数据预处理

- ※ 关于数据
- ➣ 为什么要预处理数据?
- 🌣 描述性数据汇总
- 🌣 数据清理
- >> 数据集成和变换
- 🌣 数据规约
- 🛚 小结

为什么要进行数据预处理

- △ 真实世界的数据太杂乱
 - ✓ **Incomplete**(不完整):不完整:缺少属性值,缺少感兴趣对象的确切属性,或者只有汇总数据
 - e.g., occupation=""
 - ✓ Noisy (有噪音):有错误或者是离散点
 - e.g., Salary="-10"
 - ✓ Inconsistent (不一致):编码或者名称有冲突:
 - e.g., Age="42" Birthday="03/07/1997"
 - e.g., Was rating "1,2,3", now rating "A, B, C"
 - e.g.,重复记录中的冲突

为什么数据预处理很重要?

- ≥ 没有高质量的数据,就没有高质量的挖掘结果
 - ✓ 有质量的决定必须基于有质量的数据
 - e.g., 重复的或者遗漏的数据可能导致不真实的甚至误导性的统计结果
 - ✓ 数据仓库需要一致的高质量的数据集成
- ≥ 数据抽取,清理和转换构成了建造数据仓库的大部分工作

数据预处理的主要任务

> 数据清理

✓ 填充缺失数据,平滑有噪音的数据,确认或者去除离散点,解决不一致问题

> 数据集成

✓ 多个数据库,多维数据,或者是文档的整合

> 数据转换

✓ 归一化与聚合

☎ 数据规约

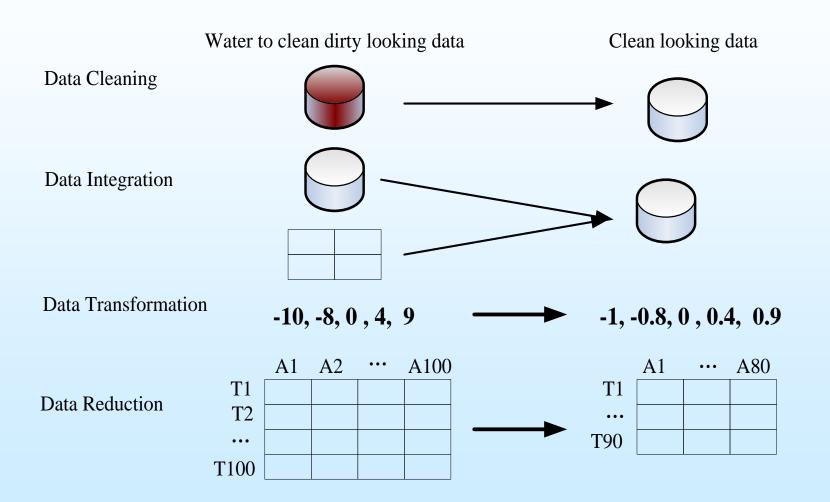
✓ 获得容量上简化的表示法,但是产生相同或者相似的分析结果

≥ 数据离散化

✓ 数据规约的一部分但却有相当的重要性,特别是对于数值的数据

数据预处理的形式

≥ 关键步骤: 掌控并理解数据



5