Chapter 2

2.1

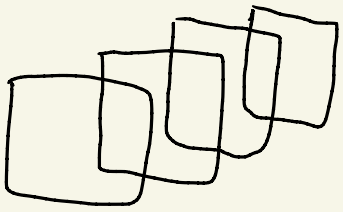
Dataที่ใช้เป็นเมทริกซ์

1
2
1
0
-1
1

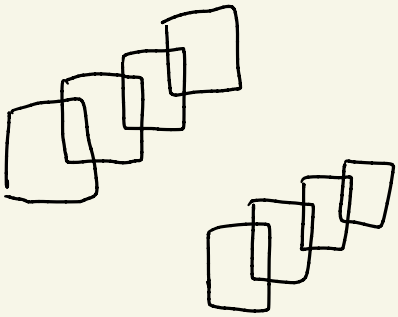
1D

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

2D



3D



4D

ที่ Attribute ชนิดเดียวกัน = จุดเดียวกัน

Attribute 1 Attribute 2 ... Attributes

Record 1
Record 2
⋮
Record b

Chapter 2

รูปแบบตาราง

Types of Data Sets: (1) Record Data

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

Crosstab มี Attribute 2 ตัว

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
Influx Crochet Glove	3.00	6.00	8.00		132.00	149.00
Influx Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

Person:

Pers_ID	Surname	First_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:

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

- 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	pla y	ball	score	game	wi n	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

นับจำนวนคำ เพื่อแปลงเป็นตัวเลขให้คอมเข้าใจได้

→ term-frequency matrix → Text

นับจำนวนเพื่อบอกถึงบทความ

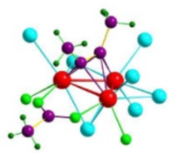
- Document data: Term-frequency vector (matrix) of text documents

Normalization → เพื่อไม่ให้ตารางซ้ำซ้อนกัน

Types of Data Sets: (2) Graphs and Networks

กราฟ

- Transportation network
- World Wide Web



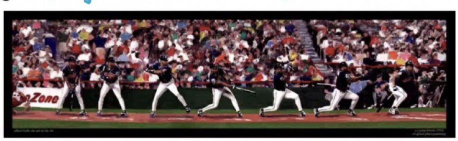
- Molecular Structures

- Social or information networks

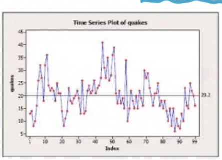
5

Types of Data Sets: (3) Ordered Data

- Video data: sequence of images ข้อมูลที่เวลาเข้ามาเกี่ยวข้อง



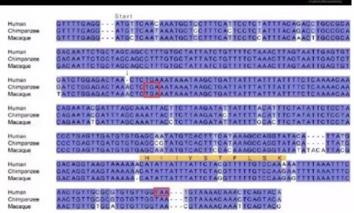
- Temporal data: time-series



- Sequential Data: transaction sequences

↳ DNA สลับลำดับไม่ได้

- Genetic sequence data



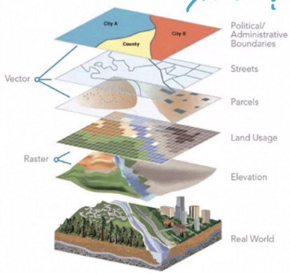
Types of Data Sets: (4) Spatial, image and multimedia Data

เชิงพื้นที่

- Spatial data: maps

→ แต่ละจุด แทน X, Y

- Image data: รูปภาพ



- Video data: spatio – temporal

เชิงพื้นที่ เวลา

ค่าที่เอาไว้เรียกข้อมูลที่เป็น Vedio spatio-temporal มีทั้งพื้นที่ทั้งเวลา

คุณสมบัติสำคัญที่ต้องรู้

Important Characteristics of Structured Data

- Dimensionality → มิติ
  - Curse of dimensionality
- Sparsity → ส่วนใหญ่แทบจะที่มีข้อมูล Ex: ตามองข้อมูลที่ไม่เอา
  - Only presence counts
- Resolution → ความละเอียดในการเก็บ
  - Patterns depend on the scale กว้าง x ยาว กี่ Pixels
- Distribution → การกระจายตัวของdata
  - Centrality and dispersion

↳ Ex ค่ากลางของ Data โดยค่าเฉลี่ย

Data Objects

- Data sets are made up of data objects. → data ที่รวบรวมมา = data set
- A **data object** represents an entity. แทนของ ↳ ข้อมูลแต่ละตัว
- Examples:
  - sales database: customers, store items, sales ← ตาม sell เก็บข้อมูลของ
  - medical database: patients, treatments
  - university database: students, professors, courses ได้ศัพท์ใหม่ ไปหาต่อ
- Also called *samples , examples, instances, data points, objects, tuples*. Data 1 จุด
- Data objects are described by **attributes**. แนวตั้ง
- Database rows -> data objects; columns ->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?

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

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*

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

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
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- 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

Measuring the Central Tendency

- 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:
  - Trimmed mean: chopping extreme values

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

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

  $$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}}\right) width$$

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

  $$mean - mode = 3 \times (mean - median)$$