



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

|    |    |    |    |
|----|----|----|----|
| 1  | 12 | 2  | 5  |
| 2  | 1  | 15 | 9  |
| 1  | 0  | 10 | 1  |
| 0  | -1 | 20 | 12 |
| -1 | 1  | 19 | 6  |
| 1  | 12 | 2  | 5  |

|    |    |    |    |
|----|----|----|----|
| 1  | 12 | 2  | 5  |
| 2  | 1  | 15 | 9  |
| 1  | 0  | 10 | 1  |
| 0  | -1 | 20 | 12 |
| -1 | 1  | 19 | 6  |
| 1  | 12 | 2  | 5  |

|    |    |    |    |
|----|----|----|----|
| 1  | 12 | 2  | 5  |
| 2  | 1  | 15 | 9  |
| 1  | 0  | 10 | 1  |
| 0  | -1 | 20 | 12 |
| -1 | 1  | 19 | 6  |
| 1  | 12 | 2  | 5  |

4D

### Data

|    |    |    |    |
|----|----|----|----|
| 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  | 1  | 15 | 9  |
| 1  | 0  | 10 | 1  |
| 0  | -1 | 20 | 12 |
| -1 | 1  | 19 | 6  |
| 1  | 12 | 2  | 5  |

3D

1 มิติ = กว้าง / ยาว  
2 มิติ = กว้าง & ยาว  
3 มิติ = 2 มิติซ้อนกัน  
4 มิติ = 3 มิติมาเรียงต่อกัน

ถ้า Data มี Attribute ซักกัน ข้อมูลจะเป็นชุดเดียวกัน

|          | 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          |

## 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    |
| 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       |

| Term       | Count | TF | DF | Score | TF | DF | Score | TF | DF | Score |
|------------|-------|----|----|-------|----|----|-------|----|----|-------|
| 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     |

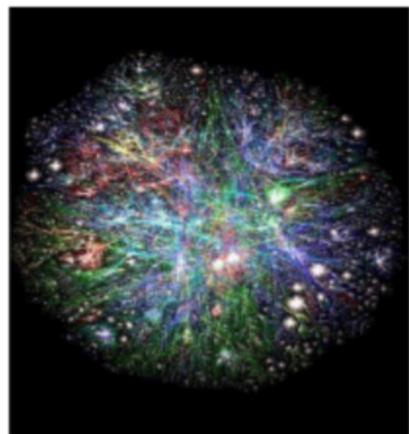
↑↑↑  
↑↑↑  
↑↑↑

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

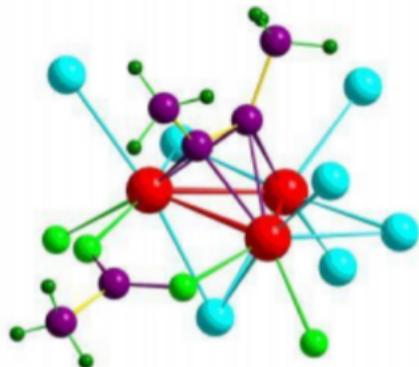
ตาราง Term-Frequency เป็นตารางที่ใช้เก็บข้อมูลที่เป็น Text

## Types of Data Sets: (2) Graphs and Networks

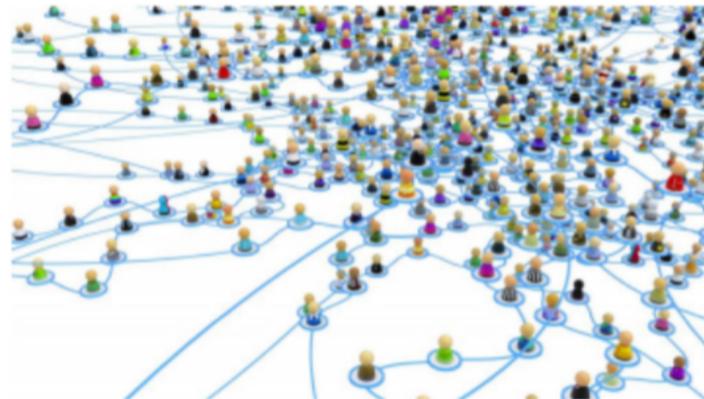
- Transportation network



- World Wide Web



- Molecular Structures

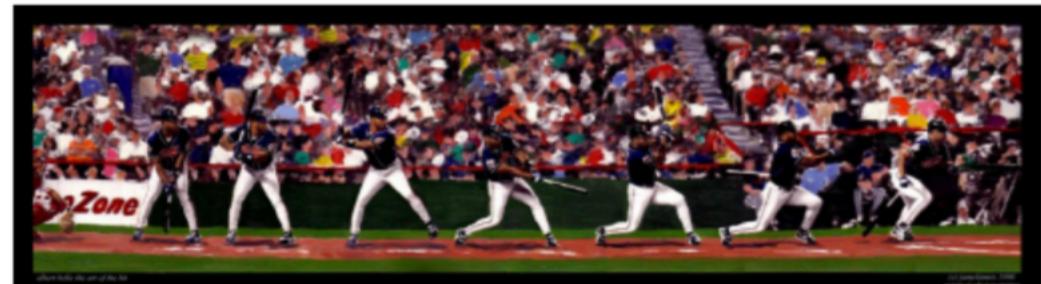


- Social or information networks

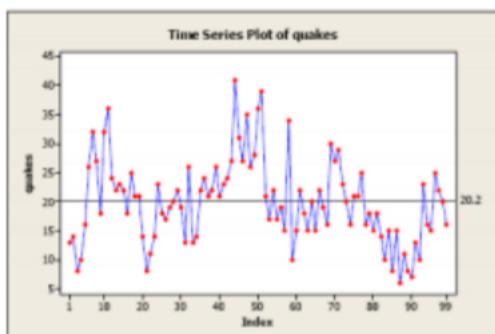
จะเป็นการยกว่าจุดนี้เชื่อมกับอะไร อย่างเช่น ข้อมูลในโซเชียล เว็บไซต์ต่างๆ แผนที่รถไฟฟ้า ที่มีการเชื่อมต่อกันไปอย่างไรบ้าง

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

- Video data: sequence of images



- Temporal data: time-series



- Sequential Data: transaction sequences

Start

Human  
Chimpanzee  
Macaque

GTTTTGAGG -- ATGTCAACAAATGCTCCTTCATTCCTCTATTACAGACCTGCCCGCA  
GTTTTGAGG -- ATGTCATAAATGCTCCTTCACCCCTCTATTACAGACCTGCCCGCA  
GTTTGAGG -- ATGTCATAAATGCTCCTTCATTCCTCTATTACAGACCTGCCCGCA

SACAACTCTGCTAGCAACCTTGTGCATTATCTGTTTCCTAAACTTAGTAATTGAGGT  
SACAACTCTGCTAGGAGCCCTTGTGCATTATCTGTTTCCTAAACTTAGTAATTGAGGT  
SACAACTCTGCTAGCAACCTTGTGCATTATCTGTTTCCTAAACTTAGTAATTGAGGT

Human  
Chimpanzee  
Macaque

GATCTGGAGACTAA-DTC TGAATATAAGCTGATATTATTTATTTCTCAAAACAA  
GATCTGGAGACTAAACCTGAAATAAAATAAGCTGATATTATTTATTTCTCAAAACAA  
TATCTGGAGACTAAACCTGAAATAAAATAAGCTGATATTATTTATTTCTCAAAACAA

CAGAATACGATTTAGCAAAATTACTCTTAAAGATAATTATTTACATTCTATATTCCTCA  
CAGAATACGATTTAGCAAAATTACTCTTAAAGATAATTATTTACATTCTATATTCCTCA  
CAGAATATQTTAACAAATTACCTCTTAAAGATAATTATTTACATTCTATATTCCTCA

Human  
Chimpanzee  
Macaque

CCCCTGGATTGATGTTGAGCAATATGCACTTTCATTAAGCCAGGTATACA---TTATG  
CCCCTGATTGATGTTGAGCAATATGCACTTTCATTAAGCCAGGTATACA---TTATG  
CCCCTGGATTGATGTTGAGCAATATGCACTTTCATTAAGCCAGGTATACA---TTATG

Human  
Chimpanzee  
Macaque

GACAGDTAAATAAAAACATATTATTTCTACATTCTTGTCCAAGAAATTTAAATTTC  
GACAGDTAAATAAAAACATATTATTTCTACATTCTTGTCCAAGAAATTTAAATTTC  
GACAGDTAAATAAAAACATATTATTTCTACATTCTTGTCCAAGAAATTTAAATTTC

Human  
Chimpanzee  
Macaque

H I Y S F L S K  
GACAGDTAAATAAAAACATATTATTTCTACATTCTTGTCCAAGAAATTTAAATTTC  
GACAGDTAAATAAAAACATATTATTTCTACATTCTTGTCCAAGAAATTTAAATTTC  
GACAGDTAAATAAAAACATATTATTTCTACATTCTTGTCCAAGAAATTTAAATTTC

Human  
Chimpanzee  
Macaque

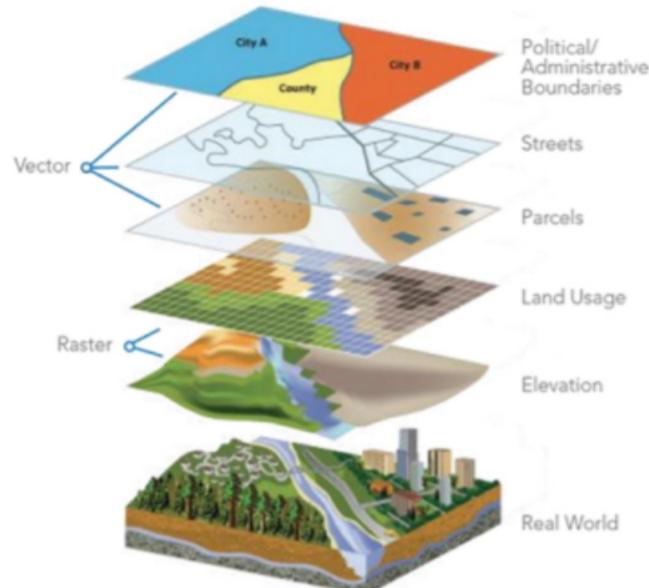
AAC TGT TGC GCG GTG TGT TGG TAA --- TGT AARAC AAC TC AGT GAA  
AAC TGT TGC GCG GTG TGT TGG TAA --- TGT AARAC AAC TC AGT GAA  
AAC TGT TGC TGT GTG TGT TGG TAA --- C TAAAAC AAC TC AGT GAA

- Genetic sequence data

เป็นข้อมูลที่มีเวลาเข้ามาเกี่ยวข้องเหมือนกัน ทำการแปลง Data เป็นตัวเลข - Time-Series เช่น ราคาหุ้น แต่ละวันที่มีราคาขึ้นลงไปมา - Sequential เช่น ข้อมูลที่ไม่สามารถสลับตำแหน่งหรือลำดับกันได้

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

- ## Spatial data: maps ព័ត៌មានអនុវត្ត



- Image data: ຜົນຖານເພື່ອປະກາດ

- ## □ Video data:

เป็นข้อมูลที่มีเวลาเข้ามาเกี่ยวข้อง

Vedio เช่น วีดิโอ โดยทำการเอารูปหลายรูปมาซ้อนกัน เป็นรูปที่ 1 ทับรูปที่ 2 ทับรูปที่ 3 ไปเรื่อยๆ (เหมือนรูปสามารถเคลื่อนไหวได้ (เรียกว่า Cartoon Making))

Image เช่น รูปภาพ ซึ่งจะมีด้านกว้างกับยาว ซึ่งอาจจะกำหนดเป็นพิกัด  $x,y$  โดยแต่ละตัวจะมีสีของจุด

Spatial เช่น แผนที่ ซึ่งจะมีการกำหนดจุดพิกัด x,y โดยอาจจะกำหนดสีให้แต่ละพิกัดจังหวัด

## Important Characteristics Of Structured Data

## Dimensionality

Curse Of dimensionality ด่าวา Data มี Dimension เท่าไรเป็นอย่างไร

## Sparsity

Only Presence Counts สนใจแค่ที่ที่มีข้อมูล

## Resolution ດຽວເກີບຂໍ້ມູນໃຫຍ່ກາຊະເກົ່າ

Patterns Depend On The Scale ความละเอียดในการเก็บข้อมูล

## Distribution នគរបាល

Centrality And Dispersion การกระจายตัวของ Data โดยการวัดค่ากลางของ Data

Digitized by srujanika@gmail.com

- Nominal រំនោគវេជ្ជកម្មនៃទីតាំង

### - Binary

- Ordinal ຈົດໜູນເວັບຈຳກັດ

- Numeric : quantitative សង្គមដែលអាចបង្កើតឡើង