# Data

จะเป็นรูปแบบ Matrix ก็มิติก็ได้

| |
|---|
| 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

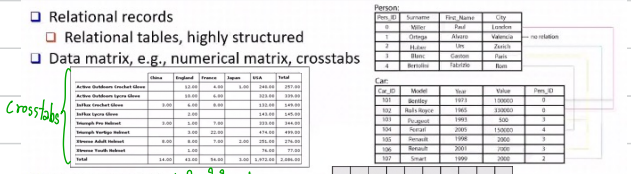
| | Attribute 1 | Attribute 2 | Attribute 3 | Attribute 4 |
|---|---|---|---|---|
| Record 1 | 1 | 12 | 2 | 5 |
| Record 2 | 2 | 11 | 7 | 2 |
| Record 3 | 1 | 15 | 9 | 3 |
| Record 4 | 0 | 10 | 1 | -3 |
| Record 5 | -1 | 20 | 12 | -2 |
| Record 6 | 1 | 19 | 6 | -5 |

## 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

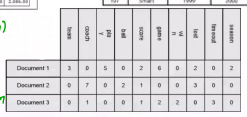
## Types of Data Sets: (1) Record Data

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




- Transaction data (ส่วนใหญ่ใช้ในการซื้อขาย)

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

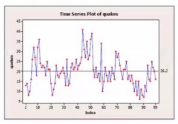


- Document data: Term-frequency vector (matrix) of text documents

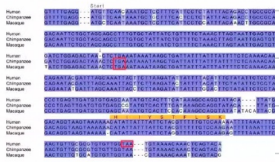
ใช้สรุปข้อมูลที่เป็น text ให้เข้าใจง่ายขึ้น เช่น วิเคราะห์ว่าพูดถึงอะไร

## Types of Data Sets: (3) Ordered Data

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



- Sequential Data: transaction sequences



- Genetic sequence data

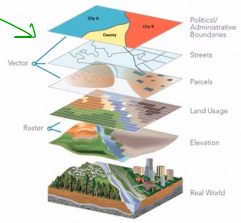
เป็นการสร้างวิดีโอจากภาพ ซึ่งมาต่อกันๆ

Data ที่ไม่สามารถสลับที่ได้ → VDO

→ DNA

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

- Spatial data: maps




- Image data:
- Video data: Spatio - temporal

- Spatial data คือ เชิงพื้นที่ (กว้าง, ยาว) (x, y)

## Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

| | 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 |

# 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)
  - Binary (e.g., {true, false})
  - Ordinal (e.g., {freshman, sophomore, junior, senior})
  - 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?

Nominal : ข้อมูลชนิดบัญญัติ (แบ่งโดยลักษณะรวม)
Binary : มีแค่ 2 ค่า เช่น T, F
Ordinal : แบ่งโดยอันดับ การเรียงลำดับ เช่น freshmen, sophomore, junior, senior
Numeric : ข้อมูลที่ใช้แทนจำนวนที่สามารถนำไปคำนวณได้

# Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - 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**
  - 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)
- **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*

0 แท้ → มีเงิน 0 บาท = ไม่มีเงินสักบาท
0 ไม่แท้ → อุณหภูมิ 0 °C ≠ ไม่มีอุณหภูมิ แต่ = เย็น

# 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 แรงจูงใจ
  - To better understand the data: central tendency, variation and spread ค่ากลาง
- Data dispersion characteristics ลักษณะการกระจายข้อมูล
  - 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
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

การวิเคราะห์การกระจายของการวัดที่คำนวณได้