



Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview 

- Data Quality
- Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Major Tasks in Data Preprocessing

- **Data cleaning**
 - Fill in missing values, 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 Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
- Accuracy: correct or wrong, accurate or not *ความถูกต้อง*
- Completeness: not recorded, unavailable, ... *ความสมบูรณ์*
- Consistency: some modified but some not, dangling, ... *ไม่ซ้ำซ้อนกัน*
- Timeliness: timely update? *ความทันสมัย*
- Believability: how trustable the data are correct? *ความน่าเชื่อถือ*
- Interpretability: how easily the data can be understood?

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Cleaning ทำความสะอาดข้อมูล

ข้อมูลสกปรก, บั๊ก, ผิด

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data ไม่สมบูรณ์
- e.g., *Occupation*=" " (missing data)
- noisy: containing noise, errors, or outlier- กรรมาผิด
- e.g., *Salary*="−10" (an error)
- inconsistent: containing discrepancies in codes or names. e.g., ไม่เหมือนกัน
- *Age*="42", *Birthday*="03/07/2010" → ข้อมูลไม่สอดคล้องกัน
- 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? default

ค่าที่ไม่สมบูรณ์

Incomplete (Missing) Data

เกิดจากทั้งบั๊กโปรแกรม ค่าใช้ไม่สมบูรณ์

- 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 เครื่องเสีย
 - inconsistent with other recorded data and thus deleted ไม่สอดคล้องกัน
 - data not entered due to misunderstanding
 - certain data may not be considered important at the time of entry
 - 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: ^{น่าเบื่อ}tedious + ^{ไปคำนวณแทนไม่ได้}infeasible?
- Fill in it automatically with
 - a global constant : e.g., “unknown”, ^{สร้าง class ใหม่}a new class?!
 - the attribute mean
 - the attribute mean ^{แทนด้วย mean แต่ต้องเป็นที้อยู่ในคลาสเดียวกัน}for all samples belonging to the same class: smarter
- the most probable value: inference-based such as Bayesian formula or decision tree