# 数据库系统概论新技术篇

# 大数据思维和方法

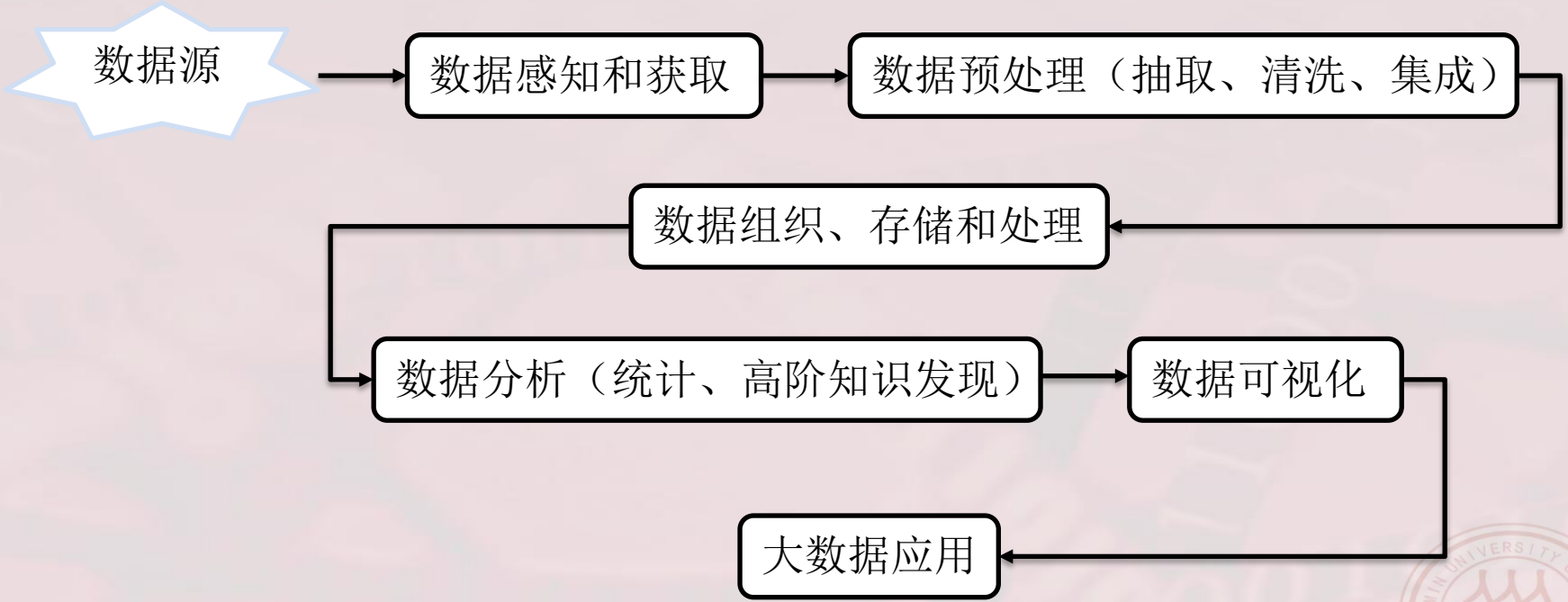
文继荣

中国人民大学信息学院

2017年4月

# 大数据管理的生命周期

# 大数据的生命周期


数据源
数据感知和获取
数据预处理（抽取、清洗、集成）
数据组织、存储和处理
数据分析（统计、高阶知识发现）
数据可视化
大数据应用

# 需求分析

- 从问题出发
- 确定需要的数据类型和来源
- 计算样本空间的大小

# 数据感知和获取

- 目标
  - 收集到尽可能多（或足够多）的相关数据来覆盖样本空间
- 技术
  - 互联网数据采集
  - 物联网数据采集
  - 物理世界感知
- 挑战
  - 大数据体量大（Volume）、速度快（Velocity）

# 数据预处理

- 目标
  - 将来自多个数据源的数据整合成高质量数据
- 技术
  - 数据抽取
  - 数据清洗
  - 数据集成
- 挑战
  - 大数据模式多样（Variety）、真伪难辨（Veracity）

# 数据组织、存储和处理

- 目标
  - 根据数据分析和应用的需求将数据适当地组织、存储并提供处理平台
- 技术
  - 批处理系统
  - NoSQL数据库
  - 流数据处理系统
  - 大数据分析系统
- 挑战
  - 大数据体量大（Volume）、速度快（Velocity）、模式多样（Variety）

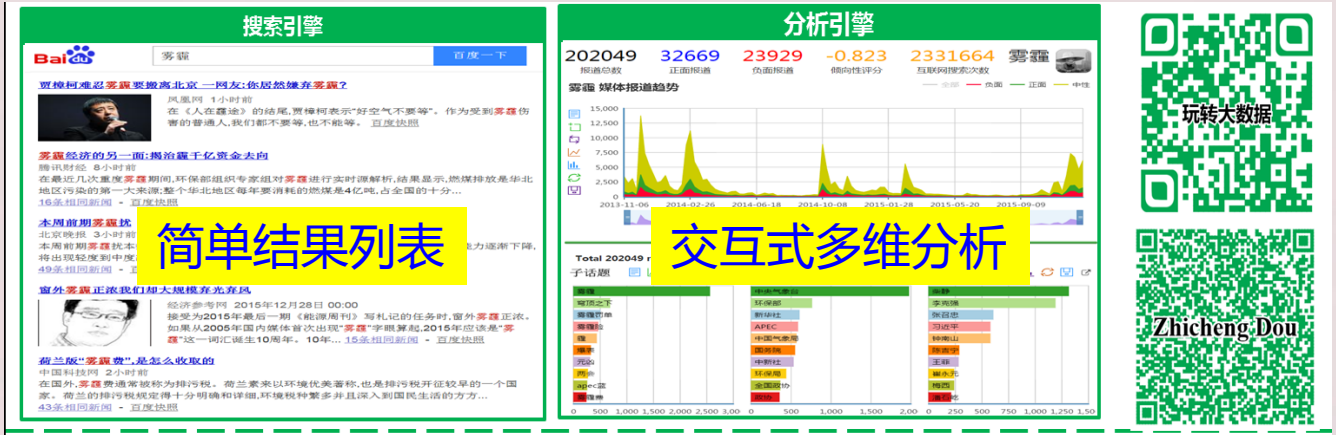
# 数据分析

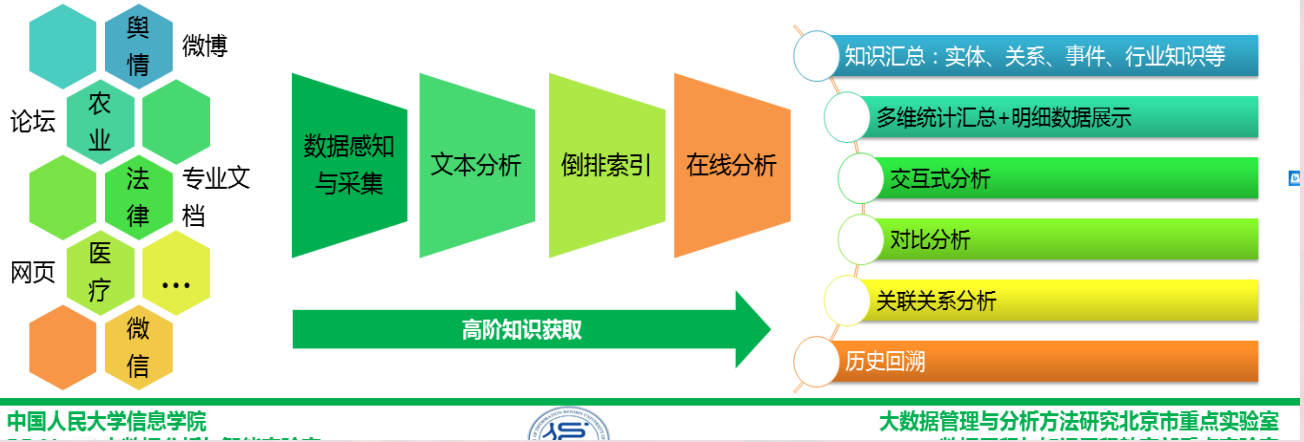
- 目标
  - 从数据中发现规律、知识和价值
- 技术
  - 统计分析
  - 数据挖掘和知识发现
  - 机器学习
  - 高效大数据算法
- 挑战
  - 大数据体量大（**Volume**）、速度快（**Velocity**）、模式多样（**Variety**）

# 数据可视化

- 目标
  - 借助于图形化手段，实现对于海量复杂数据的深入洞察
- 技术
  - 海量数据可视化
  - 高维数据可视化
  - 复杂数据可视化
  - 交互式可视化分析
- 挑战
  - 大数据体量大（**Volume**）、速度快（**Velocity**）、模式多样（**Variety**）

# 例子：互联网分析引擎


搜索引擎
简单结果列表
分析引擎
交互式多维分析
玩转大数据
Zhicheng Dou


舆情
微博
论坛
农业
专业文档
法律
网页
医疗
…
微信
数据感知与采集
文本分析
倒排索引
在线分析
高阶知识获取
知识汇总：实体、关系、事件、行业知识等
多维统计汇总+明细数据展示
交互式分析
对比分析
关联关系分析
历史回溯

# 谢谢!