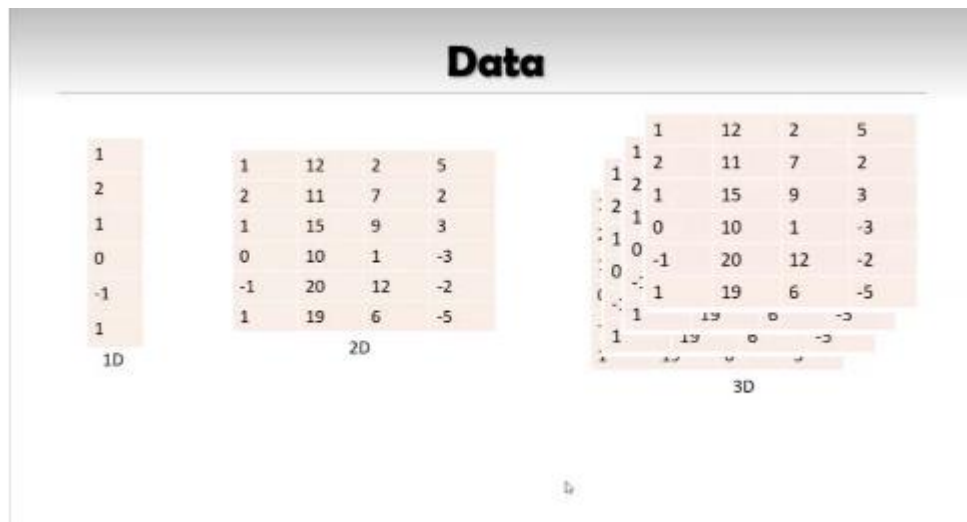


บทที่ 2 Getting to know your data

-ขนาดของข้อมูลในแต่ละมิติ ประกอบด้วย 1 มิติ 2 มิติ และ 3 มิติ

ถ้าเป็น 4 มิติลักษณะก็จะเป็นรูปแบบ 3 มิติเรียงซ้อนๆกัน



ตัวอย่างข้อมูลที่เป็นตารางตัวเลข เรียกว่า ค้าตาเซตหรือกลุ่มของข้อมูล

The table, titled "Data", shows a dataset with 6 records and 4 attributes. The first column lists the records, and the subsequent columns list the attribute values for each record.

	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

5

จะใช้ database มาช่วยในการจัดเก็บข้อมูลที่ให้ซับซ้อนน้อยลง เพื่อประหยัดพื้นที่ในการจัดเก็บข้อมูล

Types of Data Sets: (1) Record Data

- ❑ Relational records
- ❑ Relational tables, highly structured
- ❑ Data matrix, e.g., numerical matrix, crosstabs

	Black	Greyhound	Indian	Japanese	Other	Total
Black Mountain Incident Data	100,000	0	0	0	0	100,000
Black Mountain Agency Data	100,000	0	0	0	0	100,000
Indian Incident Data	0	100,000	0	0	0	100,000
Indian Agency Data	0	100,000	0	0	0	100,000
Japanese Incident Data	0	0	100,000	0	0	100,000
Japanese Agency Data	0	0	100,000	0	0	100,000
Other Incident Data	0	0	0	100,000	0	100,000
Other Agency Data	0	0	0	100,000	0	100,000
Total	100,000	100,000	100,000	100,000	0	400,000

Case ID	Student	Year	Status	Perf. ID
001	Student	1973	Student	0
002	Student	1974	Student	0
003	Student	1975	Student	0
004	Student	1976	Student	0
005	Student	1977	Student	0
006	Student	1978	Student	0
007	Student	1979	Student	0
008	Student	1980	Student	0

❑ Transaction data

TID	Items
1	Bread, Coke, Milk
2	Bread, Bread
3	Bread, Coke, Diaper, Milk
4	Bread, Bread, Diaper, Milk
5	Coke, Diaper, Milk

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




	term	1	2	3	4	5	6	7	8	9	10
Document 1		1	0	1	0	2	0	1	2	0	2
Document 2		0	1	0	1	1	0	0	0	0	0
Document 3		0	1	0	0	1	1	1	0	2	0

จากรูปในตารางจะบอกรายละเอียด การเชื่อมต่อของข้อมูลต่อตาราง

ตัวอย่าง Data มราเป็นกราฟ (นอกจากตาราง)

Types of Data Sets: (2) Graphs and Networks

- ❑ Transportation network
- ❑ World Wide Web
- ❑ Molecular Structures
- ❑ Social or information networks

ตัวอย่าง Data ที่เป็นรูปภาพ/วิดีโอ(อาจมีการบอกพิกัดด้วย)

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

- Spatial data: maps
- Image data:

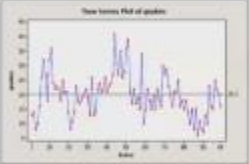
- Video data:





10

ตัวอย่าง data ที่เป็นรูปภาพในแต่ละวินาทีมาเชื่อมต่อเป็นวิดีโอ 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



9

คุณสมบัติต่างๆ

Important Characteristics of Structured Data

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

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

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?

รายละเอียด เพิ่มเติม

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

0 แท้ และ 0 ไม่แท้ คืออะไร

0 (ศูนย์) แท้ เช่น น้ำหนัก ความสูง อายุ เป็นต้น

ศูนย์ของข้อมูลระดับนี้ไม่ได้หมายความว่าไม่มี แต่เป็นศูนย์ที่เกิดจากการสมมติขึ้น เช่น การวัดอุณหภูมิ 0 องศาเซลเซียสไม่ได้หมายความว่าไม่มีอุณหภูมิ

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 เพื่อให้เข้าใจมากขึ้น

เช่น คนไทย หุ่นนี้อายุ 20 ปี (ใช้ฐานนิยม)

ค่ากลาง ได้แก่ ค่าเฉลี่ย มัธยฐาน และฐานนิยม

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

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

