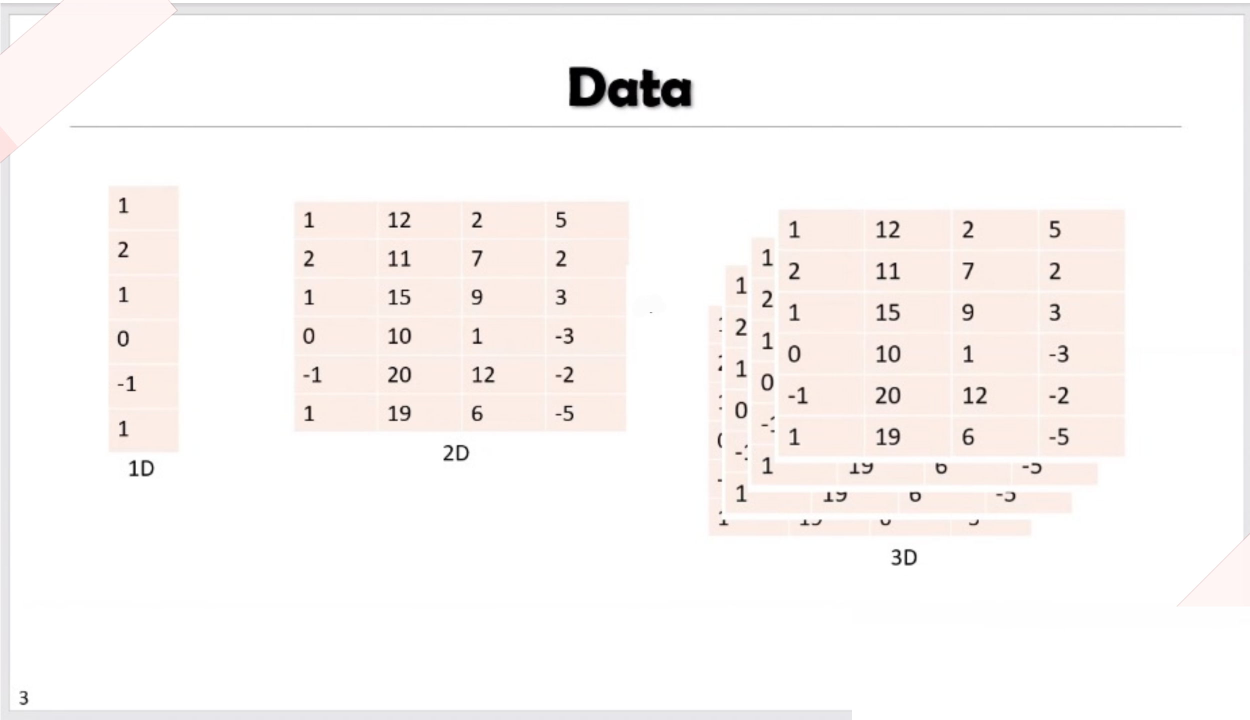
CS 412 Intro. to Data Mining

Chapter 2. Getting to Know Your Data

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

Data



Data

1
2
1
0
-1
1

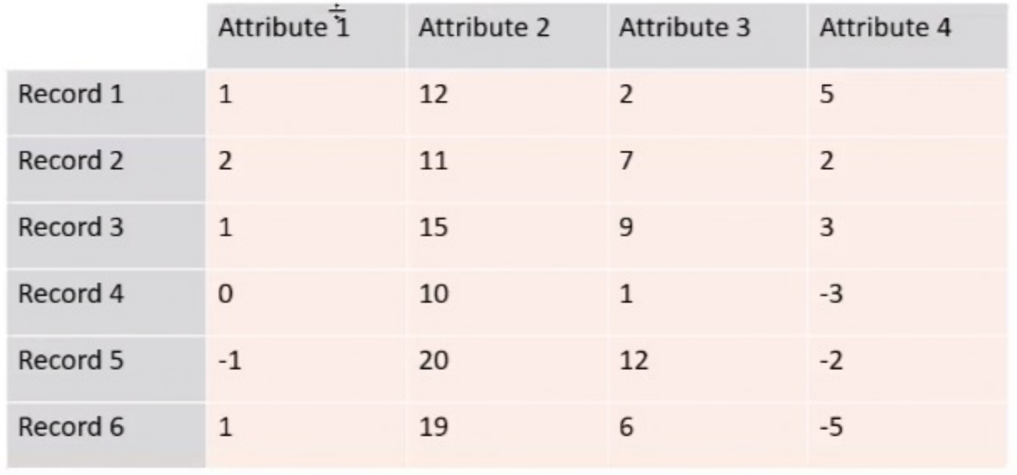
1D

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

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

3D



	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

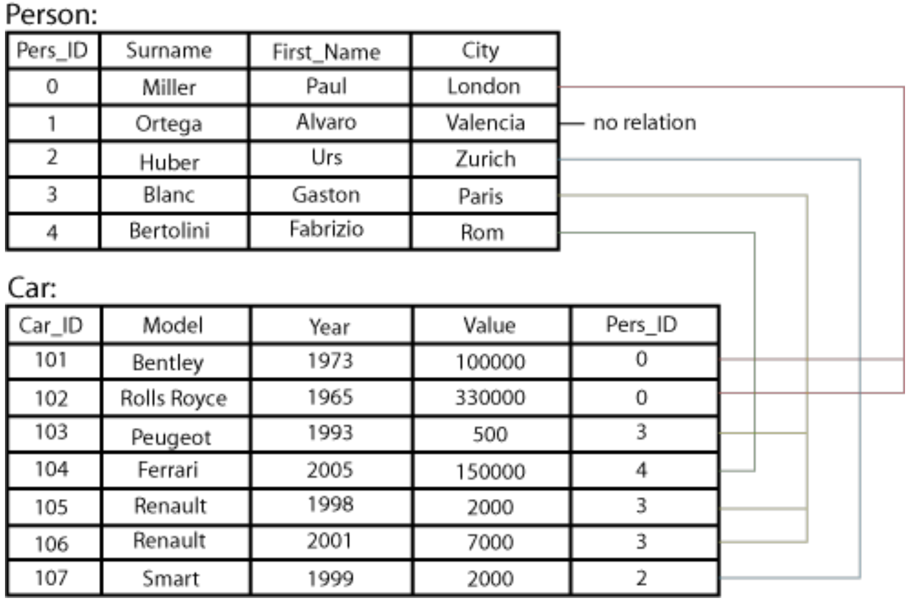
Chapter 2. Getting to Know Your Data

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

หลายตารางที่มีความสัมพันธ์กัน

Types of Data Sets: (1) Record Data

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



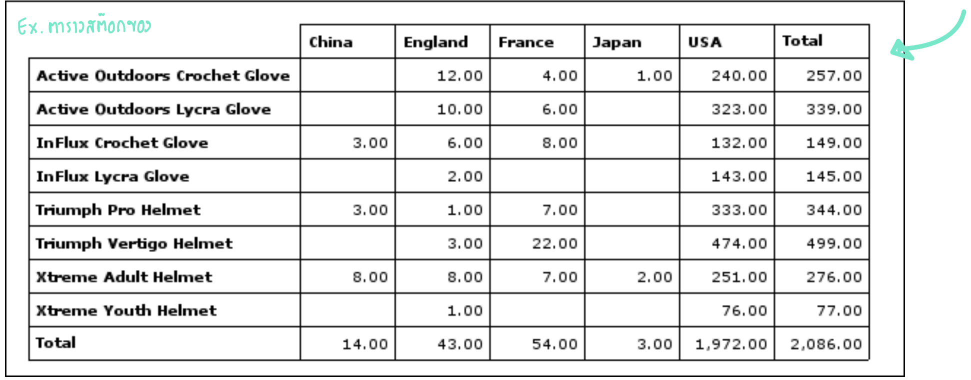
Person:

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



Ex. ตารางสต๊อกของ

	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

- 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

ผ้าอ้อม

คำ →

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
บทความ 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

→ จำนวนคำ

→ Data ใช้แทน ข้อความ

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

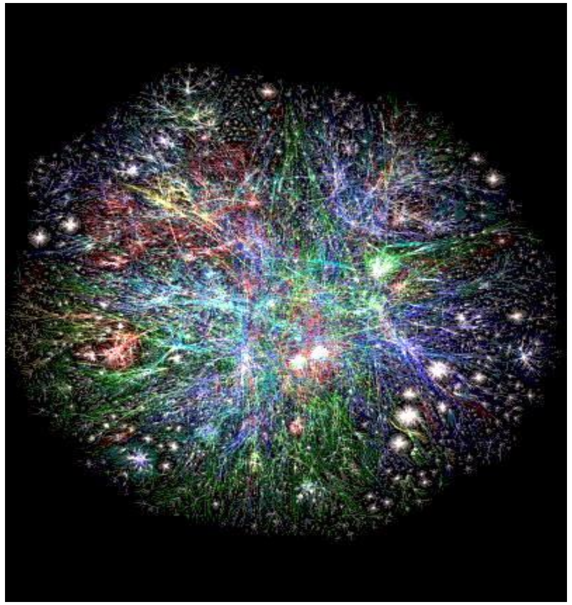
กราฟ

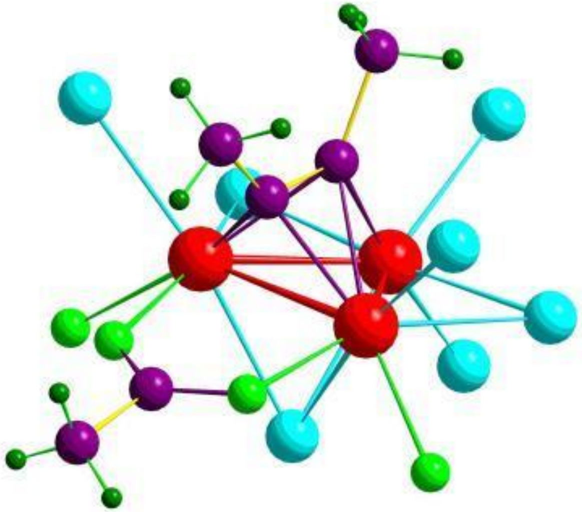
Types of Data Sets: (2) Graphs and Networks

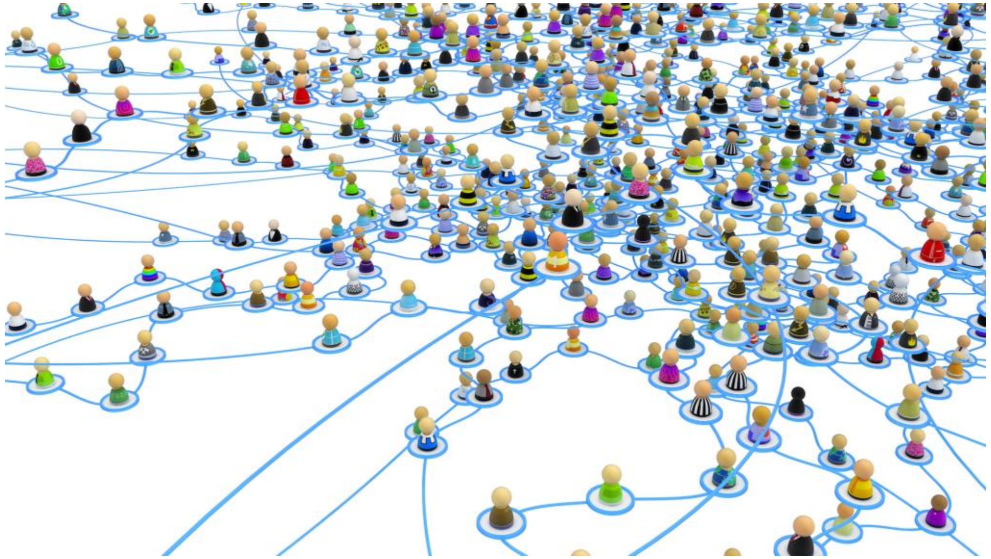
- ❑ Transportation network

Shady Grove
Rockville
Twinbrook
White Flint
Grosvenor-Strathmore
Medical Center
Bethesda
Friendship Heights
Cleveland Park
Woodley Park
Dupont Circle
Farragut North
Glenmont
Wheaton
Forest Glen
Silver Spring
Takoma
Greenbelt
College Park-U of Md
Prince George's Plaza
West Hyattsville
Fort Totten
Columbia Heights
Brookland-CUA
Rhode Island Ave
NoMa-Gallaudet U
Union Station
New Carrollton
Landover
Cheverly
Deanwood
Minnesota Ave
Rosslyn
Arlington Cemetery
Pentagon
Pentagon City
Ronald Reagan Washington National Airport
Braddock Rd
King St-Old Town
Eisenhower Ave
Huntington
Franconia-Springfield
Metro is accessible.

- ❑ World Wide Web







- ❑ Molecular Structures

- ❑ Social or information networks

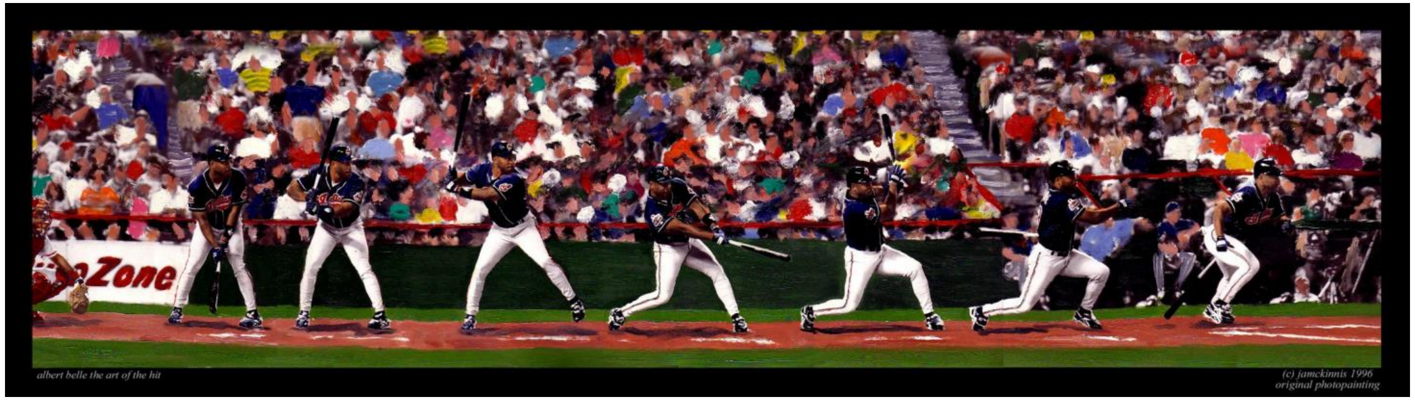
สำคัญ !! ห้ามสลับตำแหน่ง

มีเวลามาเกี่ยวข้อง, time series, หุ้น, มีลำดับมีความสำคัญ

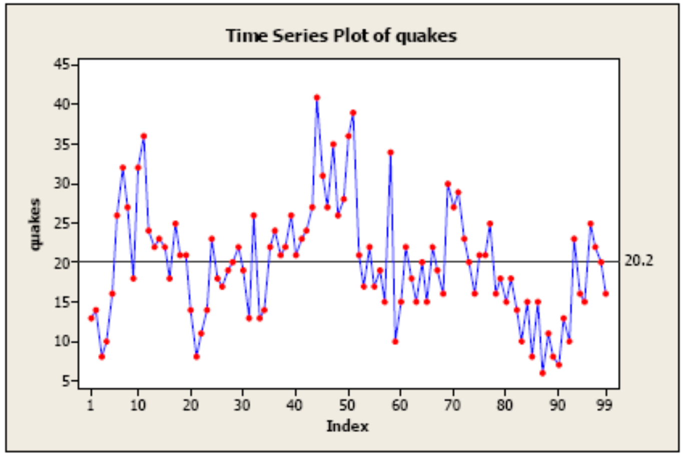
Types of Data Sets: (3) Ordered Data

ข้อมูลวิดีโอมีเวลา เข้ามาเกี่ยวข้อง, นำรูปมาซ้อนๆกัน (การ์ตูนมิกกี้เมาส์ในอดีต)

- ❑ Video data: sequence of images

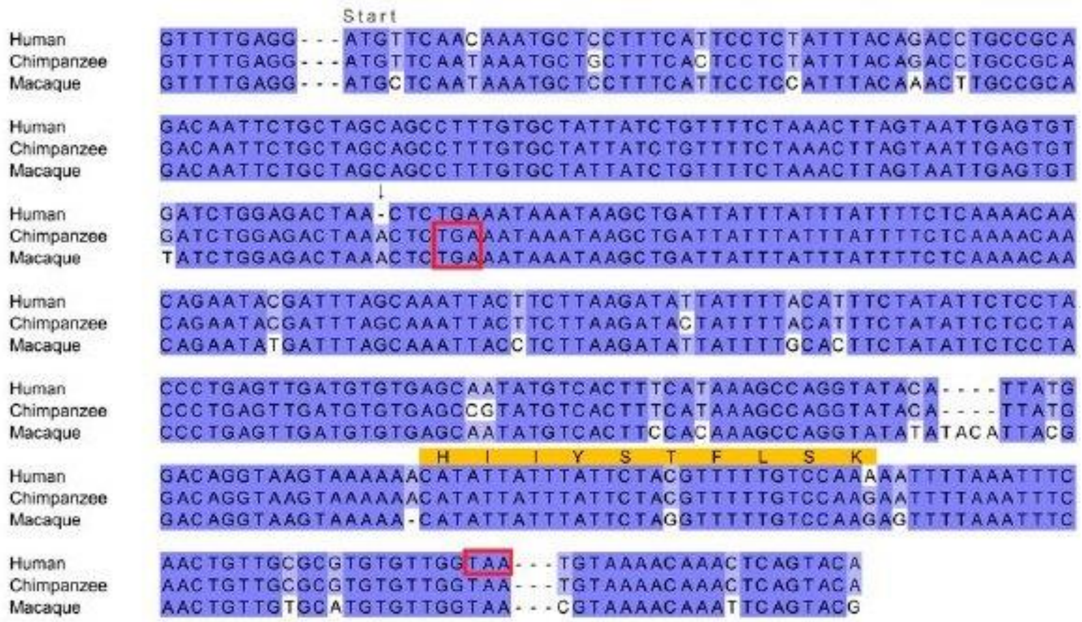


- ❑ Temporal data: time-series



- ❑ Sequential Data: transaction sequences

- ❑ Genetic sequence data

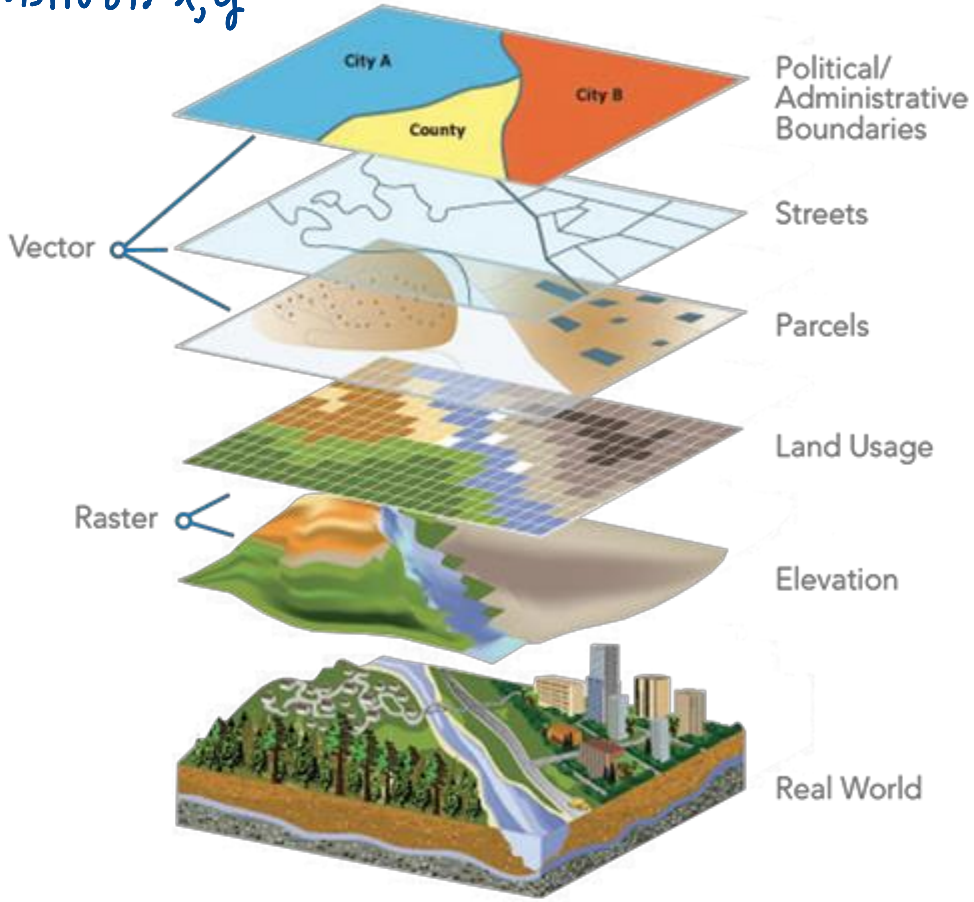


```
Human       GTTTTGAGG---ATGTTCAACAAATGCTCCTTTCATTCCTCTATTTACAGACCTGCCGCA
Chimpanzee  GTTTTGAGG---ATGTTCAATAAATGCTGCTTTCACTCCTCTATTTACAGACCTGCCGCA
Macaque     GTTTTGAGG---ATGCTCAATAAATGCTCCTTTCATTCCTCCATTTACAAACTTGCCGCA

Human       GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Chimpanzee  GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Macaque     GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT

Human       GATCTGGAGACTAA-CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Chimpanzee  GATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Macaque     TATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA

Human       CAGAATACGATTTAGCAAATTACTTCTTAAGATATTATTTTACATTTCTATATTCTCCTA
Chimpanzee  CAGAATACGATTTAGCAAATTACTTCTTAAGATACTATTTTACATTTCTATATTCTCCTA
Macaque     CAGAATATGATTTAGCAAATTACCTCTTAAGATATTATTTTGCACTTCTATATTCTCCTA

Human       CCCTGAGTTGATGTGTGAGCAATATGTCACTTTCATAAAGCCAGGTATACA----TTATG
Chimpanzee  CCCTGAGTTGATGTGTGAGCCGTATGTCACTTTCATAAAGCCAGGTATACA----TTATG
Macaque     CCCTGAGTTGATGTGTGAGCAATATGTCACTTCCACAAAGCCAGGTATATATACATTACG

Human       GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCAAAAATTTTAAATTTC
Chimpanzee  GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCAAGAATTTTAAATTTC
Macaque     GACAGGTAAGTAAAAA-CATATTATTTATTCTAGGTTTTTGTCCAAGAGTTTTAAATTTC

Human       AACTGTTGCGCGTGTGTTGGTAA---TGTAAAACAAACTCAGTACA
Chimpanzee  AACTGTTGCGCGTGTGTTGGTAA---TGTAAAACAAACTCAGTACA
Macaque     AACTGTTGTGCATGTGTTGGTAA---CGTAAAACAAATTCAGTACG
```

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

ข้อมูลเชิงพื้นที่มีด้าน กว้างกับยาว x,y

- ❑ Spatial data: maps



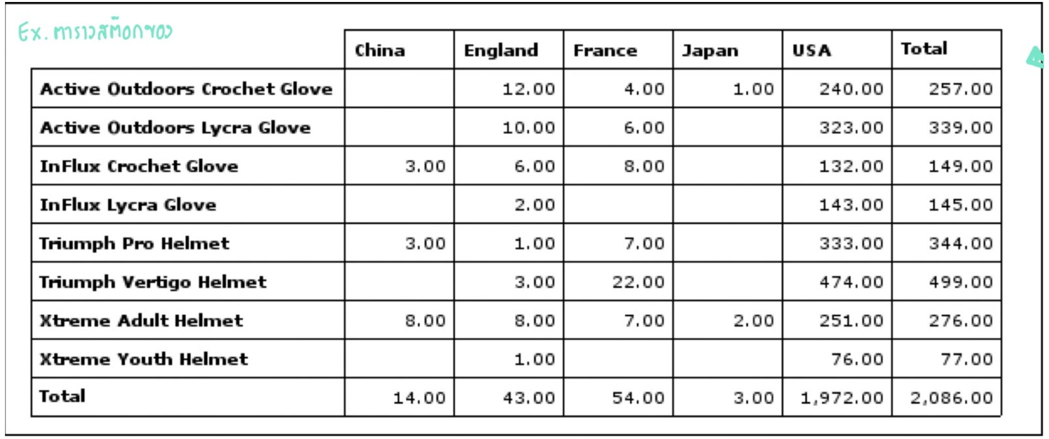
- ❑ Image data:

- ❑ Video data: spatio-temporal

Important Characteristics of Structured Data

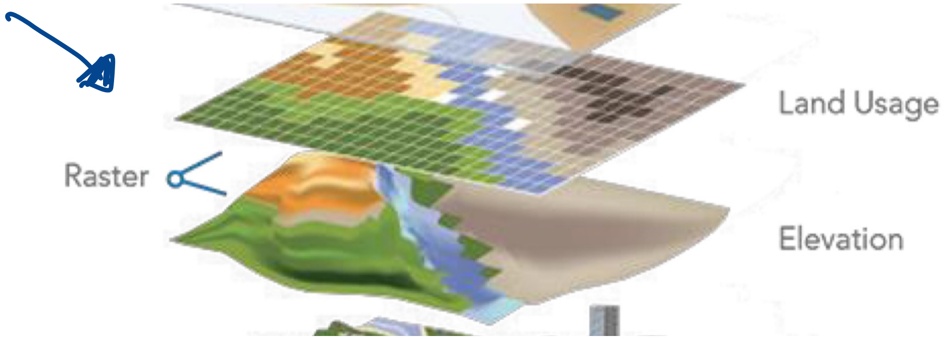
- ❑ Dimensionality → มีไดเมนชันก์ไดเมนชันก์ 2,3,4,5
  - ❑ Curse of dimensionality
- ❑ Sparsity → สนใจตรงที่มีข้อมูล, ข้อมูลเป็น 0 เยอะไม่สนใจ
  - ❑ Only presence counts
- ❑ Resolution → เก็บข้อมูลใส่สีเป็นจุดๆ
  - ❑ Patterns depend on the scale
- ❑ Distribution → หาค่ากลางที่สีนี้สว่างหรือมืด
  - ❑ Centrality and dispersion

Sparsity



Ex. ตารางสต็อกของ

	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



Data Objects

- ❑ Data sets are made up of data objects → กลุ่มของข้อมูล ประกอบด้วยหลายๆ data

ข้อมูลแต่ละตัว

- ❑ A **data object** represents an entity
- ❑ Examples:
  - ❑ sales database: customers, store items, sales
  - ❑ medical database: patients, treatments
  - ❑ university database: students, professors, courses

ชื่อ data

- ❑ Also called *samples , examples, instances, data points, objects, tuples*
- ❑ Data objects are described by **attributes** → ข้อมูลจะถูกอธิบายเรียก attributes
- ❑ Database rows → data objects; columns → attributes

คุณสมบัติที่ใช้เรียกข้อมูลแต่ละตัว

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) → ชื่อของกลุ่มชื่อของชนิดไม่ใช่ตัวเลข
  - ❑ Binary (e.g., {true, false}) → ข้อมูลมีแค่สองค่า
  - ❑ Ordinal (e.g., {freshman, sophomore, junior, senior}) → ข้อมูลเรียงลำดับ
  - ❑ Numeric: quantitative → +, -, ×, ÷ ได้แล้วมีความหมาย
    - ❑ 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?

→ Numeric

→ ชนิดของ attribute

Attribute Types

- **Nominal:** categories, states, or “names of things”
  - สีผม
    Hair_color = {auburn, black, blond, brown, grey, red, white}
  - สถานะ
    marital status, occupation, ID numbers, zip codes
- **Binary** → เหมือน Nominal แต่มีแค่ 2 เช่น 0 กับ 1 , ใช่ กับ ไม่ใช่
  - Nominal attribute with only 2 states (0 and 1)
  - สมมาตร
    Symmetric binary: both outcomes equally important
    - e.g., ~~gender~~, Left / Right-handed, Coke / Pepsi, Hot/Cold
  - Asymmetric binary: outcomes not equally important. → มี 2 ค่าความสำคัญไม่เท่ากัน เช่น ตรวจ covid มีคนเป็น 2 คน ไม่เป็น 1000 คน ในมหาลัย
    - 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

Numeric Attribute Types

↳ มีศูนย์แท้กับไม่แท้

- ❑ Quantity (integer or real-valued)
- ❑ **Interval** 0 แท้ เรามีดินสอ 0 แท่ง คือ 0 แท้
  - ❑ 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*

ไม่ต่อเนื่อง

Discrete vs. Continuous Attributes

- **Discrete Attribute** → ระหว่างค่า 2 ค่า ไม่มีค่าที่อยู่ตรงกลาง เช่น หมอ พยาบาล ไม่มีค่าไหนอยู่ตรงกลาง
  - 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** → มีค่ากลางระหว่าง 2 ค่า เช่น สูง 180 กับ สูง 181 แต่มีคนสูง 180.5
  - 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 data ของเราเพื่อความน่าเชื่อถือ

Basic Statistical Descriptions of Data

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

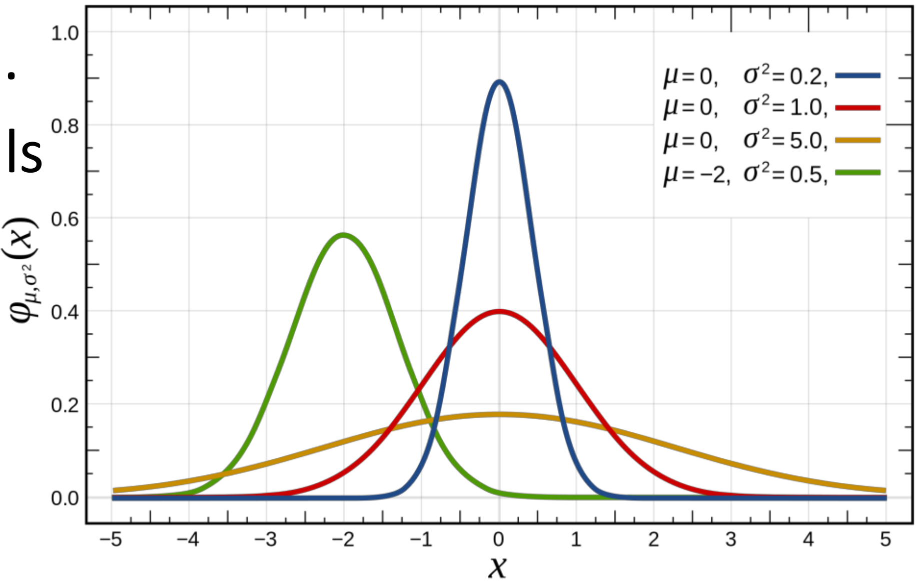
เบี่ยงเบนจากค่ากลางมากเท่าไหร่

แตกต่างจากค่ากลางมากน้อยแค่ไหน

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

มัธยฐาน

- ❑ 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



คนส่วนใหญ่ในห้องนี้อายุ 20 ปี

- ฐานนิยม (ซ้ำเยอะ)