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
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

เมทริก 1 มิติ ถ้าไม่แต่ความกว้างหรือยาวอย่างเดียว
ส่วนมากจะเรียก matrix

	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Record 1	1	12	2	5
Record 2	2	11	7	2
Record 3	1	15	9	3
Record 4	0	10	1	-3
Record 5	-1	20	12	-2
Record 6	1	19	6	-5

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

หลายตารางที่มีความสัมพันธ์กัน

- 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

จำนวนต่ำ

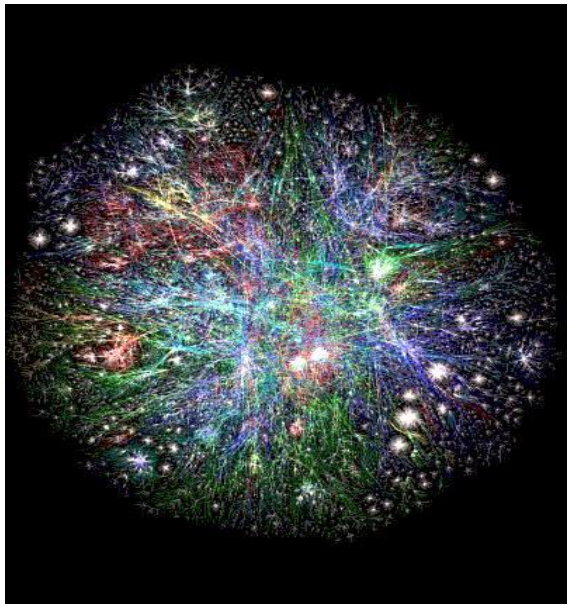
Data ใช้แทนข้อความ

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

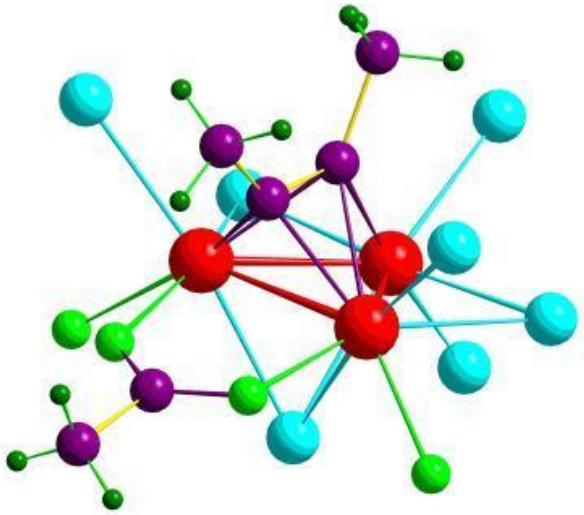
normalization ทำให้ข้อมูลซับซ้อนน้อยลง

Types of Data Sets: (2) Graphs and Networks

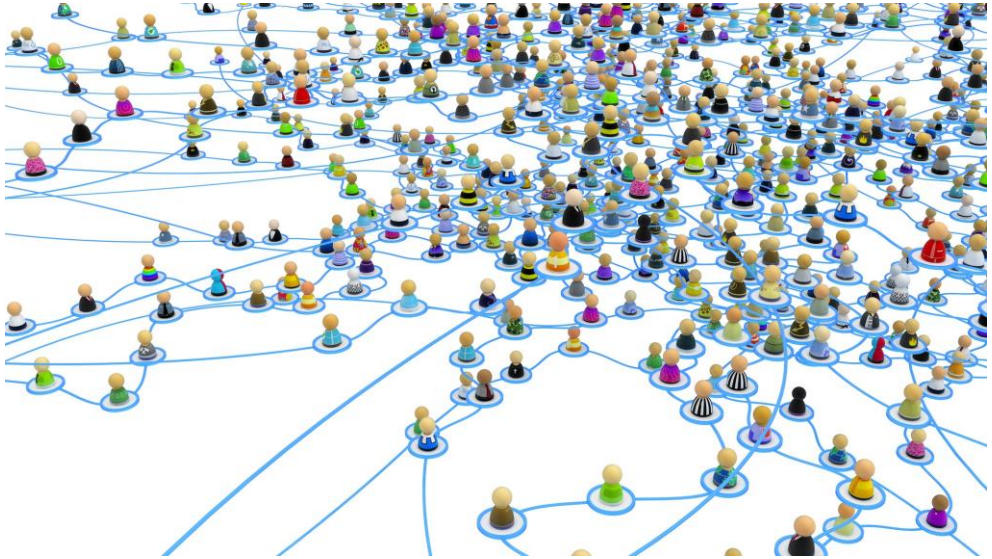
- Transportation network



- World Wide Web



- Molecular Structures



- Social or information networks

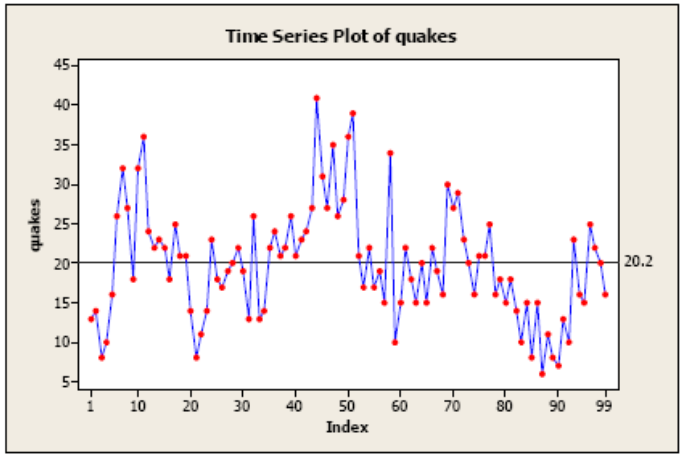
ห้ามสลับตำแหน่ง !!

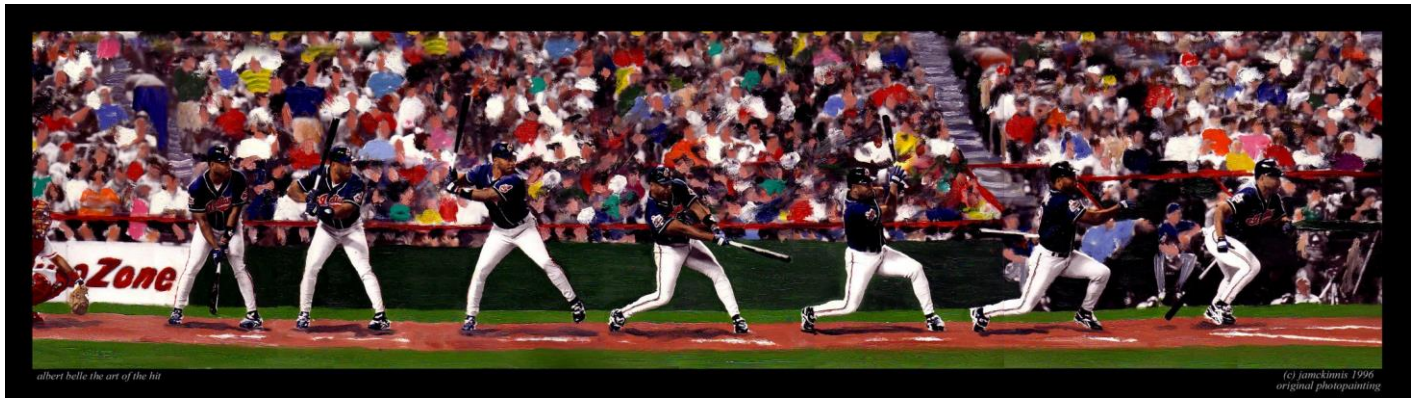
มีเวลามาเกี่ยวข้อง, time series, หุ้น, มีลำดับมีความสำคัญ

Types of Data Sets: (3) Ordered Data

ข้อมูลวิดีโอ มีเวลาเข้ามาเกี่ยวข้อง
นำรูปมาซ้อนกัน

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





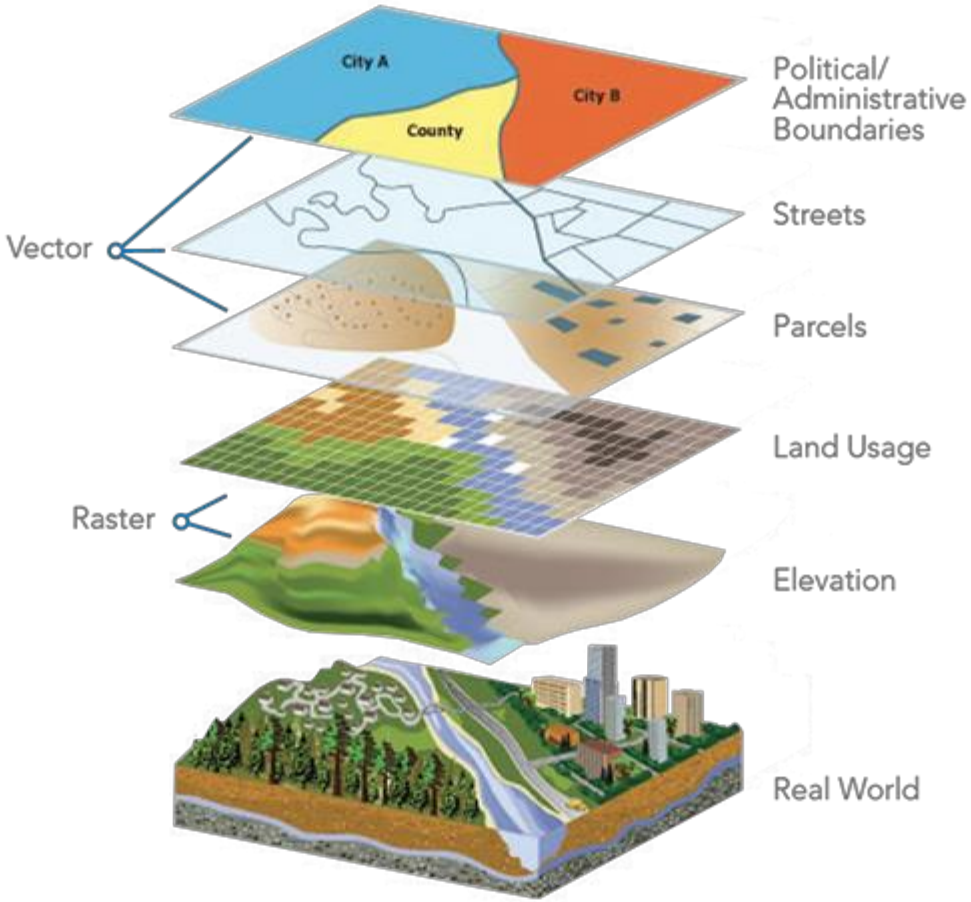
- Sequential Data: transaction sequences

- Genetic sequence data

ข้อมูลเชิงพื้นที่มีด้านกว้างกับยาว x,y

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

- Spatial data: maps



- Image data:

- Video data:

Important Characteristics of Structured Data

- Dimensionality มิติ เมนชั่น ดัด เมนชั่นส์ 2,3,4,5
  - Curse of dimensionality
- Sparsity สนใจตรงที่มีข้อมูล
  - Only presence counts
- Resolution เก็บข้อมูลใส่สี่เป็นจุดๆ
  - Patterns depend on the scale
- Distribution วัดค่าเฉลี่ย ค่าสว่างหรือสี
  - Centrality and dispersion

Data Objects

- Data sets are made up of data objects กลุ่มของข้อมูลประกอบด้วยหลาย data
- ข้อมูลแต่ละตัว
  A **data object** represents an entity
- Examples:
  - ตารางเช่น
    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 described by **attributes** ข้อมูลจะถูกอธิบายเรียก attributes
- Database rows → data objects; columns → attributes

Attributes คุณสมบัติที่ใช้เรียกข้อมูลแต่ละตัว

- **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) ชื่อของกลุ่ม ของชนิดที่ไม่ใช่ตัวเลข
  - Binary (e.g., {true, false}) ข้อมูลที่มีแค่สองค่า
  - Ordinal (e.g., {freshman, sophomore, junior, senior}) ข้อมูลเรียงลำดับ
  - Numeric: quantitative +, -, x, ÷ ได้แล้วจะมีความหมาย
    - 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

ชนิดของ attribute

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary** เหมือน Nominal แต่มีแค่ 2 เช่น 0 กับ 1 ,ใช่กับไม่ใช่
  - Nominal attribute with only 2 states (0 and 1)
  - <u>Symmetric binary</u>: both outcomes equally important
    - e.g., ~~gender~~, Left/Right-handed, coke/Pepsi, Hot/cold
  - <u>Asymmetric binary</u>: outcomes not equally important. มี 2 ค่าค่าความสำคัญไม่เท่ากัน
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal** ไม่สามารถนำมารวมกันได้ ,สามารถเรียงลำดับได้
  - Values have a meaningful order (ranking) but magnitude between successive values is not known
  - *Size* = {*small, medium, large*}, grades, army rankings

ข้อมูลที่เป็นตัวเลข

มีศูนย์แท้สัมไม่แท้

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** ระหว่างค่า 2 ค่าไม่มีค่าที่อยู่ตรงกลาง เช่น พยาบาล หมอ ไม่มีคนที่อยู่ตรงกลาง
  - 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** มีค่าตรงระหว่าง 2 ค่า ตัวอย่างเช่น สูง 160 หรือ สูง 161 แต่มีคนสูง 160.5
  - 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

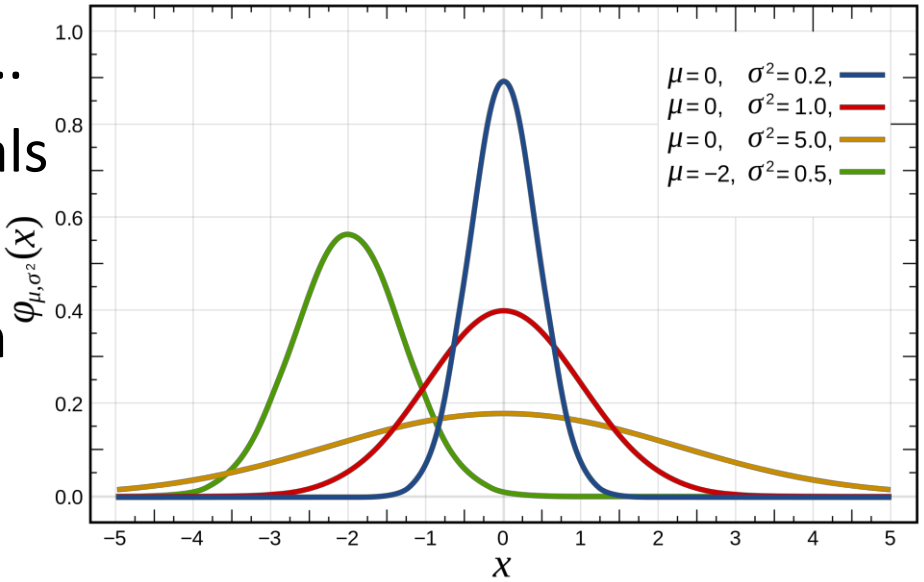
Chapter 2. Getting to Know Your Data

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

เลือกค่าทางสถิติมาใช้แสดง ลวง เพื่อตามฝูงเชื่อถือ

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



คนส่วนใหญ่ในช่วงนี้อายุ 20 ปี
ฐานนิยมค่าเยอะ