数据挖掘及应用

莫同
motong@ss.pku.edu.cn

北京大学

第1讲 数据挖掘概述

莫同
motong@ss.pku.edu.cn

北京大学

内容提要

1. 课程简介
2. 数据分析概述
3. 基本方法
4. 数据挖掘方法论

内容提要

1. 课程简介
2. 数据分析概述
3. 基本方法
4. 数据挖掘方法论

课程简介

- 任课教师

Name: 莫同

Work Unit: 北京大学软件与微电子学院

Phone: +86-10-61273661

微博： moris_pku

微信： motong_pku

Fax: +86-10-61273670

E-mail: motong@ss.pku.edu.cn

课程简介

- 课程目标
  - 数据挖掘的基本方法与基本流程
  - 数据挖掘基本理论与基本算法

通过数据挖掘等方法进行数据分析
以解决实际业务问题

课程简介

- 相关课程
  - 高等数学
  - 概率论/统计学/运筹学
  - 数据结构
  - 算法分析与设计
  - 数据库

课程简介

- 课程基本信息：
  - 名称：数据挖掘及应用
  - 面向对象：研究生、高年级本科生
  - 课程类别：专业课（专业核心课）
  - 学时：48学时，3学时一次，共16次
  - 教学方法：课堂授课+实验项目
  - 成绩评估：平时成绩10%+大作业成绩40%+考试成绩50%

实验目标

- 锻炼通过数据挖掘方法解决实际问题的能力
- 加深对数据挖掘技术和经典挖掘算法的理解

实验说明

- 以组队形式
  - 5人一组，5人为上限，可以少于5人甚至自己组队
- 针对实际数据
  - 中国裁判文书网的行政案件数据
  - 政府领导干部数据
  - 自行获取，整理
  - 数据的结构化及存储形式
  - 锻炼数据获取能力
  - 组间共享，注意与合作伙伴的分工

数据说明

- 中国裁判文书网数据
  - 行政案件
  - 判决书
  - 案由为行政复议

- 政府领导干部简历数据
  - 获取渠道为政府网站及公开平台（如百度等）
  - 基本信息（性别、年龄、籍贯、党派、学历等）
  - 履历信息（何时何地在何处担任何职位）

问题说明

- 中国裁判文书网数据分析
  - 何种情况容易导致行政复议及诉讼
  - 何种情况容易导致政府在行政复议及诉讼中败诉

- 政府领导干部简历数据分析
  - 何种干部具有较高的培养潜力（提拔的速度快）

验收方式

- 项目报告(word/PDF/...)
  - 时间：考试前提交
  - 项目报告内容：
    - 使用了什么数据，数据源，数据解读，数据量
    - 要解决什么问题
    - 如何解决的问题，采用了哪些方法，具体如何做的
    - 效果分析（尤其是与数据量的关系，至少3次实验）
    - 结果展示
    - 分析总结
  - 电子版即可，无需打印

考核标准

- 合适的方法
- 清晰的解决思路和过程
- 恰当的结果展示
- 表达和报告

内容提要

1. 课程简介
2. 数据分析概述
3. 基本方法
4. 数据挖掘方法论

数据分析概述

- 今年，你抢红包了吗？

- 那么，问题来了
  - 通过微信红包数据我们能分析出什么？

北京大学

数据分析概述

- 店小二江湖传闻
  - 数据分析师是业界急需人才
  - 高大上职业
  - 多金
  - 受领(mei)导(zhi)关注

- 如何训练自己成为数据分析人员?

北京大学

数据分析概述

- 小测验：你的数据分析技能掌握的怎么样？

1、针对业务目标，需要分析哪些数据？

2、采用什么分析方法？

4、如何改进得到更好的结果？

3、如何确定分析结果的正确性和有效性？

数据分析概述

- 课堂学习与企业实践的区别
  - 算法很容易学，课堂教学+书本资料+网络教程
  - 学会≠会用
  - 企业实践以目标为导向，方法不限

给你一堆数据，能做什么	给你一个目标，如何提高
理解业务，理解数据	方法的选择
设定目标，确定方案	方法的改进（算法、参数）
实现目标，持续改进	结果的解读

数据分析概述

- 案例：
  - 张三是一个流失用户
  - 李四、王五谁更可能流失？

姓名	性别	年龄	身高	体重	月收入	学历	籍贯
张三	男	23	170cm	90kg	2.8万	硕士	贵州
李四	女	24	170cm	63kg	1.3万	本科	北京
王五	男	35	185cm	88kg	1.0万	硕士	辽宁

  - 怎么做？
    - 什么方法？分类、聚类、频繁模式挖掘？
    - 怎么改进？效果不好的时候怎么办？

数据分析概述

- 课程目标

传统方式

- 掌握一堆算法、会考试不会用
- 会算法，不会改进
- 只会做作业，不会出作业
- 会什么用什么，无法给出新思路

我们的课

- 什么情况下该用什么方法
- 深入了解算法的思路，知道改进方向
- 会做命题作文，也会自拟题目
- 触类旁通，掌握创新套路

数据分析概述

- 必备技巧
  - 数学！数学！数学！
  - IT基础

- 方法是重点，实现手段容易！
  - Excel
  - KNIME、SPSS、WEKA
  - Spark ML
  - Python、R、GO
  - …

数据分析概述

- 为什么现在数据分析这么火?
- 大数据 vs. 大忽悠
- 在大数据火起来之前，数据分析火不火?

- 数据分析是支撑业界生产经营活动的必备手段
  - 了解现状
  - 预测未来
  - 模拟
  - …

数据分析概述

- 大数据对数据分析的贡献
  - 更加准确
  - 大数据增加了更多以前被忽略了的相关因素
  - 剔除了更多的“不利”影响

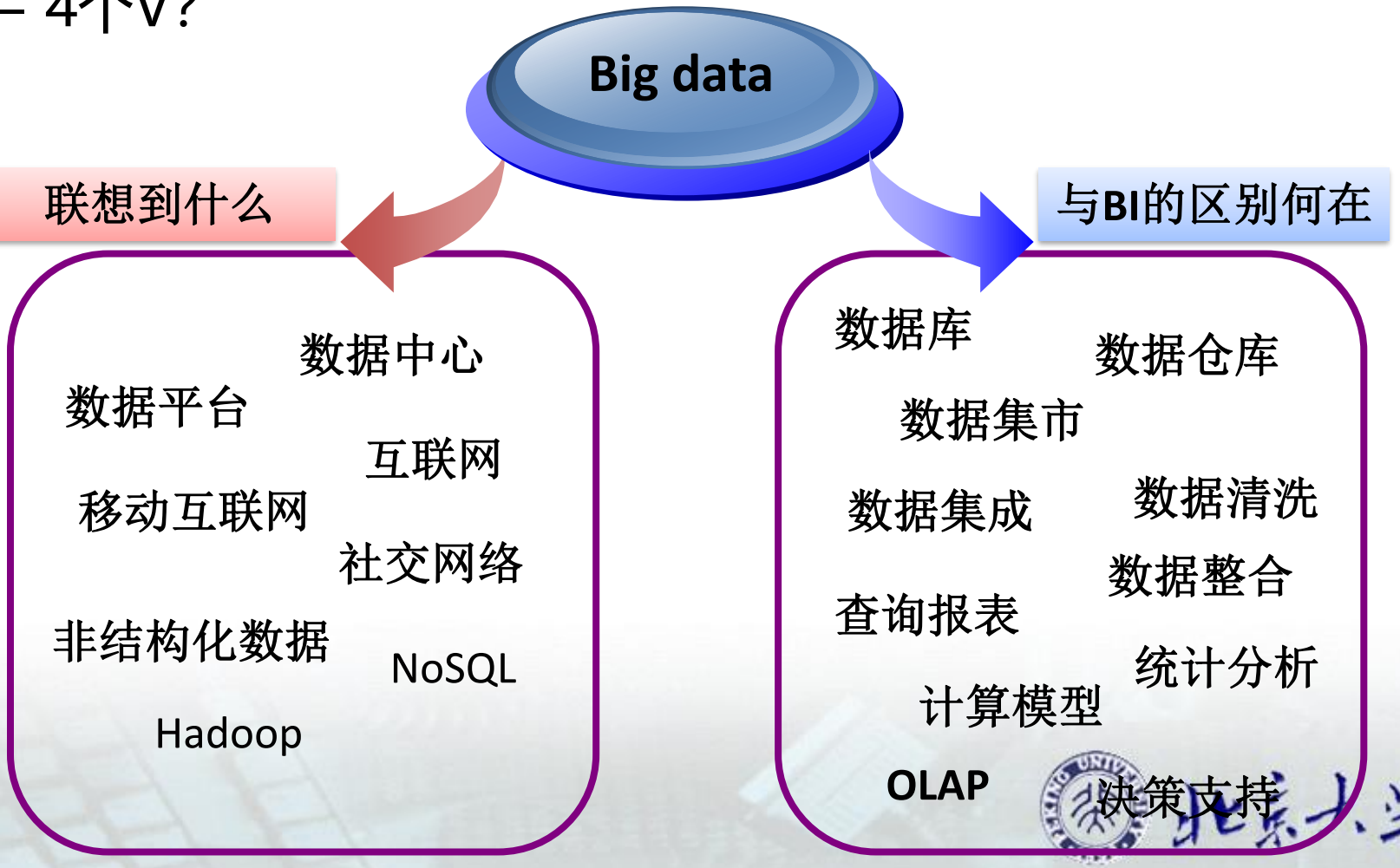
数据分析概述

- 对于数据分析而言，大数据真的是一个正确的方向吗
  - 数据量是越大越好吗?
  - 考虑的因素真的是越多越好吗?
  - DL是不是能够包治百病?

- 潜在的问题
  - 增加了分析成本（时间、金钱...）
  - 提高了数据收集难度
  - 降低了反应速度（冷启动问题等）
  - 引入更多的因素导致关联过多或冲淡了显著因素的影响

数据分析概述

- 什么是大数据
  - 数据“量”大?
  - 4个V?



数据分析概述

- 什么是大数据

大数据是一种 思想，让人们通过 科学计算 而不是凭借 主观臆断 来探寻事务本质

大数据是一套 方法论

通过间接的 数据，推测事物 难以客观表达 的 特性，揭示事物间内在的本质的 必然联系，帮助人们更加准确的认识事物 运作规律，进而指导行为。

‘真相’难以 测准，但可 无限逼近

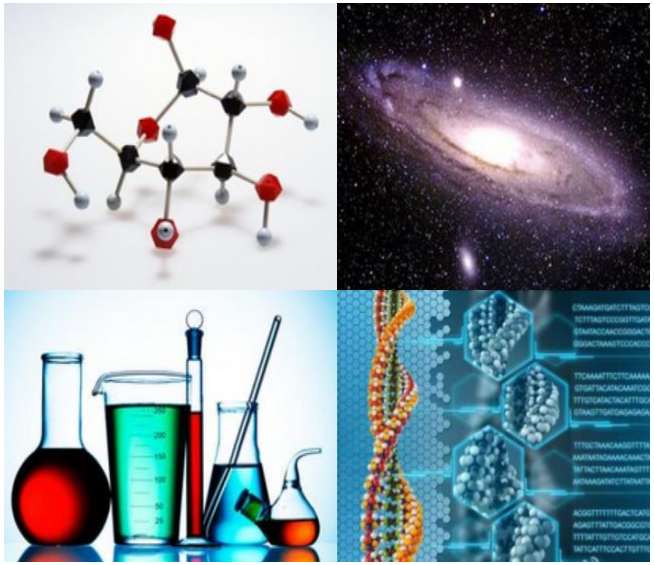
概念外延包括相关 数据，方法，技术 和 问题

数据分析概述

- 如何指导人们的行为
  - 拍脑袋
  - 扔鞋
  - 为什么是科学
  - 希望得到想要的结果

  - 种瓜得瓜、种豆得豆
  - 要想得豆，需要种豆

  - 如何发现规律?



科学——反映现实世界各种现象的客观规律

北京大学

数据分析概述

- 科学范式1：Empirical

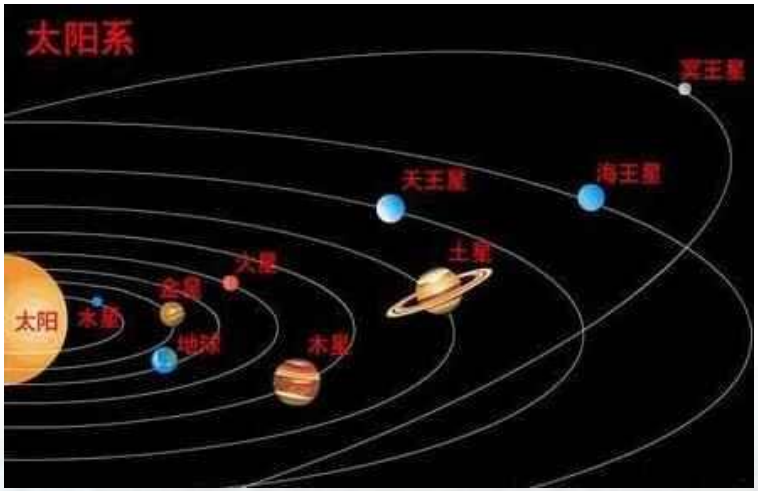
- 观察实验为基础
  - 牛顿和苹果
  - 阿基米德洗澡
  - 伽利略两个铁球同时落地

北京大学

数据分析概述

- 科学范式2：Theoretical

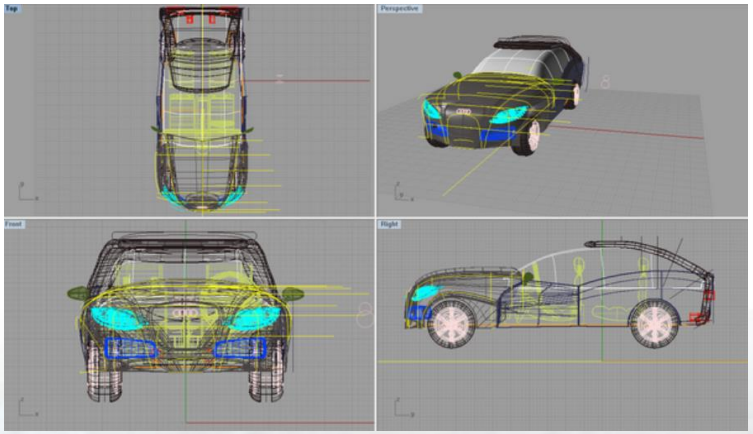
- 逻辑推理为基础
  - 冥王星的发现
  - 相对论



北京大学

- 科学范式3：Modeling

- 抽象建模为基础
  - 系统模型
  - 过程模型
  - 仿真模型
  - 模拟实验
  - 虚拟现实



数据分析概述

- 科学范式4：Data centric

- 数据为基础
  - 统计分析
  - 数据挖掘
  - 机器学习
  - 大数据

北京大学

- 群体智能（Swarm/collection intelligence）
  - 来自对自然界中昆虫群体的观察，群居性生物通过协作表现出的宏观智能行为特征被称为群体智能。

- 迪斯尼世界修路案例

沃尔特·格罗培斯
(Walter Gropius)

既是艺术的又是科学的，既是设计的又是实用的
艺术是人性化的最高体现。最人性的，就是最好的

数据分析概述

- 数据的就是科学的?
  - 哪个更合理?

	J1	J2	J3	SC	D		J1	J2	J3	RC	C
d1	9.6	9.7	9.8	29.1	2		5	3	3	11	3
d2	9.8	9.2	9.9	28.9	3		3	8	2	13	4
d3	9.7	9.9	10	29.6	1		4	2	1	7	1
d4	9.5	9.3	9.7	28.5	6		6	7	4	17	7
d5	9.9	9.4	9.5	28.8	4		2	6	6	14	5
d6	9.4	9.6	9.6	28.6	5		7	4	5	16	6
d7	9.3	9.5	9.4	28.2	7		8	5	7	20	8
d8	10	10	7	27	8		1	1	8	10	2

- 数据的就是科学的?

软件与微电子学院录取率

	软件	微电子	整体录取率
男性	4/ 20 = 0.200	25/200 = 0.125	29/220 = 0.132
女性	35/200 = 0.175	2/ 20 = 0.100	37/220 = 0.168
	软件方向男生高	微电子方向男生高	整体女生高

为什么?

数据分析概述

- 数据的就是科学的?
  - 置信度与支持度
    - 支持度s是指事务集D中包含$A \cup B$ 的百分比

$$\sup port(A \Rightarrow B) = P(A \cup B)$$

    - 置信度c是指D中包含A的事务同时也包含B的百分比

$$confidence(A \Rightarrow B) = P(B \mid A) = P(A \cup B) / P(A)$$

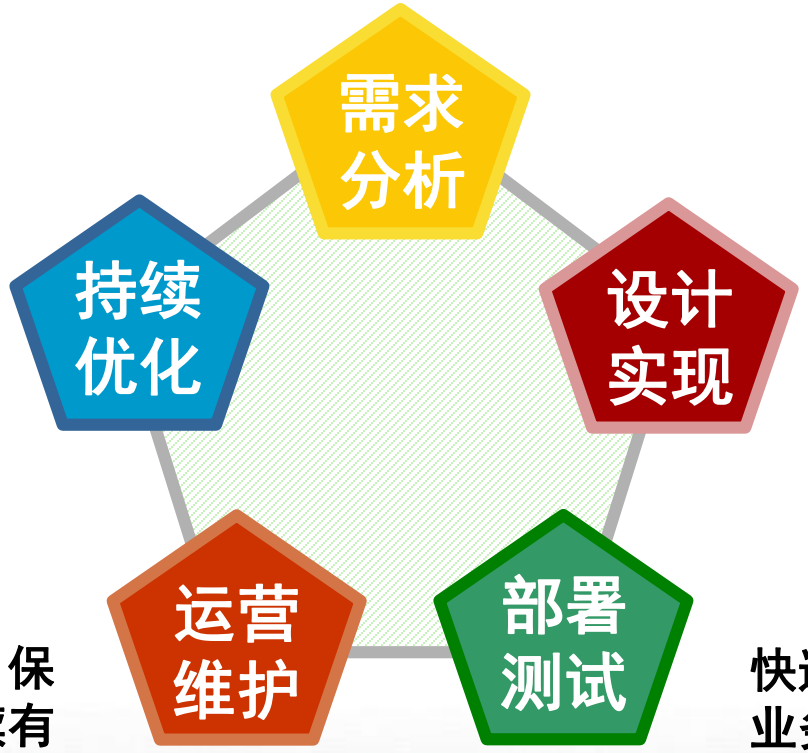
  - 前置条件、后置条件、条件完整性
  - 充分条件、必要条件、充要条件

数据分析概述

- 正确的看待数据分析的结果
  - 评价指标体系/计算体系不同导致结果不同
  - 局部推测整体
  - 分析逻辑

- 大数据应用生命周期

捕获用户需求，策划数据项目。



拿到数据，加工处理，得到数据分析结果。

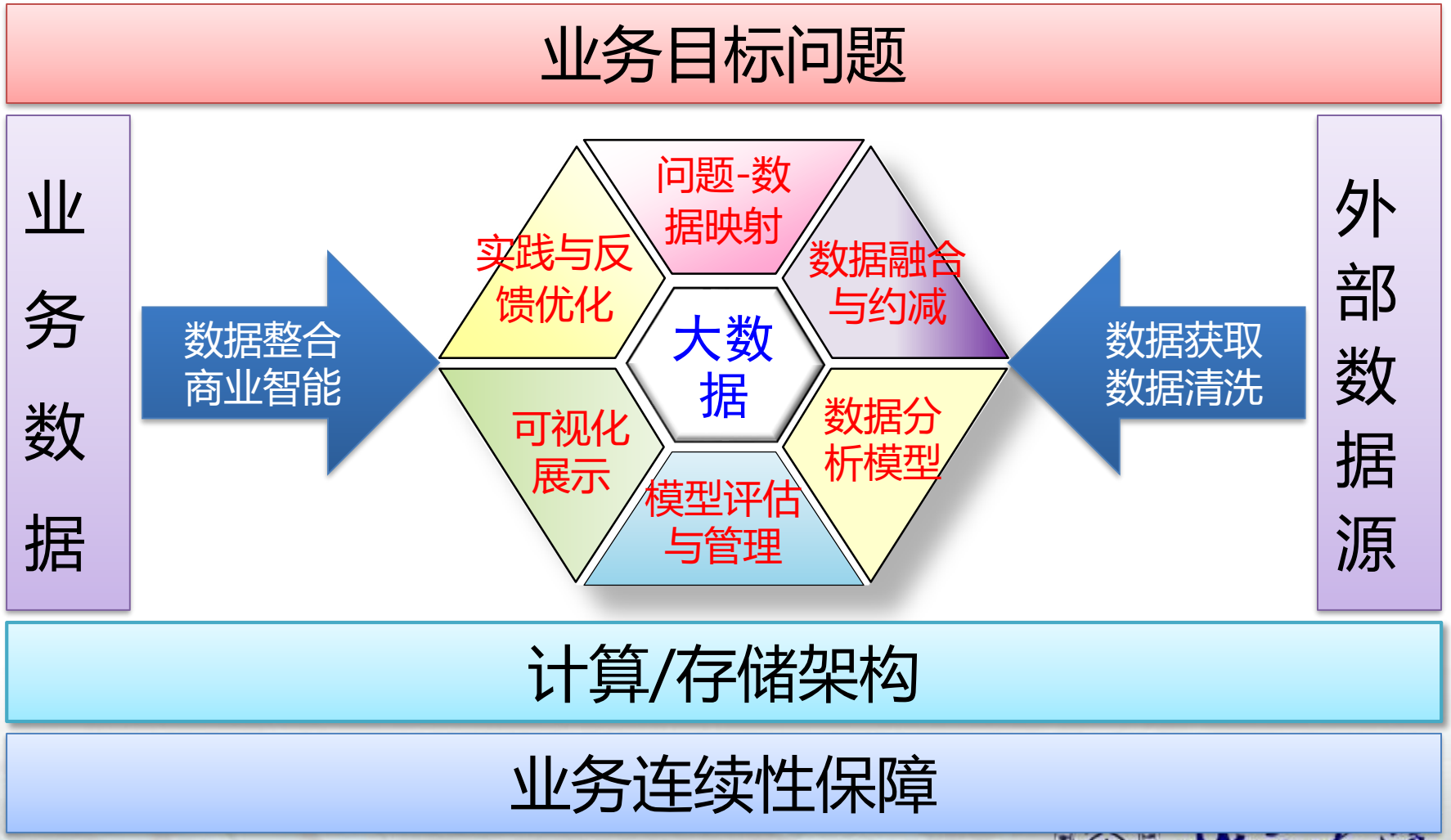
快速部署，测试数据及业务案例，DevOps，保障分析质量

数据的持续获取，保障分析应用的持续有效可用

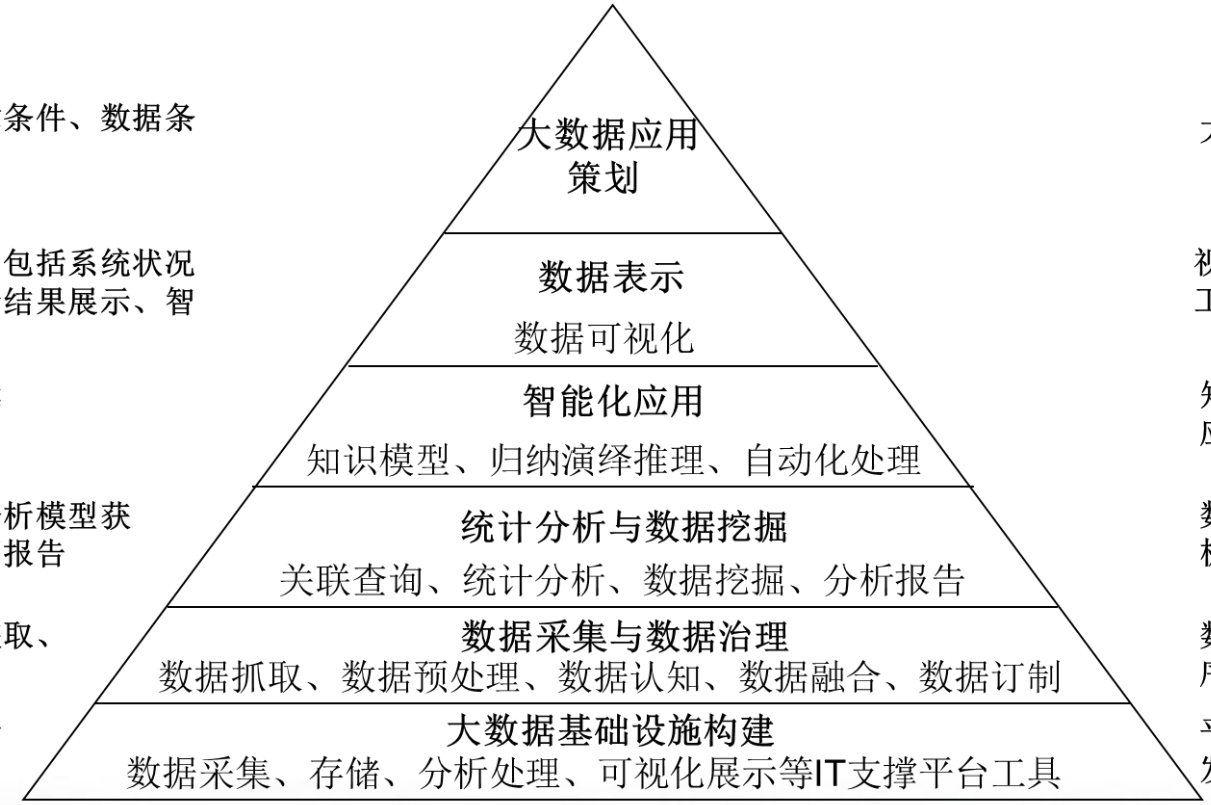
更丰富的数据，新的业务情况，分析的持续滚动优化

数据分析概述

- 大数据核心业务



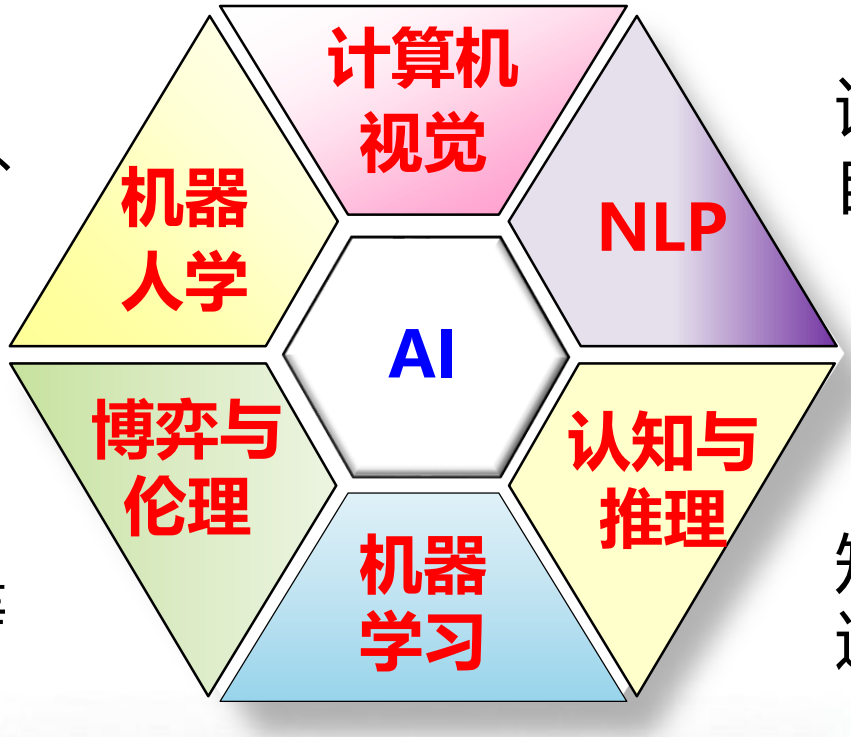
- 大数据职业分工及关键任务



根据用户需求、软硬件支撑条件、数据条件等策划大数据应用	大数据应用策划师
提供大数据相关工作展示，包括系统状况展示、数据内容展示、分析结果展示、智能化应用展示等	视觉效果总监、美工、可视化研发师
通过知识建模、逻辑推理等形成智能化应用	知识工程师、知识应用工程师
通过数据分析方法，建立分析模型获取分析结果，解读结果撰写报告	数据挖掘师、数据分析师
提供数据支撑，包括数据获取、质量保障和数据提供	数据架构师、爬虫程序员、数据开发师
提供软硬件基础支撑，包括部署应用环境和开发环境	平台架构师、工具研发程序员

- 人工智能

图像识别、模式识别等



控制、机械、
任务规划等

语音识别、
自动问答等

多agent、
对抗合作等

知识、表示、
逻辑等

建模、算法，包括深度学习等

内容提要

1. 课程简介
2. 数据分析概述
3. 基本方法
4. 数据挖掘方法论

基本方法

- 什么是数据挖掘?
  - Data Mining, Also called Knowledge Discovery in Database (KDD)
  - Wiki：从大量的数据中自动搜索隐藏于其中的有着特殊关系性的信息的过程
  - 通过分析每个数据，从大量数据中寻找其规律的技术

数据+挖掘

基本方法

- 数据挖掘 (数据库中知识发现):
  - 从大型数据库中提取有趣的 (非平凡的, 蕴涵的, 先前未知的并且是潜在有用的) 信息或模式
- 其它叫法和“inside stories”:
  - 数据挖掘: 用词不当?
  - 数据库中知识发现(挖掘) (Knowledge discovery in databases, KDD), 知识提取(knowledge extraction), 数据/模式分析 (data/pattern analysis), 数据考古(data archeology), 数据捕捞 (data dredging), 信息收获(information harvesting), 商务智能 (business intelligence), 等.

基本方法

- 什么是数据?
  - Wiki：指描述事物的符号记录，它涉及到事物的存在形式，是关于事物的一组离散且客观的事实描述。
  - 描述对象的一组属性及属性值
  - 相关概念：数据类型、数据处理...
  - 一段视频算不算数据?

基本方法

- 数据：对象+属性
- 对象：
  - 被描述的单元
- 属性：
  - 描述对象某一方面的特征

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

基本方法

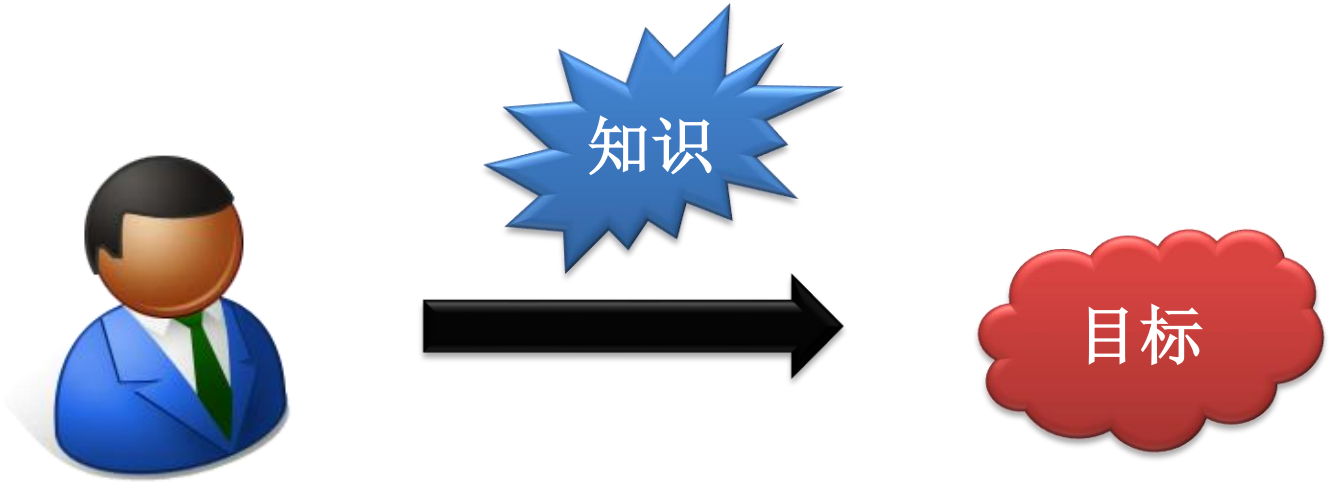
- 数据和知识有什么区别?

日期	金额
1.1	232
1.2	227
1.3	159
…	…
2.1	280
2.2	289
…	…
3.1	311
3.2	332
…	…
12.1	169
…	…

月份	金额
1	0.8w
2	1.3w
3	1.8w
4	2.1w
5	2.4w
6	2.5w
7	2.6w
8	2.5w
9	2.3w
10	1.9w
11	1.4w
12	0.9w

卖什么的?

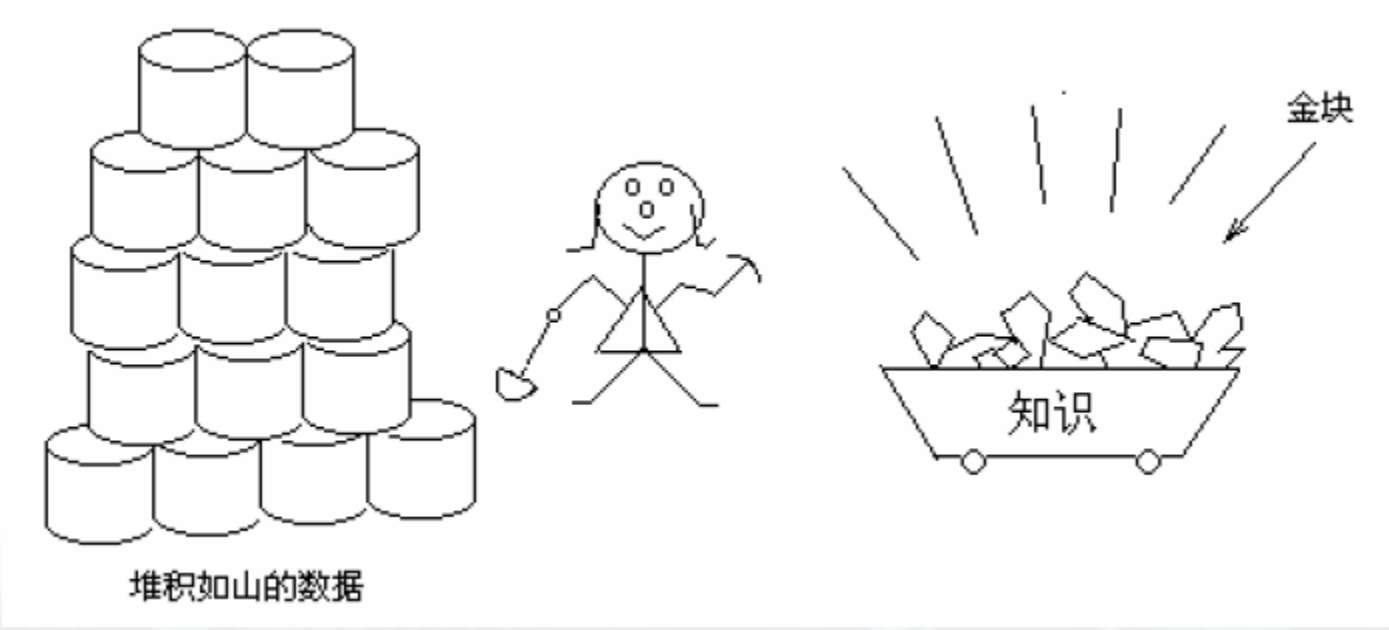
- 为什么需要知识?



知识保障我们能够达成既定目标

基本方法

- 如何获取知识?
  - 通过一些现象总结而出
  - 数据记录了这些现象
  - “总结”至关重要——mining



基本方法

- 挖掘的基础——数据
  - 传感数据（卫星、微传感器等）
  - 天体/空间物理数据
  - 生物/化学数据（基因序列、分子结构等）
  - 科研仿真数据（离子加速器等）
  - 地理、气象、水文数据
  - 企业生产、经营、销售数据
  - 日常生活数据
  - …

基本方法

- 数据需要进行组织
  - 数据库
    - 表
    - 视图
    - 索引
  - 数据仓库
    - 数据方
    - 索引
  - 文本
    - 结构化

基本方法

- 挖掘前的处理方法
  - 数据清理
  - 数据变换
  - 数据归约
  - 采样
  - 统计
  - 预计算

基本方法

- 所有的数据处理方法都是挖掘吗?
  - 张三一个月都给谁打了多少个电话?
  - 张三的电话簿里有多少个人?
  - 张三的电话簿中第10个人的电话是多少?
  - 张三的打电话习惯是什么?

挖掘vs统计vs搜索

基本方法

- 小建议
  - 数据挖掘——大量数学+算法
  - 不要惧怕数学
  - 管理、金融一样需要数学（甚至更多）

基本方法

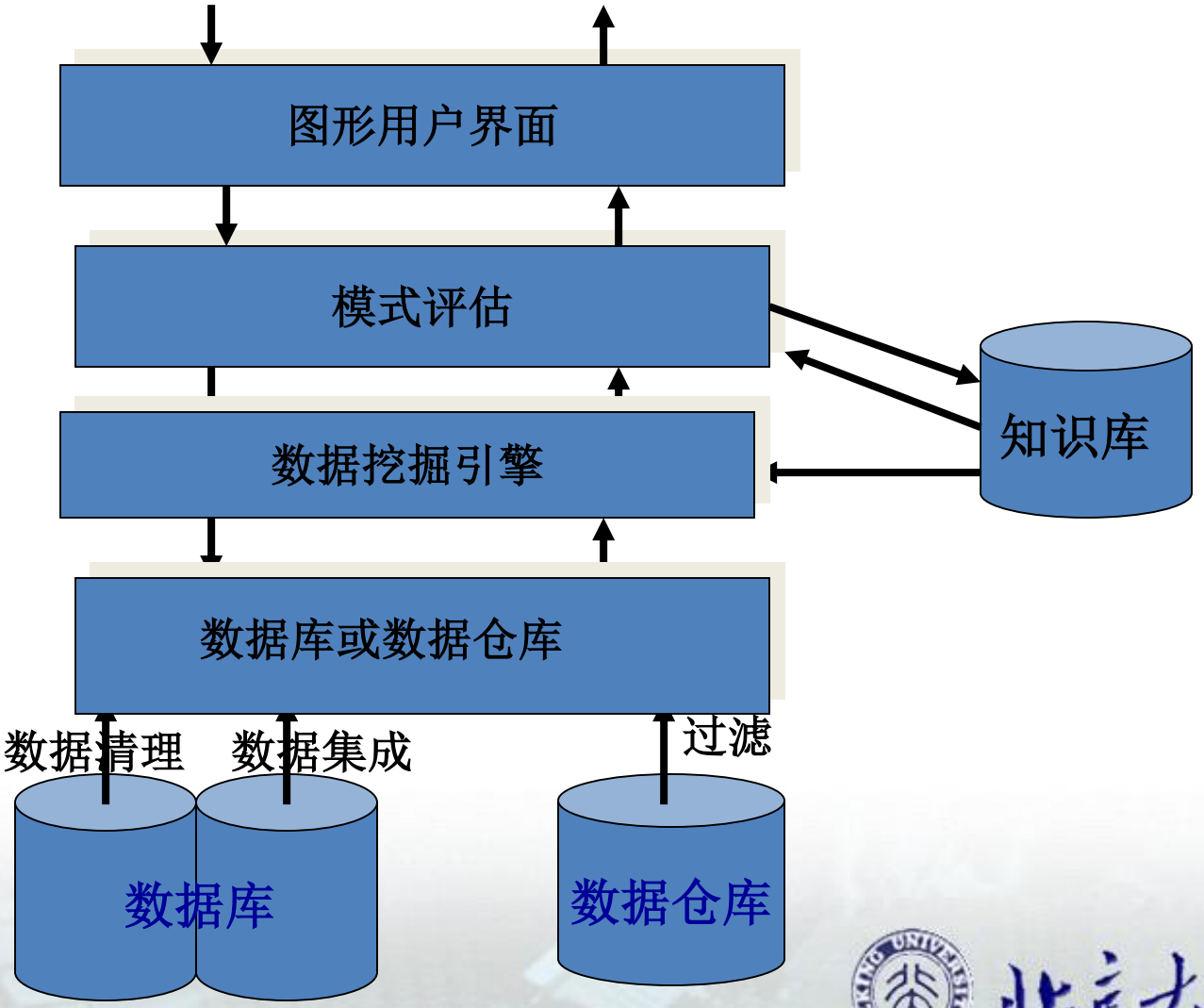
- 数据爆炸问题
  - 自动的数据收集工具和成熟的数据库技术导致大量数据存放在数据库, 数据仓库, 和其它信息存储中
- 我们正被数据淹没,但却缺乏知识
- 解决办法: 数据仓库与数据挖掘
  - 数据仓库与联机分析处理(OLAP)
  - 从大型数据库的数据中提取有趣的知识(规则, 规律性, 模式, 限制等)

基本方法

- 1960s:
  - 数据收集, 数据库创建, IMS 和网状 DBMS
- 1970s:
  - 关系数据库模型, 关系 DBMS 实现
- 1980s:
  - RDBMS, 先进的数据模型 (扩充关系的, OO, 演绎的, 等.) 和面向应用 的 DBMS (空间的, 科学的, 工程的, 等.)
- 1990s—2008s:
  - 数据挖掘和数据仓库, 多媒体数据库, 和 Web 数据库
- 2008s—201x:
  - 云计算技术，大数据

基本方法

典型的数据挖掘系统结构



基本方法

- 关系数据库
- 数据仓库
- 事务(交易)数据库
- 先进的数据库和信息存储
  - 面向对象和对象-关系数据库
  - 空间和时间数据
  - 时间序列数据和流数据
  - 文本数据库和多媒体数据库
  - 异种数据库和遗产数据库
  - WWW

数据挖掘:在什么数据上进行?

基本方法

- 关键方法
  - 概念描述: 特征和区分
    - 概化, 汇总, 和比较数据特征, 例如, 干燥和潮湿的地区
  - 关联 (相关和因果关系)
    - 多维和单维关联
    - *age*(*X*, “20..29”) ^ *income*(*X*, “20..29K”) ⇒*buys*(*X*, “PC”)

      [*support* = 2%, *confidence* = 60%]
    - *contains*(*T*, “computer”) ⇒ *contains*(*T*, “software”)

      [*support* = 1%, *confidence* = 75%]

基本方法

- 关键方法
  - 分类和预测
    - 找出描述和识别类或概念的模型(函数), 用于将来的预测
    - 例如根据气候对国家分类, 或根据单位里程的耗油量对汽车分类
    - 表示: 判定树(decision-tree), 分类规则, 神经网络
    - 预测: 预测某些未知或遗漏的数值值
  - 聚类分析
    - 类标号(Class label) 未知: 对数据分组, 形成新的类. 例如, 对房屋分类, 找出分布模式
    - 聚类原则: 最大化类内的相似性, 最小化类间的相似性

基本方法

- 关键方法
  - 孤立点(Outlier)分析
    - 孤立点: 一个数据对象, 它 与数据的一般行为不一致
    - 孤立点可以被视为例外, 但对于欺骗检测和罕见事件分析, 它是相当有用的
  - 趋势和演变分析
    - 趋势和偏离: 回归分析
    - 序列模式挖掘, 周期性分析
    - 基于相似的分析
  - 其它基于模式或统计的分析

挖掘出的所有模式都是有趣的吗?

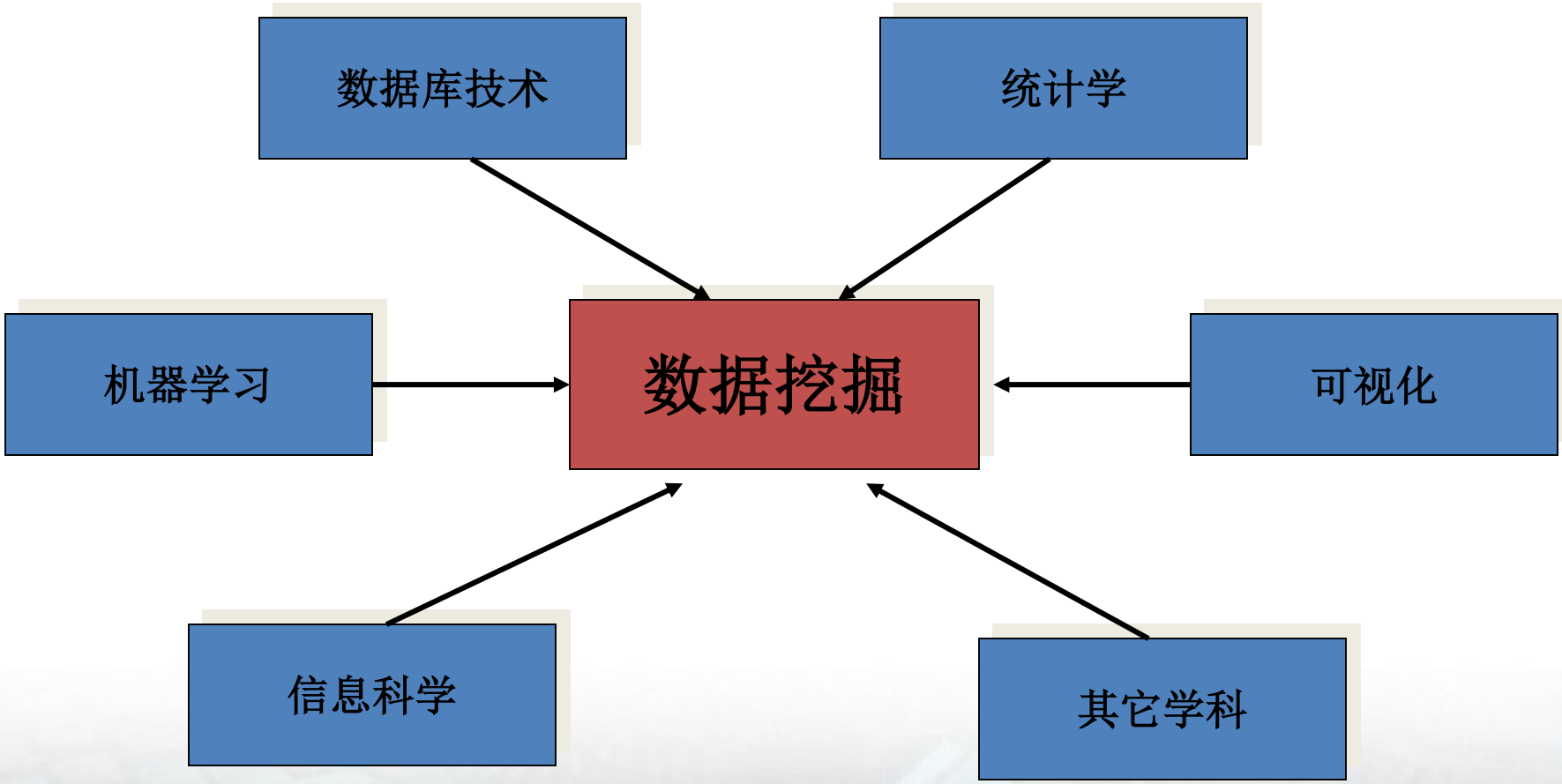
- 一个数据挖掘系统/查询可以挖掘出数以千计的模式, 并非所有的模式都是有趣的
  - 建议的方法: 以人为中心, 基于查询的, 聚焦的挖掘
- 兴趣度度量：一个模式是 有趣的 如果它是 易于被人理解的, 在某种程度上在新的或测试数据上是有效的, 潜在有用的, 新颖的, 或验证了用户希望证实的某种假设
- 客观与主观的兴趣度度量：
  - 客观: 基于模式的统计和结构, 例如, 支持度, 置信度, 等.
  - 主观: 基于用户对数据的确信, 例如, 出乎意料, 新颖性, 可行动性 (actionability), 等.

基本方法

- 发现所有有趣的模式: 完全性
  - 数据挖掘系统能够发现所有有趣的模式吗?
  - 关联 vs. 分类 vs. 聚类
- 仅搜索有趣的模式: 优化
  - 数据挖掘系统能够仅发现有趣的模式吗?
  - 方法
    - 首先找出所有模式, 然后过滤掉不是有趣的那些.
    - 仅产生有趣的模式— 挖掘查询优化

基本方法

数据挖掘：多学科交叉



基本方法

- 一般功能
  - 描述式数据挖掘
  - 预测式数据挖掘
- 不同的角度,不同的分类
  - 待挖掘的数据库类型
  - 待发现的知识类型
  - 所用的技术类型
  - 所适合的应用类型

基本方法

- 主要问题
  - 挖掘方法和用户交互
    - 在数据库中挖掘不同类型的知识
    - 在多个抽象层的交互式知识挖掘
    - 结合背景知识
    - 数据挖掘语言和启发式数据挖掘
    - 数据挖掘结果的表示和可视化
    - 处理噪音和不完全数据
    - 模式评估: 兴趣度问题
  - 性能和可伸缩性(scalability)
    - 数据挖掘算法的性能和可伸缩性
    - 并行, 分布和增量的挖掘方法

基本方法

- 数据类型的多样性问题
  - 处理关系的和复杂类型的数据
  - 从异种数据库和全球信息系统 (WWW)挖掘信息
- 应用和社会效果问题
  - 发现知识的应用
    - 特定领域的数据挖掘工具
    - 智能查询回答
    - 过程控制和决策制定
  - 发现知识与已有知识的集成: 知识融合问题
  - 数据安全, 完整和私有的保护

- 数据挖掘作用小结
  - 有大量数据的地方就需要数据挖掘
  - 统计是初级阶段，挖掘是进阶
  - 大数据时代将发挥更大作用

内容提要

1. 课程简介
2. 数据分析概述
3. 基本方法
4. 数据挖掘方法论

数据挖掘方法论

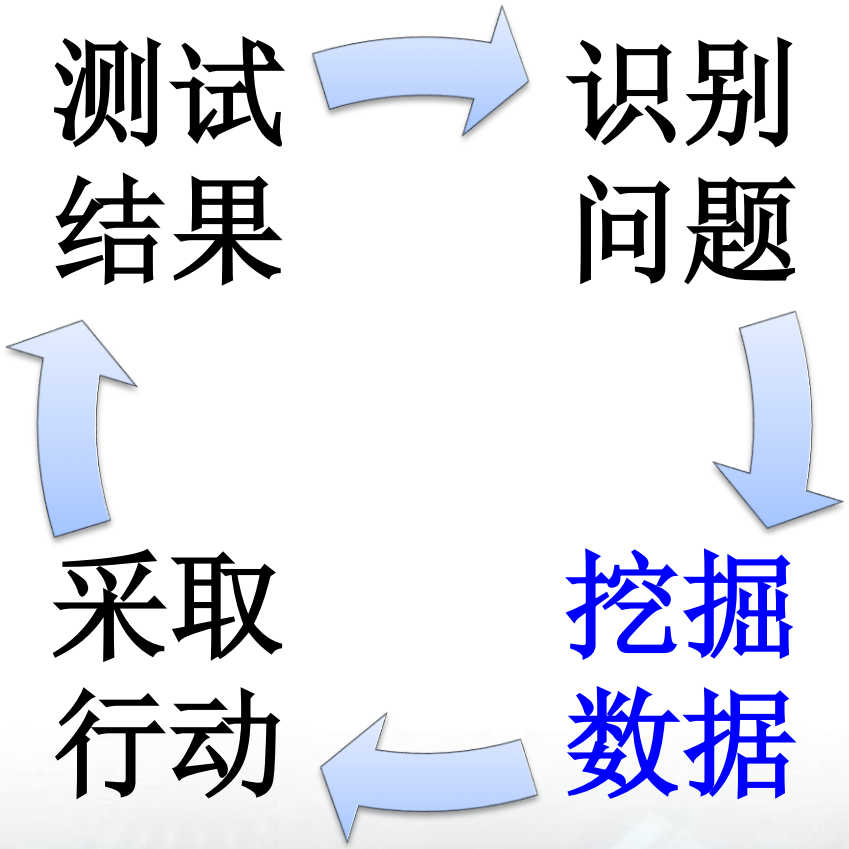
- 方法论
  - 为达成某个目标所使用的一套方法
- 软件工程
  - 需求分析-系统设计-系统实现-测试部署-实施优化

数据挖掘方法论

- 方法论
  - 为达成某个目标所使用的一套方法
- 软件工程
  - 需求分析-系统设计-系统实现-测试部署-实施优化

数据挖掘方法论

- 广义数据挖掘方法论



数据挖掘方法论

- 为什么需要数据挖掘方法论
  - 达成目标需要若干方法的支持
  - 避免方法获得“坏”的结果
    - 不真实的知识
    - 真实但无用的知识

- 不真实的知识
  - 模式可能不代表任何底层规则
  - 模型集可能不能正确反映相关人群的总体状况
  - 数据位于错误的详细层次

具有更大的危险性

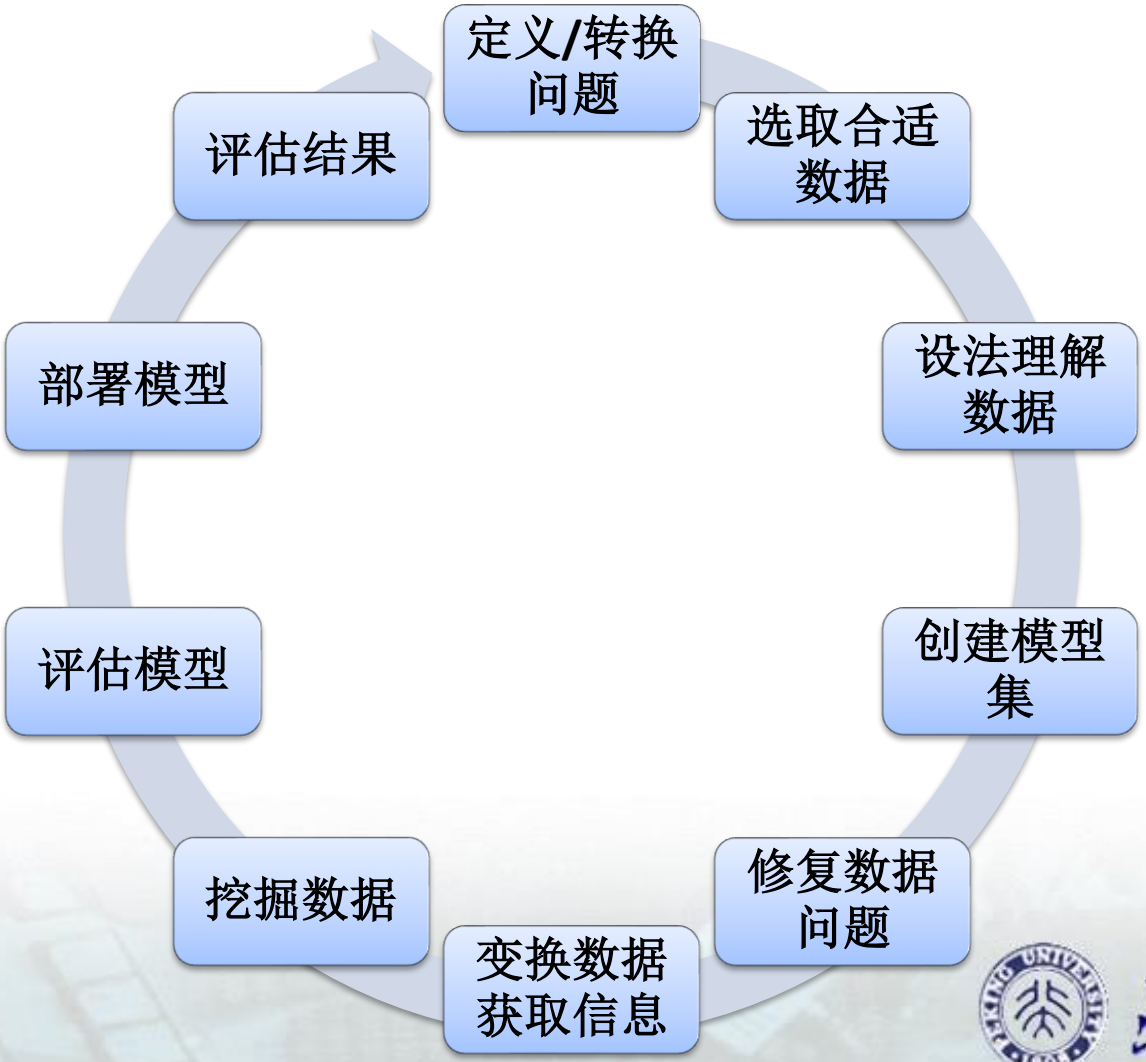
- 不真实知识的产生原因
  - 科学的方法一定能获得正确的知识吗?
    - 数据本身是不正确的
    - 数据与当前的问题没有关系
    - 发现的模式可能只反映了过去
    - 一些数据处理可能破坏或隐藏了一些重要的信息

开好车的一定是好人吗?

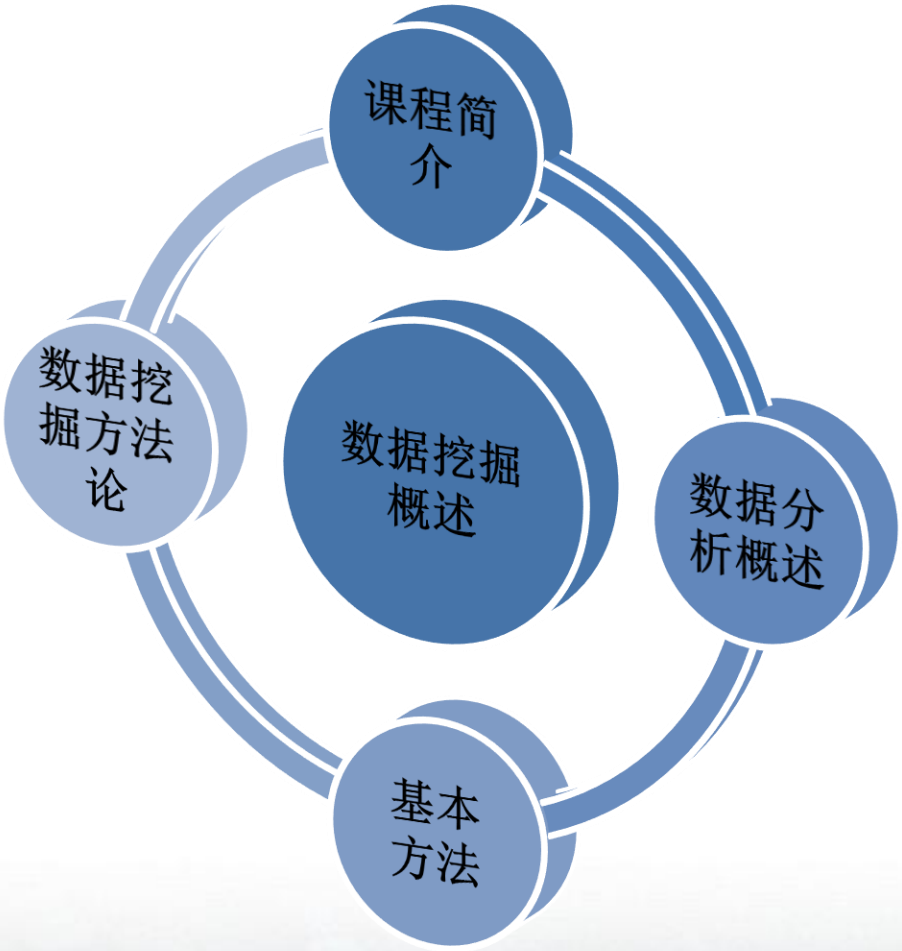
- 真实但无用的知识
  - 过于模糊或宏观
  - 缺少操作层的指导意义
  - 与目标问题无关
  - 不能使用的知识
  - “废话”

知识越多越好吗？

- 数据挖掘方法论


定义/转换问题
选取合适数据
设法理解数据
创建模型集
修复数据问题
变换数据获取信息
挖掘数据
评估模型
部署模型
评估结果

小结



- 课程简介
- 数据分析概述
  - 大数据
  - 科学范式
  - 数据和知识
- 基本方法
  - 各种方法的应用
- 数据挖掘方法论

课后作业

- 思考题：
  - 数据与知识的区别与联系？
  - 列举几项你所知道的数据挖掘应用，并论述数据挖掘在其中的作用？
  - 数据挖掘方法过程是什么？
  - 数据挖掘与统计的区别与联系？
  - 数据挖掘与数据管理的区别与联系？