CS 412 Intro. to Data Mining

Chapter 2. Getting to Know Your Data

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

DATA

1 มิติ – กว้าง ยาว

2 มิติ – กว้าง xยาว

3 มิติ – 2 มิติซ้อนกัน

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

	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

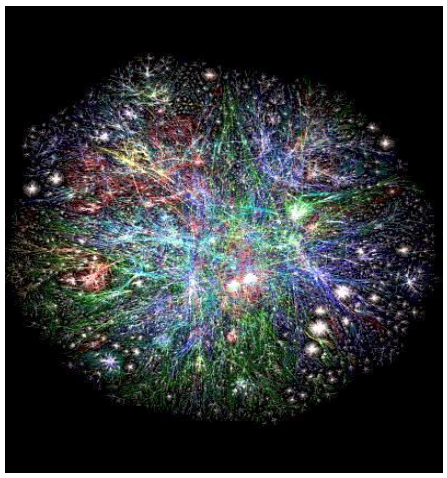
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

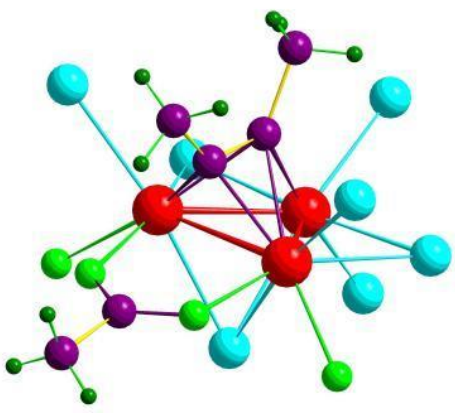
- Document data: Term-frequency vector (matrix) of text documents

Types of Data Sets: (2) Graphs and Networks

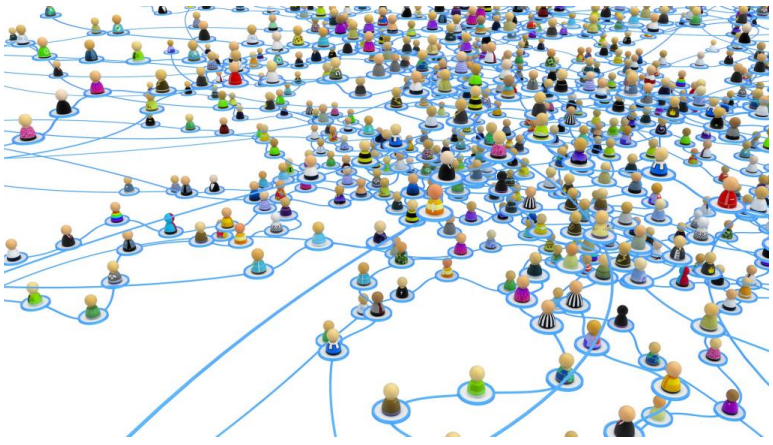
กราฟ

- Transportation network
- World Wide Web





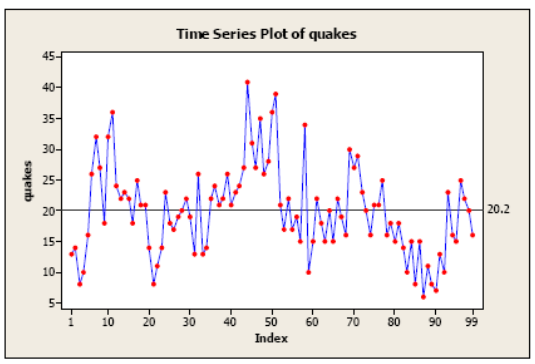
- Molecular Structures
- Social or information networks



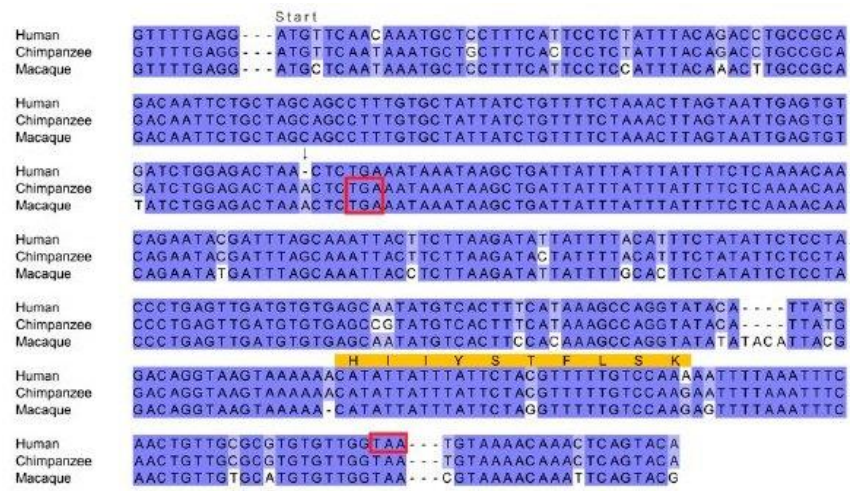
Types of Data Sets: (3) Ordered Data

- Video data: sequence of images

- Temporal data: time-series



- Sequential Data: transaction sequences



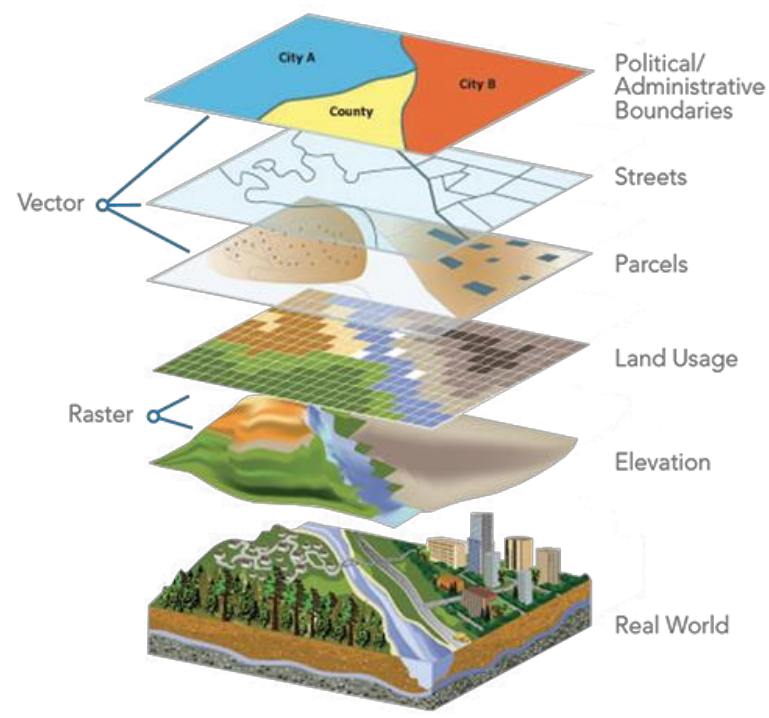
- Genetic sequence data

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

ที่ด้านกว้างและยาว (x, y)

ข้อมูลเชิงพื้นที่

- Spatial data: maps



ข้อมูลเชิงรูปภาพ

- Image data:

- Video data: Spatio-temporal

เชิงเวลา

Important Characteristics of Structured Data

มีกี่ Dimension

- Dimensionality
  - Curse of dimensionality

สนใจเฉพาะตรงที่มีข้อมูล

- Sparsity
  - Only presence counts

เก็บข้อมูลใช่ไม่ใช่

- Resolution
  - Patterns depend on the scale

จัดตำแหน่ง (ต่างไม่ได้)

- Distribution
  - Centrality and dispersion

Data Objects

ประกอบด้วยหลายๆ ชิ้นมูล

- Data sets are made up of data objects

ข้อมูลแต่ละตัว จะมี entity

- 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*

อธิบายด้วย

- Data objects are described by **attributes**
- Database rows → data objects; columns → attributes

Attributes

คุณสมบัติ

- **Attribute (**or **dimensions, features, variables**)
  - A data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - ชื่อของข้อมูลที่ไม่ใช่ตัวเลข
    Nominal (e.g., red, blue)
  - มี 2 ค่า
    Binary (e.g., {true, false})
  - เรียงค่ากัน
    Ordinal (e.g., {freshman, sophomore, junior, senior})
  - +, -, x, ÷ แล้วมีความหมาย
    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?

Numric

Attribute Types

ชนิด

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary** โค้ก, เปปซี่ ใช่,ไม่ใช่ ซ้าย,ขวา

  มี 2 ทางเลือก
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender

    มี 2 ค่าความสำคัญไม่เท่ากัน
  - 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** ค่ามาเรียงกันได้ แต่ไม่รู้ว่าต่างกันเท่าไหร่
  - Values have a meaningful order (ranking) but magnitude between successive values is not known
  - *Size* = {*small, medium, large*}, grades, army rankings

ข้อมูลที่เป็นตัวเลข

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval** ไม่มี 0 แท้ เช่น เกรด, อุณหภูมิ, ความยาว, เงิน

  *ช่วงการวัดจะมีระยะทางเท่ากัน

  - 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** ทุกอย่างที่วัดเป็นตัวเลขได้ เช่น น้ำหนัก ระยะทาง ความสูง อายุ ยอดขาย
  - 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 จำนวนสิ่งของ, จำนวนคน
  - Has only a finite or countably infinite set of values
    - E.g., zip codes (รหัสไปรษณีย์), 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** มีค่าได้ทุกค่าในช่วงที่กำหนด
  - Has real numbers as attribute values ราย ได้, 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
- ด้านสถิติเชิงพื้นฐาน
  Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Basic Statistical Descriptions of Data

แสดงภาพ

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

ศึกษาหลัก การกระจาย ขอบเขตข้อมูล

- Data dispersion characteristics
  - Median, max, min, quantiles, outliers, variance, ...

มัธยฐาน ค่าผิดปกติ ค่าเบี่ยงเบน

- Numerical dimensions correspond to sorted intervals
  - Data dispersion:
    - Analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals

การวิเคราะห์การกระจายของการจัดกลุ่มต่างๆ ได้

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