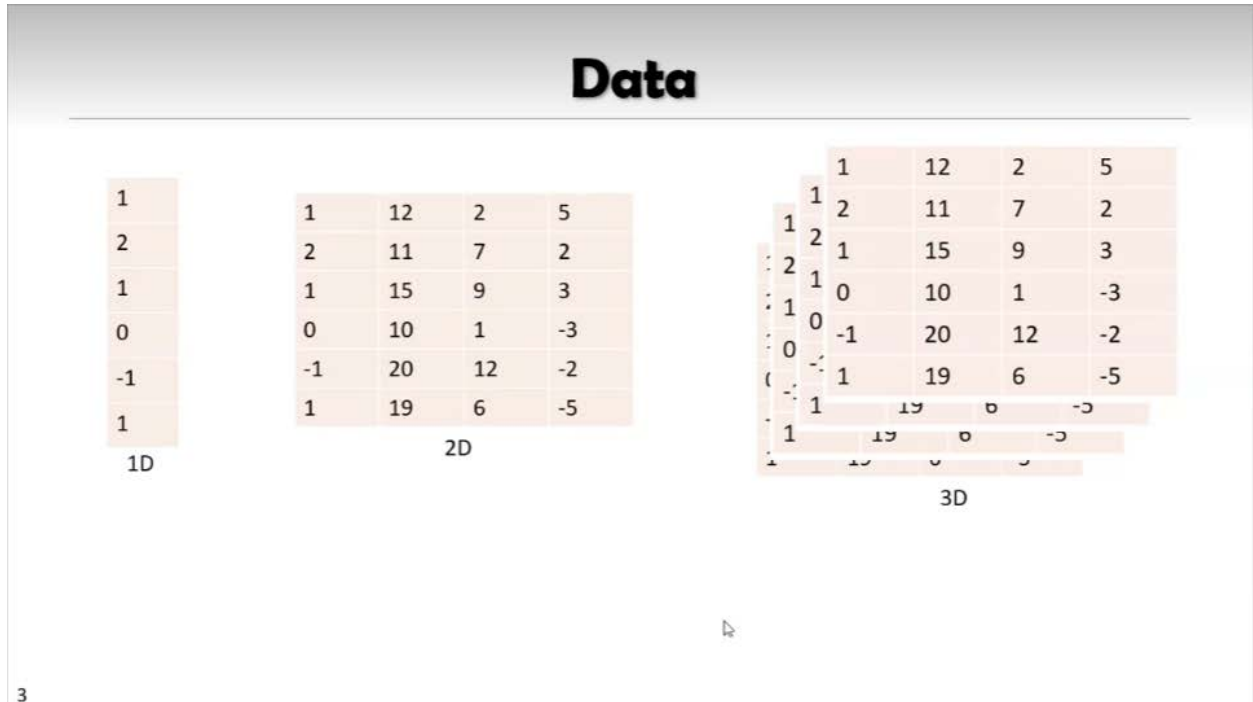


Chapter 2 Getting to know your data

- ขนาดของข้อมูลแต่ละมิติ



ตัวอย่าง Data ที่เป็นตารางตัวเลข (เรียกว่า ดาต้าเซต = กลุ่มของข้อมูล)

Data

	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

จะอาศัยการใช้ Database มาช่วยในการจัดเก็บข้อมูลให้ซับซ้อนน้อยลง เพื่อประหยัดพื้นที่ในการจัดเก็บข้อมูล

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	5.00	240.00	237.00	
Active Outdoors Igroa Glove		10.00	5.00		333.00	338.00
Infuse Crochet Glove	3.00	6.00	9.00		132.00	149.00
Infuse Igroa Glove		2.00			143.00	145.00
Intangly Pro Helmet	3.00	1.00	7.00		233.00	344.00
Intangly Vortigo Helmet		3.00	22.00		474.00	499.00
Stromer Adult Helmet	9.00	6.00	1.00	2.00	251.00	274.00
Stromer Youth Helmet		1.00			74.00	77.00
Total	34.00	43.00	54.00	3.00	1,972.00	2,084.00

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	1973	100000	0
102	Rolls Royce	1965	130000	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

	beer	coke	diaper	milk	bread	beer	coke	diaper	milk	bread
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

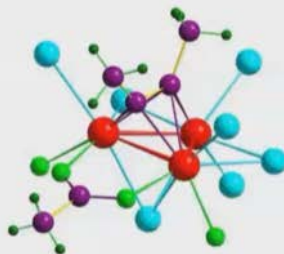
7

- ในตารางจะบอกรายละเอียด การเชื่อมต่อของข้อมูลแต่ละตาราง

ตัวอย่าง Data ที่เป็นกราฟ (นอกจากตาราง)

Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web




- Molecular Structures
- Social or information networks

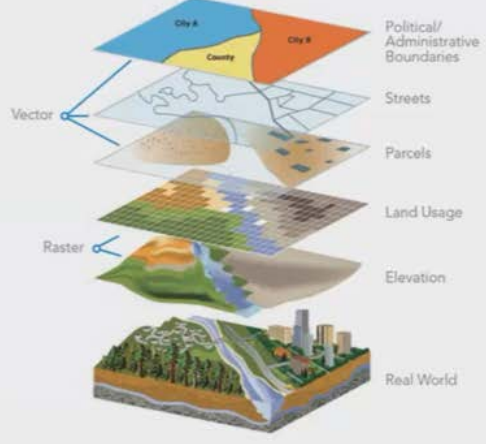


8

ตัวอย่าง Data ที่เป็นรูปภาพ/วิดีโอ (อาจมีการบอกพิกัดด้วย)

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

- Spatial data: maps
- Image data:
 
- Video data:

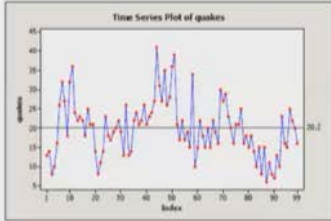



10

ตัวอย่าง Data ที่เป็นรูปภาพในแต่ละวินาที (เรื่องของเวลาเข้ามาเกี่ยว) มาเชื่อมต่อกันเป็นวิดีโอ เช่น ข้อมูลราคาหุ้น, DNA

Types of Data Sets: (3) Ordered Data

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data

	Human	Chimpanzee	Macaque
1	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
2	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
3	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
4	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
5	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
6	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
7	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
8	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
9	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
10	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
11	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
12	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
13	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
14	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
15	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
16	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
17	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
18	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
19	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
20	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
21	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
22	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
23	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
24	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
25	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
26	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
27	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
28	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
29	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
30	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
31	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
32	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
33	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
34	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
35	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
36	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
37	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
38	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
39	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
40	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
41	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
42	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
43	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
44	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
45	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
46	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
47	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
48	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
49	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
50	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
51	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
52	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
53	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
54	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
55	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
56	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
57	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
58	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
59	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
60	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
61	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
62	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
63	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
64	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
65	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
66	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
67	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
68	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
69	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
70	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
71	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
72	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
73	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
74	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
75	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
76	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
77	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
78	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
79	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
80	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
81	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
82	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
83	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
84	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
85	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
86	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
87	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
88	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
89	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
90	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
91	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
92	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
93	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
94	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
95	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
96	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
97	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
98	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC
99	GTCTTGGAGG	ATGTTCCAC	AAATGCTCTTTCATTCCTGATATTAGAGC

9

คุณสมบัติที่เราต้องรู้

Important Characteristics of Structured Data

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

11

Data Objects

- Data sets are made up of data objects
- A **data object** represents an entity
- Examples:
 - sales database: customers, store items, sales
 - medical database: patients, treatments
 - university database: students, professors, courses
- Also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples*
- Data objects are described by **attributes**
- Database rows → data objects; columns → attributes

12

ชนิดของข้อมูล

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?

13

รายละเอียด คำอธิบายเพิ่มเติม

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*

14

0 แท้ และ 0 ไม่แท้ คืออะไร?

= 0 ไม่แท้ ยกตัวอย่างได้คือ อุณหภูมิ 0 องศาเซลเซียส ที่ไม่ได้แปลว่าไม่มีอุณหภูมิ แต่แปลว่า หนาวมาก
/ 0 แท้ ยกตัวอย่างคือ เรามีดินสออยู่ 0 แท่ง แปลว่าเราไม่มี

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

15

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

16

การใช้สถิติมาอธิบาย Data เบื้องต้น เพื่อให้เข้าใจในมากขึ้น

เช่น คนในห้องนี้อายุ 20 ปี (ใช้ฐานนิยม)

Chapter 2. Getting to Know Your Data

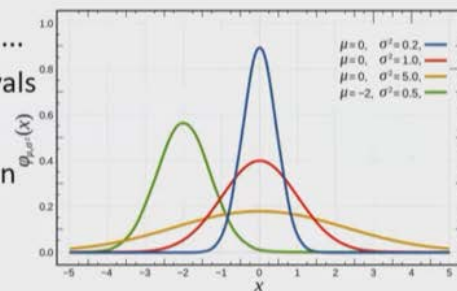
- ☐ Data Objects and Attribute Types
- ☐ Basic Statistical Descriptions of Data 
- ☐ Data Visualization
- ☐ Measuring Data Similarity and Dissimilarity
- ☐ Summary

17

----- คาบต่อไป -----

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



18