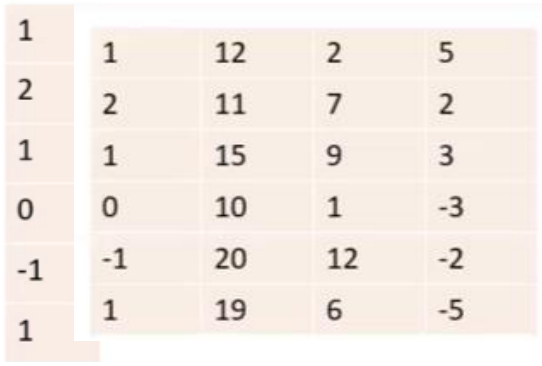
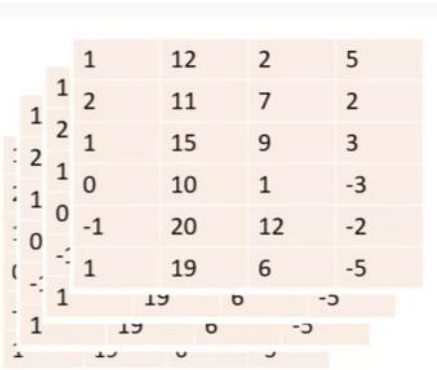
ข้อมูลที่ใช้ในวิชานี้ จะเป็นข้อมูลที่อยู่ในรูปของเมทริกซ์ คือตารางที่มีตัวเลขเรียงกันอยู่ เรียกแถว กับหลัก



1 มิติ

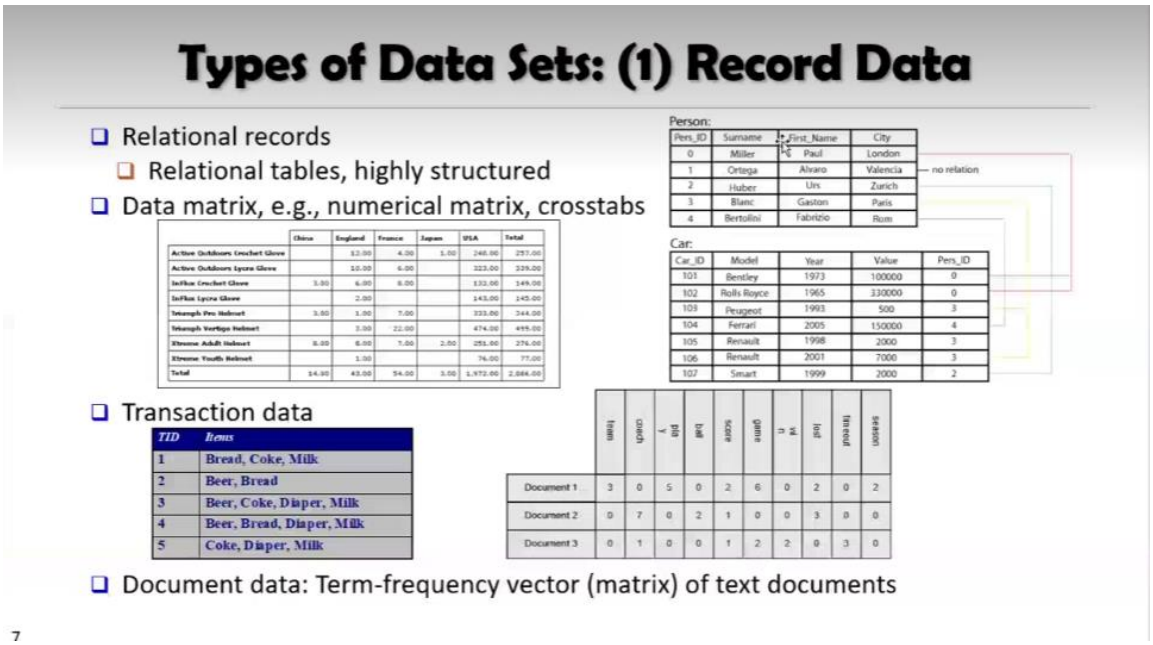


2 มิติ

3 มิติ

แนวนอน Record คือ ข้อมูลแต่ละจุด , แนวตั้ง Attribute คือ คุณสมบัติใช้อธิบายข้อมูลแต่ละจุด

4 มิติ

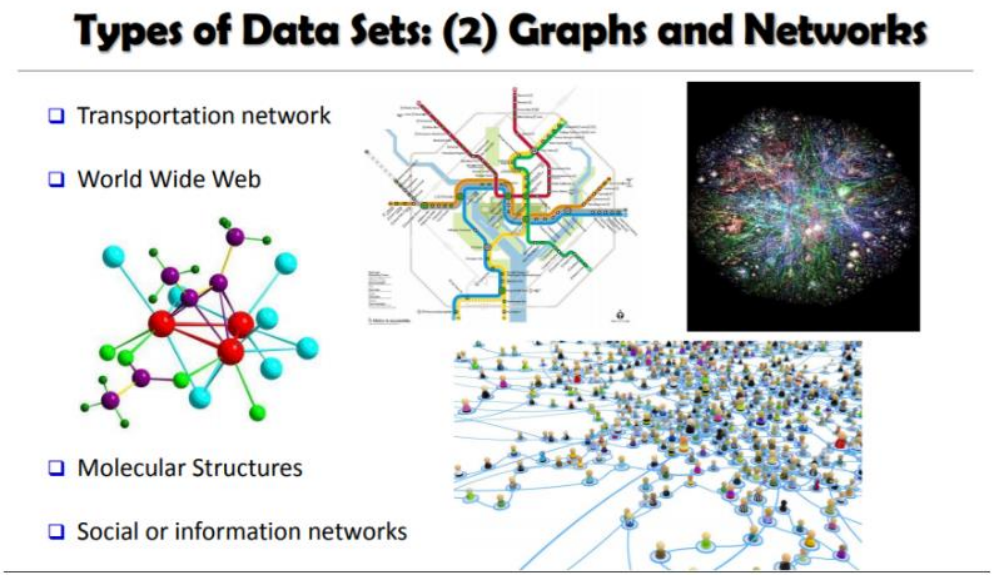


Relational records มีหลายตารางที่มีความสัมพันธ์กัน

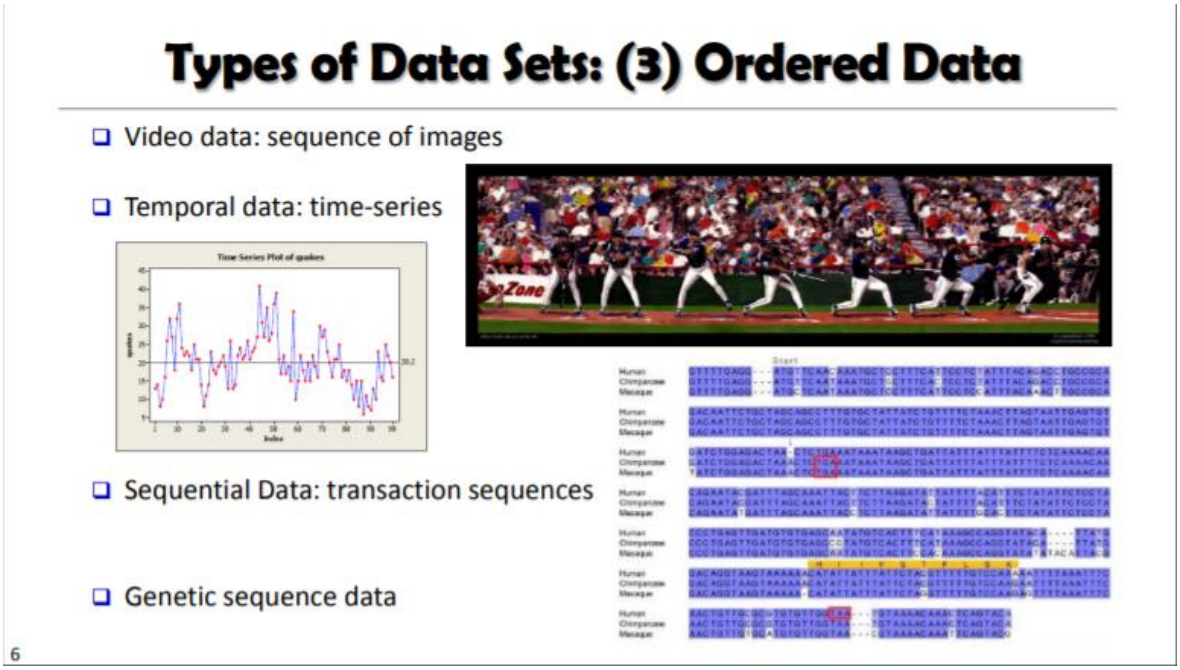
Crosstabs ต้องอ่านค่าทั้งสองฝั่ง

Transaction Data หารูปแบบของข้อมูล

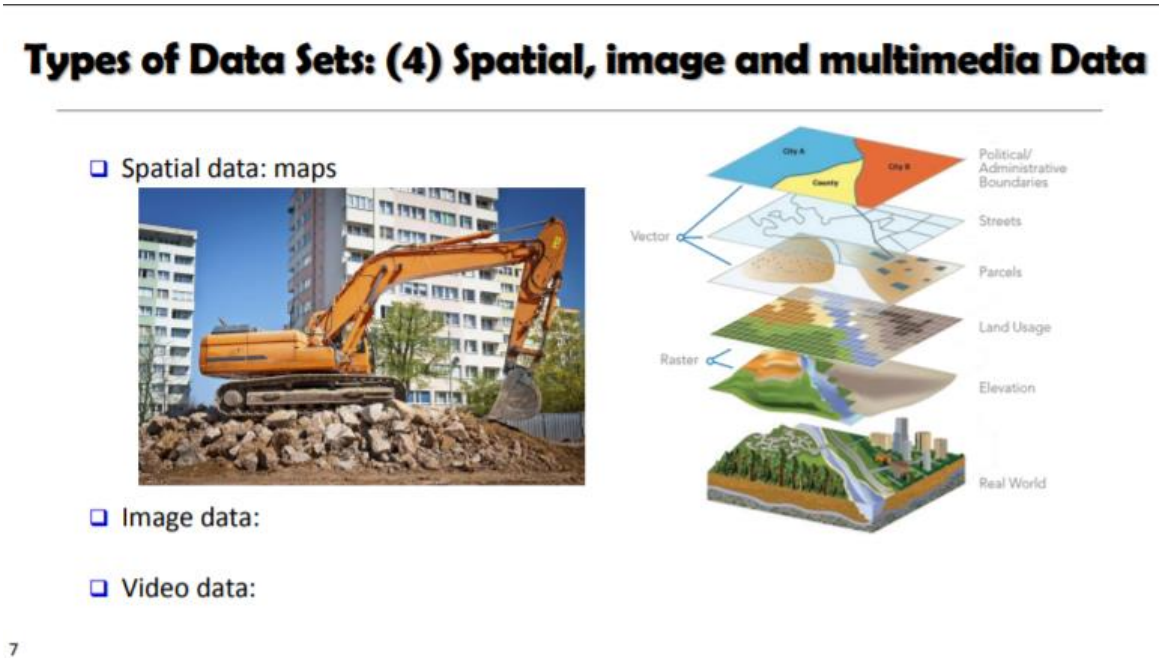
Term – frequency ตารางที่ใช้สรุปข้อมูลที่เป็นข้อความ ให้อยู่ในรูปที่คอมพิวเตอร์สามารถแปลผลได้



ข้อมูลที่เป็นกราฟ จะเป็นการบอกว่า แต่ละจุด เชื่อมโยงกับอะไร มีการเชื่อมต่อกันอย่างไร



เป็นข้อมูลที่มีเวลาเข้ามาเกี่ยวข้อง แปลงข้อมูลเป็นตัวเลข Time – Series Sequential ข้อมูลที่ไม่สามารถสลับตำแหน่งได้

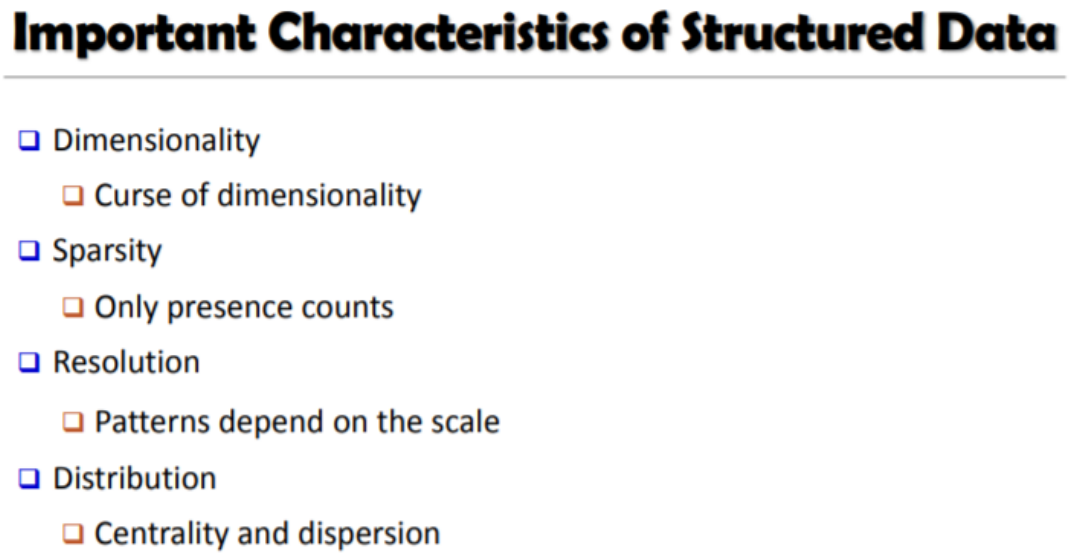


เป็นข้อมูลที่มีเวลาเข้ามาเกี่ยวข้องเช่นกัน

Video Data คือ การเอารูปหลายๆรูปมาซ้อนๆกัน ทำให้เปรียบเสมือนว่าเคลื่อนไหวได้

Image Data คือ มีด้านกว้าง และด้านยาว กำหนดพิกัดเป็น x, y โดยแต่ละตัวจะมีสีบอกจุด

Spatial Data คือ เช่น แผนที่ จะมีการกำหนดจจุดพิกัด x, y โดยจะกำหนดสีให้แต่ละพิกัด



8

คุณลักษณะสำคัญของข้อมูล ได้แก่

ข้อมูลมี Dimension มีกี่ Dimension

สนใจแค่จุดที่มีข้อมูล ตรงที่ไม่มีข้อมูลจะไม่สนใจ

ความแออัดในการเก็บข้อมูล

การกระจายตัวของข้อมูล วัดค่ากลางของมูล

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

9

Data Sets ประกอบไปด้วย Data Object หลายๆ อันมารวมกัน Data Set คือ กลุ่มของข้อมูลหลายๆ ตัวมารวมกัน

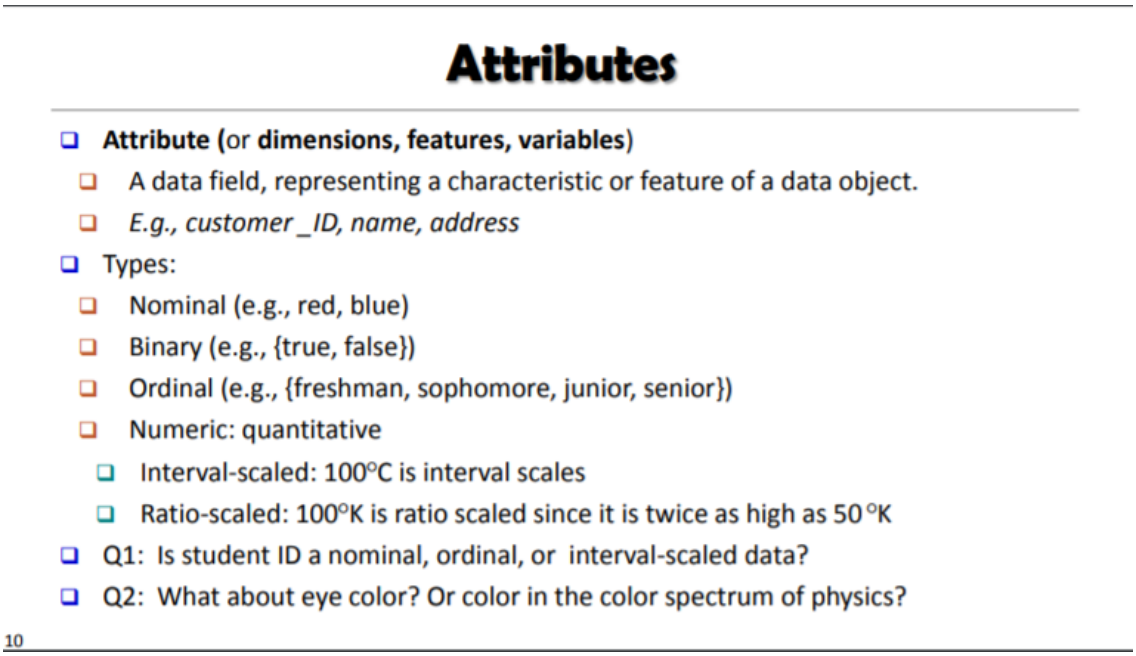
Data Object หมายถึง ข้อมูลแแต่ละตัวประกอบด้วย Entity ของแต่ละอัน

สำคัญ* Also called sample, examples, instances, data points, objects, tuples Record

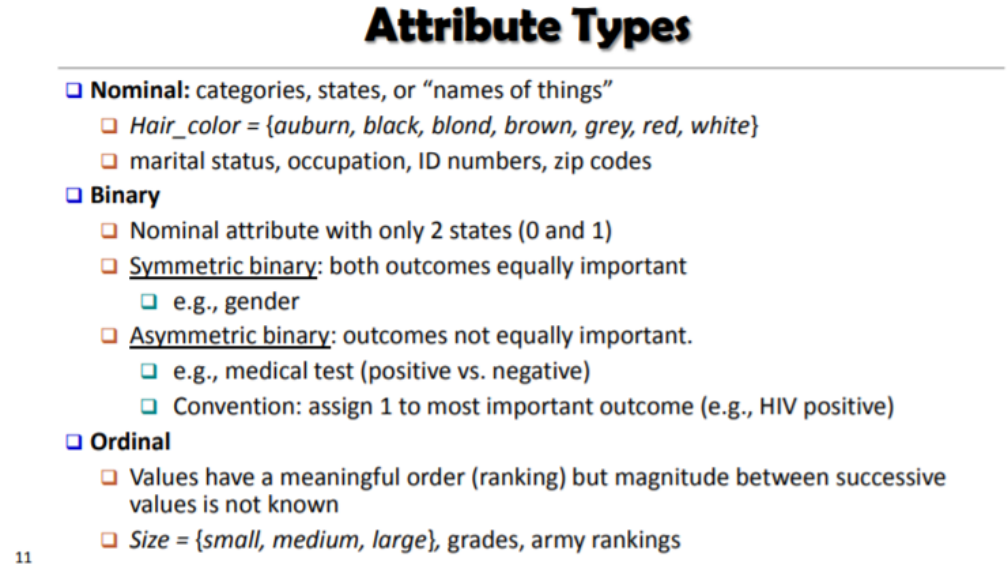
มีความหมายเดียวกันทั้งหมด คือ ข้อมูลที่เป้น Data 1 จุดที่อยู่ในแนวตั้ง Data 1 จุด คือ Data point

ข้อมูลจะถูกอธิบายด้วย Attributes

Database ที่จะนำมาทำคือ Rows คือ Data Object, Column คือ Attributes



คือ คุณสมบัติที่ใช้อธิบายข้อมูลแต่ละตัว อาจจะเรียกว่า Dimension, Features, Variables, field



ชนิดของ Attribute

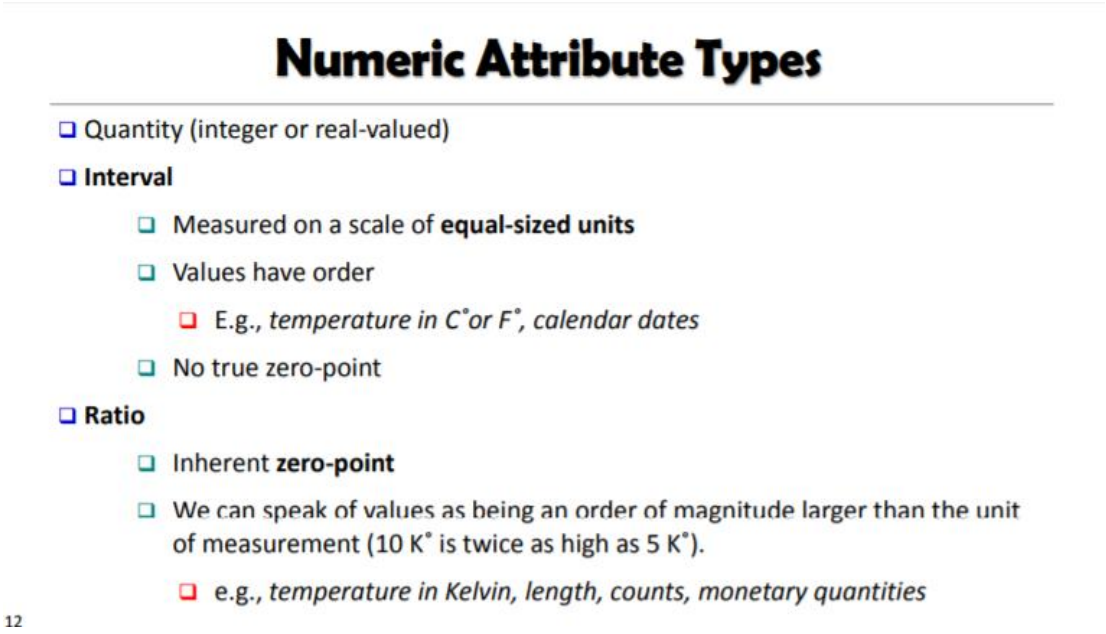
Nominal ข้อมูลที่เป็นการบอกกลุ่ม บอกสถานะ เป็นชื่อของสิ่งของ เช่น สีผม, สถานะการแต่งงาน, อาชีพ, ID

Binary คือ ข้อมูล Nominal เพียงแต่มี 2 สถานะ เช่น Yes/No

Symmetric Binary คือสมมาตรกัน เช่น ชอบกินโค้กหรือเปปซี่ ถนัดซ้ายหรือขวา

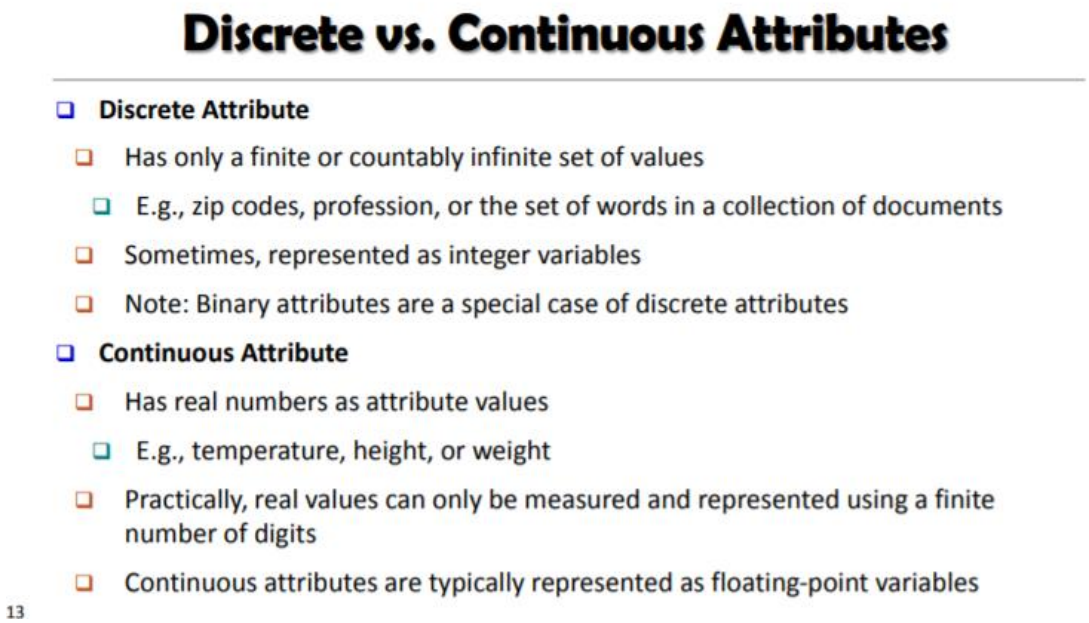
Asymmetric Binary คือไม่สมมาตรกัน เช่น เป็นเอดส์ไหม

Ordinal คือ ข้อมูลที่มีความหมายสามารถนำมาเรียงลำดับได้ แต่จะไม่รู้ว่ามีค่าต่างกันมากน้อยแค่ไหน เช่น ขนาดใหญ่ ขนาดกลาง ขนาดเล็ก เกรด



Interval ข้อมูลที่ไม่มี 0 แท้ เช่น อุณหภูมิ

Ratio ข้อมูลที่มี 0 แท้ เช่น ความยาว เงิน



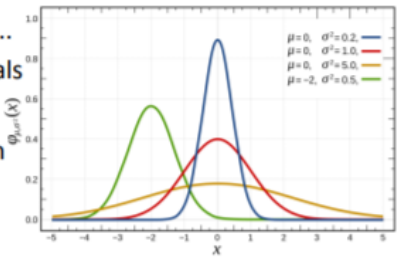
การแบ่ง Attributes อีกแบบ

Discrete Attribute คือ ข้อมูลที่ไม่ต่อเนื่องกัน เช่น 1,2,3 ซึ่งไม่มีค่าอยู่ระหว่าง 1-2 เช่น อาชีพ ครู กับ พยาบาล ก็จะไม่มีอาชีพอื่นอยู่ระหว่าง สองค่านี้

Continuous Attribute คือ ข้อมูลที่ต่อเนื่องกัน ส่วนใหญ่แทนด้วยจำนวนจริง เช่น ส่วนสูง 180,180.5,181

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



15

เป็นการดูค่าทางสถิติ โดยการใช้ค่ากลางของข้อมูลแทน ค่ากลางทางสถิติมีทั้งหมด 3 ชนิด คือ ค่าเฉลี่ย ค่ามัธยฐาน ค่าฐานนิยม

Measuring the Central Tendency: (1) Mean

- Mean (algebraic measure) (sample vs. population):

Note: *n* is sample size and *N* is population size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

- Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Trimmed mean:
  - Chopping extreme values (e.g., Olympics gymnastics score computation)

16