

- Chapter 1 Introduction
  - 大数据定义
  - 大数据特征
  - 数据挖掘定义
  - KDD流程
  - 数据挖掘主要任务
  - 数据挖掘的挑战

## Chapter 1 Introduction

---

### 大数据定义

🔖 是指大容量的、难以用传统的数据库和软件技术处理的数据

### 大数据特征

🔖 Big Volume——大容量

🔖 Big Velocity——流速、流量(数据流)

🔖 Big Variety——多样性(数据形式)

🔖 Big Veracity——不确定性

### 数据挖掘定义

🔖 从大量的数据中挖掘那些令人感兴趣的、有用的、隐含的、先前未知的和可能有用的模式或知识

### KDD流程

🔖 KDD(Knowledge Discovery in Database) 知识发现，是数据挖掘更广义的说法

1. 数据清理
2. 数据集成
3. 数据选择
4. 数据变换
5. 数据挖掘(核心)
6. 模式评估
7. 知识表示

### 数据挖掘主要任务

🔖 关联分析(关联规则挖掘/亚组发现)

🔖 聚类分析

最大化类内相似性，最小化类间相似性

🔖 分类/预测

📁 孤立点(离群点)分析

## 数据挖掘的挑战

- 数据容量
- 数据实时性
- 数据多样性
- 数据不确定性