

สรุปเนื้อหา Chapter 2

Data คือข้อมูลชุดๆหนึ่งที่เรานำมาใช้งาน โดยวิชานี้จะใช้งาน Data ด้วยรูปแบบของ Matrix โดยมีเงื่อนไขว่า หาก Data มี Attribute 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

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

3D

ตัวอย่างข้อมูลแบบ 3 มิติ

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

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

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

ตัวอย่างข้อมูลแบบ 4 มิติ

Type of Data sets : Record Data

เงื่อนไขคือพยายามหาข้อมูลให้เป็นตัวเลข ลักษณะเด่นๆคือตารางหลายๆตารางต่างมีความสัมพันธ์กัน (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

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Sports Glove		20.00	6.00		120.00	226.00
Infant Crochet Glove	3.00	6.00	6.00		132.00	147.00
Infant Sports Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	3.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Stromo Adult Helmet	6.00	6.00	7.00	6.00	256.00	276.00
Stromo Youth Helmet		4.00			76.00	77.00
Total	14.00	43.00	34.00	6.00	1,172.00	2,286.00

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

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1978	100000	0
102	Rolls Royce	1965	330000	0
103	Pugeot	1998	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

	bread	coke	beer	milk	diaper	beer	coke	diaper	milk
Document 1	3	0	5	0	2	6	0	2	0
Document 2	0	7	0	2	1	0	0	3	0
Document 3	0	1	0	0	1	2	2	0	3

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

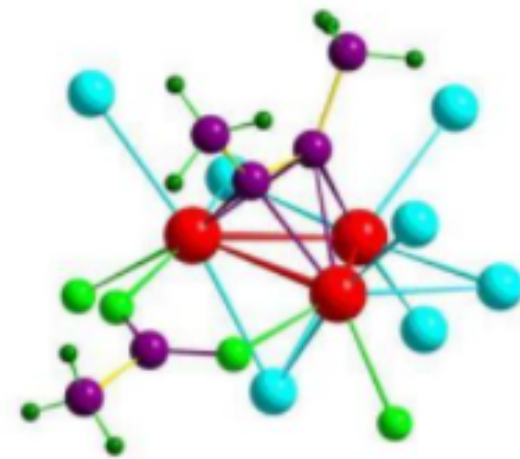
Type of Data sets : Graphs and Networks

เหมือนเครือข่ายที่เราคุ้นเคยกันเช่น อินเทอร์เน็ต เครือข่ายการขนส่ง โครงสร้างโมเลกุล เป็นต้น

Types of Data Sets: (2) Graphs and Networks

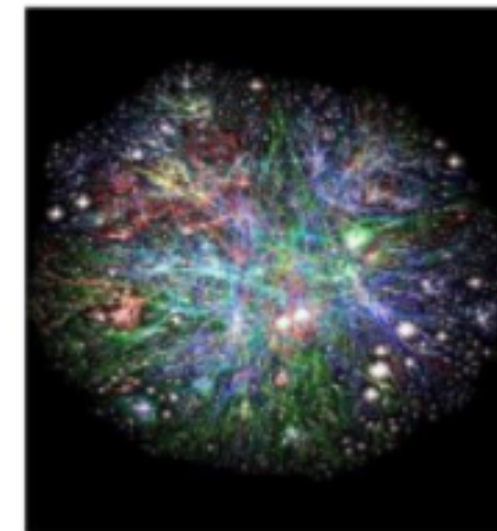
- ❑ Transportation network

- ❑ World Wide Web



- ❑ Molecular Structures

- ❑ Social or information networks



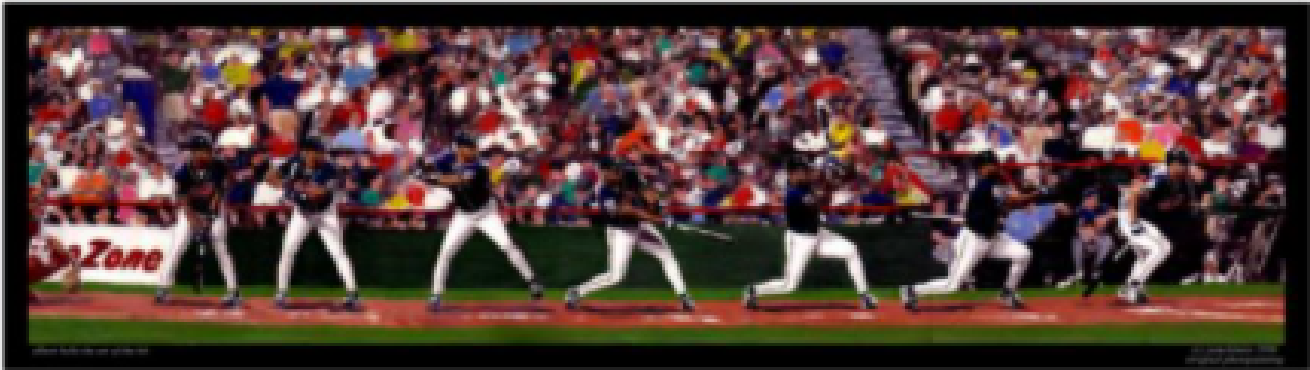
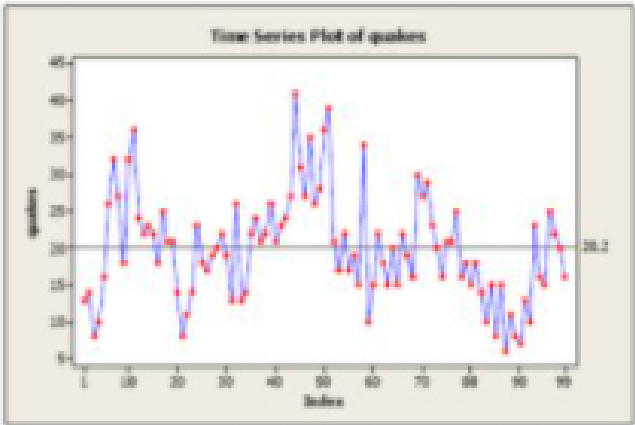
Type of Data sets : Ordered Data

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

	Start
Human	GTTTTGAGG...ATGTTCAACAAATGCTCCTTTTCATTCCTCTATTTACAGACC TGCCGCA
Chimpanzee	GTTTTGAGG...ATGTTCAATAAATGCTCCTTTTCATTCCTCTATTTACAGACC TGCCGCA
Mouse	GTTTTGAGG...ATGTTCAATAAATGCTCCTTTTCATTCCTCTATTTACAGACC TGCCGCA
Human	SACAAATTCGCTAGCAGCCCTTTGTGCTATTATCTGTTTTCTAAACCTTAGTAATTGAGTGT
Chimpanzee	SACAAATTCGCTAGCAGCCCTTTGTGCTATTATCTGTTTTCTAAACCTTAGTAATTGAGTGT
Mouse	SACAAATTCGCTAGCAGCCCTTTGTGCTATTATCTGTTTTCTAAACCTTAGTAATTGAGTGT
Human	GATCTGGAGACTAA...CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Chimpanzee	GATCTGGAGACTAA...CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Mouse	GATCTGGAGACTAA...CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Human	CAGAATACGATTTAGCAAAATTACCTTTAAGATAATTTTACATTTCTATATTCTGCTA
Chimpanzee	CAGAATACGATTTAGCAAAATTACCTTTAAGATAATTTTACATTTCTATATTCTGCTA
Mouse	CAGAATACGATTTAGCAAAATTACCTTTAAGATAATTTTACATTTCTATATTCTGCTA
Human	CCCTGAGTTGATGTGTGAGGCAATATGTCACCTTCTAAAGCCAGGTATAC...TTATG
Chimpanzee	CCCTGAGTTGATGTGTGAGGCAATATGTCACCTTCTAAAGCCAGGTATAC...TTATG
Mouse	CCCTGAGTTGATGTGTGAGGCAATATGTCACCTTCTAAAGCCAGGTATAC...TTATG
Human	SACAGGTAAAGTAAAAACATATTATTATTCTACCTTTTGTCCAA...AATTTAAATTTT
Chimpanzee	SACAGGTAAAGTAAAAACATATTATTATTCTACCTTTTGTCCAA...AATTTAAATTTT
Mouse	SACAGGTAAAGTAAAAACATATTATTATTCTACCTTTTGTCCAA...AATTTAAATTTT
Human	AACGTGTTGCGGTGTGTGTGTA...TGTAAGCAAACTCAGTAC
Chimpanzee	AACGTGTTGCGGTGTGTGTGTA...TGTAAGCAAACTCAGTAC
Mouse	AACGTGTTGCGGTGTGTGTGTA...TGTAAGCAAACTCAGTAC

Type of Data sets : Spatial, image and multimedia Data

คือข้อมูลที่เป็นประเภท รูปภาพหรือวิดีโอ

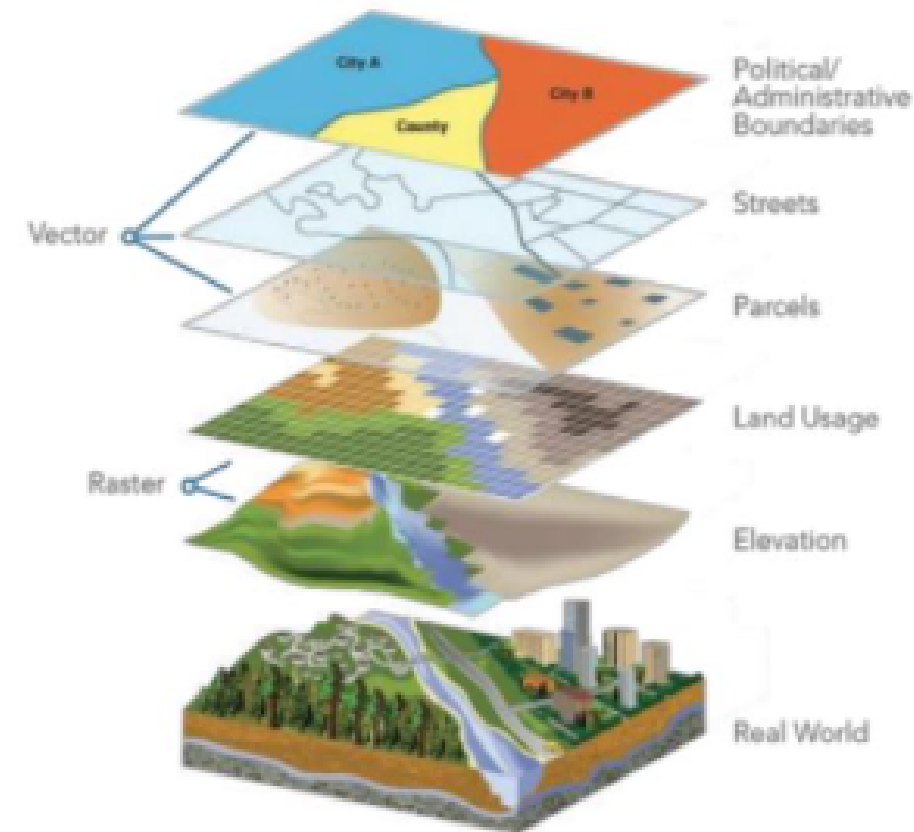
Types of Data Sets: (4) Spatial, image and multimedia Data

❑ Spatial data: maps



❑ Image data:

❑ Video data:



ลักษณะสำคัญข้อมูล

1. ข้อมูลมีมิติเป็นอย่างไร (Dimensionality)
2. สparse แค่ที่มีข้อมูล (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

Attribute 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 ศูนย์แท้กับศูนย์ไม่แท้

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

Discrete VS Continuous Attributes

แบ่งเป็น 2 ลักษณะใหญ่ๆคือ

- Discrete 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

- Motivation คือการเอาค่ากลางทั้ง 3 ตัวมาวาง plot ลงกราฟว่าแตกต่างกันน้อยเพียงใด
- Data Dispersion Characteristics 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

