genetic algorithms**  
模拟退火和遗传算法
- Weakness
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify  $k$ , the number of clusters, in advance
  - Unable to handle noisy data and outliers
  - Not suitable to discover clusters with *non-convex shapes*  
非凸

# Variations of the *K-Means* Method

---

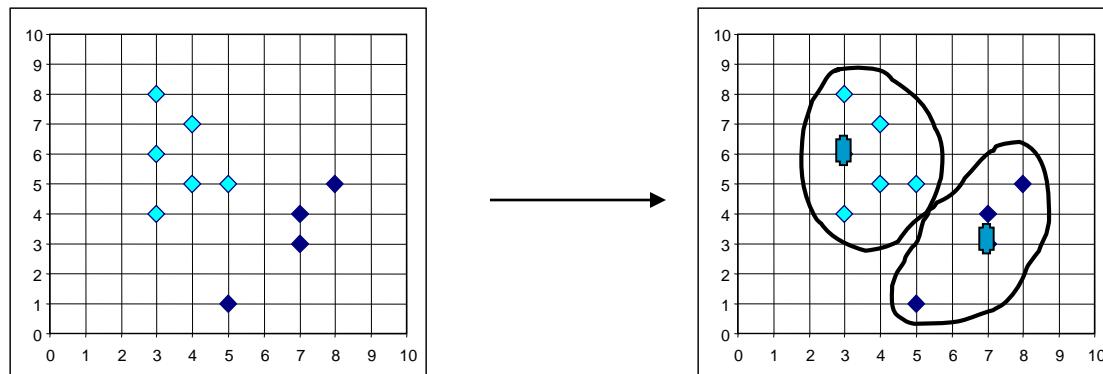
- Handling categorical data: *k-modes* modes -众数
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method 原型
- Expectation Maximization: an extension to k-means
  - Assign each object to a cluster according to a weight (prob.)
  - New means are computed based on weighted measures

# What Is the Problem of the K-Means Method?

- K-means algorithm is sensitive to outliers!
  - Since an object with an extremely large value may substantially distort the distribution of the data
- K-Medoids: Instead of taking the **mean** value of the objects in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster

枚举算法

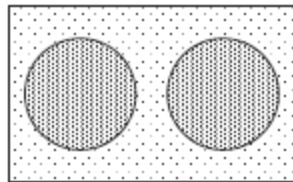
复杂度:  $O(n^2)$



# Exercise

---

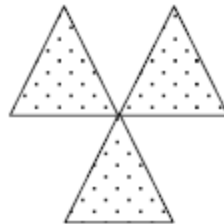
1. Identify the clusters using the K-means (using squared error as the objective function). Note that darkness or the number of dots indicates density.



(a)



(b)



(c)



(d)

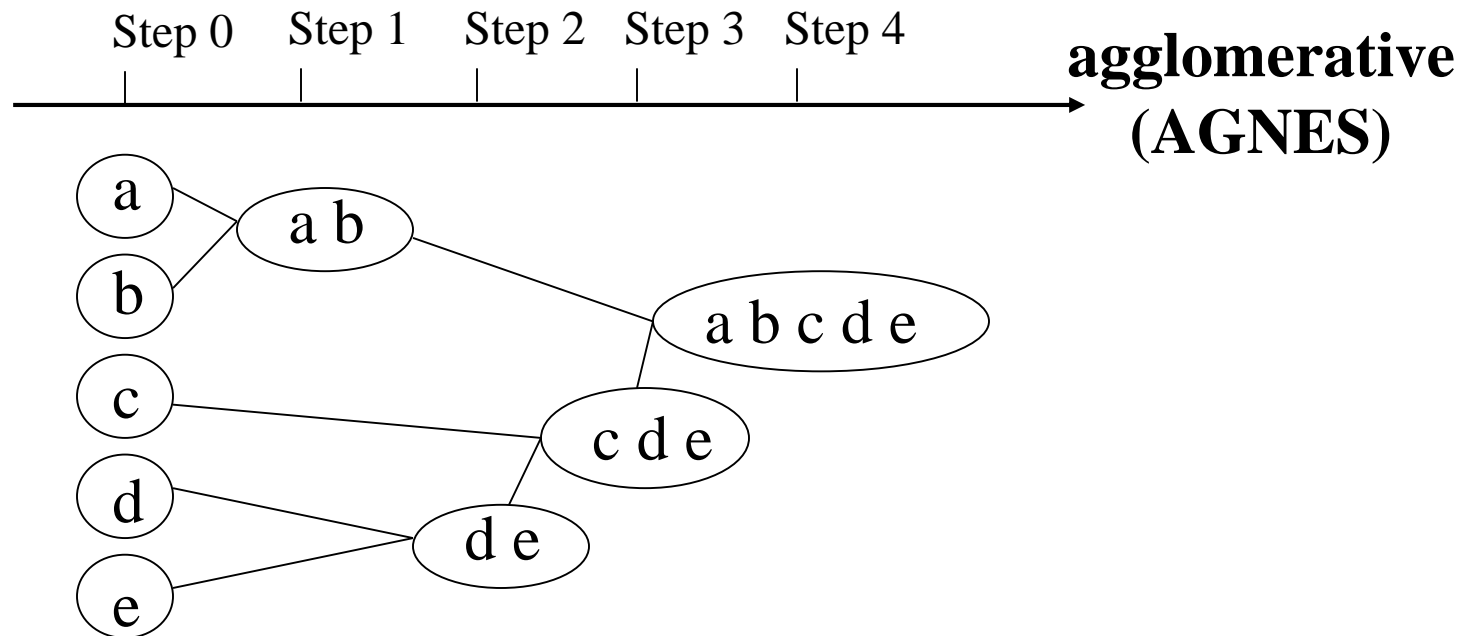
# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Outlier Analysis
- Summary

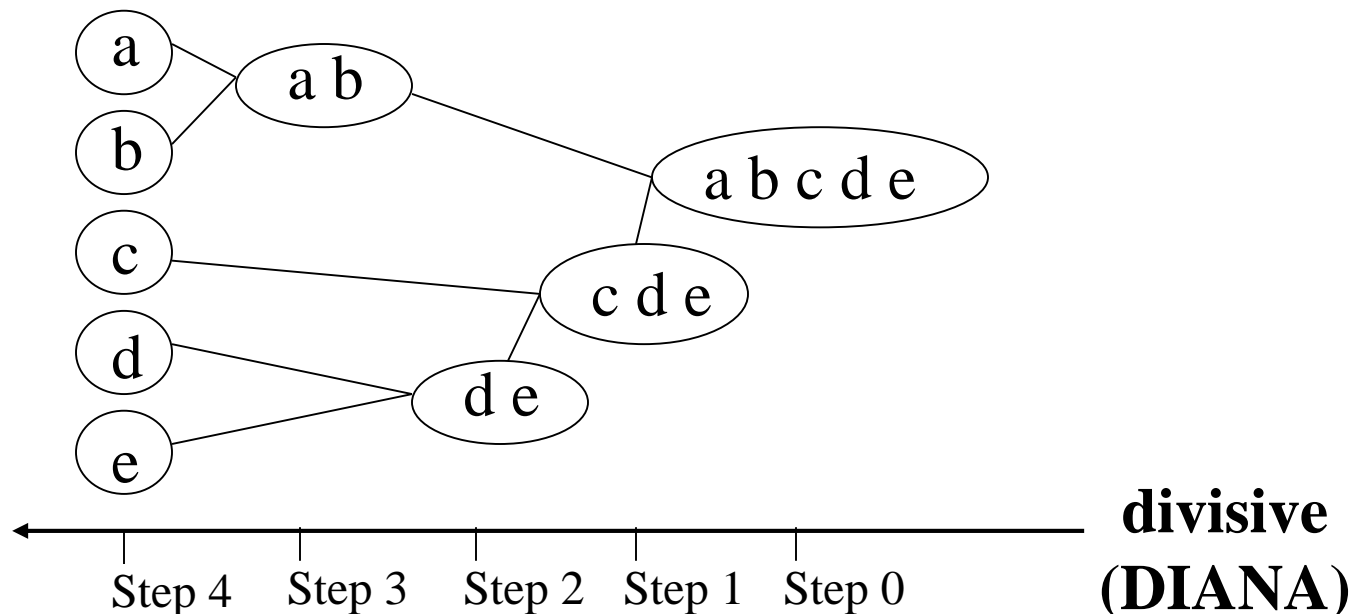
# Hierarchical Clustering

- This method does not require the number of clusters  $k$  as an input, but needs a termination condition



# Hierarchical Clustering

- This method does not require the number of clusters  $k$  as an input, but needs a termination condition

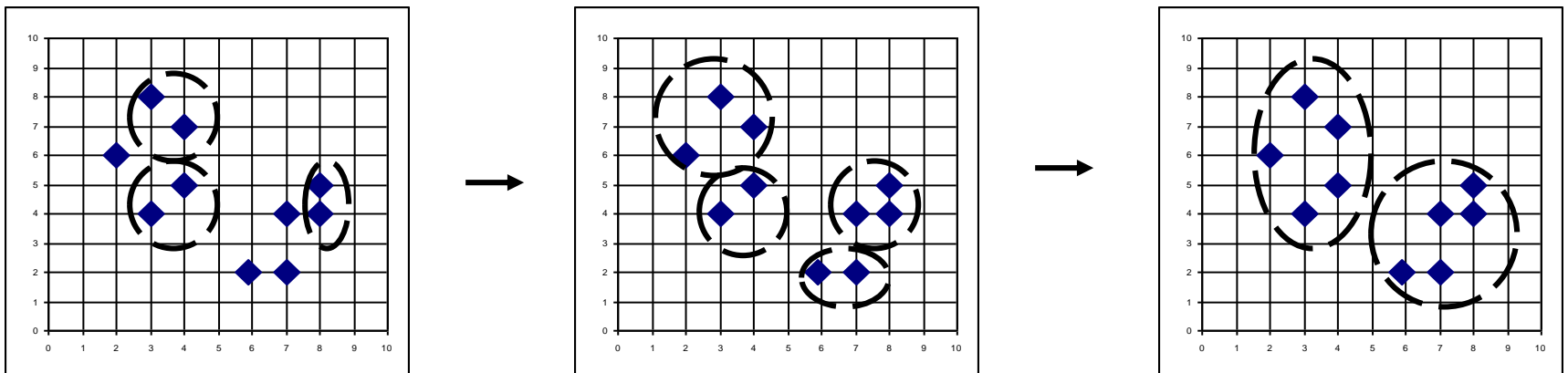




# AGNES (Agglomerative Nesting)

---

- Introduced by Kaufmann and Rousseeuw (1990)
- Use the Single-Link method and the dissimilarity matrix
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

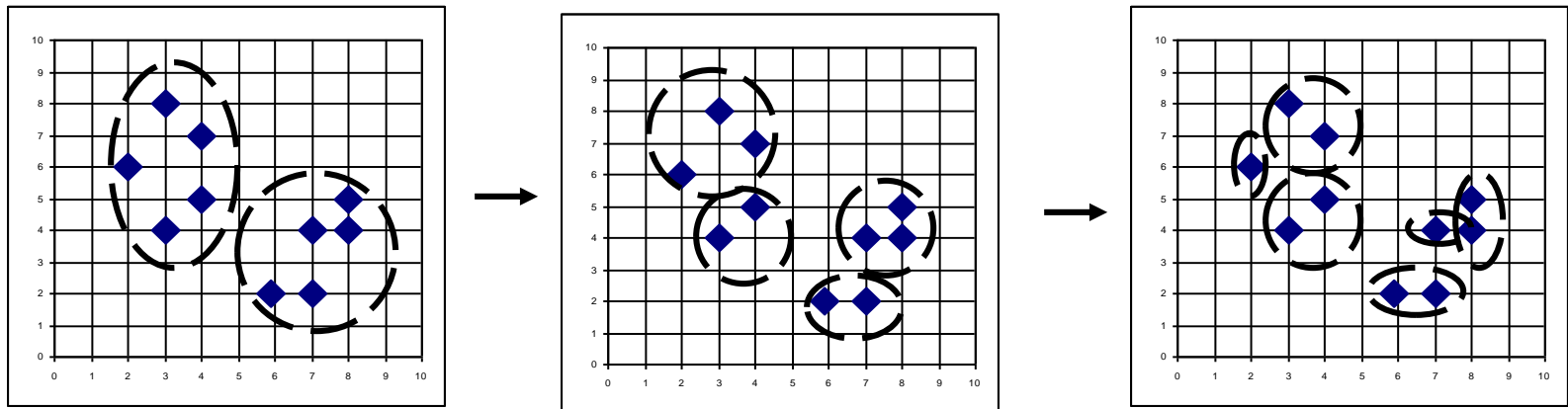


# DIANA (Divisive Analysis)

- Introduced by Kaufmann and Rousseeuw (1990)
- Inverse order of AGNES
- Eventually each node forms a cluster on its own

如何去切割是一个难点

复杂度至少是 $n^2$



# Recent Hierarchical Clustering Methods

---

- Major weakness of hierarchical clustering methods
  - **do not scale** well: time complexity of at least  $O(n^2)$ , where  $n$  is the number of total objects
  - can never undo what was done previously
- Integration of hierarchical with distance-based clustering
  - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
  - ROCK (1999): clustering categorical data by neighbor and link analysis
  - CHAMELEON (1999): hierarchical clustering using dynamic modeling

# BIRCH (1996)

---

- BIRCH: integrated hierarchical clustering
- Clustering feature, Clustering feature tree
- Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering
  - 阶段 Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)
  - Phase 2: use a clustering algorithm to cluster the leaf nodes of the CF-tree

# Clustering Feature Vector in BIRCH

**Clustering Feature:**  $CF = (N, \overrightarrow{LS}, SS)$ , summarize the cluster members

**$N$ : Number of data points**

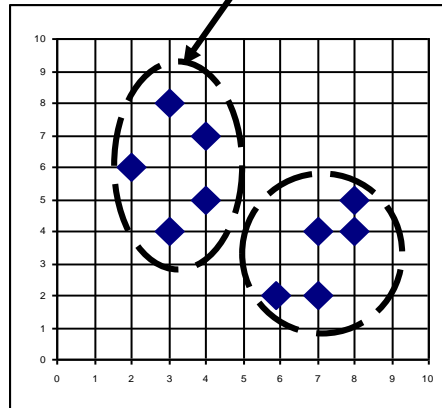
$$LS: \sum_{i=1}^N \overrightarrow{X_i}$$

线性和：所有横纵坐标加起来

$$SS: \sum_{i=1}^N \overrightarrow{X_i^2}$$

平方和：横纵坐标的平方加起来

$$CF = (5, (16,30), (54,190))$$



(3,4)

(2,6)

(4,5)

(4,7)

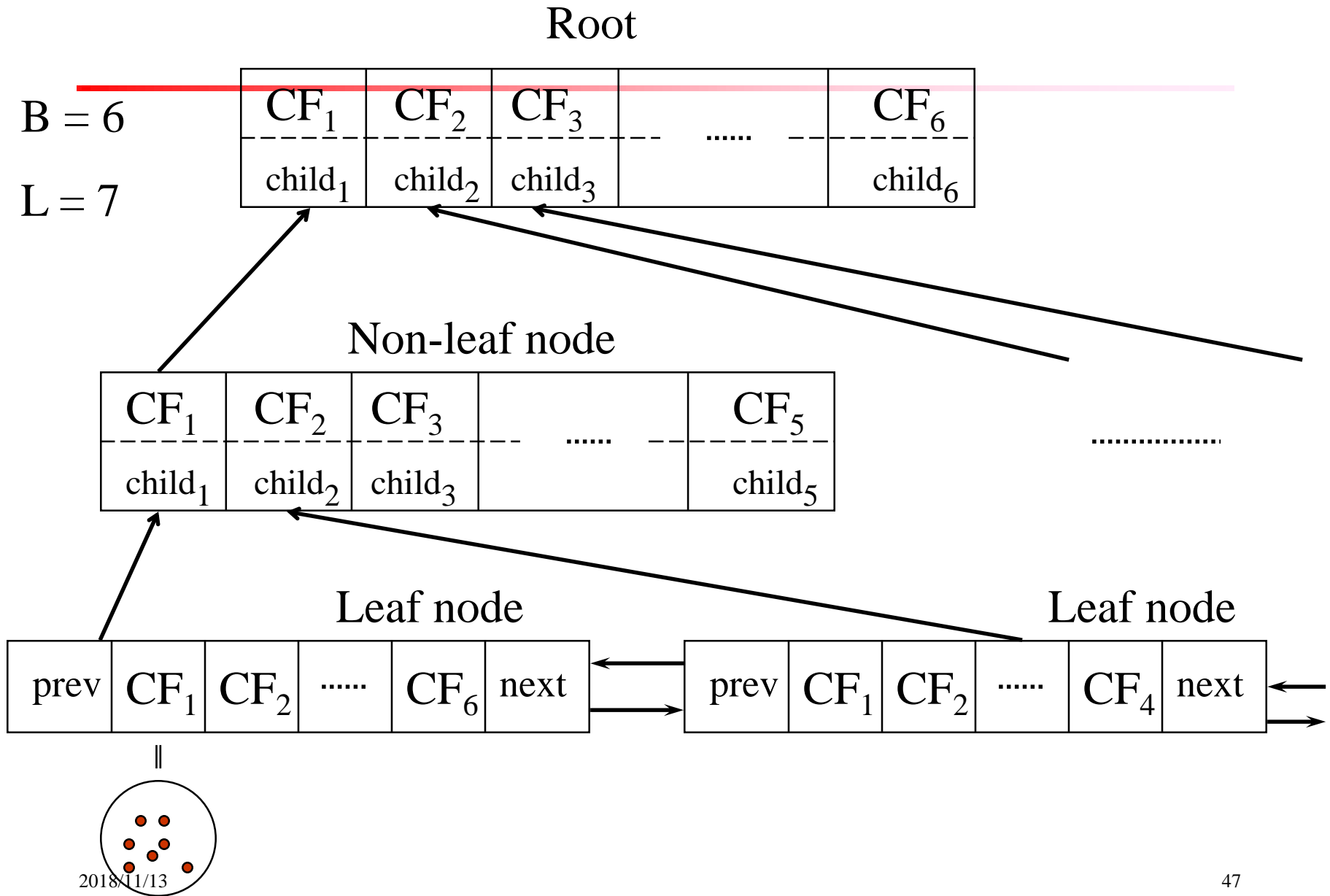
(3,8)

# CF-Tree in BIRCH

---

- Clustering feature:
  - summary of the statistics for a given subcluster
  - registers crucial measurements for computing cluster and utilizes storage efficiently 为计算集群注册重要的度量并有效地利用存储
- A CF tree is a height-balanced tree that stores the clustering features for a hierarchical clustering
  - A nonleaf node in a tree has descendants or “children”
  - The nonleaf nodes store sums of the CFs of their children
- A CF tree has two parameters
  - Branching factor: specify the maximum number of children
  - Threshold: max diameter of sub-clusters stored at the leaf nodes

# The CF Tree Structure



# BIRCH

---

## ■ Phase 1

- Insert each object to its closest leaf entry
- If the diameter of a leaf is larger than a threshold, the leaf will be split
- Update the CF and its ancestor's CF
- If the size of the CF tree is too big, re-build the tree from the leaf node, no re-scan the original objects
- Two parameters (branching factor, threshold), control the size of the tree



# BIRCH

---

## ■ Scales linearly

- Complexity:  $O(n)$
- Scalable for large database
- Incremental clustering
- Finds a good clustering with a single scan, I/O cost small

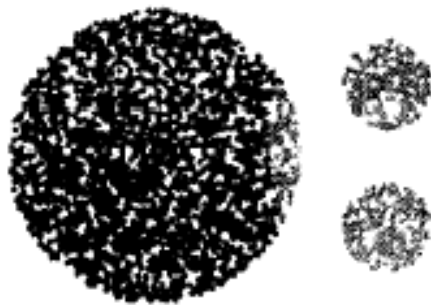
## ■ Weakness

- Handles only numeric data, and sensitive to the order of the data record
- Not good at arbitrary shaped cluster

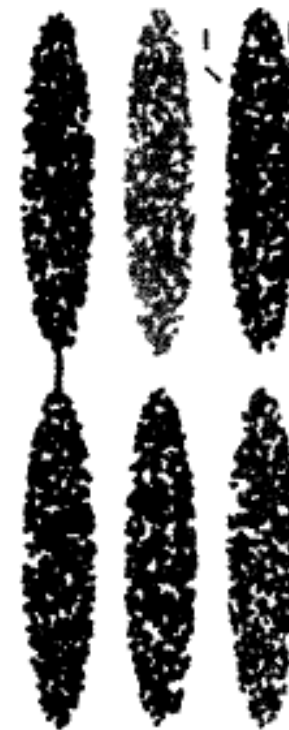
# CURE: Clustering Using Representatives

---

Centroid-based clustering      All-points agglomerative clustering



一个中心点



一对中心点

CURE: middle ground between centroid-based clustering and all-points agglomerative clustering

# CURE: Clustering Using Representatives

---

- Start with each individual point as a separate cluster
- Merge closest clusters till each cluster contains more than  $c$  points
- For each cluster, use  $c$  scattered points as representatives
- If more than  $k$  clusters
  - Clusters with the closest pair of representative points are merged
  - Update the representative points of merged clusters

# CURE: Clustering Using Representatives

---

## ■ Choose representatives

- the point farthest from the mean of the cluster
- for 2 to  $c$  do
  - the point farthest from the previously chosen point
- Shrink the scattered points toward the mean by a fraction  $\alpha$ 
  - for each scattered point  $p$  do
    - representative =  $p + \alpha * (\text{mean} - p)$

## ■ Merge

- Euclidian distance 欧式距离
- Closest clusters -- minimum distance between representative points from two clusters

$$dist(u, v) = \min_{p \in u.rep, q \in v.rep} dist(p, q)$$

# CURE: Clustering Using Representatives

---

- Multiple representative points allow CURE to discover arbitrary shaped clusters
- Less sensitive to outliers
  - Shrink scattered points toward the mean, weaken the effects of outliers
- Time complexity  $O(n^2 \log(n))$
- For large-scale database, do sampling and partitioning

# CURE: Clustering Using Representatives

---

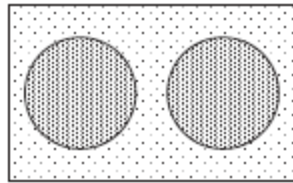
- Draw a random sampling  $S$  from original objects
- Cluster the sampled objects (basic *CURE*)
- Eliminate outliers
- Each unsampled original object is assigned to the cluster containing the closest representative point to it

???

# Exercise

---

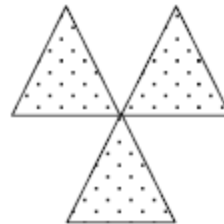
1. Identify the clusters using the Single-Link method. Note that darkness or the number of dots indicates density.



(a)



(b)



(c)



(d)

# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Outlier Analysis
- Summary



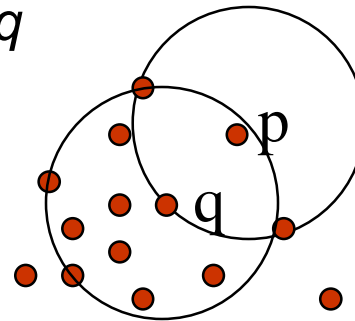
# Density-Based Clustering Methods

---

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - Need density parameters as termination condition
  - Complexity is  $O(n^2)$

# Density-Based Clustering: Basic Concepts

- Two parameters:
  - $\varepsilon$ -neighborhood: neighborhood within a radius  $\varepsilon$  of a point
  - $MinPts$ : Min number of points in  $\varepsilon$ -neighborhood of a point
- core object: If the number of points in  $\varepsilon$ -neighborhood of point  $p$  exceeds  $MinPts$
- **Directly density-reachable**: A point  $p$  is directly density-reachable from a point  $q$  w.r.t. ,  $\varepsilon$  ,  $MinPts$  if
  - $p$  belongs to  $\varepsilon$ -neighborhood of  $q$
  - $q$  is core object



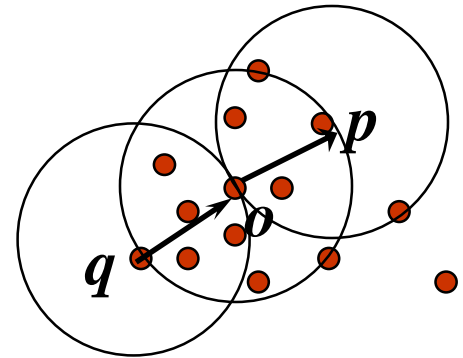
$MinPts = 5$

$\varepsilon = 1 \text{ cm}$

# Density-Reachable and Density-Connected

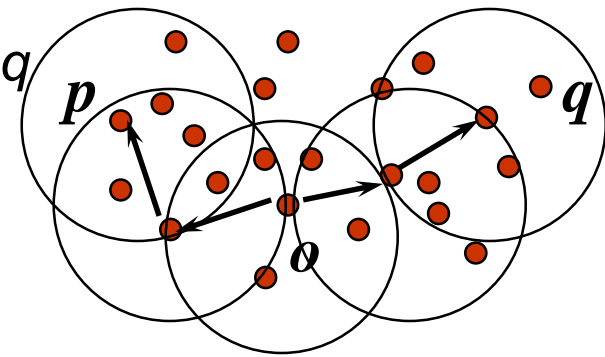
## ■ Density-reachable:

- A point  $p$  is **density-reachable** from a point  $q$  w.r.t.  $\varepsilon$ ,  $MinPts$  if there is a chain of points  $p_1, \dots, p_n$ ,  $p_1 = q$ ,  $p_n = p$  such that  $p_{i+1}$  is directly density-reachable from  $p_i$



## ■ Density-connected

- A point  $p$  is **density-connected** to a point  $q$  w.r.t.  $\varepsilon$ ,  $MinPts$  if there is a point  $o$  such that both  $p$  and  $q$  are density-reachable from  $o$  w.r.t.  $\varepsilon$  and  $MinPts$

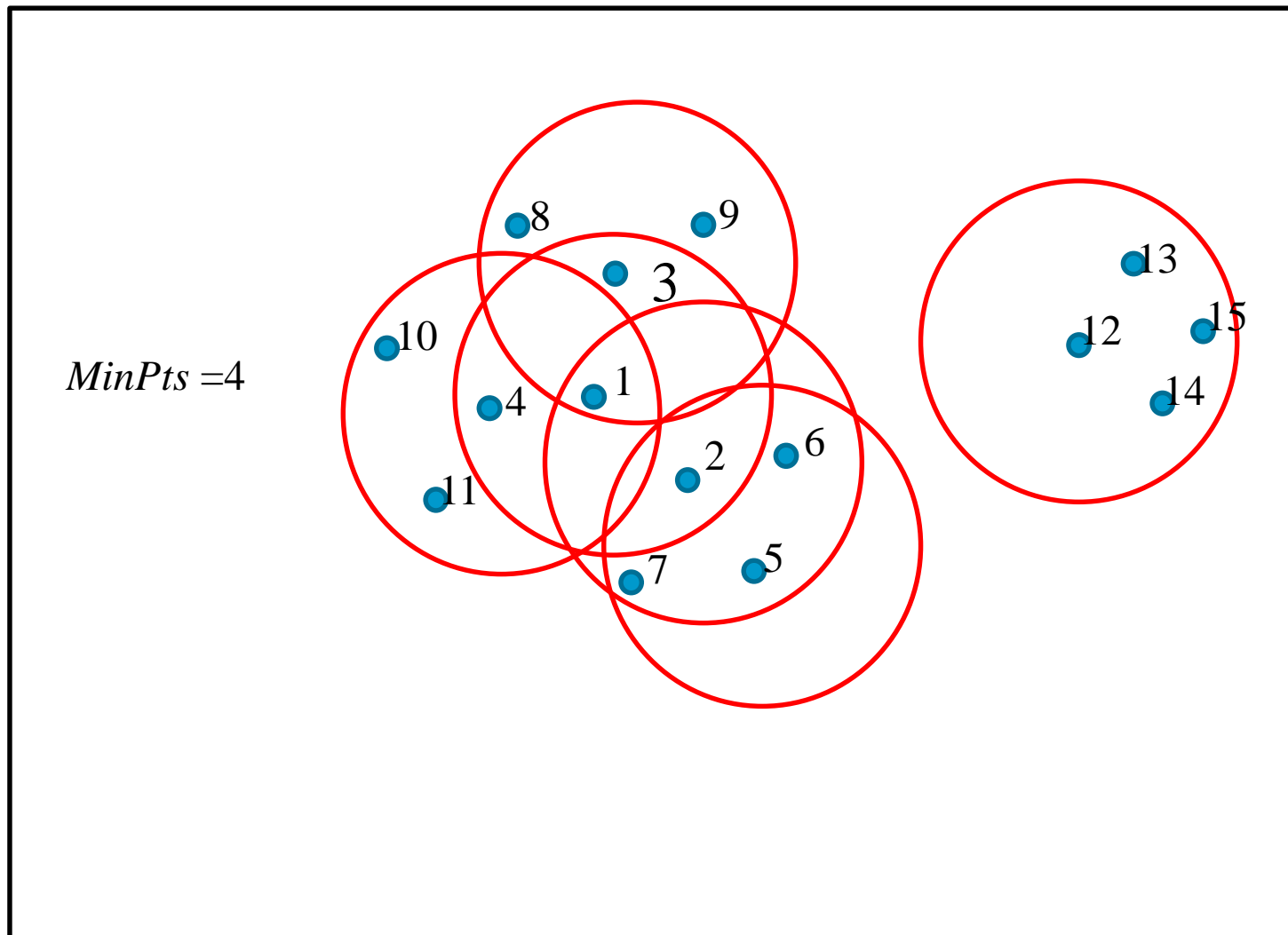


# DBSCAN: The Algorithm

---

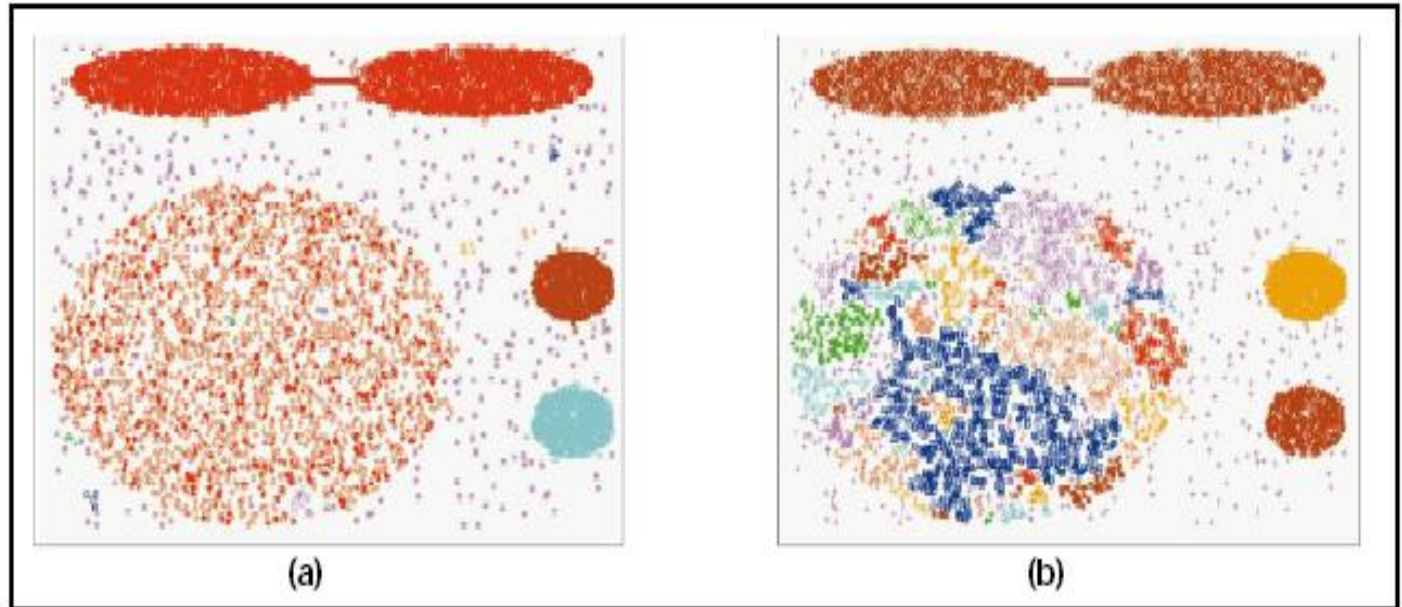
- (1) mark all the objects as ***unvisited***;
- (2) **do**
- (3)     randomly select an unvisited object ***p***;
- (4)     mark ***p*** as ***visited***;
- (5)     **if** the  $\varepsilon$ -neighborhood of ***p*** has at least *MinPts* objects
- (6)         create a new cluster ***C***, and add ***p*** to ***C***;
- (7)         let *N* be the set of objects in the  $\varepsilon$ -neighborhood of ***p***;
- (8)         **for** each point ***p'*** in *N*
- (9)             **if** ***p'*** is ***unvisited***
- (10)                 mark ***p'*** as ***visited***;
- (11)                 **if** the  $\varepsilon$ -neighborhood of ***p'*** has at least *MinPts* points, add those points to *N*;
- (12)                 **if** ***p'*** is not yet a member of any cluster, add ***p'*** to *C*;
- (13)         **end for**
- (14)         output *C*;
- (15)     **else** mark ***p*** as ***noise***;
- (16) **until** no object is ***unvisited***.

# DBSCAN: The Algorithm

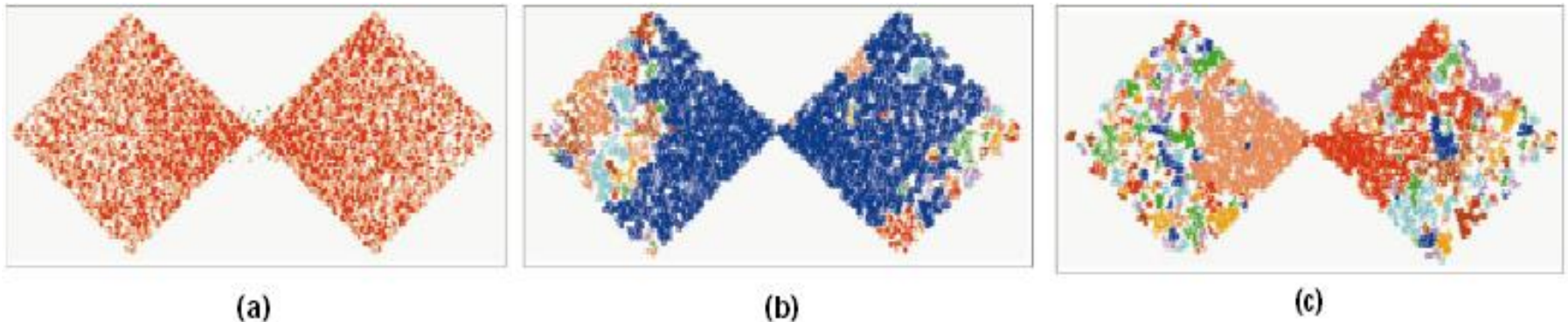


# DBSCAN: Sensitive to Parameters

DBScan results  
with MinPts = 4  
and  $\epsilon =$  (a) 0.5, (b)  
0.4



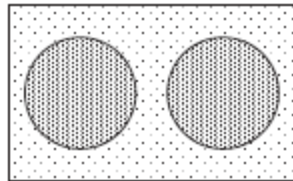
DBScan results  
with MinPts = 4  
and  $\epsilon =$  (a) 5.0, (b)  
4.0, (c) 3.0



# Exercise

---

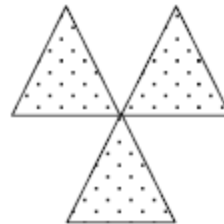
1. Identify the clusters using DBSCAN. Note that darkness or the number of dots indicates density.



(a)



(b)



(c)



(d)

# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- **Grid-Based Methods**
- Outlier Analysis
- Summary



# Grid-Based Clustering Method

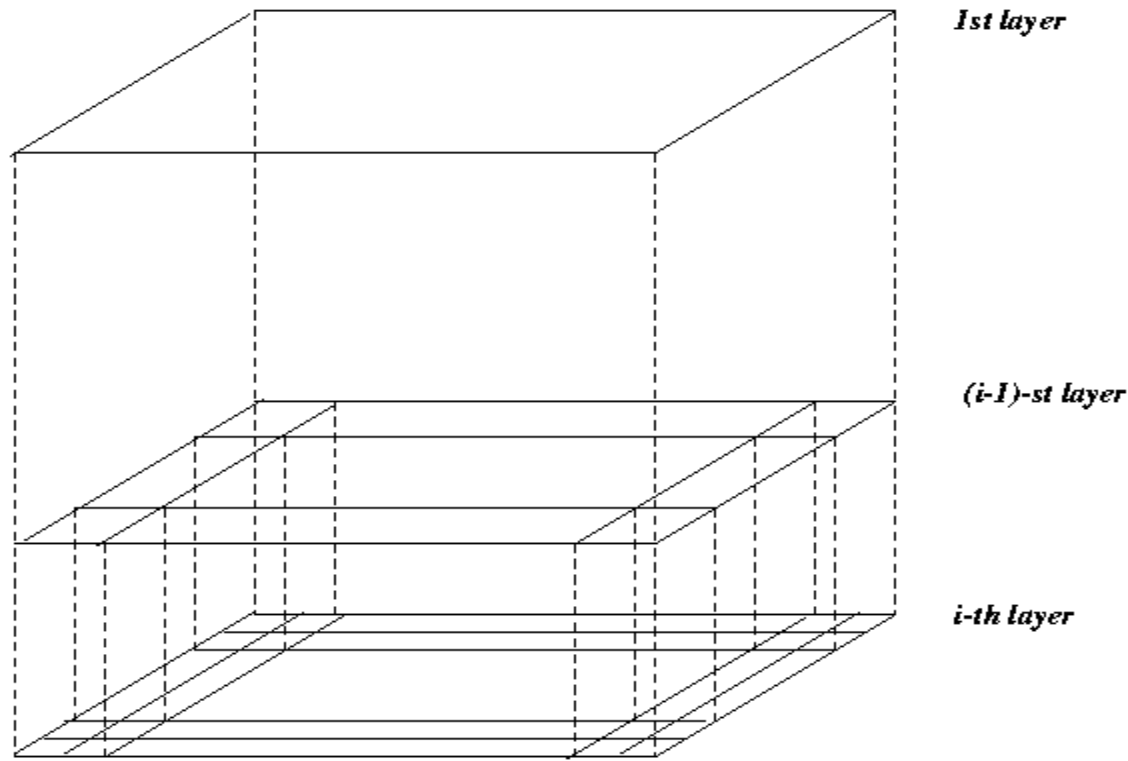
---

- Using multi-resolution grid data structure
- Several interesting methods
  - **STING** (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - **WaveCluster** by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
    - A multi-resolution clustering approach using wavelet method
  - **CLIQUE**: Agrawal, et al. (SIGMOD'98)
    - On high-dimensional data

# STING: A Statistical Information Grid Approach

---

- The spatial area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution
- Each cell at a high level is partitioned into a number of smaller cells in the next lower level



# The STING Clustering Method

---

- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
  - *count, mean, s, min, max*
  - type of distribution—normal, uniform, NONE, etc.
- Parameters of higher level cells can be easily calculated from parameters of lower level cells
- Clusters are identified based on *count, cell size, etc.*

$$n = \sum_i n_i$$
$$m = \frac{\sum_i m_i n_i}{n}$$
$$s = \sqrt{\frac{\sum_i (s_i^2 + m_i^2) n_i}{n} - m^2}$$
$$\min = \min_i (\min_i)$$
$$\max = \max_i (\max_i)$$

# The STING Query Method

---

- Use a top-down approach to answer spatial data queries
- Start from a pre-selected layer — typically with a small number of cells
- For each cell in the current level compute the confidence interval
- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level of the relevant cells
- Repeat this process until the bottom layer is reached

# Exercise

---

1. Please give some comments on STING in the following aspects:
  - (1) cluster shape
  - (2) computational complexity
  - (3) cluster quality
  - (4) incremental clustering

# Comments on STING Clustering

---

## ■ Advantages:

- Query-independent
- incremental update
- $O(K)$  for query, where  $K$  is the number of grid cells at the lowest level
- $O(n)$  for generating clusters

## ■ Disadvantages:

- All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected
- Processing time depends on the size of each grid

# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Outlier Analysis
- Summary

# What Is Outlier Discovery?

---

- What are outliers?
  - The set of objects are considerably dissimilar from the remaining of the data
  - Caused by
    - Measurement or execution errors
    - Result of inherent variability
- Mining outliers is valuable
- Applications:
  - Credit card fraud detection
  - Customer segmentation
  - Medical analysis



# Outlier Detection

---

## ■ Visualization

- Weak in data with categorical data, high dimensional data
- Good at numerical data of 2 or 3 dimensions

## ■ Clustering

- Byproduct of clustering may be outliers

## ■ Computer-based methods

- Statistical-based outlier detection
- Distance-based outlier detection
- Deviation-based outlier detection

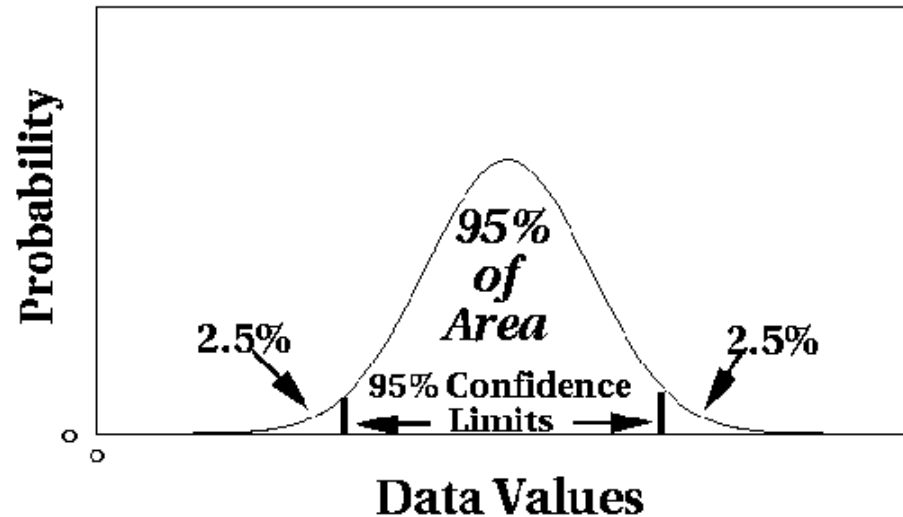
# Outlier Detection: Statistical Approaches

---

- Assume a distribution (e.g. normal distribution) for the data set and then use discordancy test to find outliers
- Discordancy tests depends on knowledge
  - data distribution
  - two hypothesis
  - distribution parameter (e.g., mean, variance)
  - number of expected outliers

# Outlier Detection: Statistical Approaches

---



## ■ Drawbacks

- Most tests are for single attribute
- In many cases, data distribution may not be known
- Require input parameters

# Outlier Detection: Distance-Based Approach

---

- Introduced to overcome the main limitations imposed by statistical methods
  - We need multi-dimensional analysis without knowing data distribution, no statistical test
- Distance-based outlier: A  $DB(p, d)$ -outlier is an object  $O$  in a dataset  $T$  such that at least a fraction  $p$  of the objects in  $T$  lies at a distance greater than  $d$  from  $O$
- Algorithms for mining distance-based outliers
  - Index-based algorithm
  - Cell-based algorithm

# Index-based Algorithm

邻域搜索

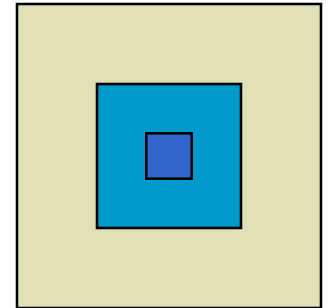
- Search for neighbors of each object  $O$  within radius  $d$  around the object
- Multi-dimensional index structure, e.g. kd tree
- Max number of objects within  $d$ -neighborhood of each outlier
- The worst case  $O(kn^2)$   
 $k$  dimensionality,  $n$  number of objects
- Drawbacks:
  - Tree building is computational intensive

# Cell-based Algorithm

---

## ■ Cell partition

- Partition data space into cells, side length  $d/2k^{1/2}$
- Each cell has two layers around it
  - First layer one cell thick
  - Second layer  $(2k^{1/2} - 1)$  cells thick



## ■ Outlier detection

- If count of the first layer  $>M$ , no outlier in this cell
- If count of the second layer  $\leq M$ , all objects are outliers
- Otherwise, examine every object in the cell

## ■ Good for large-scale data set

# Cell-based Algorithm

---

2	2	2	2	2	2	2
2	2	2	2	2	2	2
2	2	1	1	1	2	2
2	2	1	<b>C</b>	1	2	2
2	2	1	1	1	2	2
2	2	2	2	2	2	2
2	2	2	2	2	2	2

- $k=2$ ,  $k$  dimensionality
- Layer-1 property: given any point  $x$  in cell  $C$ , and any point  $y$  in layer-1 cell,  $\text{dist}(x,y) \leq d$

# Outlier Detection: Deviation-Based Approach

---

- Identifies outliers by examining the main characteristics of objects in a group
- Objects that “deviate” from this description are considered outliers
- Sequential exception technique
  - Simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects
  - A sequence of subsets,  $\{S_1, S_2, \dots, S_m\}$ ,  $S_{j-1} \subset S_j$
  - Calculate the dissimilarity difference between the current subset with the proceeding subset in the sequence



# Cluster Analysis

---

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Outlier Analysis
- **Summary**

# Summary

---

- **Cluster analysis** groups objects based on their **similarity** and has wide applications
- Measure of similarity can be computed for **various types of data**
- Clustering algorithms can be **categorized** into partitioning methods, hierarchical methods, density-based methods, grid-based methods
- **Outlier detection** and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches

# Summary

---

- Considerable progress has been made in scalable clustering methods
  - Partitioning: k-means, k-medoids, CLARANS
  - Hierarchical: BIRCH, ROCK, CHAMELEON
  - Density-based: DBSCAN, OPTICS, DenClue
  - Grid-based: STING, WaveCluster, CLIQUE
  - Model-based: EM, Fuzzy K-Means
  - Frequent pattern-based: pCluster
  - Constraint-based: COD, constrained-clustering