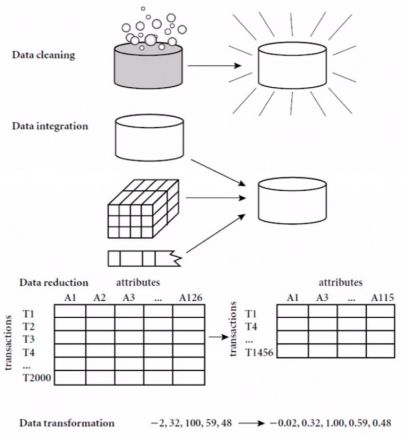
Chapter 3: Data Preprocessing

เป็น 40% ของวิชานี้

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



ตัวแปลง Data ให้เป็นตัวเลข

What is Data Preprocessing? — Major Tasks

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

เปลี่ยนแปลงข้อมูลให้สามารถนำมาประมวลผลได้

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? Data เท่าที่อ่านแล้วน่าเชื่อถือมั้ย
  - Interpretability: how easily the data can be understood?

ถ้าหมอไร้ความรับผิดชอบ ไม่ดูแลทำให้เกิด Inconsistency ได้

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation* = " " (missing data)
  - Noisy: containing noise, errors, or outliers
    - e.g., *Salary* = "−10" (an error)
  - Inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age* = "42", *Birthday* = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

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

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**