Chapter 2 Getting to Know Your Data

Data คืออะไร มีกี่ชนิด จัดการหรือวิเคราะห์เกี่ยวข้องอะไรกับข้อมูลได้บ้าง !!

หน่วย Data

1			
2			
1			
0			
-1			
1			

10

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

20

30

40

	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Record 1				
⋮				
Record 6				

Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Types of Data Sets: (1) Record Data

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

Sailor	Name	Birthdate	Prof_Name	City
0	Miller		Paul	London
1	Ortega		Alvaro	Valencia
2	Huber		Urs	Zurich
3	Blanc		Gaston	Paris
4	Bertolini		Fabrizio	Rom

— no relation

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

	China	England	France	Japan	USA	Total
Reflex Outdoor Crochet Glove		12.00	4.00	5.00	240.00	257.00
Reflex Outdoor Lycra Glove		10.00	6.00		283.00	299.00
Reflex Crochet Glove	3.00	6.00	8.00		132.00	149.00
Reflex Lycra Glove		3.00			142.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		332.00	344.00
Triumph Vertigo Helmet		2.00	22.00		474.00	498.00
Xtreme Adult Helmet	8.00	8.00	1.00	2.00	256.00	275.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	2.00	1,472.00	1,586.00

- 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

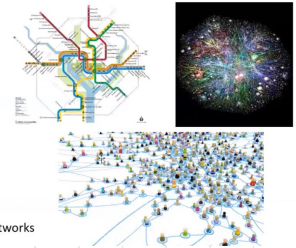
	team	coach	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

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

Normalization → เพื่อไม่ให้ตารางซ้ำกัน

Types of Data Sets: (2) Graphs and Networks

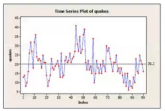
- Transportation network
- World Wide Web
- Molecular Structures
- Social or information networks



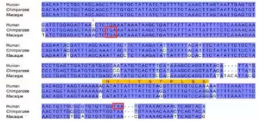
กราฟเชื่อมโยง

Types of Data Sets: (3) Ordered Data

- Video data: sequence of images ข้อมูลที่มีเวลาเข้ามาเกี่ยวข้อง → วิดีโอ
- Temporal data: time-series



- Sequential Data: transaction sequences
  สลับลำดับไม่ได้ เช่น DNA

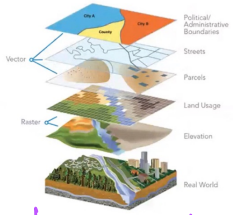


- Genetic sequence data

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

ข้อมูลเชิงพื้นที่

- Spatial data: maps
- Image data:



- Video data: Spatio-temporal ** ข้อมูลที่มีเวลาเข้ามาเกี่ยวข้อง

ข้อมูลเชิงวิดีโอ

Important Characteristics of Structured Data

คุณสมบัติสำคัญ

- Dimensionality → มิติ
  - Curse of dimensionality
- Sparsity → สนใจตรงที่มีข้อมูลเท่านั้น!! เพราะบางชนิดมี 0 เยอะ/ช่องว่าง
  - Only presence counts
- Resolution → ความละเอียดในการเก็บ เช่น ภาพ กว้าง x ยาว กี่พิกเซล
  - Patterns depend on the scale
- Distribution → การกระจายของข้อมูล
  - Centrality and dispersion

Data Objects

ข้อมูลที่ได้มา

ประกอบด้วย Data Objects หลายๆตัวมารวมกัน

- Data sets are made up of data objects
- A **data object** represents an entity ข้อมูลแต่ละตัว
- 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* ข้อมูล 1 ชุด
- Data objects are described by **attributes** ข้อมูลถูกอธิบายด้วย attributes
- Database rows → data objects; columns → attributes

แนวตั้ง

Attributes → คุณสมบัติที่อธิบายข้อมูลแต่ละตัว

ชื่อเรียก

- **Attribute** (or **dimensions, features, variables**)
  - A data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types: ชนิด
  - Nominal (e.g., red, blue)
  - Binary (e.g., {true, false}) มีแค่ 2 ค่า
  - Ordinal (e.g., {freshman, sophomore, junior, senior}) ข้อมูลที่เรียงลำดับ
  - Numeric: quantitative ข้อมูลตัวเลข บวกลบคูณหารแล้วมีความหมาย
    - Interval-scaled: 100°C is interval scales +, -, x, ÷ ได้
    - 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?

ชนิดของ Attribute Types

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

ตำแหน่ง / ยศ

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval** 0 ไม่แท้ คือ มีความหมาย
  - 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** 0 แท้ คือ ไม่มีความหมาย
  - Inherent **zero-point** มีตัวเลข 0 แท่ง = ไม่มีตัวเลข
  - 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˚). 0 K = ไม่มีอุณหภูมิ
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

0 ฿ = ไม่มีเงิน

Discrete vs. Continuous Attributes

ไม่ต่อเนื่อง

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

Chapter 2. Getting to Know Your Data

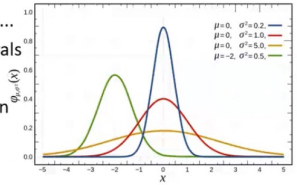
- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Basic Statistical Descriptions of Data

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

ค่ากลาง เปลี่ยนจากค่ากลางไปค่าไหน

- 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



ฐานนิยม

Measuring the Central Tendency: (1) Mean

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

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

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

- 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)