CS 412 Intro. to Data Mining

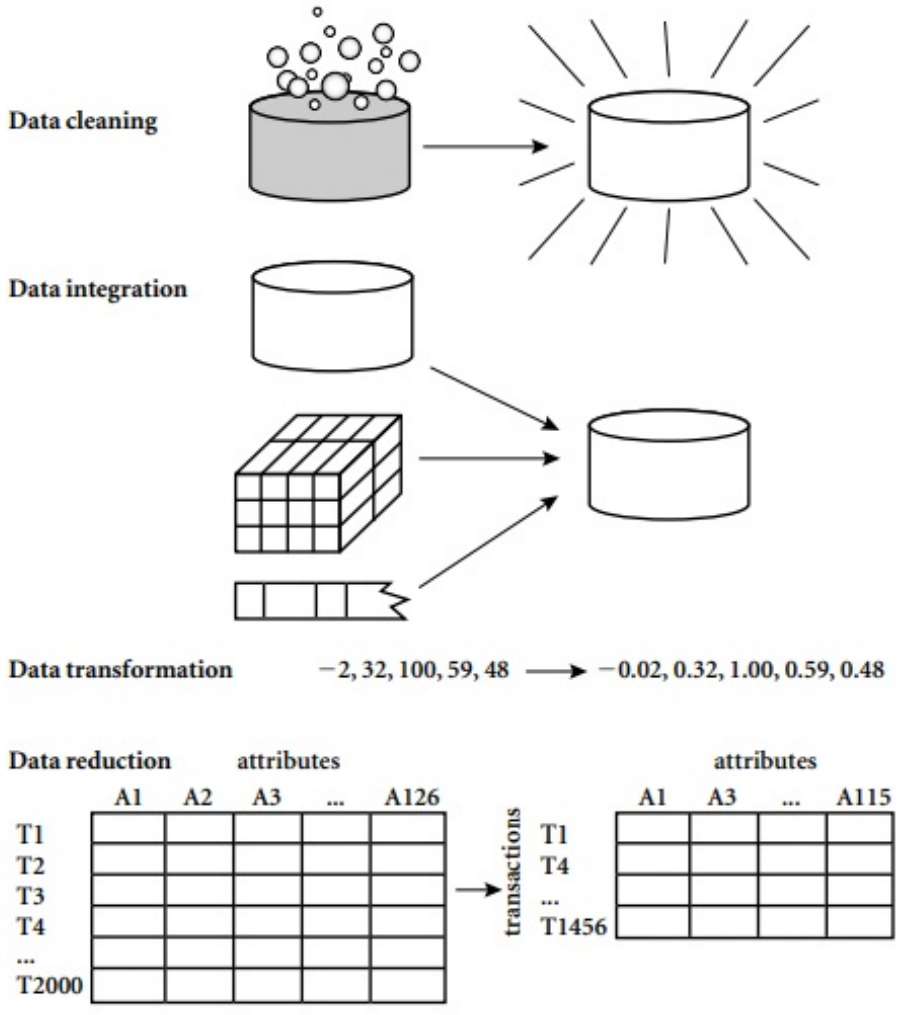
Chapter 3. Data Preprocessing

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

การเตรียมข้อมูล, การประมวลผลข้อมูล

Chapter 3: Data Preprocessing

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



What is Data Preprocessing? — Major Tasks

- **Data cleaning** เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป
  - Handle missing data (จัดการ missing), smooth noisy data (กำจัด noisy), identify or remove outliers, and resolve inconsistencies (จัดการ inconsistencies คือ ตรวจว่าเกิดที่นี่ว่า ตายุอีกที่นึง คือการกรองตายุไม่ควรเป็นปี เช่น พ.ศ. 2544 ตายุ 18 ปี)
- **Data integration** รวมตาต่างเข้าด้วยกัน, รวมข้อมูลหลายแห่งมารวมไว้ที่เดียวกัน
  - Integration of multiple databases, data cubes, or files
- **Data reduction** → ลดมิติข้อมูล
  - Dimensionality reduction เลือกตาต่างที่สำคัญ
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization** → แปลงข้อมูลในขั้นตอนการคัดเลือก ให้เหมาะสมสำหรับขั้นตอนการทำเหมืองข้อมูล
  - Normalization
  - Concept hierarchy generation

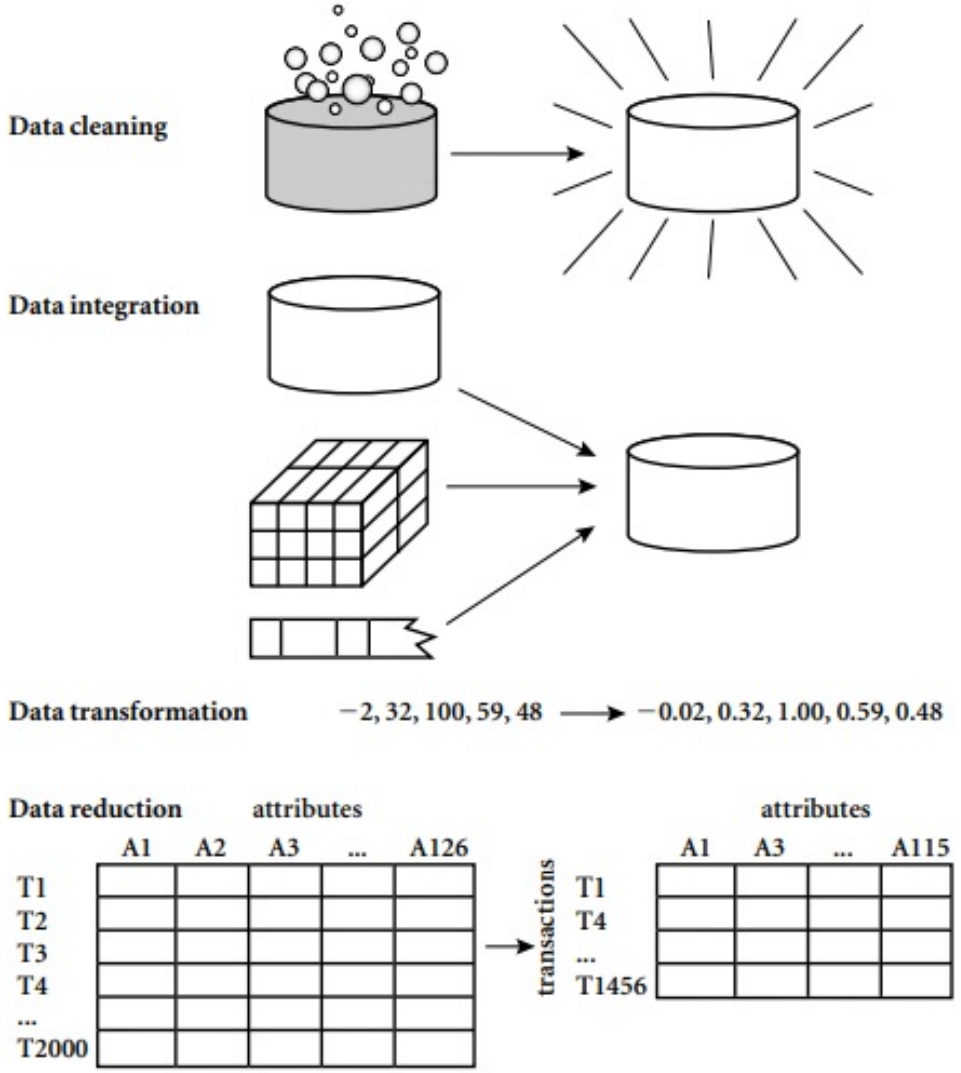
Why Preprocess the Data? — Data Quality Issues

ทำไมต้องทำ Data Preprocess

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not คนกรอกหลายคน , เปลี่ยนเครื่อง
  - Completeness: not recorded, unavailable, … กรอกข้อมูลไม่ครบถ้วน
  - Consistency: some modified but some not, dangling, ... ทำ Normalization ไม่ดี ข้อมูลสิ่งใดอยู่ในหลายที่
  - Timeliness: timely update? ตารางที่กรอกเวลาหนึ่ง ในเวลาต่อมามีการเปลี่ยนแปลง
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Chapter 3: Data Preprocessing

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



Data Cleaning

การทำความสะอาดข้อมูล

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error ข้อมูลที่มันผิดไม่ว่าจากเครื่องคอม, คนกรอก
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data data ไม่สมบูรณ์ ส่วนชะ
    - e.g., *Occupation* = "~" (missing data)
  - Noisy: containing noise, errors, or outliers
    - e.g., *Salary* = "−10" (an error) เช่น เงินเดือนติดลบ
  - Inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age* = "42", *Birthday* = "03/07/2010" อายุกับวันเกิดไม่ตรงกัน
    - Was rating "1, 2, 3", now rating "A, B, C" data เก่ากับใหม่ไม่ตรงกัน
    - 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?

เราจะจัดการกับ missing ยังไง

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably

ลบออกเลย แต่ถ้า data มีเยอะมันจะหายมาก

- Fill in the missing value manually: tedious + infeasible?

เติม ลบออก

- Fill in it automatically with กรอกแทนโดยสมัย classใน
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean เอาค่ากลางมาแทนค่า missing
  - the attribute mean for all samples belonging to the same class: smarter แทนด้วย mean ที่อยู่ใน class เดียวกัน
  - **the most probable value: inference-based such as Bayesian formula or decision tree**

Noisy Data ข้อมูลรบกวน

- **Noise:** random error or variance in a measured variable
- **Incorrect attribute values** may be due to
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention
- **Other data problems**
  - Duplicate records
  - Incomplete data
  - Inconsistent data

How to Handle Noisy Data?

- Binning → ทำให้ค่าข้อมูลที่เรียงลำดับเป็นไปอย่างราบรื่นโดยพิจารณาจากค่าข้อมูลรอบๆ
  - First sort data and partition into (equal-frequency) bins
  - Then one can **smooth by bin means, smooth by bin median, smooth by bin boundaries**, etc.
- Regression → ข้อมูลสามารถปรับให้เรียบได้โดยการปรับข้อมูลให้เหมาะสมกับฟังก์ชัน
  - Smooth by fitting the data into regression functions
- Clustering → ค่าผิดปกติอาจถูกตรวจพบโดยการจัดกลุ่ม โดยที่ค่าที่คล้ายคลึงกันจะถูกจัดเป็นกลุ่ม
  - Detect and remove outliers
- Semi-supervised: Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)