CS 412 Intro. to Data Mining

Chapter 2. Getting to Know Your Data

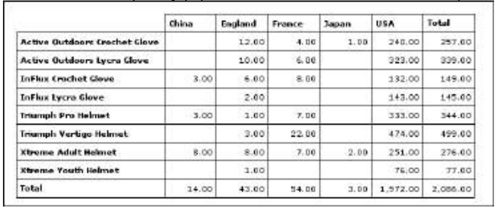
Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

Chapter 2. Getting to Know Your Data

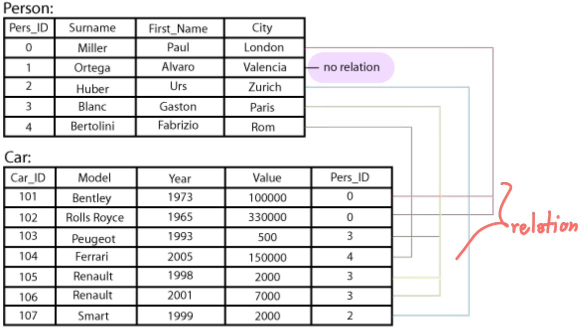
- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Types of Data Sets: (1) Record Data

- Relational records
  - สัมพันธ์ โครงสร้าง
    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	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	1,572.00	2,086.00



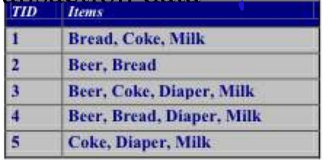
Person:

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

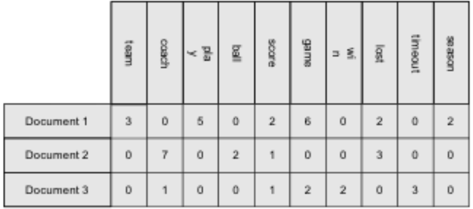
Car:

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 ข้อมูลรายการเปลี่ยนแปลง



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

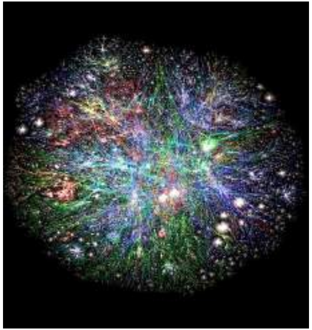


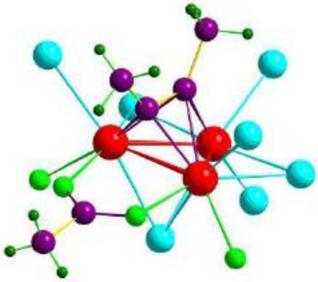
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

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

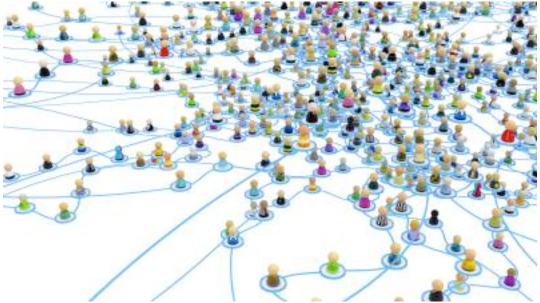
Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web





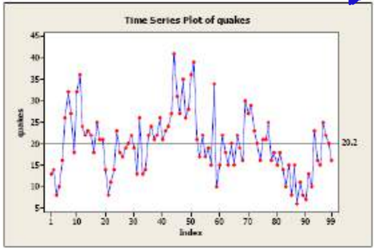
- Molecular Structures
- Social or information networks



Types of Data Sets: (3) Ordered Data

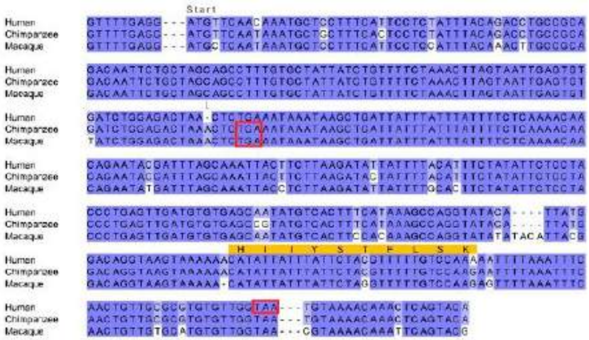
- Video data: sequence of images
  ลำดับ

- Temporal data: time-series
  ข้อมูลเชิงเวลา



- Sequential Data: transaction sequences

- Genetic sequence data



พื้นที่ < กว้าง ยาว

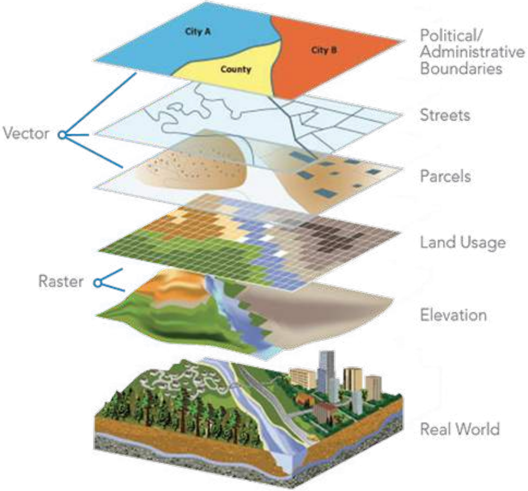
ภาพ

ข้อมูลมัลติมีเดีย

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

ข้อมูลเชิงพื้นที่

- Spatial data: maps



- Image data:

ข้อมูลเชิงภาพ

- Video data:

ข้อมูลเชิงวิดีโอ

Important Characteristics of Structured Data

- Dimensionality → มีกี่ Dimension
    - Curse of dimensionality
- Sparsity → วัดที่แน่นอน สนใจเฉพาะที่มีข้อมูล
    - Only presence counts
- Resolution ความละเอียด
    - Patterns depend on the scale
- Distribution การกระจาย
    - Centrality and dispersion
      ค่ากลาง การกระจายตัว

Data Objects

- Data sets ประกอบ 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 อธิบาย <u>described</u> by **attributes**
- Database rows แถว → data objects; columns คอลัมน์ → attributes

Attributes คุณสมบัติ

- **Attribute** (or **dimensions**, **features**, **variables**)
  - เขตข้อมูล คุณสมบัติ/ลักษณะเฉพาะ ลักษณะเฉพาะ
    A data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - เช่น ชื่อ สี
    Nominal (e.g., red, blue)
  - มีเพียง 2 ค่า ถูก ผิด
    Binary (e.g., {true, false})
  - เรียงลำดับ
    Ordinal (e.g., {freshman, sophomore, junior, senior})
  - ตัวเลข เชิงปริมาณ พวก +,-,x,÷
    Numeric: quantitative
    - Interval-scaled: 100○C is interval scales ไม่มีศูนย์แท้
    - 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

ประเภท ชื่อของสิ่งต่างๆ

- **Nominal:** categories, states, or "names of things"

  สีผม

  - *Hair color* = {*auburn, black, blond, brown, grey, red, white*}

  สถานภาพการสมรส อาชีพ

  - marital status, occupation, ID numbers, zip codes
- **Binary** มี 2 ทางเลือก
  - Nominal attribute with only 2 states (0 and 1)
  - ★ Symmetric binary: both outcomes equally important

    เพศ

    - 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**

  เรียง แต่ ไม่รู้ขนาดระหว่างอันดับ

  - 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)
  จน.เต็ม ค่าจริง
- **Interval** → สามารถนำค่ามาเรียงลำดับ
  - Measured on a scale of **equal-sized units**
  - Values have order บอกค่าความต่างของตัวเลขได้
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point → ค่า 0 ไม่ได้แปลว่าไม่มีข้อมูลแต่หมายถึงข้อมูลมีค่าเท่ากับ 0
- **Ratio** → เรียงลำดับ, หาค่าความต่าง
  - 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

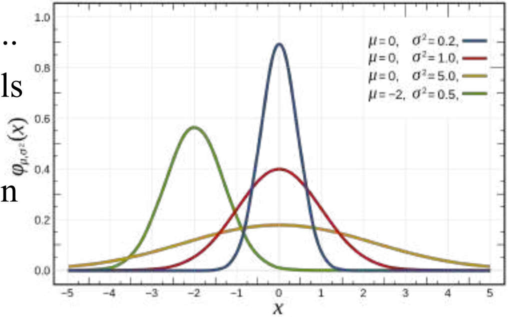
- **Discrete Attribute** ไม่ต่อเนื่อง แบ่งชัดเจน
  - Has only a finite or countably infinite set of values
    - รหัสปณ. อาชีพ
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes *
- **Continuous Attribute** ต่อเนื่อง มีจุดย่อยๆลงไปได้อีกเรื่อยๆ
  - Has real numbers as attribute values
    - อุณหภูมิ ส่วนสูง น้ำหนัก
    - 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

- แรงจูงใจ
  Motivation
  - แนวโน้มส่วนกลาง การเปลี่ยนแปลง การแพร่กระจาย
    To better understand the data: central tendency, variation and spread
- ลักษณะการกระจายข้อมูล
  Data dispersion characteristics
  - ค่ามัธยฐาน มาก น้อย ปริมาณ ค่าผิดปกติ ความแปรปรวน
    Median, max, min, quantiles, outliers, variance, ...
- สอดคล้องกับช่วงเวลาที่จัดเรียง
  Numerical dimensions correspond to sorted intervals
  - การกระจายข้อมูล
    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


$\mu=0, \sigma^2=0.2$
$\mu=0, \sigma^2=1.0$
$\mu=0, \sigma^2=5.0$
$\mu=-2, \sigma^2=0.5$
$\varphi_{\mu,\sigma^2}(x)$
x