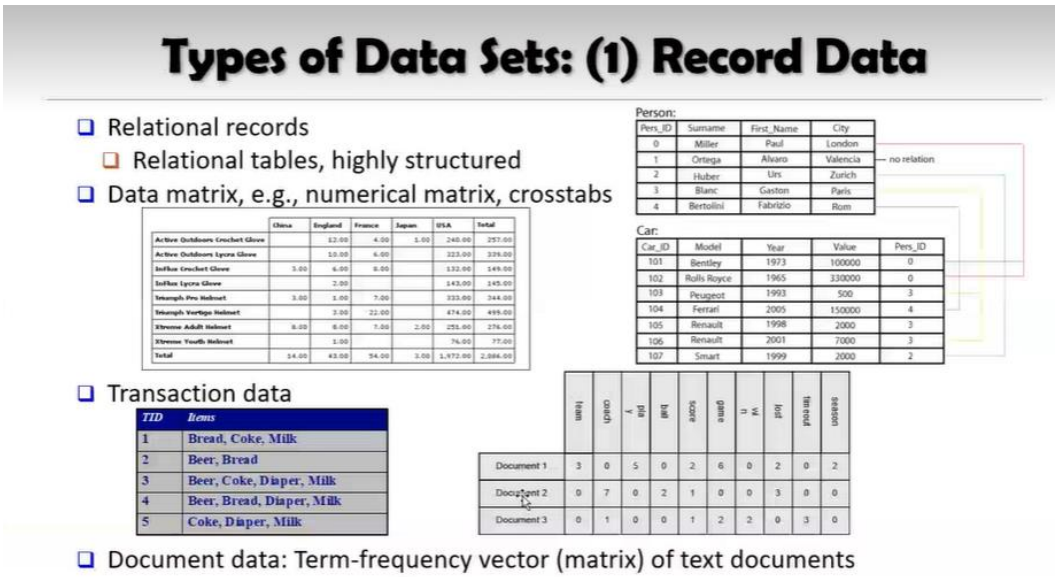
ชื่อ นางสาวชนัญชิดา เมธีกุลมานิต รหัสนักศึกษา 623020516-1

สรุปบทที่2.1

Chapter2 Getting to know Your Data

Data



Data Relational คือดาต้าที่มีหลายตารางและมีความสัมพันธ์กัน

Important Characteristics of Structured Data

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

-มีกี่ Dimension

-สนใจเฉพาะช่องที่มีข้อมูล

-ความแอดอัดในการเก็บข้อมูล เช่นรูปภาพว่ามีขนาดภาพกี่พิกเซล ถ้าเกิดเรสรูต่ำเราจะแทนค่าสีของแต่ละหลายๆพิกเซลด้วย1สี จะได้เปลืองพื้นที่เก็บน้อย

-วัดค่ากลางของข้อมูล (สว่าง/มืด)

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

แนวนอนเรียกว่า data objects

แนวตั้ง เรียกว่า attributes

Data ที่เราได้ ว่าจะเรียกว่า Data sets ก็คือประกอบไปด้วยข้อมูลหลายๆข้อมูล

ดาต้า ออฟ คือข้อมูลแต่ละตัว เช่น ตารางเซลล์ก็จะมีการเก็บข้อมูลของคอสตอเมอร์ เซลล์ เป็นต้น

การเรียกดาต้า

Also called *samples , examples, instances, data points, objects, tuples*

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?

คือ คุณสมบัติที่ใช้อธิบายข้อมูลแต่ละตัว

ชนิดของข้อมูล

-Nominal ข้อมูลที่ไม่มีตัวเลข เช่น สีแดง สีน้ำเงิน

-Binary เช่น มีข้อมูล 2 ค่า เช่น ถูกกับผิด

-Ordinal เช่น ข้อมูลที่มีการเรียงลำดับ เช่น เฟรชชี่ หม่อ จูเนียร์ ซีเนียร์ เป็นต้น

Numeric

Interval บวกลบคูณหารแล้วมีความหมาย

Ratio ความสัมพันธ์ของจำนวนทั้งสองจำนวน

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

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*

ไม่มีศูนย์แท้ เช่น อุณหภูมิ เกรด ความยาว เงิน เป็นต้น

มีศูนย์แท้ เช่น น้ำหนัก ระยะทาง ความสูง ยอดขาย เป็นต้น

Discrete vs. Continuous Attributes

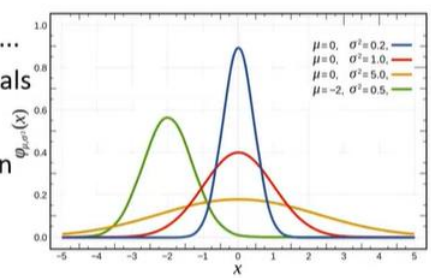
- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - 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

Discrete ความไม่ต่อเนื่อง เช่นจำนวนของสิ่งของ รหัสไปรษณีย์

Continuous มีค่าได้ทุกค่าในช่วงที่กำหนด

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