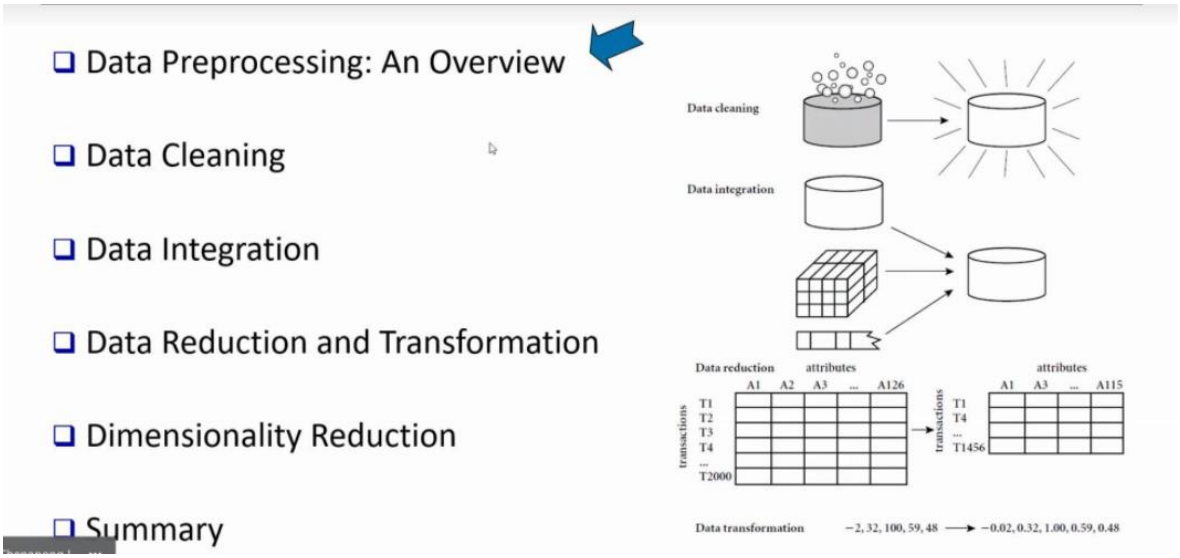
ชื่อ นางสาวชนัญชิดา เมธีกุลมานิต รหัสนักศึกษา 623020516-1

Chapter3 Data Preprocessing (การจัดการเดต้าก่อนที่จะไปประมวณผล)



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 integration

-การรวมดาต้ามาจากหลายๆแหล่ง ให้เป็นข้อมูลที่เข้ากันสามารถนำมาประมวลผลได้

Data transformation and reduction

-การลดจำนวนของข้อมูล

Why Preprocess the Data?

- เพราะว่าข้อมูลมีทั้งถูกและผิด

-มีความสมบูรณ์ของข้อมูลไม่ครบ เช่น มีข้อมูลเดียวกันให้กรอกหลายๆที่

-ดาต้าที่นำเอามาประมวลผลคือ กรอก ณ เวลานั้น พอเวลาเปลี่ยนข้อมูลอาจมีความเปลี่ยนไป

Data Cleaning

-เกิดจากข้อมูลที่ใส่มาผิดพลาด

-กรอกข้อมูลใส่ผิด หรือ เกิดจาก คอมพิวเตอร์เออเร่อได้

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?

Missing data กรอกดาต้าไม่สมบูรณ์

inconsistencies เช่น การกรอกอายุกับวันเดือนปีเกิดไม่ตรงกัน

Missing Data ดาต้าไม่สมบูรณ์

-เกิดจากเครื่องเสีย เช่น การวัดค่าของเครื่องสองเครื่อง ไม่ตรงกัน

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**

วิธีการจัดการข้อมูลที่หายไป

-การกรอกข้อมูลที่หายไปด้วยตัวเอง เช่น เป็นไปได้ หรือ เป็นไปไม่ได้

-ถ้า Data ไหน มี missing ควรจะทำการลบออกไป