

CSCI 5408 Data Analytics: DM and DW Tech

(Apr 4, Week 13)

- Ass6 Due: Apr 11
 - Read Assignment 6 & Ass6-Tutorial slides
 - Help Hours: Fri, 1:00-2:30 PM, CS 233
- Final Exam: Apr 20, 3:30-5:30PM, DALPLEX
- Write answers for review questions
- **MACS Co-op Orientation:**
 - April 11, 10:00-11:00AM in Rm 430 in the CS building (pizza provided)

6. Clustering Mining

(Textbooks 3rd: 10.1-10.3; 2nd: 7.1-4, 7.5.1, 7.12)

- Clustering problem overview
- Data types for clustering
- K-means method
- K-medoids method
- Hierarchical methods
- Text clustering
- Summary

Concepts of Clustering DM

- **Goal: To find accurate clusters from data, i.e. groups of data objects representing different object concepts**
 - Similar to one another within the same cluster
 - Dissimilar to the objects in other clusters
 - E.g., Customer groups, News discussion groups, Twits topic groups, Web page theme groups, etc.
- **Clustering DM:**
 - Methods/algorithms of dividing dataset into groups based on **similarity measure** (based on generic criteria)
 - No need of labeled training data (**unsupervised learning**)

Find customer groups based on
Customer (id, age, credit, income)

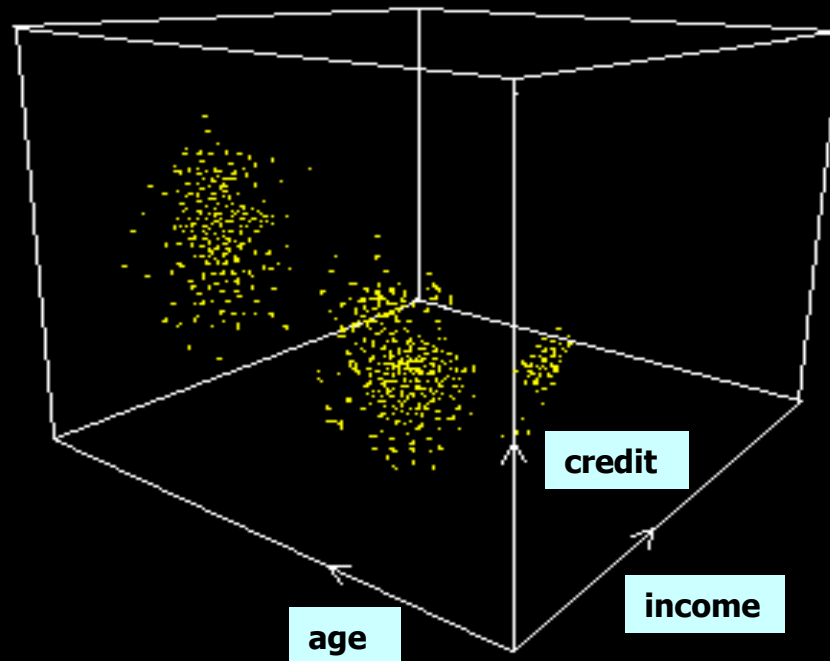
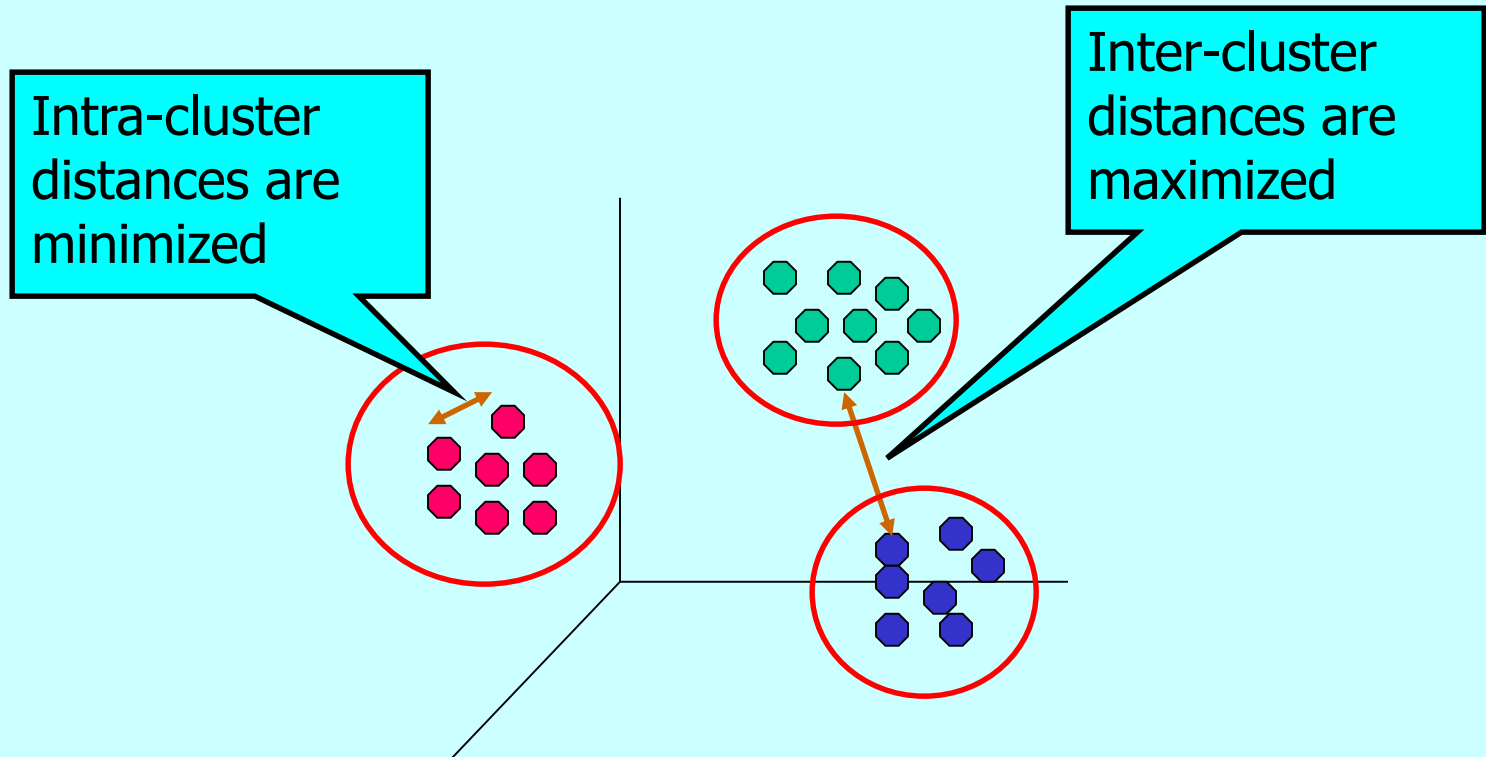


Illustration: Cluster Analysis

- Finding groups of objects which are similar in a group and dissimilar from different groups measured by distance criteria



What are the differences between
Clustering and Classification?

Clustering vs. Classification

- **Classification – Supervised learning on a specific target concept**
 - Training data, i.e. samples with known answers of the target concept are required.
 - What happens when concepts/answers are hidden, i.e. don't have any specific target concept?
- **Clustering – Unsupervised learning, no specific target concept**
 - No teacher/training data, which requires that the learner form and evaluate concepts on its own.
 - Clustering may be also viewed as unsupervised generic classification: the target is the data set itself, based on multiple attributes, there are no predefined classes, but rather to discover the classes.

Typical Usages

- When DM objective is less precise
 - Clustering mining differs from other data mining techniques in that its objective is generally far less precise than the objectives of other predictive (e.g. classification) and descriptive (e.g. association) mining.
 - We are always being told to "look at the big picture". But the fact is, sometimes the big picture is too confusing to be understood, so we may need to apply “divide and conquer” strategy to analyze the data.
 - E.g., Customer behaviors analysis, customer profile DM, etc.
- As a stand-alone tool to get insight into hidden data concepts, and data distribution to the concepts
- As a preprocessing step for other algorithms

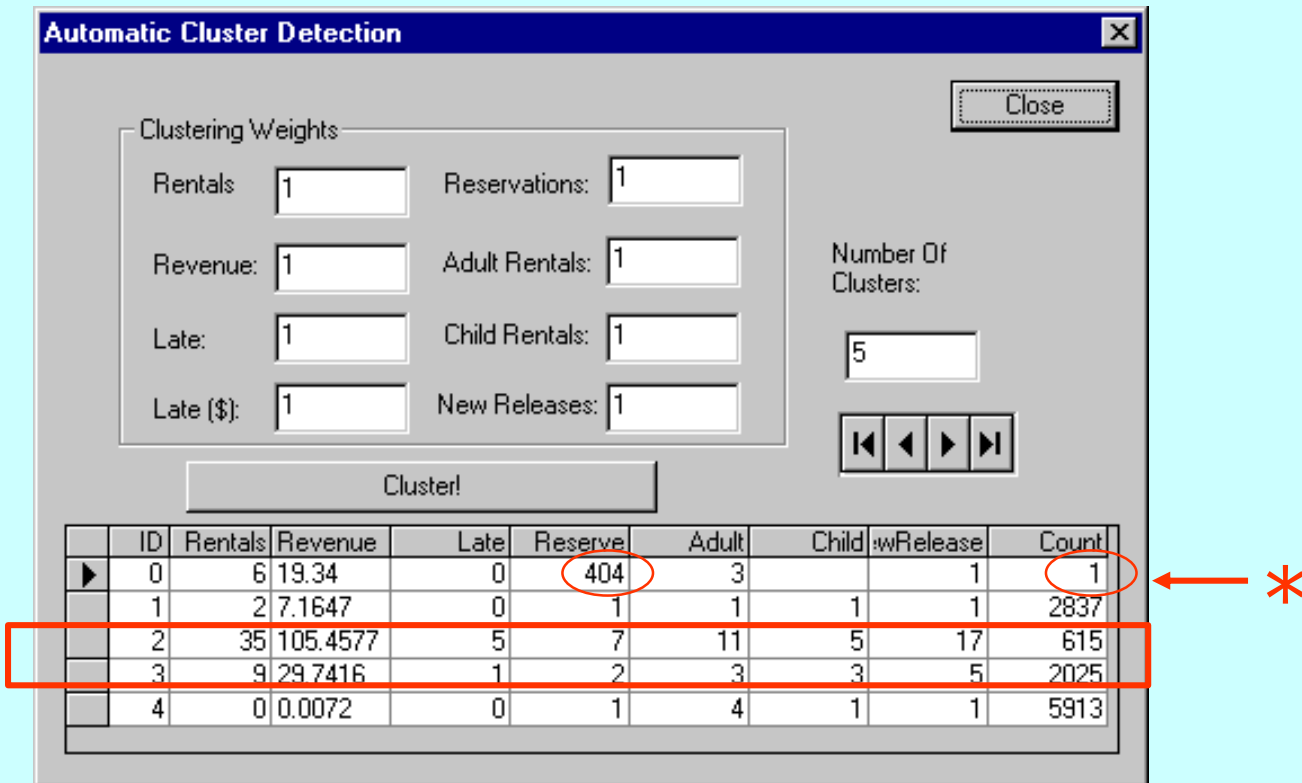
E.g.1, Video store data clustering for customer group analysis. (Doc/Thesis/HBthesisPothier.pdf)

Background: This project utilized data originating from a video outlet in Yarmouth, Nova Scotia containing transactional and customer information for a summer of 6 months. This database consists of approximately 11,000 customer records, approximately 50,000 fully detailed transactions, and information on the items being rented and sold. The **objective** is to find customer groups for improving marketing practices.

An instance of the DB (after data preprocessing):

CustID	nRentals	Revenue	nLate	LateChg	ReserveCount	Adult	Child	NewRel
49163	29	\$58.99	1	\$2.47	19	1	9	19
49164	14	\$31.22	10	\$37.70	0	3	7	4
49165	0	\$0.00	0	\$0.00	0	0	0	0
49166	22	\$123.86	7	\$22.23	6	0	1	15

E.g. The result when choosing $k=5$ with unit weights:



* An anomaly since it contains only one data instance but that one customer has reserved 404 movies. This would seem to indicate a problem that occurred during the initial phases of the data summarization and coding that led to this cluster.

E.g., Analytical Interpretation

Cluster	Description	# Of Customers
4	“Zero” cluster – customers that either <u>never or very rarely</u> purchase/rent from the store (in this period)	5913
1 *	Customers that <u>seldom rent</u> from the store (Averaging 2-3 transactions in a 6 month period)	2837
<u>3</u> **	Customers with <u>more frequent transactions</u> (Ave: 9 transactions in the 6 month period)	2025
<u>2</u> ***	Customers with an <u>extremely high frequency</u> of transactions (Ave: 35) and sales volume. (This cluster would form the business’ most valuable customers)	615
Total Customers		11,390

E.g. 2, Clustering newsgroup emails for supporting information retrieval application (Doc/Thesis/MCStthesisGuo00.pdf)

A sample of Email:

Return-path: <LUS.MD.GOOJU@lpch.stanford.edu>

Received: from DIRECTORY-DAEMON by SYSWRK.UCIS.DAL.CA (PMDf V4.3-13 #6307)
id <01JEAINJT3NK001RMV@SYSWRK.UCIS.DAL.CA>; Mon, 02 Aug 1999 12:26:06 -0300

Received: from 171.65.56.137 by SYSWRK.UCIS.DAL.CA (PMDf V4.3-13 #6307)
id <01JEAINBMNQ8001NW4@SYSWRK.UCIS.DAL.CA>; Mon, 02 Aug 1999 12:25:55 -0300

Received: from meditech.Stanford.EDU by LPCH.Stanford.Edu with SMTP
(Microsoft Exchange Internet Mail Service Version 5.0.1458.49)
id P98FYL31; Mon, 2 Aug 1999 08:25:29 -0700

Date: Mon, 02 Aug 1999 08:27:16 -0700

From: LUS.MD.GOOJU@lpch.stanford.edu

To: PEDIATRIC-PAIN@ac.dal.ca

Message-id: <990803527.LUS475898@lpch.stanford.edu>

Content-transfer-encoding: 7BIT

To: PEDIATRIC-PAIN@ac.dal.ca

From: GOOD,JULIE

Date: August 2, 1999 08:27am

Our nurses use the Wong-Baker FACES on a ten point scale, with each of the faces divided by an extra point. We are not sure if this affects the validity or reliability of the measure - but it facilitates a smooth transition to the pure numeric scale when the patients are older.

-Julie J. Good, M.D.

Pediatric Pain Management Fellow

Packard Children's Hospital at Stanford

From raw data to clean and normalized data:

Return-path: <Drhbg@aol.com>
Received: from DIRECTORY-DAEMON by SYSWRK.UCIS.DAL.CA (PMDf V4.3-13 #6307) id <01J615F5VHLS00BCUD@SYSWRK.UCIS.DAL.CA>; Fri, 01 Jan 1999 15:42:13 -0400
Received: from imo23.mx.aol.com by SYSWRK.UCIS.DAL.CA (PMDf V4.3-13 #6307) id <01J615F12Y6O00CSAS@SYSWRK.UCIS.DAL.CA>; Fri, 01 Jan 1999 15:42:07 -0400
Received: from Drhbg@aol.com by imo23.mx.aol.com (IMOV18.1) id NVXF07005 for <pediatric-pain@ac.dal.ca>; Fri, 1 Jan 1999 14:41:54 -0500 (EST)
Date: Fri, 01 Jan 1999 14:41:54 -0500 (EST)
From: Drhbg@aol.com
Subject: Re: Management of nerve injury
To: pediatric-pain@ac.dal.ca
Message-id: <7deafb8.368d2502@aol.com>
MIME-version: 1.0
x-Mailer: AOL 2.5 for Windows
Content-type: text/plain; charset=US-ASCII
Content-transfer-encoding: 7bit

I agree with William Fenton. I think mexiletine should be used as a second line drug. I ordinarily treat patients with chronic neuropathic pain. However, on a number of occasions, I have treated patients with acute neuropathic pains such as sciatica or brachial plexopathies. I have prescribed gabapentin at the outset of the pain, and have found that patients have responded extremely well. They often require lower than anticipated dosages of opioid analgesics. I doubt there is any data on the benefits of early use of anticonvulsants, but a case-control study would be of value.

Return-path: ...

Thread 1: Opioids and Meningitis

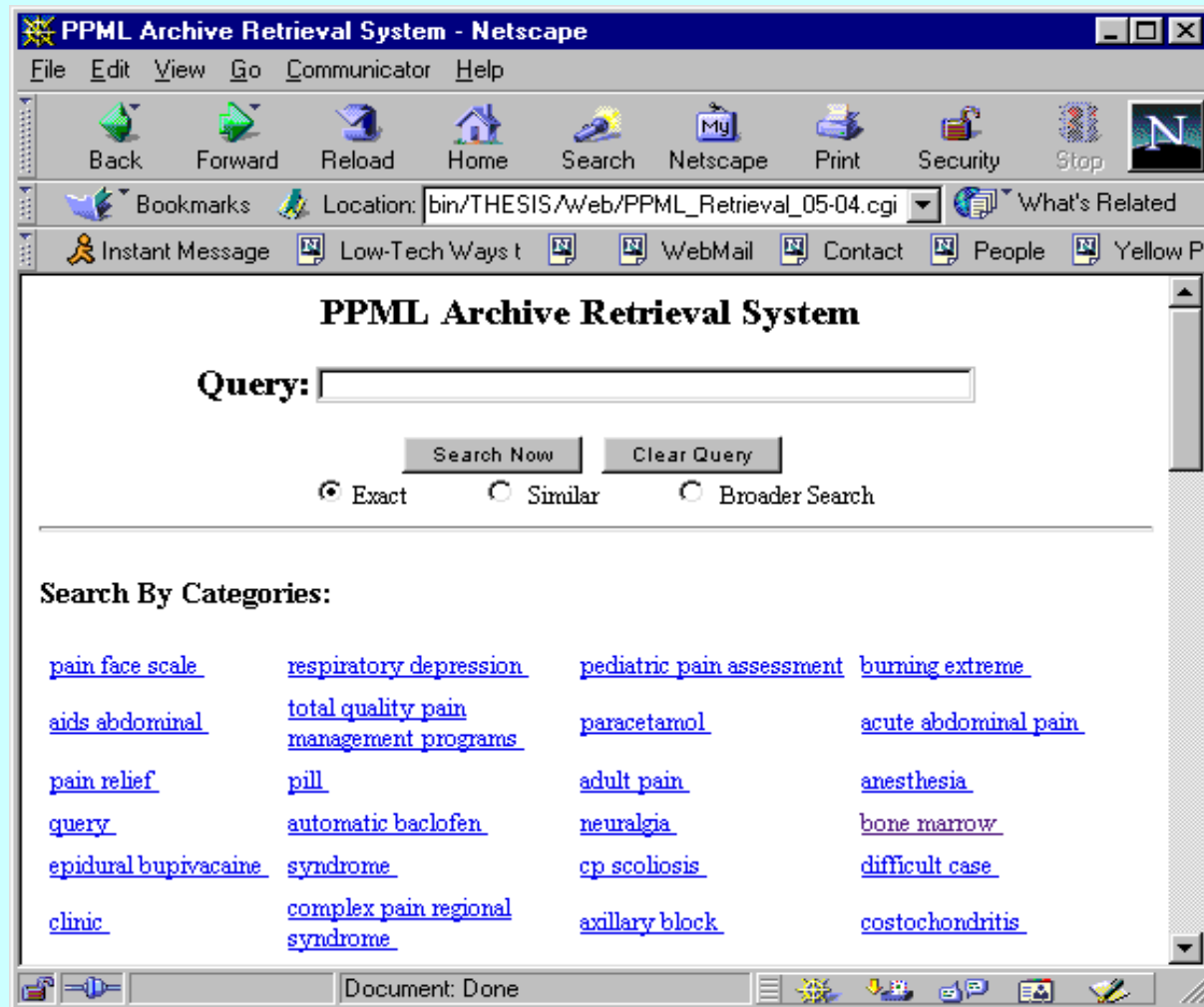
Date: Wed, 04 Jan 1995 16:54:48 -0500 (EST) From: posterSubject: opioids and meningitisX is a 13 month (9.8kg) old boy suffering from acute meningitis (pneumococque) treated with IV cefotaxime; at day three, I have been called as pediatric pain consultant to assess X; I have discovered an extreme painful state: one could not handle or touch him without producing screaming. The child was unable to move spontaneously he looked paralysed by pain and hypertonia; he also presented a neurological complication: ptosis at the right side. The pain treatment was IV acetaminophen. The first day I have prescribed IV Nalbuphine (weak opioid antagonist and agonist) 11mg/24h after a loading dose of 1.4 mg; Pain at rest has been successfully relieved but not the mobilisation pain; the dose has been increased at 14 mg/day without relieving the pain associated with moving; he has moved spontaneously limbs 2 days later; nalbuphine has been stopped 4 days later. Neurological examination and CT scan have been still normal (except ptosis) during this period. No opioid's side effects have been observed. What do you think of this case? Have you any experience with opioids and acute meningitis? Dr Poster, Pediatric pain unit, Poster Hospital

Date: Wed, 04 Jan 1995 17:27:25 -0500 (EST) From: first replySubject: re: opioids and meningitisIs there any periosteal involvement? If so an NSAID (ibuprofen or naproxen) may be much more effective than even opioid.

Date: Wed, 04 Jan 1995 19:06:32 -0400 From: second replySubject: Re: opioids and meningitisPoster writes:>X is a 13 month (9.8kg) old boy suffering from acute meningitis...>extreme painful state: one could not handle or touch him without>producing screaming....>The first day I have prescribed IV Nalbuphine ...>successfully relieved but not the mobilisation pain;...>has moved spontaneously limbs 2 days later; nalbuphine has been stopped 4>days later. Neurological examination and CT scan have been still normal...I have used IV morphine for similar severe meningitis pain, with success. I wouldn't hesitate to use a pure opioid agonist (in conjunction with acetaminophen, NSAID, and/or tricyclics). However, it sounds like you have the situation under control. Second Reply, Associate Professor, Dept and University

Date: Thu, 05 Jan 1995 18:58:32 -0800 (PST) From: Third ReplySubject: Re: opioids and meningitisI wonder if the problem is not due to severe arachnoiditis that is secondary to the inflammation. I would suggest a trial of steroids in this patient, perhaps in combination with a benzodiazepine to reduce the spasm. Narcotics may reduce the pain but I would not like to keep X on them for too long. Good luck Third Reply

E.g., The generated concepts for supporting information retrieval:



Define Clustering Mining

- Given a dataset D with n tuples: $D=\{t_1, t_2, \dots, t_n\}$, and a parameter k , the **Clustering Problem** is to find a mapping $f : D \Rightarrow \{1, \dots, k\}$ where each t_i is assigned to one cluster K_j , $1 \leq j \leq k$
- A *Cluster*, K_j , contains precisely those tuples mapped to it (with the highest similarity)
- A general similarity measure method: **Euclidean distance**
- Unlike classification problem, clusters are not known a priori.

What Is Good Clustering Method?

- 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.

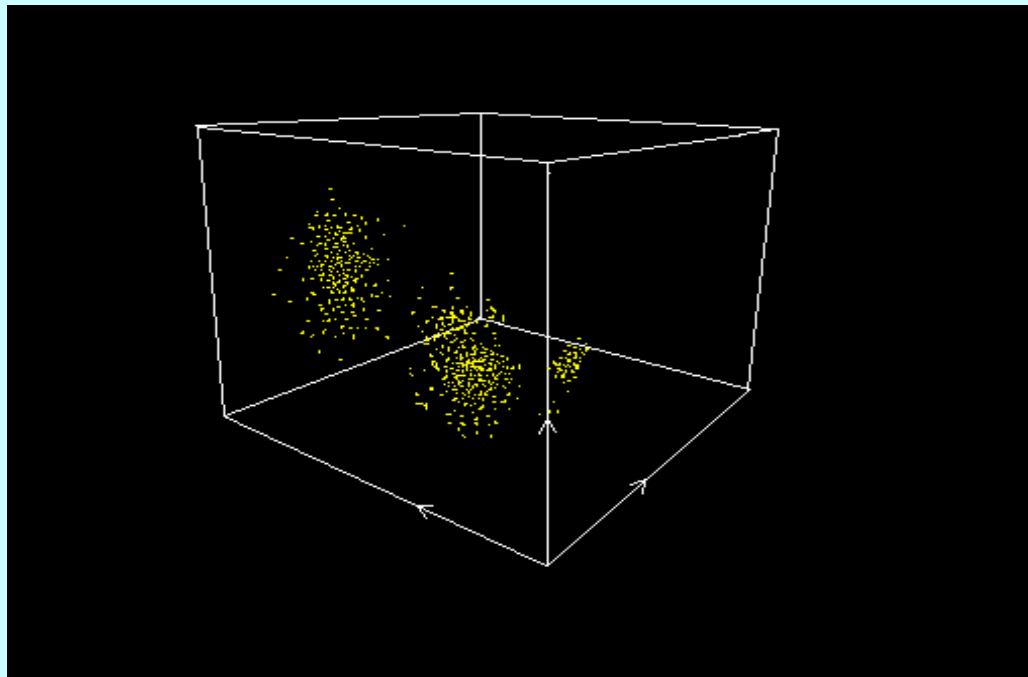
General Criteria of Clustering

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

Data Types for Clustering

(Textbook 3rd: 2.4, 2nd:7.2)

- The records in a data set need to be mapped into points (position vector) in a multidimensional space
- **How to handle features with non-numerical domains?**



How to handle variables with non-numerical domains

- Attributes can be
 - non-numerical, such as binary, nominal (categorical), ordinal variables.
 - Some numeric variables such as rankings do not have the right behavior to properly be treated as components of a position vector
- E.g., Records for purchases, phone calls, airplane trips, addresses, and many other things that have no obvious connection to the dots in a cluster diagram.

Data Type Issues for Clustering

- Data representation: (data structure)
 - The records in a data set need to be mapped into points (position vector) in a multi-dimensional space
- How to process non-numerical attributes for cluster analysis?

Data Structures for Clustering

- **With a dataset of n objects:**

- **Data matrix (object-by-variable structure)**

- Two modes: n objects by p variables
 - A form of relational table or $n \times p$ 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 (object-by-object structure)**

- One mode: n objects by n objects, and $d(i, j) = d(j, i)$, $d(i, i) = 0$
 - A form of $n \times n$ table, stores available differences for all pairs of n objects

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

- **Relationship of the two matrixes:**

- Many clustering algorithms operate on dissimilarity matrix directly, in which data matrix is used for serving data preparation process.

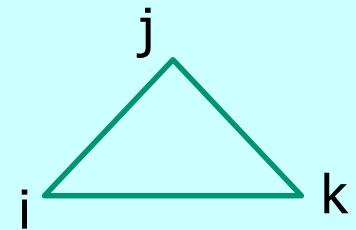
Dissimilarity Between Objects

- Euclidean distance:

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

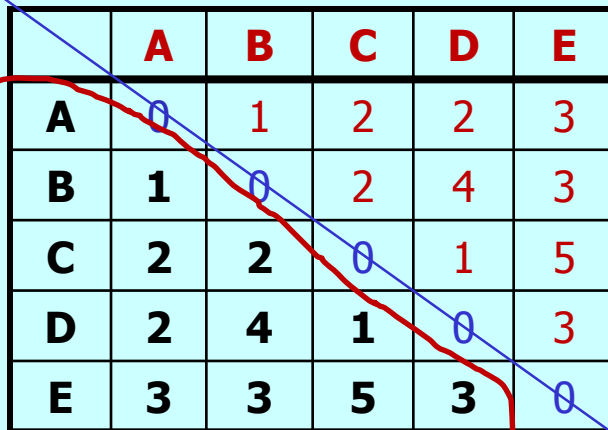
– Properties

- $d(i, j) \geq 0$
- $d(i, i) = 0$
- $d(i, j) = d(j, i)$
- $d(i, j) \leq d(i, k) + d(k, j)$



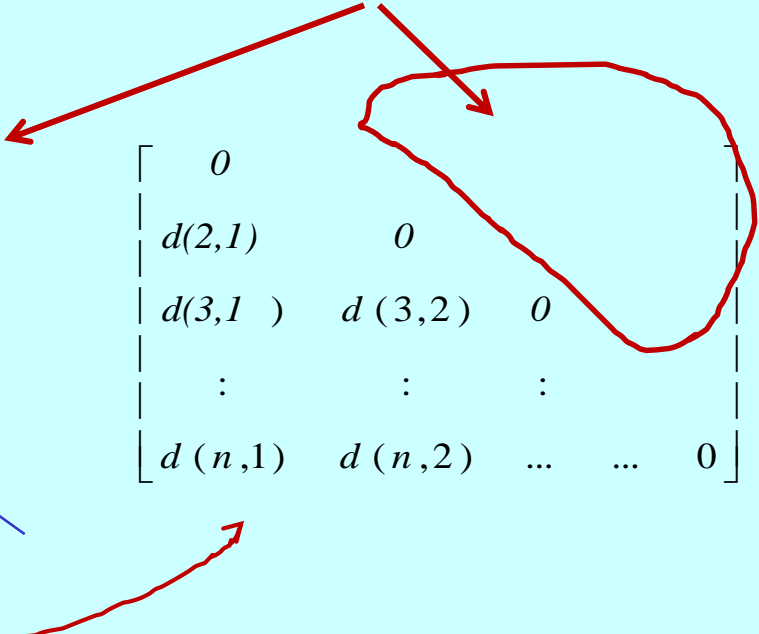
- Also, one can use weighted distance, or other dissimilarity measures

E.g. A dissimilarity matrix with available distances of objects {A,B,C,D,E}



	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

Redundant: $d(i,j) = d(j,i)$



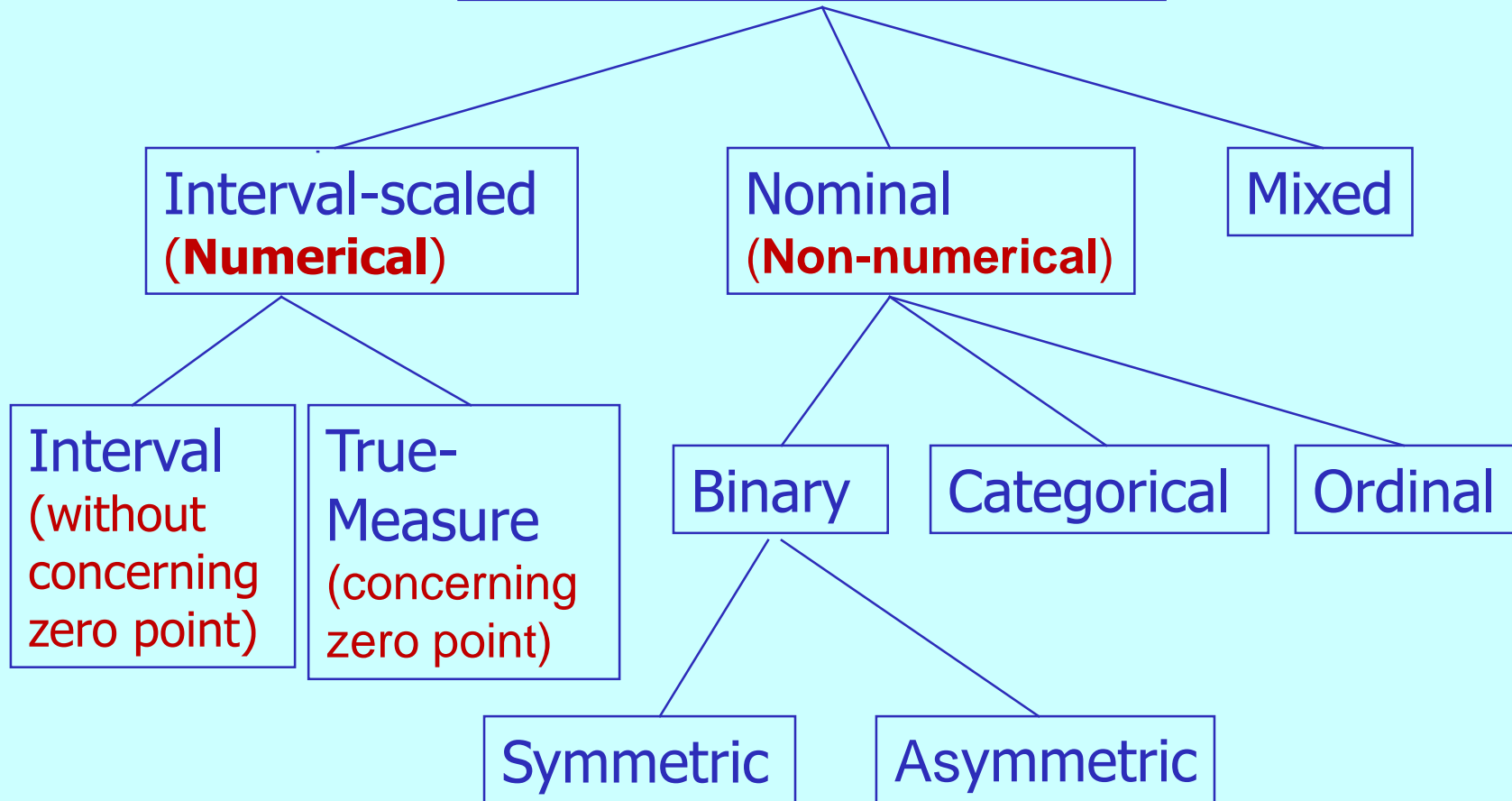
$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

What data types can be used for clustering?

Main Attribute Types for Clustering

- Interval-scaled variables
- Binary variables
- Categorical variables
- Ordinal variables
- Variables of mixed types

Attribute Data Types



Interval-valued variables

- **Interval Variables:**

- Describe the distance between two measure values without concerning a zero point, such as temperature, etc.

E.g., A day's highest temperature: 20 C° and lowest temperature is 10 C° , the highest difference = $20 - 10 = 10\text{ C}^{\circ}$.

- Ratio has no meaning for interval variables

E.g., Is a temperature 20 C° warmer twice than 10 C° ?

No. but it is true to say it is 10 C° warmer.

- **True Measure Variables:**

- An interval variables with a zero point
- The ratio of two values of the variable is meaningful

E.g., A distance 1000km is twice far of a distance 500km.

Distance, volume, age, etc. are examples of true measures.

Binary Variables

- How to measure dissimilarity between objects with binary variables?
 - E.g., Hospital patient database

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

- Two values of binary variables may not have equal impact (i.e. importance) for an application
 - E.g., In the given medical application, HIV tested as positive will have more impact to the overall outcome. If two variables of such have positive (or “1s”), the outcome would be considered more significant than of two “0s”.

Convert a set of binary variables into $d(i,j)$

- A **contingency table** for p binary variables of *object i* and *object j*

		<i>Object j</i>		
		1	0	<i>sum</i>
<i>Object i</i>	1	q	r	$q + r$
	0	s	t	$s + t$
	<i>sum</i>	$q + s$	$r + t$	p

The total number of binary variables: $p = q + r + s + t$.

- Simple matching coefficient** (invariant, if the binary variable is symmetric):

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

(Symmetric: if a variable has the same weight for its states and carry same weight, such as gender (male, female)).

- Jaccard coefficient** (non-invariant if the binary variable is asymmetric):

$$d(i, j) = \frac{r + s}{q + r + s}$$

if $1_{\text{importance}} > 0_{\text{importance}}$
("t" is unimportant and ignored)

E.g., The patient testing data set.

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	1	0	1	0	0	0
Mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

- Gender is a symmetric attribute (male and female carry on the same weight)
- The remaining attributes are asymmetric binary
- Let the values “Y” and “P” be set to 1, and the value “N” be set to 0

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

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

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

	1	0	sum
1	q	r	$q + r$
0	s	t	$s + t$
sum	$q + s$	$r + t$	p

$$d(i, j) = \frac{r + s}{q + r + s}$$

Nominal Variables

- **A generalization of the binary variable in that it can take more than 2 states**, e.g. color: {red, yellow, blue, green}
 - They only tell us to which of several unordered categories a data object belongs to. In mathematical terms, if X and Y are two categories, we can tell that $X \neq Y$, but not whether $X > Y$ or $X < Y$.
 - **Nominal variables must be transformed for dissimilarity measure.**
- **How to avoiding spurious information?**
 - E.g., If we number ice cream flavors 1 though 28, it will appear that flavors 5 and 6 are closely related while flavors 1 and 28 are far apart, but these are not necessarily true.

Convert a set of nominal variables into $d(i, j)$

- **Method 1:** Simple matching

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

Where p is the total # of nominal variables, and m is the # of nominal variables matching the object pair i and j .

- **Method 2:** Use a large number of binary variables
 - creating a new binary variable for each of the possible nominal states of each normal variable.

Ordinal Variables

- An ordinal variable resembles a nominal variable, except that the M states of the ordinal value are ordered in a meaningful sequence
 - An ordinal variable can be discrete or continuous
- In some applications, order is important, e.g., rank
 - Ranking: $r_{if} \in \{1, \dots, M_f\}$
 - E.g., For object i
 - Attribute: f is Medal, $r_{iMedle} \in$
 - Ranking: $\{\text{gold, silver, bronze}\}$, or $\{1^{\text{st}}, 2^{\text{nd}}, 3^{\text{rd}}\}$

Convert each ordinal variable into interval-scaled in [0,1]

1. Convert an ordinal value by its rank $r_{if} \in \{1, \dots, M_f\}$

E.g., For the set {gold, silver, bronze}: gold=1, silver=2, bronze=3.

2. Map each value into an interval value of [0.0, 1.0] by the following formula:

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

E.g., $Z_{1_Medal} = 0$, by $(1-1)/(3-1)$,
 $Z_{2_Medal} = 0.5$, by $(2-1)/3-1)$,
 $Z_{3_Medal} = 1$, by $(3-1)/3-1)$.

3. Compute the dissimilarity using a method for interval-scaled variables.

Attributes of Mixed Types

- **A database may contain all the major types of attributes**
 - i.e. symmetric binary, asymmetric binary, nominal, ordinal, interval, etc
- A weighted formula to combine their effects:

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

Where $d(i, j)$ is the dissimilarity between two objects i and j ; $\delta_{ij}^{(f)}$ is a indicator having value of 1 or 0 for variable f ; and $d_{ij}^{(f)}$ is the difference between f 's two values (x_{if} and x_{jf}).

- If f is interval-based:
 - use the normalized distance, $d_{ij}^{(f)} = |x_{if} - x_{jf}| / (\max_{hf} - \min_{hf})$, [0.0-1.0]
- If f is binary or categorical:
 - $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
- If f is ordinal:
 - compute ranks r_{if} and
 - and treat z_{if} as interval-scaled

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

Example:

Customer data set

Credit number	Age	Income	Credit	Car owner	House owner	Region	Car magazine	House	Sports	Music	Comic
23003	20	18.5	17.8	0	0	1	1	1	0	1	1
23009	25	36.0	26.6	1	0	1	0	0	0	0	1

Example: $d(2003, 2009) = ?$

Customer dataset

Credit number	Age	Income	Credit	Car owner	House owner	Region	Car magazine	House	Sports	Music	Comic
23003	20	18.5	17.8	0	0	1	1	1	0	1	1
23009	25	36.0	26.6	1	0	1	0	0	0	0	1

Numerical

Binary

```
max_Age = 91, min_Age = 11;  
max_Incomd = 120.0, min_Income = 20.0;  
max_Credit = 250.0, min_Credit = 40.0;
```

Example: $d(2003, 2009) = ?$

Customer dataset

Credit number	Age	Income	Credit	Car owner	House owner	Region	Car magazine	House	Sports	Music	Comic
23003	20	18.5	17.8	0	0	1	1	1	0	1	1
23009	25	36.0	26.6	1	0	1	0	0	0	0	1

min_Age = 11, max_Age = 91; 5/80

min_Income = 20.0, max_Income = 120.0; 17.5/100

min_Credit = 10.0, max_Credit = 30.0; 8.8/20

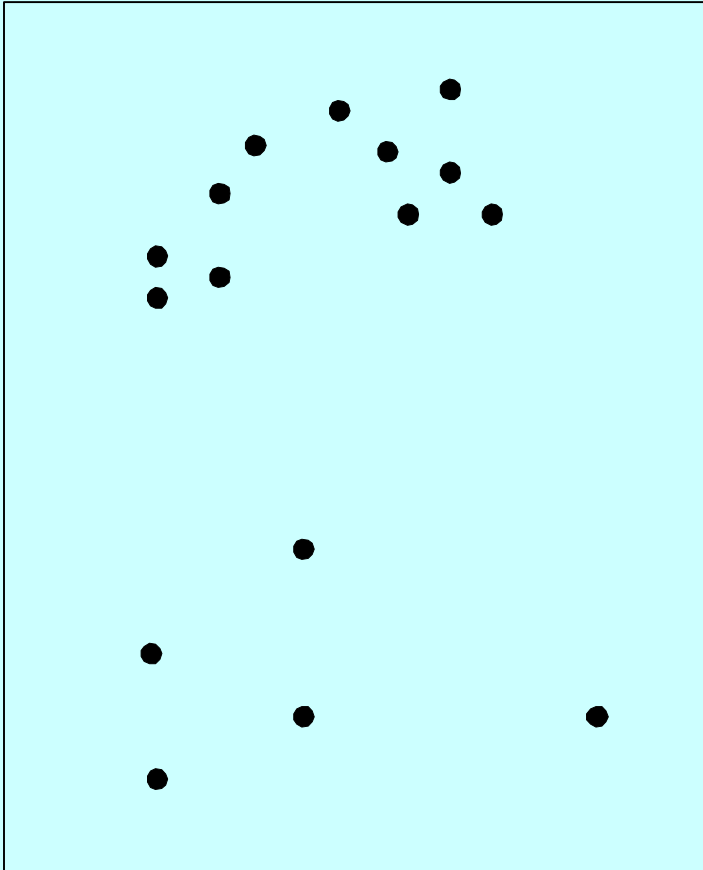
?

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

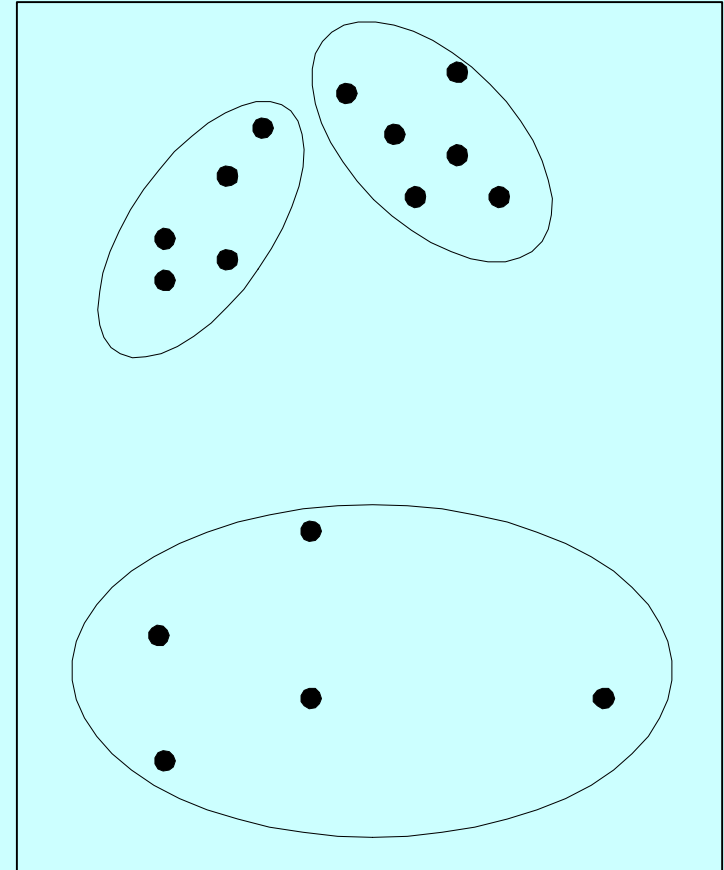
Two Major Types of Clustering

- **Partition based Clustering**
 - Divide data set 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
 - Two strategy: agglomerative and divisive

Partitional Clustering

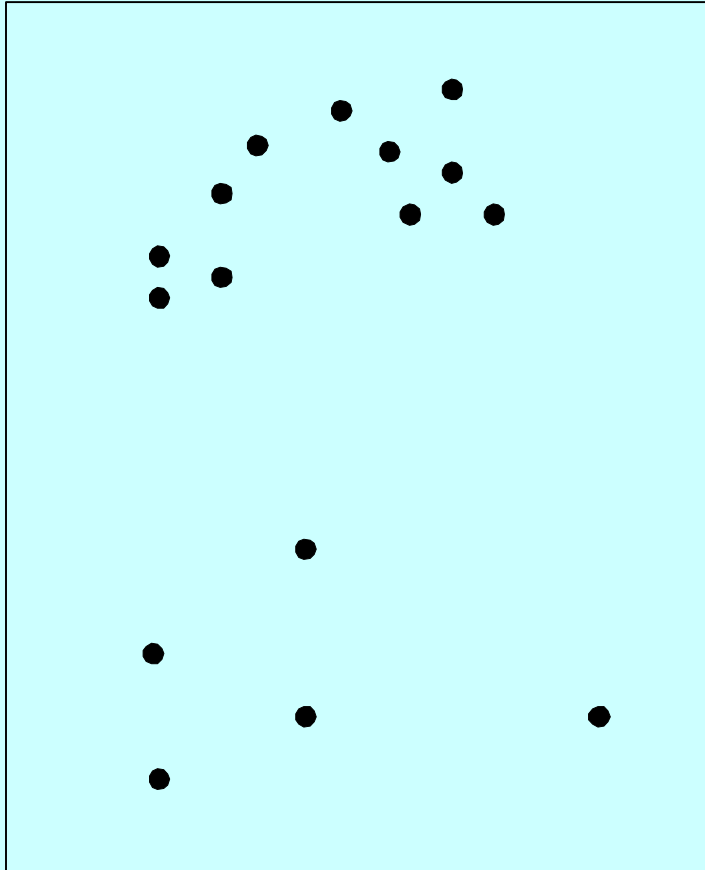


Original Points

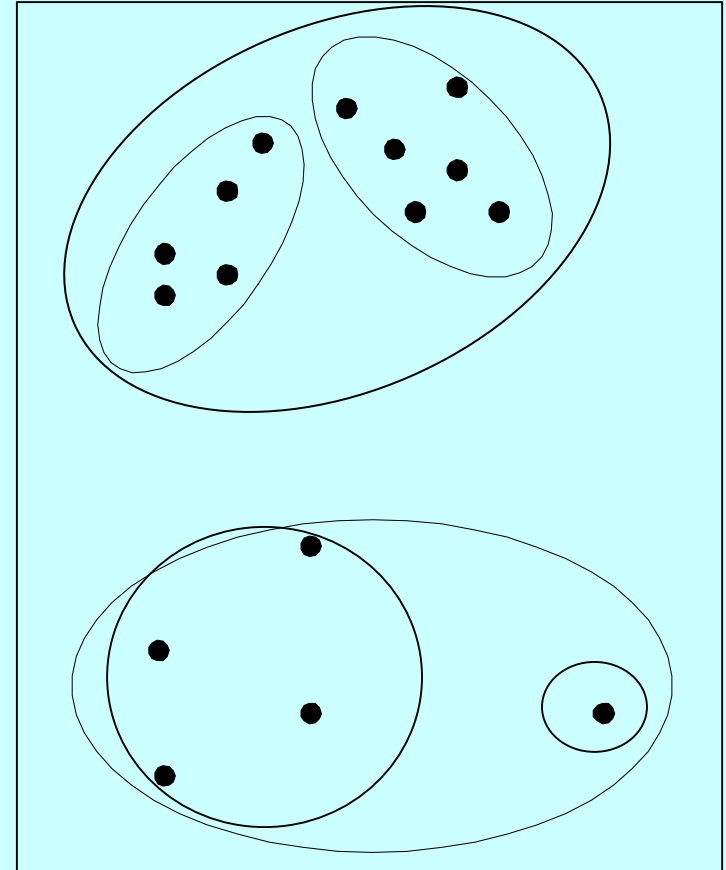


Resulted flat clusters

Hierarchical Clustering



Original Points



**Resulted Hierarchical
clusters**

Partition-based Approach

- It is a single level clustering technique: data partition without hierarchy
- It creates clusters in one stage as opposed to several stages
- Since only one set of clusters is output, the user normally has to input the desired number of clusters, k

General Partitioning Concept

- **Idea:** construct a partition of a data set D of n objects into a set of K clusters
- **Methods:**

Given an integer 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

Review Questions

1. What are the main differences between Classification and Clustering DM (list 3 from different perspectives)?
2. Provide two application examples of clustering DM, explain how the DM result may be used for supporting business decision making.
3. What are the general criteria for judging quality of clustering DM results?
4. What data attribute types can be directly applied for clustering mining? How to prepare your data with various attribute types for clustering?
5. What are the basic data structures for clustering mining?
6. How to calculate dissimilarity of object pairs for a dataset with mixed attribute types for clustering?