สรุป

ประเภทของข้อมูล

- **ชุดข้อมูล** คือ กลุ่มของ data objects ประกอบด้วย Attributes ที่บอกลักษณะ

ตารางที่ 2.1. ชุดข้อมูลนิสิต (Student Information Data Set)

Attributes

Data objects

ลำดับที่	รหัสนิสิต	ชั้นปี	เกรดเฉลี่ยสะสม
1	1034261	2	2.75
2	1034262	3	3.24
3	1034263	2	3.51
4	1034265	1	2.99
5	1034266	3	3.12

Attributes ของ Data objects สามารถ เปลี่ยน แปลงได้ตลอด
เช่น อุณหภูมิพื้นผิวโลก-มหาสมุทร

การแบ่งประเภทของ Attributes โดย Operation ของระบบจำนวน

ตารางที่ 2.2. ชนิดของแอทริบิวต์

ชนิดของแอทริบิวต์		คำอธิบาย	ตัวอย่าง	โอเปอเรชั่น
Categorical (เชิงคุณภาพ)	Nominal	ค่าของ Nominal attribute สามารถใช้ในการแยกแยะดาต้าอ็อบเจ็กต์ได้ ด้วยโอเปอเรชั่น Distinctness (=, ≠)	รหัสไปรษณีย์ รหัสพนักงาน สีตา เพศ	ฐานนิยม, entropy, contingency correlation, Chi-squared test
	Ordinal ก. binary ค. Discete จ.	มีคุณสมบัติและโอเปอเรชั่น Distinctness เช่นเดียวกันกับ Nominal attributes และค่าของ Ordinal attribute สามารถใช้ในการเรียงลำดับดาต้าอ็อบเจ็กต์ได้ ด้วยโอเปอเรชั่น Order (<, >)	ความแข็งของแร่ธาตุ, เกรด {A, B+, B, C+, C, D+, D, F},	มัธยฐาน, เปอร์เซ็นต์ไทล์, rank correlation, run tests, sign tests
Numeric (เชิงปริมาณ)	Interval	มีคุณสมบัติและโอเปอเรชั่น Distinctness และ Order เช่นเดียวกันกับ Nominal attributes และ Ordinal attributes นอกจากนี้ ความแตกต่างระหว่าง interval attributes สองค่า คำนวณได้ด้วยโอเปอเรชั่น Addition (+, -) สามารถตีความได้ กล่าวคือ interval attributes จะมีหน่วยของการวัด	อุณหภูมิในหน่วยองศาเซลเซียส หรือองศาฟาเรนไฮต์, วันที่ตามปฏิทิน	ค่าเฉลี่ย, ส่วนเบี่ยงเบนมาตรฐาน, Pearson's correlation, t-test, F-test
	Ratio ง. Continuous	มีคุณสมบัติและโอเปอเรชั่น Distinctness, Order, และ Interval เช่นเดียวกันกับ Nominal attributes, Ordinal attributes, และ Interval attributes นอกจากนี้ อัตราส่วนของ ratio attributes ซึ่งคำนวณได้โดยใช้โอเปอเรชั่น Multiplication (x, /) สามารถตีความได้	อุณหภูมิในหน่วยเคลวิน (Kelvin), อายุ, มวล, ความยาว, กระแสไฟฟ้า	ค่าเฉลี่ยเรขาคณิต, ค่าเฉลี่ยฮาร์มอนิค, เปอร์เซ็นต์ความผันแปร

การแบ่งประเภท Attributes ด้วยจำนวนของค่าที่เป็นไปได้

1. Discrete Attribute ที่มีค่าเป็นไปได้ไม่กี่ค่าหรือจำกัด แต่สามารถหนึ่งแสนแสนได้ มักใช้กับ Attribute เชิงคุณภาพ
2. Binary Attribute มีแค่ 2 ค่า 0/1, จริง/เท็จ, ใช่/ไม่ใช่, ชาย/หญิง
3. Continuous Attribute มีค่าเป็นจำนวนจริง มักใช้กับ Attributes เชิงปริมาณ

ชนิดของ Data sets มี 3 ประเภท

1. ข้อมูล Record data

ข้อมูลที่ใช้ในการทำเหมืองข้อมูล มักอยู่ในรูปแบบของกลุ่ม Record data ซึ่งประกอบด้วย Attributes หรือ ฟิลด์ข้อมูลต่างๆ

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Soda, Diapers, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

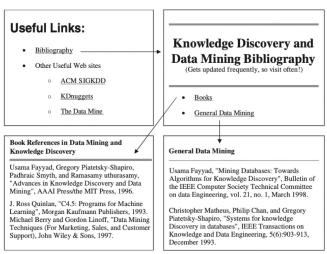
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

(d) Document-term matrix.

2. ข้อมูลกราฟ

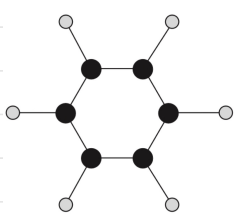
กราฟเป็นโครงสร้างข้อมูลสำหรับแสดงความสัมพันธ์ระหว่าง Data object เรามักใช้ทำกราฟ 2 กรณี

1. แสดงความสัมพันธ์ระหว่าง Data object



(a) Linked web pages.

2. Data object มีโครงสร้างแบบกราฟ



(b) Benzene molecule.

3. ข้อมูลแบบมีลำดับ

ข้อมูลบางชนิดประกอบด้วย Attributes ที่มีความสัมพันธ์กันเชิงเวลาหรือเชิงพื้นที่

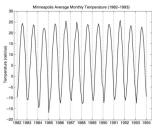
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

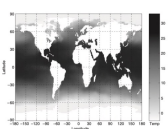
(a) Sequential transaction data.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(b) Genomic sequence data.



(c) Temperature time series.



(d) Spatial temperature data.

คุณภาพข้อมูล

การแก้ปัญหาคุณภาพข้อมูล เทคนิคการจัดการกับคุณภาพของข้อมูล แบ่งได้ 2 กลุ่ม

1. เทคนิคสำหรับ การตรวจจับและการแก้ไขปัญหา คุณภาพข้อมูล
2. การใช้อัลกอริทึมการกำหนดข้อมูลที่ทนทานต่อข้อมูลคุณภาพต่ำ

ความผิดพลาดจากการวัด และการเก็บข้อมูล

ปัญหาคุณภาพข้อมูลเกิดได้จากความผิดพลาดของมนุษย์ ข้อจำกัดของอุปกรณ์

เช่น ข้อมูลรบกวน ,ข้อมูลเที่ยม , ความลำเอียง ,ความเที่ยงตรง และความแม่นยำ

- ความผิดพลาดจากการวัด
- ความผิดพลาดจากการเก็บข้อมูล

ปัญหาคุณภาพที่เกี่ยวข้องกับการประยุกต์ใช้งาน

ข้อมูลที่มีคุณภาพสูง คือข้อมูลที่เหมาะกับการประยุกต์ใช้งาน คุณภาพข้อมูลจากมุมมองของการใช้

แต่ปัญหาทั่วไปของคุณภาพต่ำไปใช้งาน มี 3 อย่าง

1. Timeliness เวลาในการเก็บ
2. Relevance ข้อมูลทุกชิ้นที่จำเป็นต้องใช้ครบถ้วนหรือเปล่า
3. Knowledge about the Data คุณภาพเอกสาร