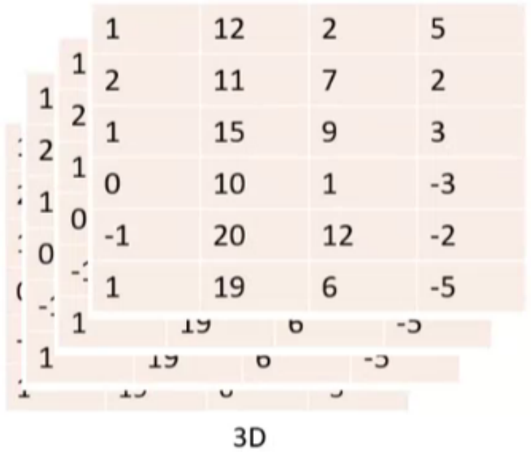
สรุปเนื้อหา Chapter 2

Data คือข้อมูลชุดๆหนึ่งที่เราจะนำมาใช้งาน โดยวิชานี้จะใช้งาน Data ด้วยรูปแบบของ Matrix โดยมีเงื่อนไขว่า หาก Data มี Attibute ที่ลักษณะเหมือนกันคือข้อมูลที่ซ้ำกัน
แนวตั้ง (column) เราเรียกว่า Attribute คือคุณสมบัติที่ใช้อธิบายข้อมูล แนวนอน (row) เรียก Record เป็นข้อมูลแบบจุด

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

ตัวอย่างข้อมูลแบบ 2 มิติ



ตัวอย่างข้อมูลแบบ 3 มิติ

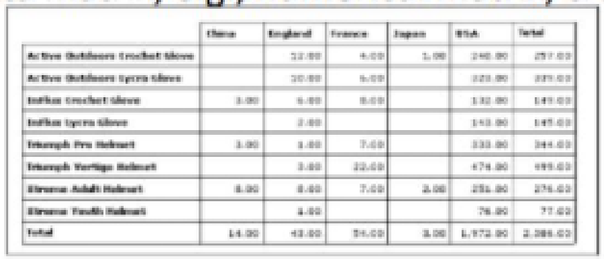
ตัวอย่างข้อมูลแบบ 4 มิติ

Type of Data sets : Record Data

เงื่อนไขคือพยายามหาข้อมูลให้เป็นตัวเลข ลักษณะเด่นๆคือตารางหลายๆตารางต่างมีความสัมพันธ์ดัน (Relationals Records)
Term-frequency คือ ตารางเก็บแบบข้อความ (Text)

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

- 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

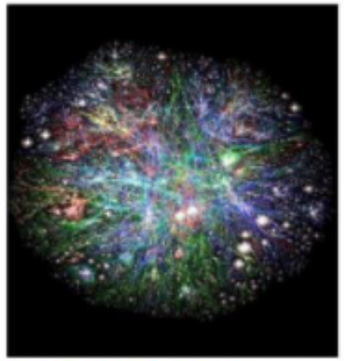
- Document data: Term-frequency vector (matrix) of text documents

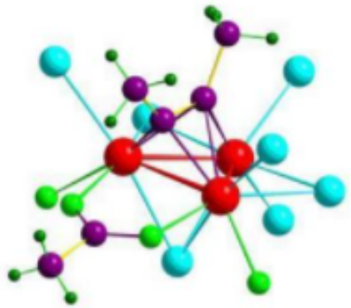
Type of Data sets : Graphs and Networks

เหมือนเครือข่ายที่เรารู้จักกันเช่น อินเตอร์เน็ต เครือข่ายการขนส่ง โครงสร้างโมเลกุล เป็นต้น

Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web





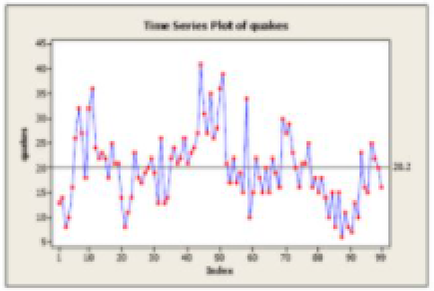
- Molecular Structures
- Social or information networks

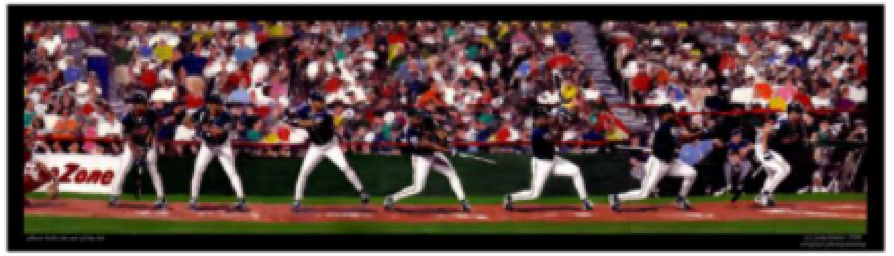
Type of Data sets : Ordered Data

เป็นข้อมูลที่มีตัวเลข หรือ เวลาเข้ามาเกี่ยวข้องเช่น ราคาหุ้น เป็นต้น

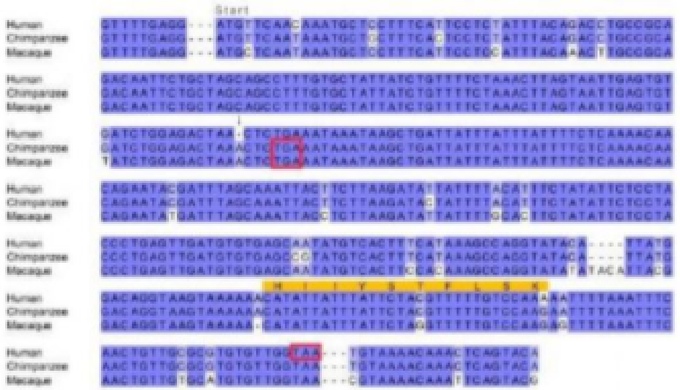
Types of Data Sets: (3) Ordered Data

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





- Sequential Data: transaction sequences
- Genetic sequence data

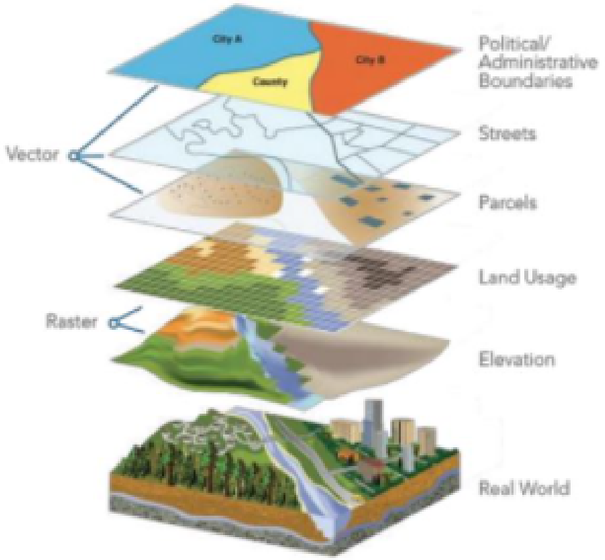


Type of Data sets : Spatial, image and multimedia Data

คือข้อมูลที่เป็นประเภท รูปภาพหรือวิดิโอ

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

- Spatial data: maps



- Image data:
- Video data:

ลักษณะสำคัญข้อมูล

1. ข้อมูลมีมิติเป็นยังไง (Dimensionality)
2. สนใจแค่ที่มีข้อมูล (Sparsity)
3. ความละเอียดในการเก็บข้อมูล (Resolution)
4. การกระจายตัวของข้อมูล (Distribution)

Data Objects

1. Data Sets คือข้อมูลหลายๆข้อมูลมารวมกัน
2. Data Object คือข้อมูลแต่ละตัวประกอบด้วย Entity
3. ข้อมูล 1 จุดที่อยู่ในแนวตั้ง (Data 1 จุด คือ Data point)
4. ข้อมูลจะถูกอธิบายด้วย Attributes
5. ใน Database ที่จะนำมาทำคือ Rows > Data, Column > Attributes

Attibute Type

- ข้อมูลแบบ Norminal คือข้อมูลที่เป็นการบอกกลุ่มเช่น เพศ อาชีพ เป็นต้น
- ข้อมูลแบบ Binary คือ ข้อมูลเพียงแต่มี 2 สถานะ เช่น ใช่-ไม่ใช่ ถูก-ผิด เป็นต้น โดยแบ่งเป็น 2 แบบได้แก่
  - Symmetric Binary คือสมมาตรกัน
  - Asymmetric Binary คือไม่สมมาตรกัน
- ข้อมูลแบบ Ordinal คือ ข้อมูลที่มีความหมายสามารถนำมาเรียงลำดับหรืออันดับได้

Numeric Attribute Types
เรื่องของ ศุนย์แท้กับศูนย์ไม่แท้
0 แท้ คือ 0 ที่ไม่มีค่าเลยหรือมีค่าเป็น 0 จริงๆ
0 ไม่แท้ คือ มีค่าเป็นอีกรูปแบบหนึ่งเช่น 00.00 น. เป็นต้น

- ข้อมูลแบบ Interval คือข้อมูลที่ 0 ไม่แท้
- ข้อมูลแบบ Ratio ข้อมูลที่มี 0 แท้

Discete VS Continuous Attributes
แบ่งเป็น 2 ลักษณะใหญ่ๆคือ

- Discete Attribute คือ ข้อมูลที่ไม่ต่อเนื่องกัน เช่น รหัสไปรษณีย์ เป็นต้น
- Continuous Attribute คือ ข้อมูลที่ต่อเนื่องกัน(มีค่าตรงกลาง) โดยข้อมูลส่วนใหญ่เป็นจำนวนจริง เช่น ส่วนสูง 180,181.5,182

Basic Statistical Descriptions Of Data

เป็นการดูค่าทางสถิติโดยใช้ค่ากลางมีทั้งหมด 3 ชนิด คือ ค่าเฉลี่ย ค่ามัธฐาน ค่าฐานนิยม

- Movation ตือการเอาค่ากลางทั้ง 3 ตัวมาวาง plot ลงกราฟว่าแตกต่างมากน้อยเพียงใด
- Data Dispersion Characteristcis คือ ค่ากลาง,มากที่สุด,ต่ำสุด,Quantile,Outlines,Varaiance

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