CS 412 Intro. to Data Mining

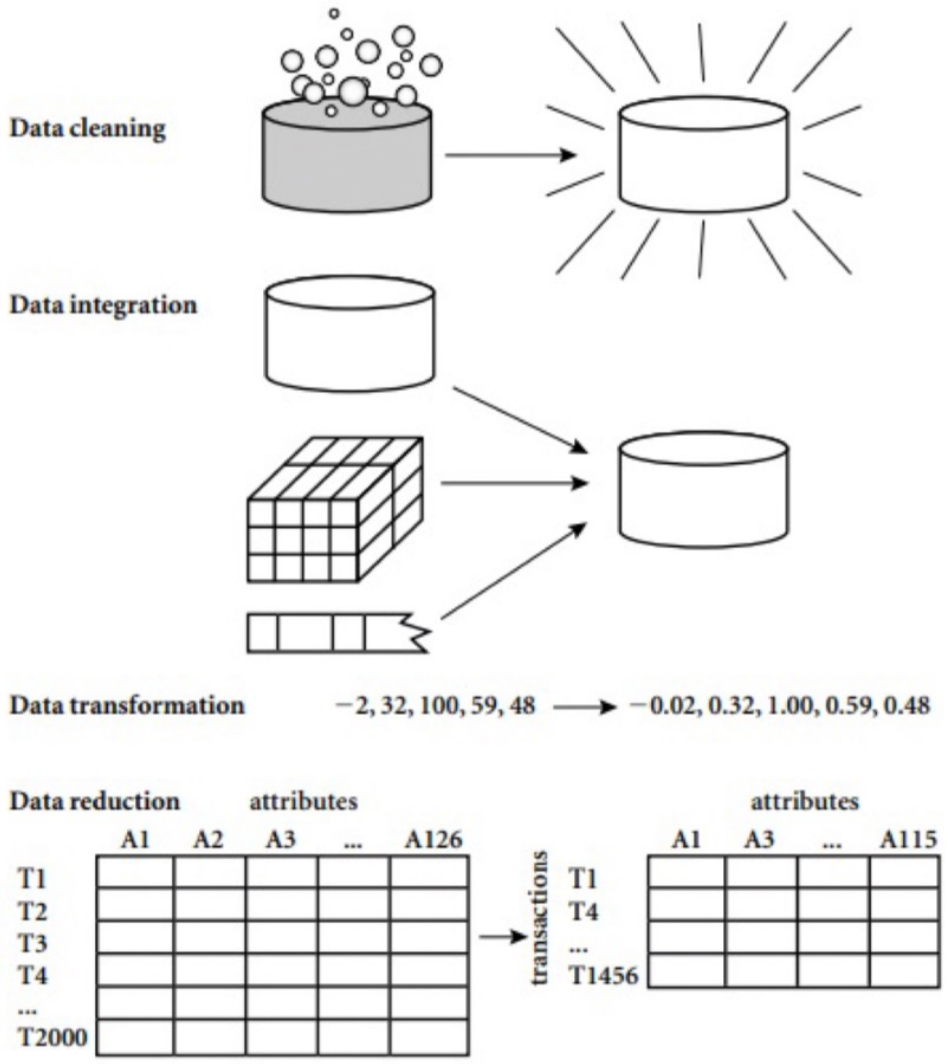
Chapter 3. Data Preprocessing

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

Chapter 3: Data Preprocessing

การเตรียมข้อมูล หรือการประมวลผลข้อมูล

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



What is Data Preprocessing? — Major Tasks

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

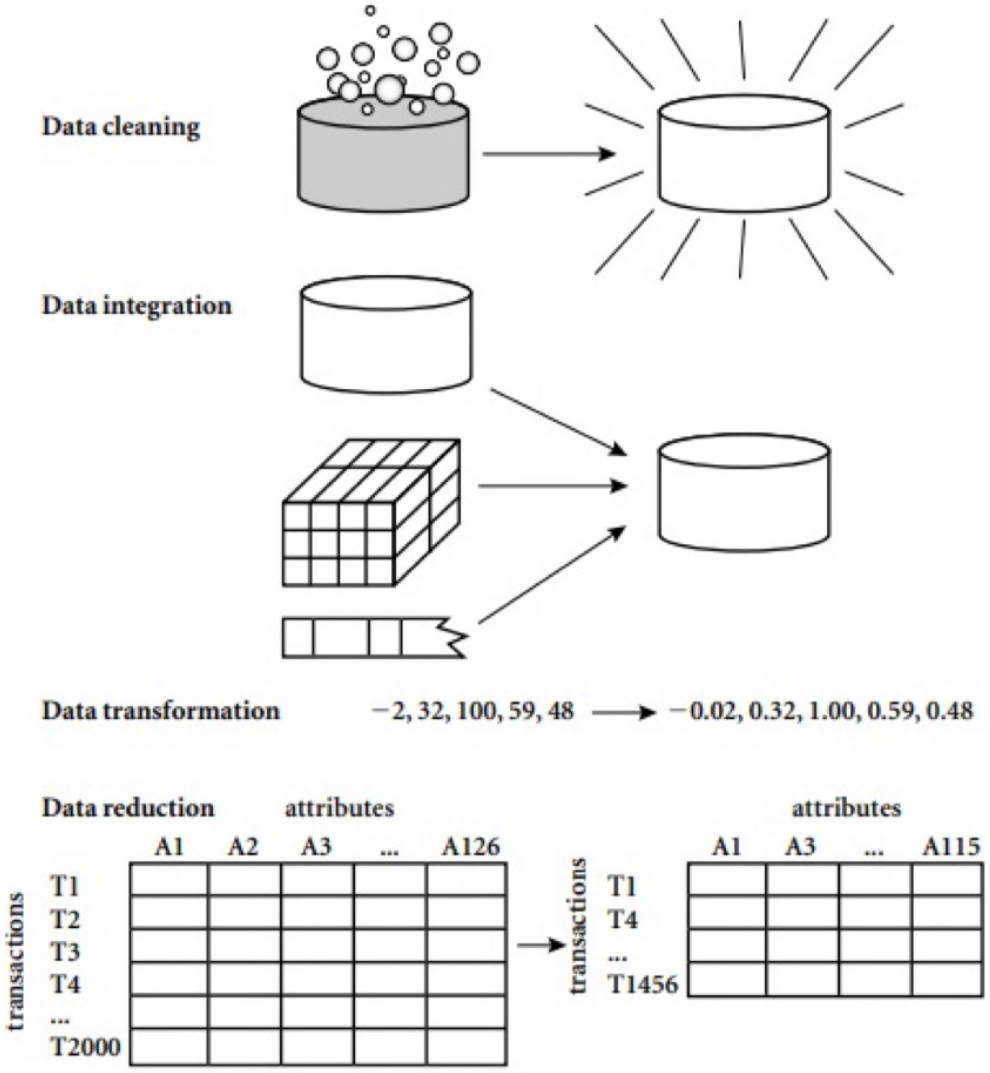
Why Preprocess the Data? — Data Quality Issues

ทำไมต้องทำ Data Preprocess

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not คนกรอกหลายคน, เปลี่ยนเสียง
  - Completeness: not recorded, unavailable, … กรอกข้อมูลไม่ครบถ้วน
  - Consistency: some modified but some not, dangling, … ทำ Normalization ไม่ดีข้อมูลจะไปอยู่ในหลายที่
  - Timeliness: timely update? Data ที่กรอกกลางหนึ่ง ในหลายช่องมีการเปลี่ยนแปลง
  - 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 ข้อมูลที่มีโอกาสผิดพลาดจากเครื่องคอม หรือคนกรอก
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data ข้อมูลไม่รวบรวม
    - e.g., *Occupation* = “ ” (missing data) ว่าง
  - 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” อายุ กับวันเกิด ไม่ตรงกัน
    - Was rating “1, 2, 3”, now rating “A, B, C” data เก่ากับ data ใหม่ ไม่ตรงกัน
    - 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**

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

Noisy Data

ข้อมูลรบกวน คือข้อมูลที่มีความผิดพลาด หรือ คลาดเคลื่อนจากการวัด สาเหตุจาก

- อุปกรณ์เก็บรวบรวมข้อมูลที่มีข้อผิดพลาด
- ปัญหาการป้อนที่เกิดจากมนุษย์
- ปัญหาการส่งข้อมูลผิดพลาด
- ข้อจำกัดทางเทคโนโลยี

How to Handle Noisy Data?

- Binning ทำให้ข้อมูลที่เรียงลำดับเป็นไปอย่างราบรื่น โดยพิจารณาจากค่าที่อยู่รอบๆ
  - 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 ค่าผิดปกติอาจถูกตรวจสอบพบโดยการจัดกลุ่ม โดยที่ค่าที่คล้ายกันจะถูกจัดเป็นกลุ่ม
  - Detect and remove outliers
- Semi-supervised: Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)