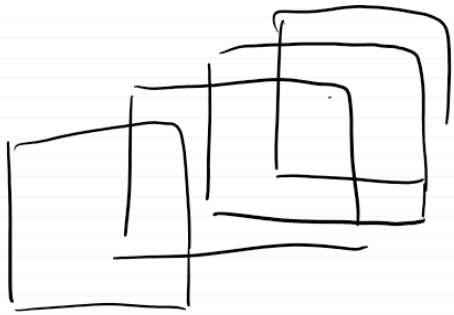
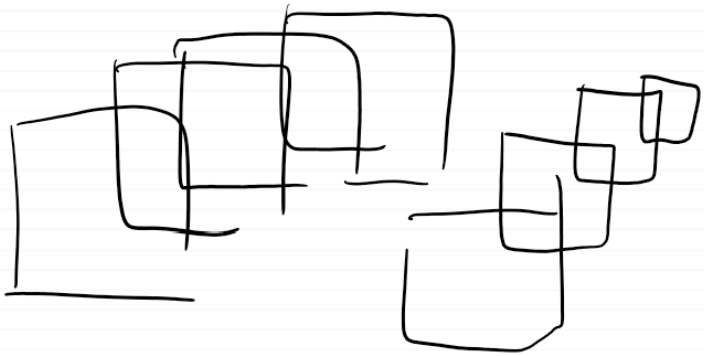
Data ที่อยู่เป็น เมทริกซ์

1D

2D



3D



4D

ถ้า Attribute เหมือนกัน => คุณสมบัติเดียวกัน

Atribute1 Atribute2 ... Attributen

Record1

Record2

⋮

Recode6

CS 412 Intro. to Data Mining

Chapter 2. Getting to Know Your Data

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

มีหลายตารางที่ สัมพันธ์กัน

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

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

Crosstab มีAttribute 2 ฝั่ง

	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
Inflex Crochet Glove	3.00	6.00	8.00		132.00	149.00
Inflex 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	0.00	7.00	2.00	251.00	274.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

- 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	play	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

term-frequency matrix ⇒ ทำกับ Text เพื่อบ่งบอกถึงบทความ

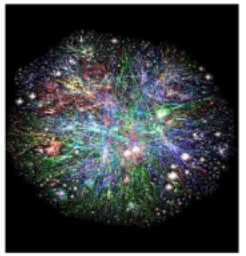
- Document data: Term-frequency vector (matrix) of text documents

Normalization → เลขไม่ให้ตารางซ้ำซ้อนกัน.

Types of Data Sets: (2) Graphs and Networks

กราฟ

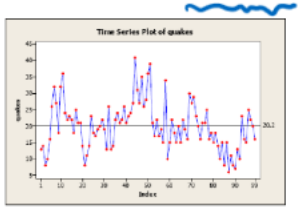
- Transportation network
- World Wide Web



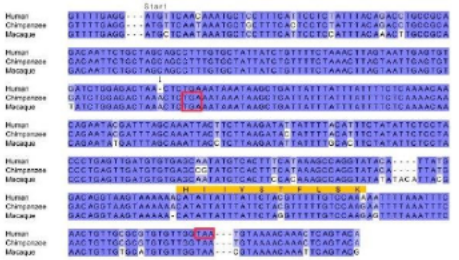
- Molecular Structures
- Social or information networks

Types of Data Sets: (3) Ordered Data

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



- Sequential Data: transaction sequences
  - ↳ DNA สลับลำดับไม่ได้

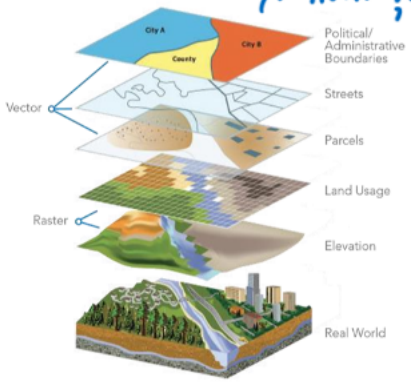


- Genetic sequence data

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

เชิงพื้นที่

- Spatial data: maps



แต่ละจุดตาม X, Y

- Image data: รูปภาพ
- Video data: Spatio - temporal
  - เชิงพท. เวลา

Important Characteristics of Structured Data

คุณสมบัติ สำคัญ ที่ต้องรู้

- Dimensionality → มิติ
  - Curse of dimensionality
- Sparsity → สนใจแต่ ตรงที่มีข้อมูล Ex ตารางที่มีช่องว่าง ไม่เอา
  - Only presence counts
- Resolution → ความละเอียดในการเก็บ
  - Patterns depend on the scale กว้าง ยาว กี่ Pixels
- Distribution → การกระจายตัว ของ ข้อมูล
  - Centrality and dispersion
    - ↳ Ex ค่ากลางของ Data โดยหาค่าเฉลี่ย

Data Objects

- Data sets are made up of data objects → data ที่รวมกัน = data set
- A **data object** represents an entity แนวนอน
  - ข้อมูลเก่าแต่ละตัว
- Examples:
  - sales database: customers, store items, sales ตาราง Sell เก็บ ข้อมูลของ
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples* Data 1จุด
- Data objects are described by **attributes** แนวตั้ง
- Database rows → data objects; columns → attributes

Attributes คุณสมบัติ ที่อธิบายแต่ละตัว

- **Attribute (**or **dimensions, features, variables)**
  - A data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types: ชนิดของข้อมูล
  - Nominal (e.g., red, blue) หมวดหมู่ / ชื่อของกลุ่ม/ชนิด ⇒ ☺,☹,☠,☹ ไม่ได้
  - Binary (e.g., {true, false}) มีแค่ 2 ค่า
  - Ordinal (e.g., {freshman, sophomore, junior, senior}) ข้อมูลเรียงลำดับ
  - Numeric: quantitative ข้อมูลตัวเลข
    - Interval-scaled: 100°C is interval scales บวก,ลบ,คูณ,หารได้ มีความหมาย
    - 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?

สีตา เป็นข้อมูลชนิดไหน → สีแดง – สีม่วง → Numeric ใส่ค่าสี

35%

ชนิดของ Attribute Types

- **Nominal:** categories, states, or "names of things" บอกกลุ่ม, สถานะ, ชื่อของสิ่งของ
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*} สีผม
  - marital status, occupation, ID numbers, zip codes
- **Binary** สถานะ อาชีพ
  - Nominal attribute with only 2 states (0 and 1) มี 2 ค่า
  - Symmetric binary: both outcomes equally important ⇒ สมมาตรกัน (ความสัมพันธ์เท่ากัน)
    - e.g., ~~gender~~ (ชนิดเมีย ชาย, มือขวา) / (ชอบโสด/เปปซี่) / (มอน,เอิน)
  - Asymmetric binary: outcomes not equally important. ⇒ ไม่สมมาตรกัน (ความสัมพันธ์ไม่เท่ากัน)
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal** เรียงลำดับได้ ไม่สามารถรู้ว่าแตกต่างมากน้อยแค่ไหน
  - Values have a meaningful order (ranking) but magnitude between successive values is not known
  - *Size* = {*small, medium, large*}, grades, army rankings

Size, เกรด, ยศทหาร

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval** → ไม่มี 0 แท้ = มีความหมาย
  - Measured on a scale of **equal-sized units**
  - Values have order 0°c = หนาวมาก
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio** → มี 0 แท้ = ไม่มี เงินเดือน 0 แท่ง ← ไม่มี เงินเดือน
  - Inherent **zero-point** เช่น 0 K = ไม่มีอุณหภูมิ
  - 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 dollars = ไม่มีเงิน
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

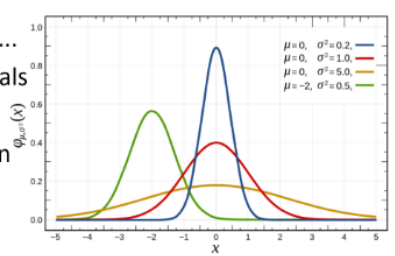
- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**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

มัธยฐาน

Measuring the Central Tendency: (2) Median

- 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

Approximate median

Sum before the median interval

Interval width ($L_2 - L_1$)

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

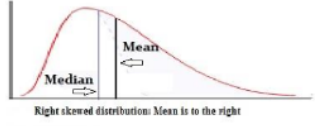
Low interval limit

ฐานนิยม

Measuring the Central Tendency: (3) Mode

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

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



Right skewed distribution: Mean is to the right

- Multi-modal
  - Bimodal
  - Trimodal