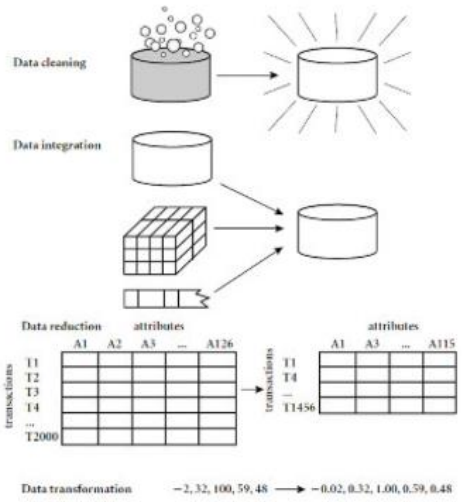
Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary



3

- Data Cleaning เป็นการทำ Cleaning data เนื่องจากเก็บข้อมูลมาหลายแห่ง เช่น แบบฟอร์มที่ให้คนอื่นกรอก แล้วมีการกรอกข้อมูลที่ผิดพลาด
- Data Integration การนำ Data จากหลายๆแหล่งมารวมกัน ลักษณะการรวมอาจจะรวมแบบเป็นตารางเพื่อนำไปทำ Data mining หรือ Data warehouse เพื่อเรียกดูข้อมูลแนวต่างๆได้
- Data Reduction and Transformation การลดจำนวนข้อมูล แปลงข้อมูลอย่างไรให้สามารถนำไปประมวลผลได้
- Dimensionality Reduction เป็นการลดจำนวนข้อมูลในแนวตั้ง

What is Data Preprocessing? — Major Tasks

- **Data cleaning** จัดการ missing กำจัด noisy ,outlier
  - Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration** รวม data จากหลายๆแหล่ง
  - Integration of multiple databases, data cubes, or files
- **Data reduction** การลดจำนวนข้อมูล
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization** ทำการเปลี่ยนแปลงข้อมูลเพื่อให้ข้อมูลเท่ากับข้อมูลอื่นๆ
  - Normalization
  - Concept hierarchy generation

4

Why Preprocess the Data? — Data Quality Issues

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

เนื่องจาก มี data มาจากหลายๆแหล่ง จึงจำเป็นต้องทำ preprocessing

ขั้นตอนที่สำคัญในการทำ preprocessing คือ data cleaning incomplete(ความไม่สมบูรณ์),Noisy,ความไม่สอดคล้อง

5

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data were not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Did not register history or changes of the data
- Missing data may need to be inferred

เช่น ให้น้องปี 1 กรอกข้อมูลทั่วไปในแบบฟอร์ม พอมาปีนี้มีการเพิ่มว่า มีการฉีดวัคซีนแล้วหรือยัง แบบนี้เรียกว่า missing data เพราะเนื่องจากว่าพึ่งให้มากรอกในปีนี้

8

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - **the most probable value: inference-based such as Bayesian formula or decision tree**

ถ้า data ไหนมี missing ก็สามารถลบออกไปได้ แต่หากข้อมูลมีมาก การลบข้อมูลออกไปอาจจะไม่ใช่เรื่องที่ง่าย แต่ก็สามารถทำได้