数据挖掘

概要

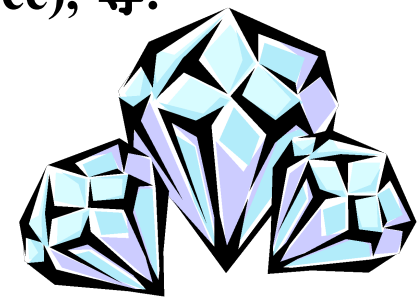
- 动机:为什么要数据挖掘?
- 什么是数据挖掘?
- 数据挖掘:在什么数据上进行?
- 数据挖掘功能
- 所有的模式都是有趣的吗?
- 数据挖掘系统分类
- 数据挖掘的主要问题

动机: 需要是发明之母

- **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)

什么是数据挖掘?

- 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.
- 什么不是数据挖掘?
  - (演绎) 查询处理.
  - 专家系统 或小型 机器学习(ML)/统计程序



为什么要数据挖掘?—可能的应用

- 数据库分析和决策支持
  - 市场分析和管理
    - 针对销售(target marketing), 顾客关系管理, 购物篮分析, 交叉销售(cross selling), 市场分割(market segmentation)
  - 风险分析与管理
    - 预测, 顾客关系, 改进保险, 质量控制, 竞争能力分析
  - 欺骗检测与管理
- 其它应用
  - 文本挖掘 (新闻组, email, 文档资料)
  - 流数据挖掘(Stream data mining)
  - Web挖掘.
  - DNA 数据分析

市场分析与管理(1)

- 用于分析的数据源在哪?
  - 信用卡交易, 会员卡, 打折优惠卷, 顾客投诉电话, (公共) 生活时尚研究
- 针对销售(Target marketing)
  - 找出顾客群, 他们具有相同特征 : 兴趣, 收入水平, 消费习惯, 等.
- 确定顾客随时间变化的购买模式
  - 个人帐号到联合帐号的转变: 结婚, 等.
- 交叉销售分析(Cross-market analysis)
  - 产品销售之间的关联/相关
  - 基于关联信息的预测

市场分析与管理(2)

- 顾客分类(Customer profiling)
  - 数据挖掘能够告诉我们什么样的顾客买什么产品(聚类或分类)
- 识别顾客需求
  - 对不同的顾客识别最好的产品
  - 使用预测发现什么因素影响新顾客
- 提供汇总信息
  - 各种多维汇总报告
  - 统计的汇总信息 (数据的中心趋势和方差)

法人分析和风险管理

- 财经规划和资产评估
  - 现金流分析和预测
  - 临时提出的资产评估
  - 交叉组合(cross-sectional) 和时间序列分析 (金融比率(financial-ratio), 趋势分析, 等.)
- 资源规划：
  - 资源与开销的汇总与比较
- 竞争:
  - 管理竞争者和市场指导
  - 对顾客分类和基于类的定价
  - 在高度竞争的市场调整价格策略

欺骗检测和管理(1)

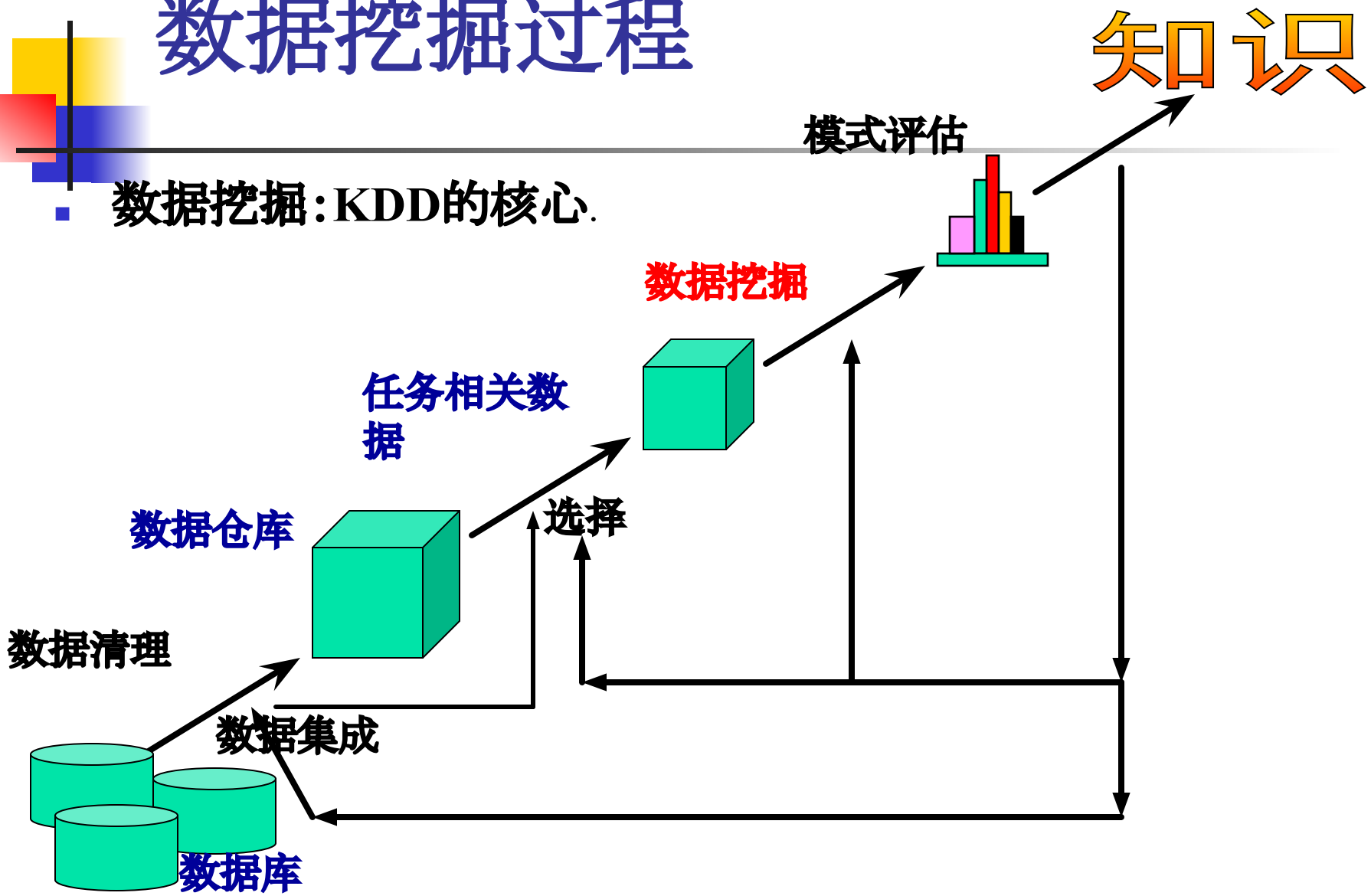
- 应用
  - 广泛用于健康照料, 零售, 信用卡服务, 电讯 (电话卡欺骗), 等.
- 方法
  - 使用历史数据建立欺骗行为模型, 使用数据挖掘帮助识别类似的实例
- 例
  - 汽车保险: 检测这样的人, 他/她假造事故骗取保险赔偿
  - 洗钱: 检测可疑的金钱交易 (US Treasury's Financial Crimes Enforcement Network)
  - 医疗保险 : 检测职业病患者, 医生和介绍人圈

其它应用

- 运动
  - IBM Advanced Scout分析NBA的统计数据 (阻挡投篮, 助攻, 和犯规) 获得了对纽约小牛队(New York Knicks)和迈艾米热队(Miami Heat) 的竞争优势
- 天文
  - 借助于数据挖掘的帮助,JPL 和 Palomar Observatory 发现了22 颗类星体(quasars)
- Internet Web Surf-Aid
  - IBM Surf-Aid 将数据挖掘算法用于有关交易的页面的Web访问日志, 以发现顾客喜爱的页面, 分析Web 销售的效果, 改进Web 站点的组织, 等.

数据挖掘过程

- 数据挖掘:KDD的核心.



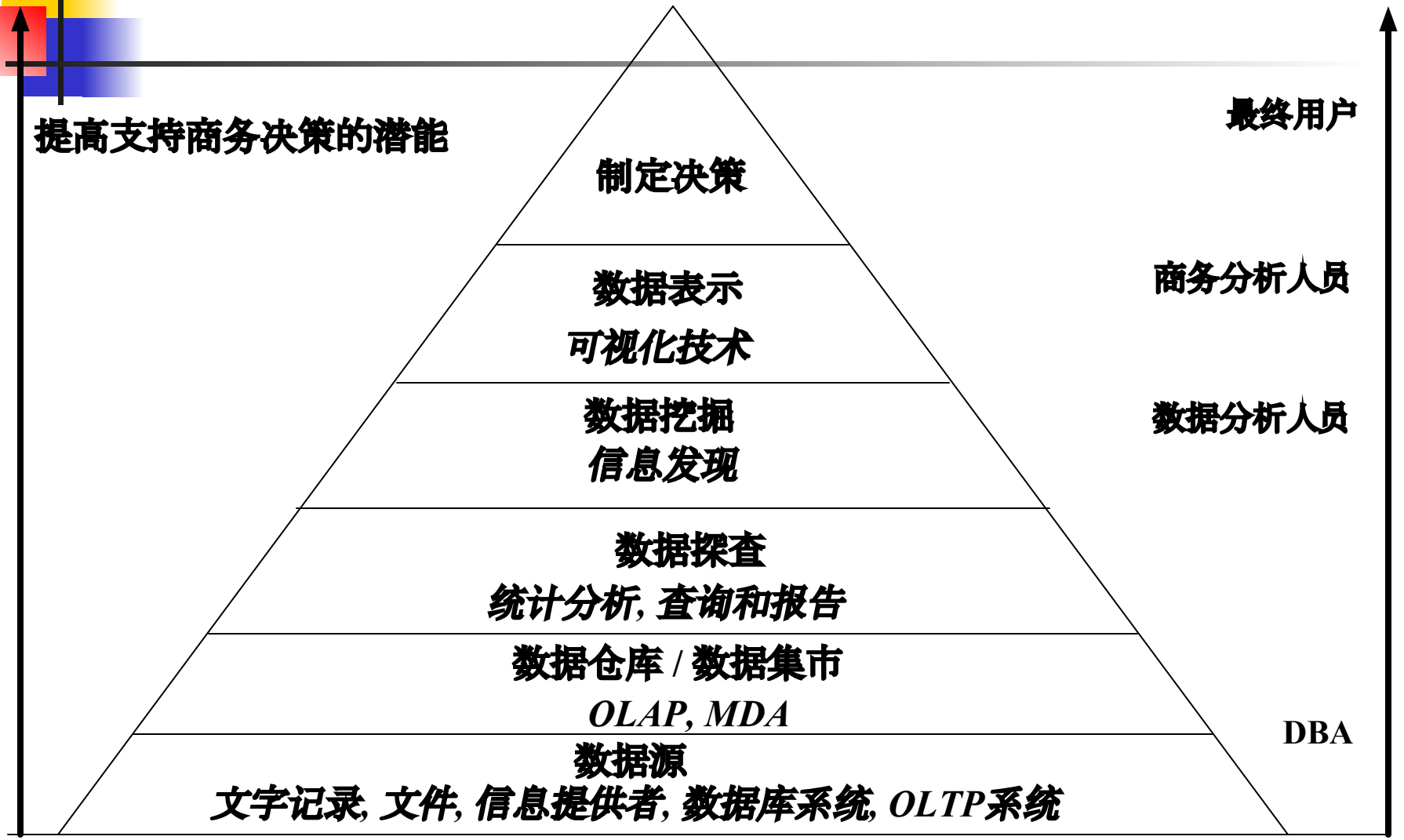
KDD过程的步骤

- 学习应用领域:
  - 相关的先验知识和应用的目标
- 创建目标数据集: 数据选择
- 数据清理和预处理: (可能占全部工作的 60%!)
- 数据归约与变换:
  - 发现有用的特征, 维/变量归约, 不变量的表示.
- 选择数据挖掘函数
  - 汇总, 分类, 回归, 关联, 聚类.

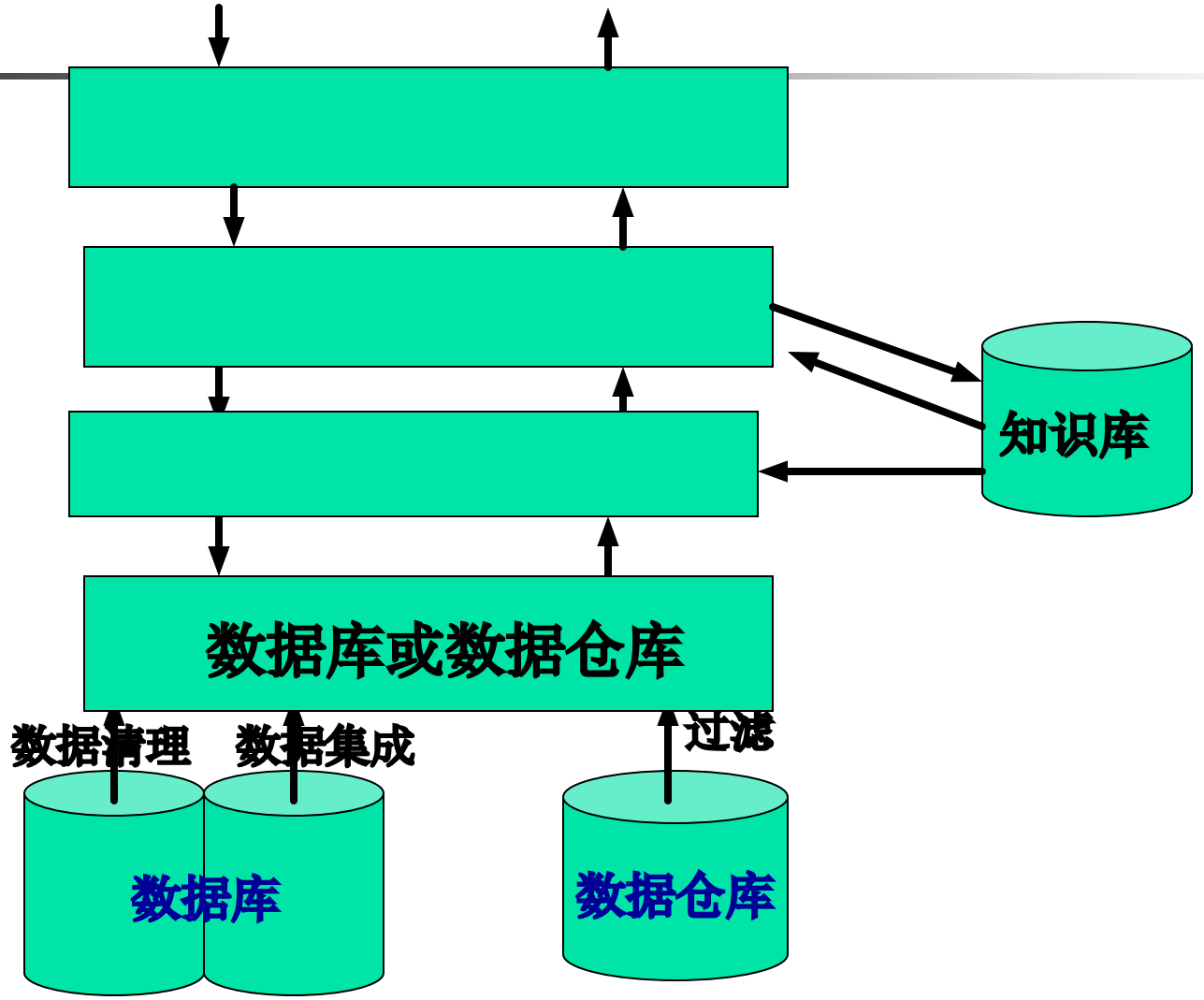
KDD过程的步骤(续)

- 选择挖掘算法
- 数据挖掘: 搜索有趣的模式
- 模式评估和知识表示
  - 可视化, 变换, 删除冗余模式, 等.
- 发现知识的使用

数据挖掘和商务智能


提高支持商务决策的潜能
最终用户
制定决策
数据表示
可视化技术
商务分析人员
数据挖掘
信息发现
数据分析人员
数据探查
统计分析, 查询和报告
数据仓库 / 数据集市
OLAP, MDA
DBA
数据源
文字记录, 文件, 信息提供者, 数据库系统, OLTP系统

典型的数据挖掘系统结构



数据挖掘:在什么数据上进行?

- 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

数据挖掘功能(1)

- **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%]

数据挖掘功能(2)

- 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

数据挖掘功能(3)

- 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

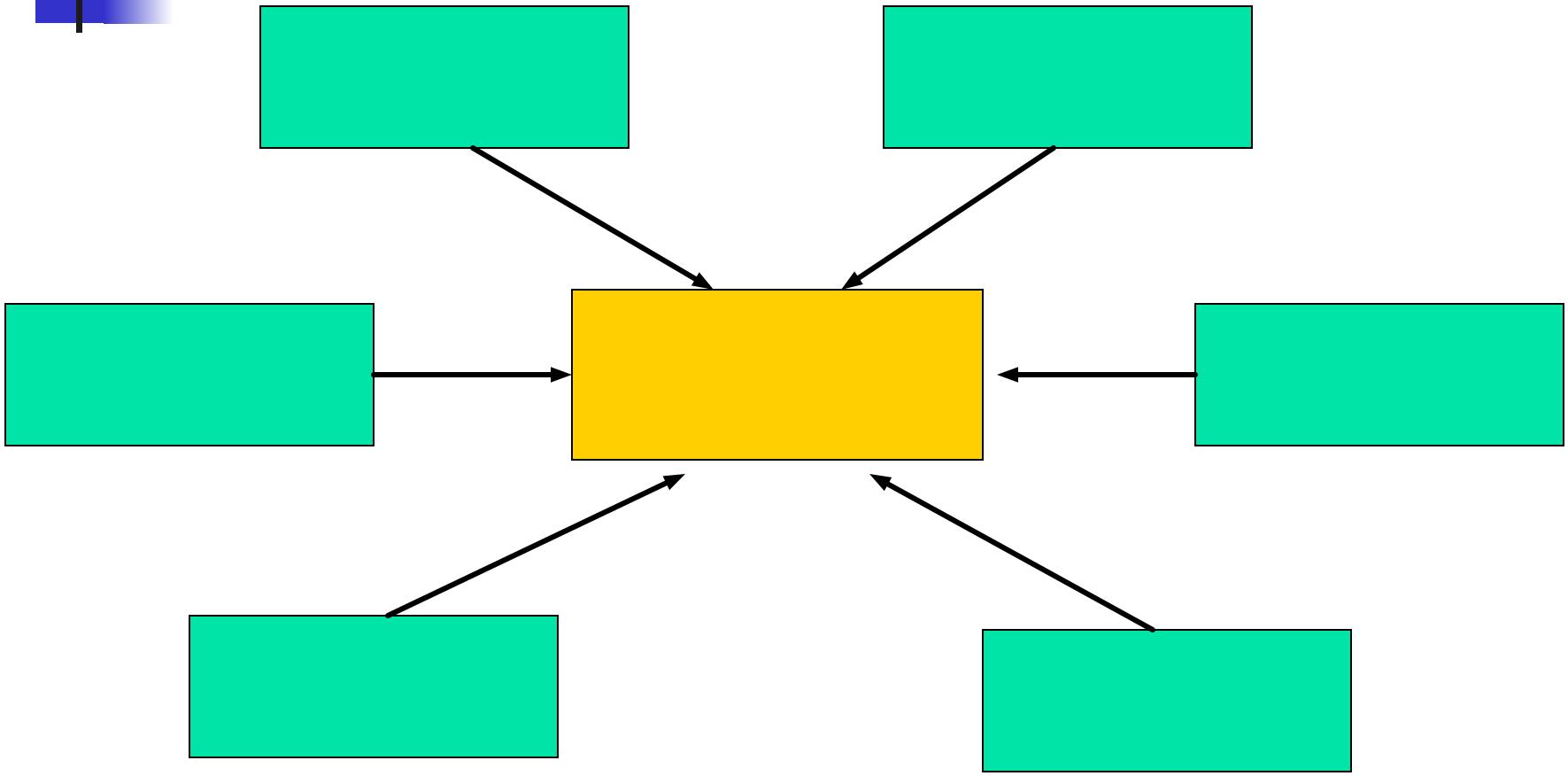
挖掘出的所有模式都是有趣的吗?

- 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.

能够只发现有趣的模式吗?

- 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

数据挖掘: 多学科交叉



数据挖掘系统分类

- 待挖掘的数据库
  - 关系的, 事务的, 面向对象的, 对象-关系的, 主动的, 空间的, 时间序列的, 文本的, 多媒体的, 异种的, WWW, 等.
- 所挖掘的知识
  - 特征, 区分, 关联, 分类, 聚类, 趋势, 偏离和孤立点分析, 等.
  - 多/集成的功能, 和多层次上的挖掘
- 所用技术
  - 面向数据库的, 数据仓库 (OLAP), 机器学习, 统计学, 可视化, 神经网络, 等.
- 适合的应用
  - 零售, 电讯, 银行, 欺骗分析, DNA 挖掘, 股票市场分析, Web 挖掘, Web日志分析, 等

数据挖掘的主要问题(1)

- 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

数据挖掘的主要问题(2)

- 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

小结

- 数据挖掘: 从大量数据中发现有趣的模式
- 数据库技术的自然进化, 具有巨大需求和广泛应用
- KDD 过程包括数据清理, 数据集成, 数据选择, 变换, 数据挖掘, 模式评估, 和知识表示
- 挖掘可以在各种数据存储上进行
- 数据挖掘功能: 特征, 区分, 关联, 分类, 聚类, 孤立点 和趋势分析, 等.
- 数据挖掘系统的分类
- 数据挖掘的主要问题

谢谢大家!