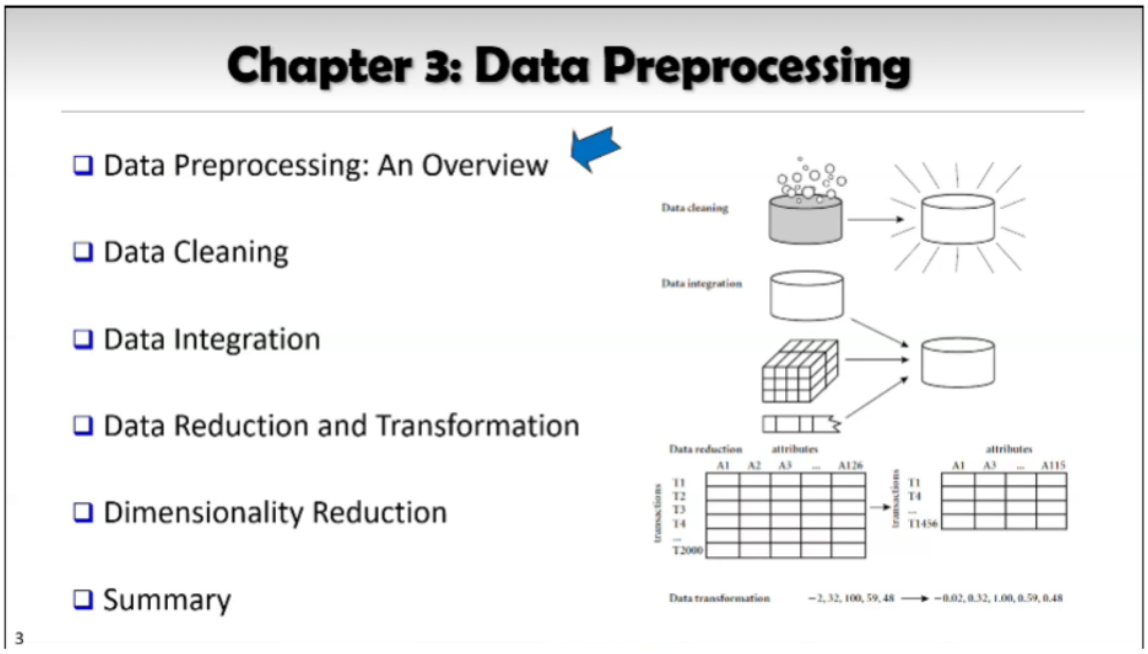
Chapter 3 Data Preprocessing (40% คือบทนี้ ตั้งใจเรียนด้วยนะจ๊ะ)



- Data Cleaning คือ ทำการ cleaning Data เนื่องจากเก็บข้อมูลมาหลายแหล่ง เช่น แบบฟอร์มให้คนอื่นกรอก แล้วเกิดเป็น noise คือ กรอกข้อมูลผิด เป็นต้น

- Data Integration คือ การนำ Data จาก หลาย แหล่งมารวมกัน ลักษณะการรวม อาจจะรวมเป็นตารางเพื่อนำไปทำ Data Mining ต่อ หรือรวมเพื่อเป็น Data Warehouse เพื่อเรียกดูข้อมูลแนวต่างๆได้

- Data Reduction and Transformation คือ การลดจำนวนข้อมูล / ลักษณะแปลงข้อมูลเข้าไปให้ประมวลผลได้

- Dimensionality Reduction คือการ ลดจำนวนข้อมูลหนึ่งตัว

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
/ กำจัด noisy, outliers

Data integration รวม Data จากหลายๆแหล่ง ไม่จำเป็น
ว่ามาจาก database

Data reduction ทำลดจำนวนข้อมูล

Data transformation เปลี่ยนแปลงข้อมูลเพื่อให้ เท่ากัน เพื่อนๆ

Why Preprocess the Data? — Data Quality Issu

- 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 ?

เพราะ Data มาจากหลายๆแหล่ง

ขั้นตอนสำคัญในการทำ Preprocessing

Data Cleaning : Incomplete (ไม่สมบูรณ์).

Noisy, Inconsistent (ไม่ลงตลงกัน),

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 กรอกข้อมูล ทั่วไปในแบบฟอร์ม พอมาถึงนี่ เราพึ่งรู้เขากรอกว่า ได้รับวัคซีนหรือยัง แบบนั้นเรียกว่า Missing Data เพราะพึ่งให้เขามากรอกในปีนี้ 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 ไหนมี Missing ก็ให้ลบมันออกไป แต่ถ้าข้อมูลไหนมีเป็นหมื่น ๆ อาจไม่ดีเท่าไหร่ ถ้าใช้วิธีนี้ แต่เราก็สามารถทำได้ ถ้าเราอยากทำ