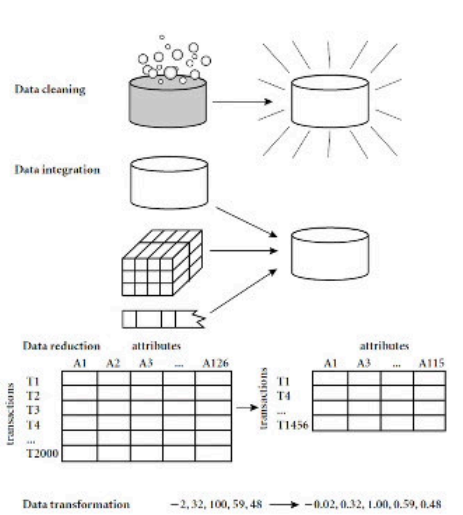
Chapter 3: Data Preprocessing

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



- Data Cleaning เป็นการทำ cleaning Data เนื่องจากเก็บข้อมูลมาหลายแหล่ง
  เช่น แบบฟอร์มให้คนอื่นๆกรอก แล้วกรอกข้อมูลผิด เป็นต้น

- Data Integration การนำ Data จากหลายแหล่งมารวมกัน ลักษณะการรวม อาจจะรวมเป็นตารางเพื่อเอาไปทำ Data Mining หรือเชื่อมรวมต่อไปใน Data Warehouse เพื่อเรียกดูข้อมูลแบบอื่นๆได้

- Data Retraction and transformation การคัดเลือกข้อมูล / เป็นการลดเวลา จากแปลงข้อมูลอย่างไร ให้ประมวลผลได้

- Dimensionality Reduction ลดจำนวนข้อมูลแนวตั้ง

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

- Data Cleaning จัดการ missing, inconsistencies หรือ noisy, outlier
- Data integration รวม Data จากหลาย แหล่งเข้าด้วยกัน database
- Data reduction ทำให้ขนาดเล็กลง
- Data transformation เปลี่ยนข้อมูลให้อยู่ในรูปแบบที่เหมาะสม

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?

- ทำไมต้องทำ Preprocess ก่อน: Data ภายนอกมักจะ ไม่ดี

ข้อมูลที่ไม่ดีในการทำ Preprocessing

Data Cleaning: Incomplete (ไม่ครบถ้วน), Noisy, Inconsistent (ไม่ตรงกันหรือซ้ำซ้อน), Intentional

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 เราให้น้องปี 1 กรอกข้อมูลทั่วไปในแบบฟอร์ม พอสักปีนึง เราพึ่งได้มาบอกว่า ไม่ได้รับรหัสนักเรียนเรียงขึ้น แบบนี้เรียกว่า Missing Data เพราะฟังไม่ทันตอนกรอกในปีนั้น เป็นต้น

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 ในนมี Missing ก็ไม่ลบออกไป แต่ถ้าข้อมูลในนมีเป็นนมีใหญ่ อาจไม่ส่งผลในการวิจัย แต่เราก็สามารถ
ที่ใช้ค่าเฉลี่ยแทน