

ข้อมูลที่ใช้ในวิชานี้ จะเป็นข้อมูลที่อยู่ในรูปของเมทริกซ์ คือตารางที่มีตัวเลขเรียงกันอยู่ เรียกแถว กับหลัก

1	1	12	2	5
2	2	11	7	2
1	1	15	9	3
0	0	10	1	-3
-1	-1	20	12	-2
1	1	19	6	-5

1 มิติ

2 มิติ

		1	12	2	5
	1	2	11	7	2
1	2	1	15	9	3
2	1	0	10	1	-3
1	0	-1	20	12	-2
0	-1	1	19	6	-5
-1	1		19	6	-5
1			19	6	-5

3 มิติ

	1	12	2	5	
1	2	11	7	2	
1	2	1	15	9	3
2	1	0	10	1	-3
1	0	-1	20	12	-2
0	-1	1	19	6	-5
1	1	19	6	-5	
1	1	19	6	-5	

	1	12	2	5	
1	2	11	7	2	
1	2	1	15	9	3
2	1	0	10	1	-3
1	0	-1	20	12	-2
0	-1	1	19	6	-5
1	1	19	6	-5	
1	1	19	6	-5	

	1	12	2	5	
1	2	11	7	2	
1	2	1	15	9	3
2	1	0	10	1	-3
1	0	-1	20	12	-2
0	-1	1	19	6	-5
1	1	19	6	-5	
1	1	19	6	-5	

แนวนอน Record คือ ข้อมูลแต่ละจุด , แนวตั้ง Attribute คือ คุณสมบัติใช้อธิบายข้อมูลแต่ละจุด

4 มิติ

Types of Data Sets: (1) Record Data

- Relational records
 - Relational tables, highly structured
- Data matrix, e.g., numerical matrix, crosstabs

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove	12.00	4.00	1.00	248.00	251.00	
Active Outdoors Igloo Glove		10.00	6.00		323.00	339.00
Active Crochet Glove	3.00	6.00	8.00		132.00	149.00
Active Igloo Glove		2.00			143.00	143.00
Triumph Pro Helmet	3.00	1.00	7.00		323.00	344.00
Triumph Youth Helmet		7.00	22.00		474.00	495.00
Altrona Adult Helmet	8.00	6.00	7.00	2.00	291.00	294.00
Altrona Youth Helmet	1.00				76.00	77.00
Total	34.00	43.00	54.00	3.00	1,973.00	2,084.00

Person_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rome

Car_ID	Model	Year	Value	Person_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Porsche	1993	500	3
104	Ferrari	2005	150000	4
105	Bentley	1998	2000	3
106	Bentley	2001	7000	3
107	Smart	1999	2000	2

- Transaction data

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

	Wine	Whisky	Tea	Coffee	Icecream	Chocolate	Pastry	Soft drink	Beer	Wine	Whisky	Tea	Coffee	Icecream	Chocolate	Pastry	Soft drink	Beer
Document 1	3	0	5	0	2	6	0	2	0	2								
Document 2	0	7	0	2	1	0	0	3	0	0								
Document 3	0	1	0	0	1	2	2	0	3	0								

- Document data: Term-frequency vector (matrix) of text documents

7

Relational records มีหลายตารางที่มีความสัมพันธ์กัน

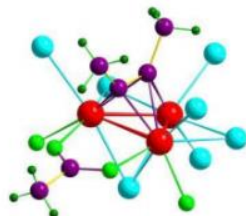
Crosstabs ต้องอ่านค่าทั้งสองฝั่ง

Transaction Data หารูปแบบของข้อมูล

Term – frequency ตารางที่ใช้สรุปข้อมูลที่เป็นข้อความ ให้อยู่ในรูปที่คอมพิวเตอร์สามารถแปลผลได้

Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web



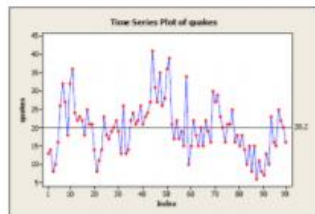
- Molecular Structures
- Social or information networks



ข้อมูลที่เป็นกราฟ จะเป็นการบอกว่า แต่ละจุด เชื่อมโยงกับอะไร มีการเชื่อมต่อกันอย่างไร

Types of Data Sets: (3) Ordered Data

- Video data: sequence of images
- Temporal data: time-series

[illegible]

- Sequential Data: transaction sequences
- Genetic sequence data

6

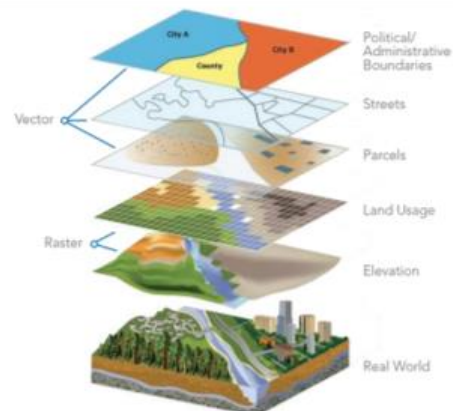
เป็นข้อมูลที่มีเวลาเข้ามาเกี่ยวข้อง แปลงข้อมูลเป็นตัวเลข Time – Series Sequential ข้อมูลที่ไม่สามารถสลับตำแหน่งได้

Types of Data Sets: (4) Spatial, image and multimedia Data

- ❑ Spatial data: maps



- ❑ Image data:
- ❑ Video data:



เป็นข้อมูลที่มีเวลาเข้ามาเกี่ยวข้องเช่นกัน

Video Data คือ การเอารูปหลายๆรูปมาซ้อนๆกัน ทำให้เปรียบเสมือนว่าเคลื่อนไหวได้

Image Data คือ มีด้านกว้าง และด้านยาว กำหนดพิกัดเป็น x, y โดยแต่ละตัวจะมีสีบอกจุด

Spatial Data คือ เช่น แผนที่ จะมีการกำหนดจุดพิกัด x, y โดยจะกำหนดสีให้แต่ละพิกัด

Important Characteristics of Structured Data

- Dimensionality
 - Curse of dimensionality
- Sparsity
 - Only presence counts
- Resolution
 - Patterns depend on the scale
- Distribution
 - Centrality and dispersion

8

คุณลักษณะสำคัญของข้อมูล ได้แก่

ข้อมูลมี Dimension มีกี่ Dimension

สนใจแค่จุดที่มีข้อมูล ตรงที่ไม่มีข้อมูลจะไม่สนใจ

ความแออัดในการเก็บข้อมูล

การกระจายตัวของข้อมูล วัดค่ากลางของมูล

Data Objects

- ❑ Data sets are made up of data objects
- ❑ A **data object** represents an entity
- ❑ Examples:
 - ❑ sales database: customers, store items, sales
 - ❑ medical database: patients, treatments
 - ❑ university database: students, professors, courses
- ❑ Also called *samples* , *examples*, *instances*, *data points*, *objects*, *tuples*
- ❑ Data objects are described by **attributes**
- ❑ Database rows → data objects; columns → attributes

9

Data Sets ประกอบไปด้วย Data Object หลายๆ อันมารวมกัน Data Set คือ กลุ่มของข้อมูลหลายๆ ตัวมารวมกัน

Data Object หมายถึง ข้อมูลแต่ละตัวประกอบด้วย Entity ของแต่ละอัน

สำคัญ* Also called sample, examples, instances, data points, objects, tuples Record

มีความหมายเดียวกันทั้งหมด คือ ข้อมูลที่เป็น Data 1 จุดที่อยู่ในแนวตั้ง Data 1 จุด คือ Data point

ข้อมูลจะถูกอธิบายด้วย Attributes

Database ที่จะนำมาทำคือ Rows คือ Data Object, Column คือ Attributes

Attributes

- ❑ **Attribute (or dimensions, features, variables)**
 - ❑ A data field, representing a characteristic or feature of a data object.
 - ❑ *E.g., customer_ID, name, address*
- ❑ **Types:**
 - ❑ Nominal (e.g., red, blue)
 - ❑ Binary (e.g., {true, false})
 - ❑ Ordinal (e.g., {freshman, sophomore, junior, senior})
 - ❑ Numeric: quantitative
 - ❑ Interval-scaled: 100°C is interval scales
 - ❑ Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50°K
- ❑ Q1: Is student ID a nominal, ordinal, or interval-scaled data?
- ❑ Q2: What about eye color? Or color in the color spectrum of physics?

10

คือ คุณสมบัติที่ใช้อธิบายข้อมูลแต่ละตัว อาจจะเรียกว่า Dimension, Features, Variables, field

Attribute Types

- ❑ **Nominal:** categories, states, or “names of things”
 - ❑ *Hair_color = {auburn, black, blond, brown, grey, red, white}*
 - ❑ marital status, occupation, ID numbers, zip codes
- ❑ **Binary**
 - ❑ Nominal attribute with only 2 states (0 and 1)
 - ❑ Symmetric binary: both outcomes equally important
 - ❑ e.g., gender
 - ❑ Asymmetric binary: outcomes not equally important.
 - ❑ e.g., medical test (positive vs. negative)
 - ❑ Convention: assign 1 to most important outcome (e.g., HIV positive)
- ❑ **Ordinal**
 - ❑ Values have a meaningful order (ranking) but magnitude between successive values is not known
 - ❑ *Size = {small, medium, large}, grades, army rankings*

11

ชนิดของ Attribute

Nominal ข้อมูลที่เป็นการบอกกลุ่ม บอกสถานะ เป็นชื่อของสิ่งของ เช่น สีผม, สถานะการแต่งงาน, อาชีพ, ID

Binary คือ ข้อมูล Nominal เพียงแต่มี 2 สถานะ เช่น Yes/No

Symmetric Binary คือสมมาตรกัน เช่น ชอบกินไก่หรือเป็ด ถนัดซ้ายหรือขวา

Asymmetric Binary คือไม่สมมาตรกัน เช่น เป็นเอดส์ไหม

Ordinal คือ ข้อมูลที่มีความหมายสามารถนำมาเรียงลำดับได้ แต่จะไม่ว่ามีค่าต่างกันมากน้อยแค่ไหน เช่น ขนาดใหญ่ ขนาดกลาง ขนาดเล็ก เกรด

Numeric Attribute Types

- Quantity (integer or real-valued)

- Interval

- Measured on a scale of **equal-sized units**

- Values have order

- E.g., *temperature in C° or F°, calendar dates*

- No true zero-point

- Ratio

- Inherent **zero-point**

- We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).

- e.g., *temperature in Kelvin, length, counts, monetary quantities*

12

Interval ข้อมูลที่ไม่มี 0 แท้ เช่น อุณหภูมิ

Ratio ข้อมูลที่มี 0 แท้ เช่น ความยาว เงิน

Discrete vs. Continuous Attributes

▣ Discrete Attribute

- ▣ Has only a finite or countably infinite set of values
 - ▣ E.g., zip codes, profession, or the set of words in a collection of documents
- ▣ Sometimes, represented as integer variables
- ▣ Note: Binary attributes are a special case of discrete attributes

▣ Continuous Attribute

- ▣ Has real numbers as attribute values
 - ▣ E.g., temperature, height, or weight
- ▣ Practically, real values can only be measured and represented using a finite number of digits
- ▣ Continuous attributes are typically represented as floating-point variables

13

การแบ่ง Attributes อีกแบบ

Discrete Attribute คือ ข้อมูลที่ไม่ต่อเนื่องกัน เช่น 1,2,3 ซึ่งไม่มีค่าอยู่ระหว่าง 1-2 เช่น อาชีพ ครู กับ พยาบาล ก็จะไม่มียาชีพอื่นอยู่ระหว่าง สองค่านี้

Continuous Attribute คือ ข้อมูลที่ต่อเนื่องกัน ส่วนใหญ่แทนด้วยจำนวนจริง เช่น ส่วนสูง 180,180.5,181

Basic Statistical Descriptions of Data

Motivation

- To better understand the data: central tendency, variation and spread

Data dispersion characteristics

- Median, max, min, quantiles, outliers, variance, ...

Numerical dimensions correspond to sorted intervals

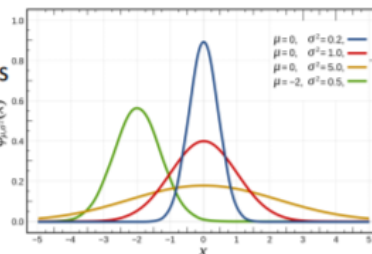
- Data dispersion:

- Analyzed with multiple granularities of precision

- Boxplot or quantile analysis on sorted intervals

Dispersion analysis on computed measures

- Folding measures into numerical dimensions
- Boxplot or quantile analysis on the transformed cube



15

เป็นการดูค่าทางสถิติ โดยใช้ค่ากลางของข้อมูลแทน ค่ากลางทางสถิติมีทั้งหมด 3 ชนิด คือ ค่าเฉลี่ย ค่ามัธยฐาน ค่าฐานนิยม

Measuring the Central Tendency: (1) Mean

Mean (algebraic measure) (sample vs. population):

Note: n is sample size and N is population size.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Trimmed mean:

- Chopping extreme values (e.g., Olympics gymnastics score computation)

16