

# 引言

## 什么是数据挖掘

---

DB派：

从海量数据中提取有兴趣的模式或知识；

数据库中的知识发现（Knowledge Discovery in Databases, KDD）

## 为什么需要数据挖掘

---

数据爆炸问题：数据自动获取技术的不断发展，导致了数据呈指数级增长。

## 针对什么样的数据

---

- 关系数据库（Relational database）
- 数据仓库（Data warehouse）
- 事务数据库（Transactional database）
- 高级数据库
  - 空间数据库（Spatial data）
  - 时间序列数据（Time-series data）
  - 多媒体数据库（Multimedia database）
  - 文本数据库与WWW（Text databases & WWW）

## 数据挖掘功能

---

- 分类与回归分析(Classification & Regression)  
对类或者概念构造模型或函数以便对未来数据进行预测表示: 决策树, 支持向量机, 概率图模型, 深度网络等。
- 聚类分析(Clustering)  
类标识符是未知的: 把数据分成不同的组。使得同一组中的元素具有极大的相似性, 不同组元素的相似性极小。
- 关联分析(Association, Correlation)  
通过数据分析事物之间的关联性。
- 强化学习(Reinforcement Learning)
- 网络分析（Network Analysis）  
链接分析与Ranking；网络传播影响力分析。

- 数据摘要 (Data Summarization)  
摘要(Summarization), 如文本摘要、视频摘要等