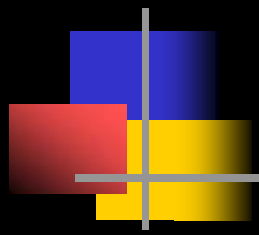


数 据 挖 掘

概 念 与 技 术

——第一章——

Jiawei Han

（加） 著

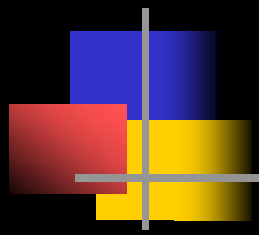
Micheline Kamber

http://www.cs.sfu.ca

幻灯片的出处

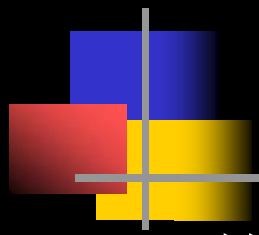
- 指南部分的幻灯片:
  - http://www.cs.sfu.ca/~han/dmbook
- 其它的会议介绍幻灯片:
  - http://db.cs.sfu.ca/ or http://www.cs.sfu.ca/~han
- 研究论文，数据库挖掘系统和其它相关信息:
  - http://db.cs.sfu.ca/ or http://www.cs.sfu.ca/~han

第一章　　引言



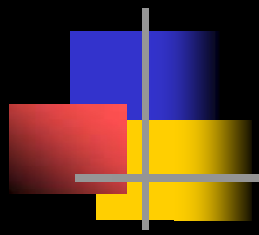
- 什么激发了数据挖掘，为什么它是重要的？
- 什么是数据挖掘？
- 在何种数据上进行数据挖掘？
- 数据挖掘功能——可以挖掘什么类型的模式
- 所有模式都是有趣的吗？
- 数据挖掘系统的分类
- 数据挖掘的主要问题

动机：“需要是发明之母”



- 数据泛滥问题
  - 自动数据收集工具和成熟的数据库技术使得大量数据存储于数据库，数据仓库和其他信息库。
- 我们数据丰富但信息贫乏
- 解决办法：数据仓库和数据挖掘
  - 数据仓库和联机分析处理
  - 大型数据库中的有趣知识（规则、模式）

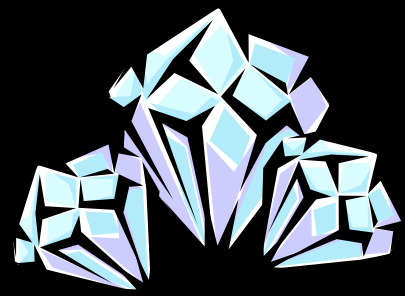
数据库技术的演化
（见图1-1）



- 20世纪60年代：
  - 数据收集，数据库创建，信息管理系统（IMS)和数据库管理系统(DBMS)
- 20世纪70年代
  - 关系数据模型，关系数据库管理系统工具
- 20世纪80年代
  - 关系数据库管理系统（RDBMS），高级数据模型 (面向对象、演绎等等)和面向应用的DBMS(空间的、科学的、工程的）
- 20世纪90年代至今
  - 数据挖掘和数据仓库，多媒体数据库和web数据库

什么是数据挖掘

- 数据挖掘（数据库中的知识发现）
  - 在大型数据库中提取有趣的（重要的，隐含的，目前未知的，潜在有用的）信息和模式
- 另外的名字和它们的“内在故事”
  - 数据挖掘: 一个错误的名字？
  - 数据库中的知识发现（挖掘）(KDD), 知识提取，数据/模式分析，数据考古，数据捕捞，信息收获和商业智能等等。
- 什么不是数据挖掘？
  - （演绎）询问过程
  - 专家系统或小型的统计程序



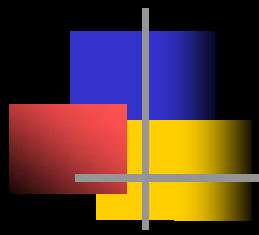
为什么进行数据挖掘——潜在应用

- 数据库分析和决定支持
  - 市场分析和管理
    - 目标市场， 用户关系管理,市场菜篮子分析,交叉销售, 市场分割。
  - 风险性分析和管理
    - 预测， 顾客保留,改善保险,质量控制, 竞争分析
  - 欺骗察觉和管理
- 其他应用
  - 文本挖掘 (新闻组，电子邮件，文件) 和WEB分析
  - 智能询问回答

市场分析和管理(1)

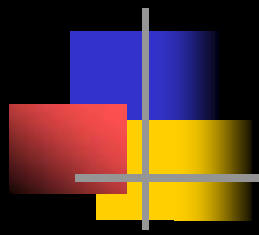
- 用于分析的数据从何来？
  - 信用卡交易， 信誉卡，折扣券，用户投诉电话，公众生活方式调查。
- 目标市场
  - 找出具有相同特征（兴趣，收入水平，消费习惯等等）的“模式”顾客群。
- 随着时间的推移决定顾客的购买方式
  - 从单独银行账户向联合银行账户的转变。例如：结婚
- 交叉市场分析
  - 不同产品之间的销售关联关系
  - 在此关联信息上进行预测

市场分析和管理(2)



- 顾客形象
  - 数据挖掘可以告诉你什麽样的顾客会买什麽样的产品（聚类或分类）
- 识别顾客需求
  - 保证为不同的顾客提供了最好的产品
  - 使用预测手段去发现什麽因素会吸引新的顾客。
- 提供汇总信息
  - 各种各样的多方位汇总信息
  - 统计的汇总信息（数据中心的趋势和变化）

公司分析和风险管理



- 财政计划和财产评估
  - 现金流分析和预测
  - 财产分析的偶发性需求分析
  - 典型性分析和时序分析（财政比率，趋势分析等等）
- 资源计划：
  - 总结和比较资源和花销
- 竞争：
  - 控制对手和市场的方向
  - 把顾客划分成许多类，依据类的划分编制价格程序
  - 把这个价格策略放到高度竞争的市场环境内

欺骗性检测和管理(1)

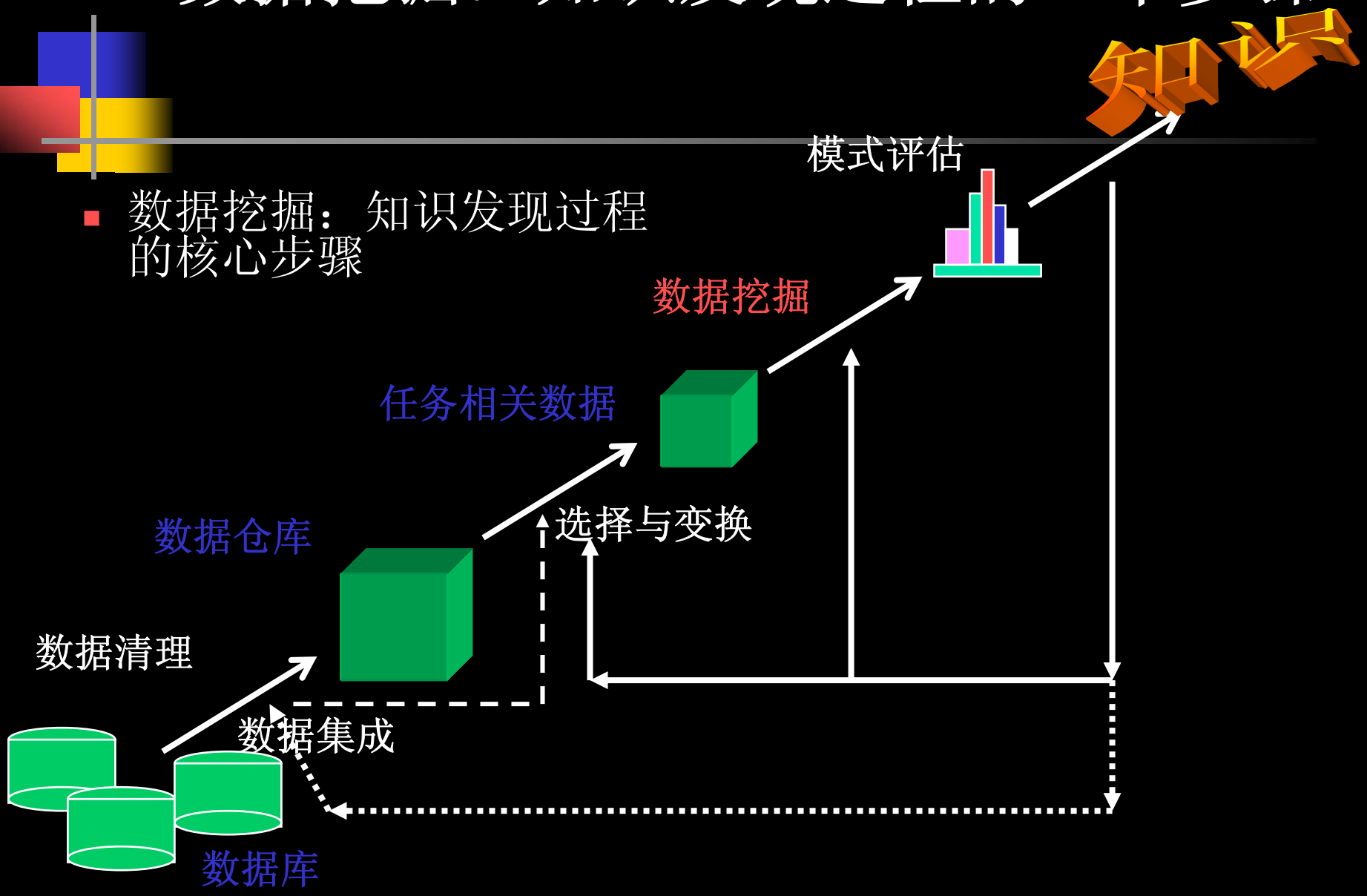
- 应用
  - 广泛应用于医疗系统, 零售系统,信用卡服务, 电信(电话卡欺骗行为), 等等.
- 实现途径
  - 利用历史性数据建立欺骗性行为模型并使用数据挖掘帮助识别同类例子
- 具体事例
  - 汽车保险：检测出那些故意制造车祸而索取保险金的人
  - 来路不明钱财的追踪: 发现可疑钱财交易(美国财政部的财政犯罪执行网)
  - 医疗保险: 检测出潜在的病人，呼叫医生和证明人

欺骗性检测和管理(2)

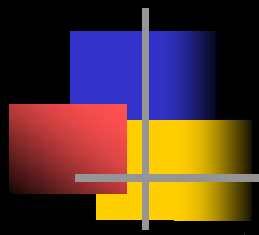
- 发现不正确的医学治疗
  - 澳大利亚医疗保险协会证明在许多情况下全面审查测试是很需要的
- 检测电话错误
  - 电话呼叫模式：呼叫目的地，持续时间，每天或每周的次数。分析与预期标准相背离的模式
- 零售
  - 分析家估计38%的零售收缩缘于雇员的不诚实。

数据挖掘：知识发现过程的一个步骤

- 数据挖掘：知识发现过程的核心步骤

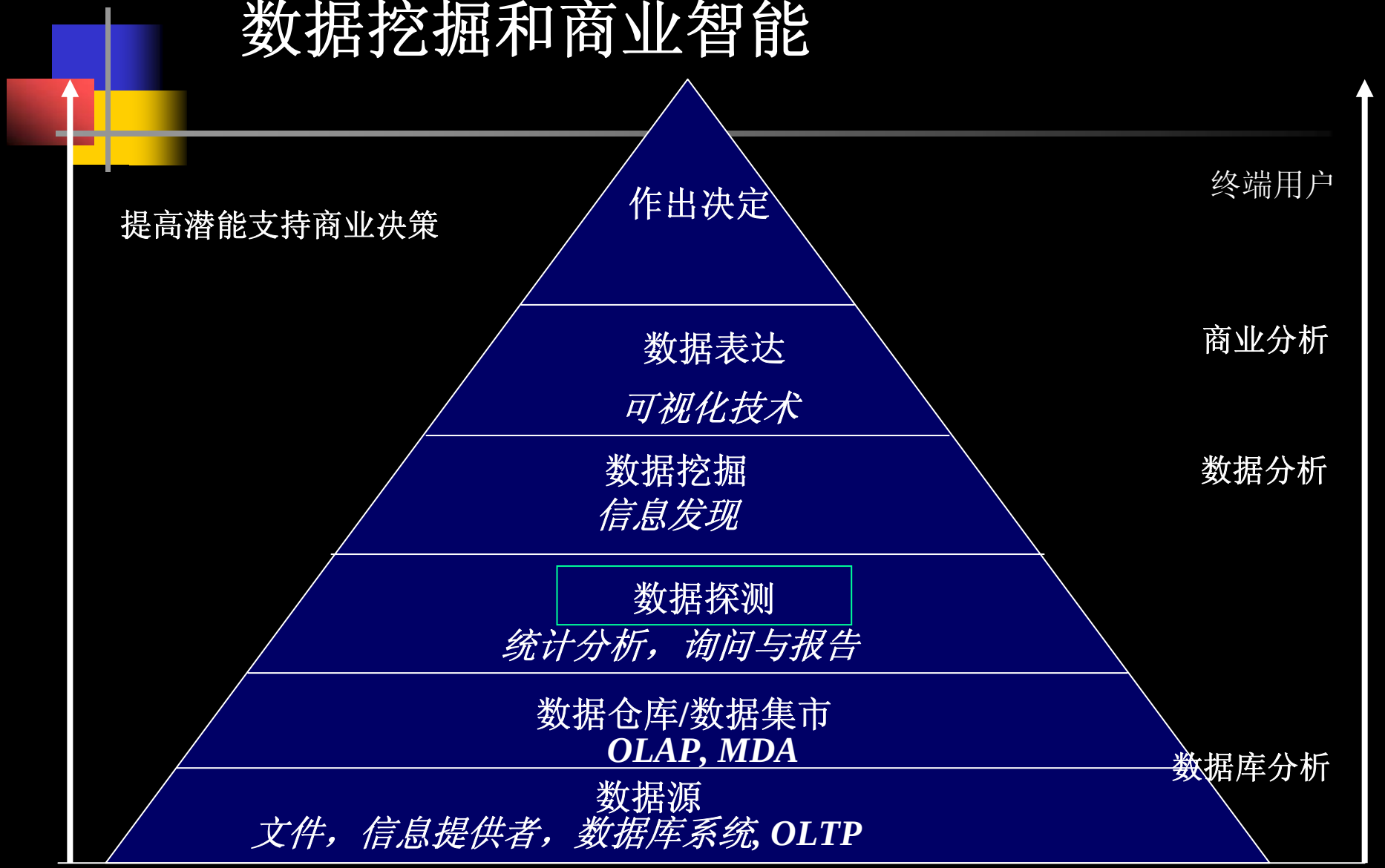


KDD过程的步骤

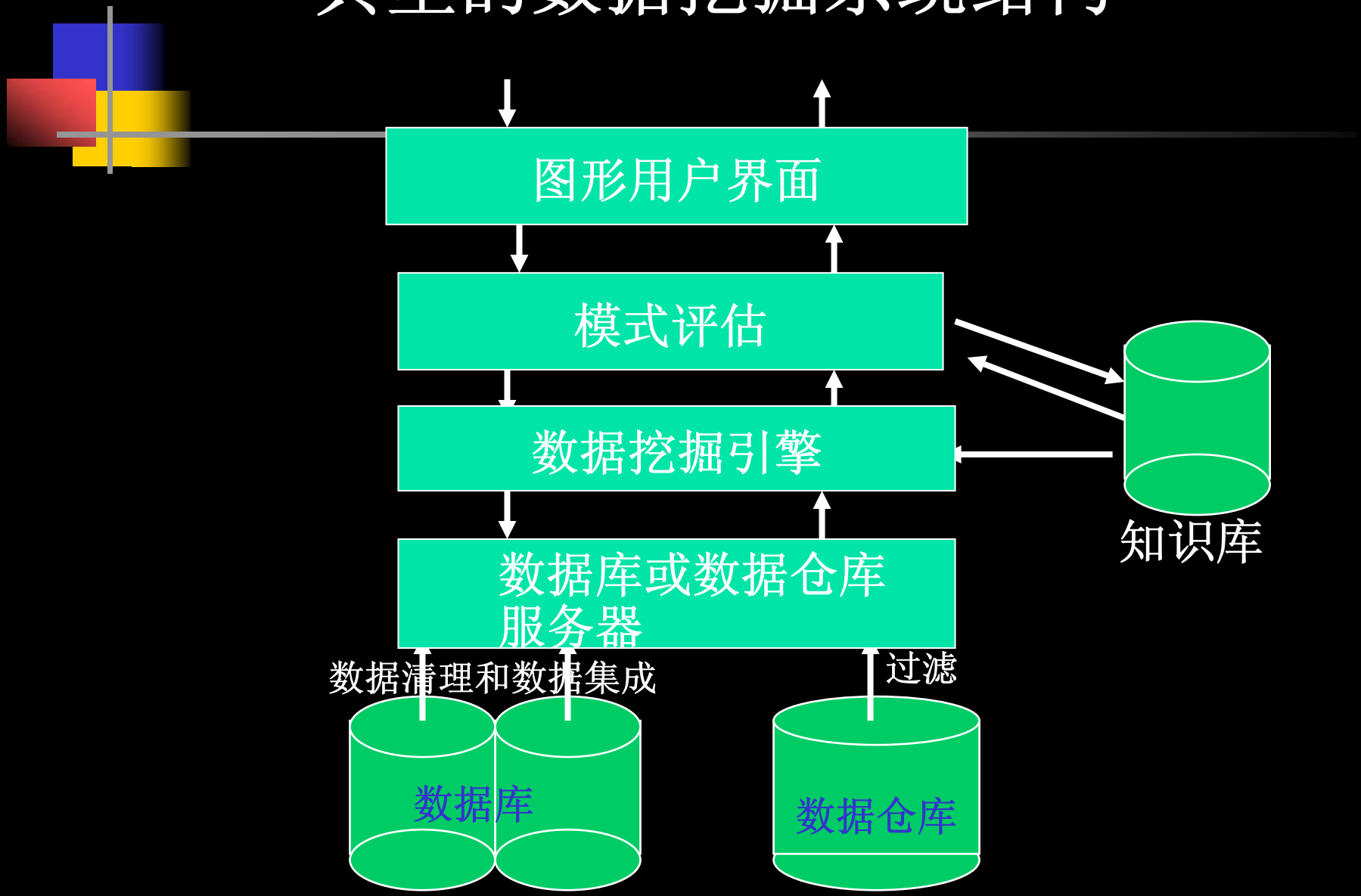


- 了解应用领域：
  - 相关的预备知识和应用目标
- 创建一个目标数据集：数据选择
- 数据清理和预加工（可能占用60%精力）
- 数据变换：
  - 发现有用的特征，维/变量的变换， 常量的表示
- 选择数据挖掘功能
  - 汇总，分类，关联，聚集
- 选择挖掘算法
- 数据挖掘：搜索兴趣模式
- 模式评估和知识表达
  - 可视化，变形，去掉冗余模式等等
- 使用发现的知识

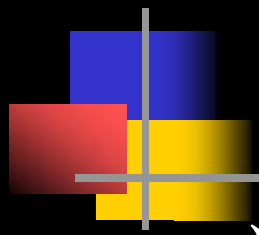
数据挖掘和商业智能


提高潜能支持商业决策
作出决定
数据表达
可视化技术
数据挖掘
信息发现
数据探测
统计分析，询问与报告
数据仓库/数据集市
OLAP, MDA
数据源
文件，信息提供者，数据库系统, OLTP
终端用户
商业分析
数据分析
数据库分析

典型的数据挖掘系统结构



在何种数据上进行数据挖掘



- 关系数据库
- 数据仓库
- 事务数据库
- 高级数据库与信息库
  - 面向对象和对象-关系数据库
  - 空间数据库
  - 时间序列数据库和暂时数据库
  - 文本数据库和多媒体数据库
  - 异源数据库和继承数据库
  - **WWW**

数据挖掘功能（1）

- 概念描述：特征化和区分
  - 归纳，概括和比较数据特征，例如，干燥地区和湿润地区
- 关联分析（相关性和因果关系）
  - 多维关联和单维关联
  - age(X, “20..29”) ^ income(X, “20..29K”) → buys(X, “PC”) [support = 2%, confidence = 60%]
  - contains(T, “computer”) → contains(x, “software”) [1%, 75%]

数据挖掘功能（2）

- *分类和预测*
  - 找出描述并区分数据类和概念的模型（或函数）以便能够使用模型预测类标记未知的对象类。
  - 例如：依据气候划分国家类型或者依据每里的耗油量划分汽车类型。
  - 表示形式：判定树，分类规则，神经网络。
  - 预测：预测某些未知的或空缺的数据值。
- *聚类分析*
  - 类标记未知：把数据聚类或分组成新的类，例如：把房子聚类来找出房子的分布模式。
  - 聚类依据以下原则：最大化类内的相似性和最小化类间的相似性。

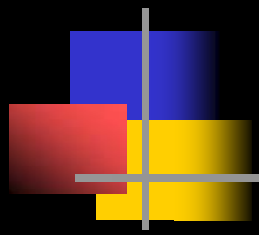
数据挖掘功能（3）

- 孤立点分析
  - 孤立点：和数据的一般行为和模型不一致的数据。
  - 它可被视为噪声或异常，但在它在欺骗检测和罕见事件的分析中是非常有用的。
- 演变分析
  - 序列模式挖掘，周期分析
  - 基于类似性的数据分析
- 其他面向模式或统计分析

所有发现模式都是有趣的吗？

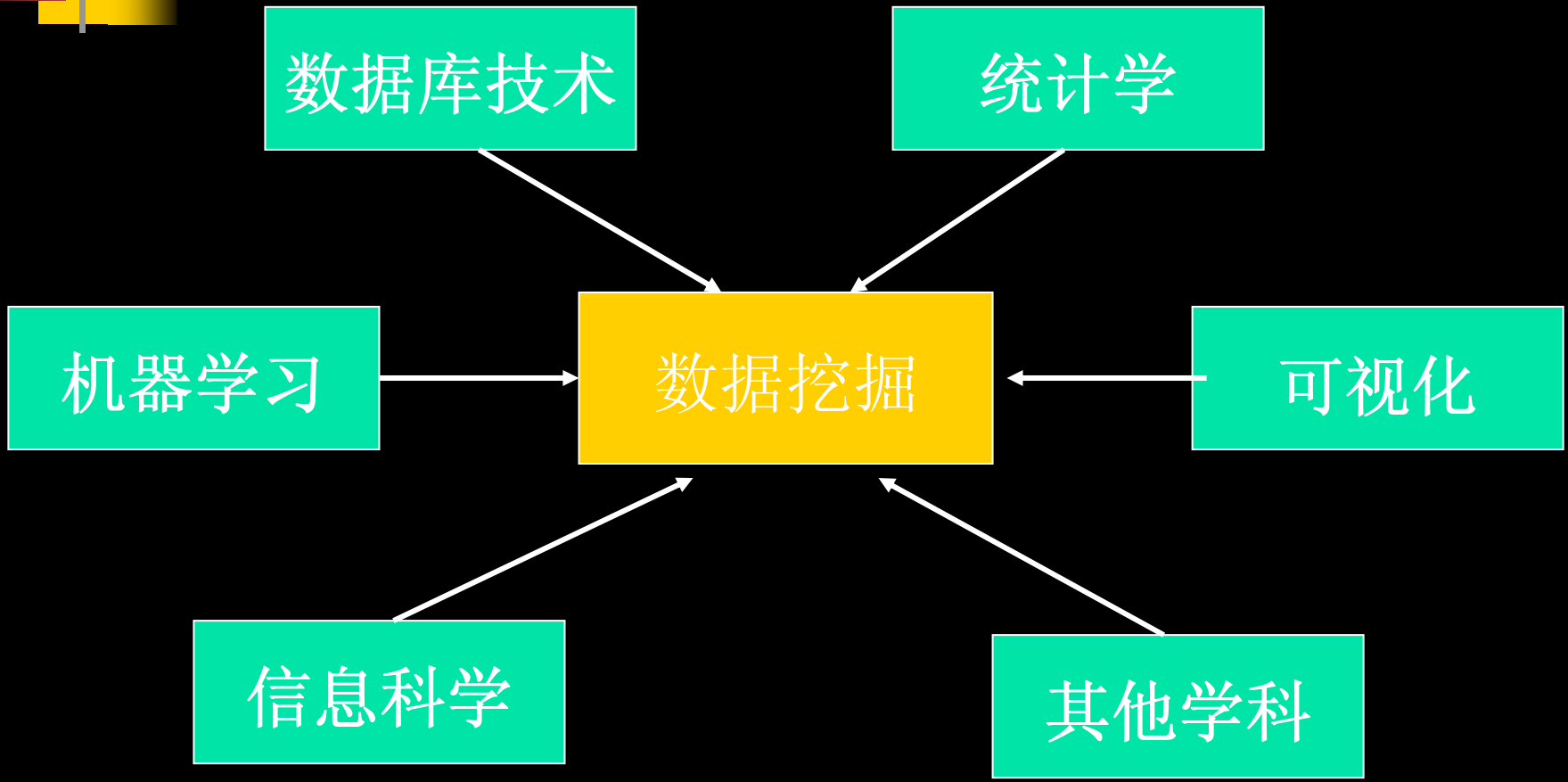
- 数据挖掘系统可以产生数以万计的模式，但不是所有的模式都是有趣的。
- 模式兴趣度度量：一个模式是有趣的如果（1）它易于被人理解；(2)在某种程度上，对于新的或测试数据是有效的；（3）是潜在有用的；（4）是新颖的或对用户正在寻求证实的假设是有效的。
- 客观兴趣度度量和主观兴趣度度量
  - 客观度量：依据统计和模式结构。例如：支持度（support)和置信度(confidence)
  - 主观度量：基于用户对数据的确信。例如：出乎意料的，新颖的，可行动的。

我们能找出所有的并只是有趣的模式吗？

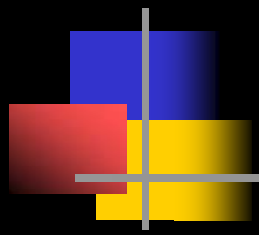


- 找出所有的有趣模式：涉及数据挖掘的完全性
  - 一个数据挖掘系统能找出所有的有趣模式吗？
  - 关联，分类，聚类
- 搜索只是有趣的模式：是数据挖掘的优化问题
  - 一个数据挖掘系统能够仅产生有趣的模式吗？
  - 方法
    - 首先概括所有的模式，接着过滤非有趣模式。
    - 仅产生有趣的模式——挖掘询问的最优化

数据挖掘受多学科的影响



数据挖掘的分类表

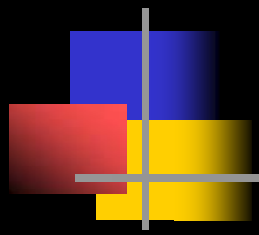


- 根据一般功能分类
  - 描述性数据挖掘
  - 预测性数据挖掘
- 不同的观点，不同的分类
  - 根据挖掘的数据库类型分类
  - 根据挖掘的知识类型分类
  - 根据所用的技术分类
  - 根据应用分类

数据挖掘分类的多维视图

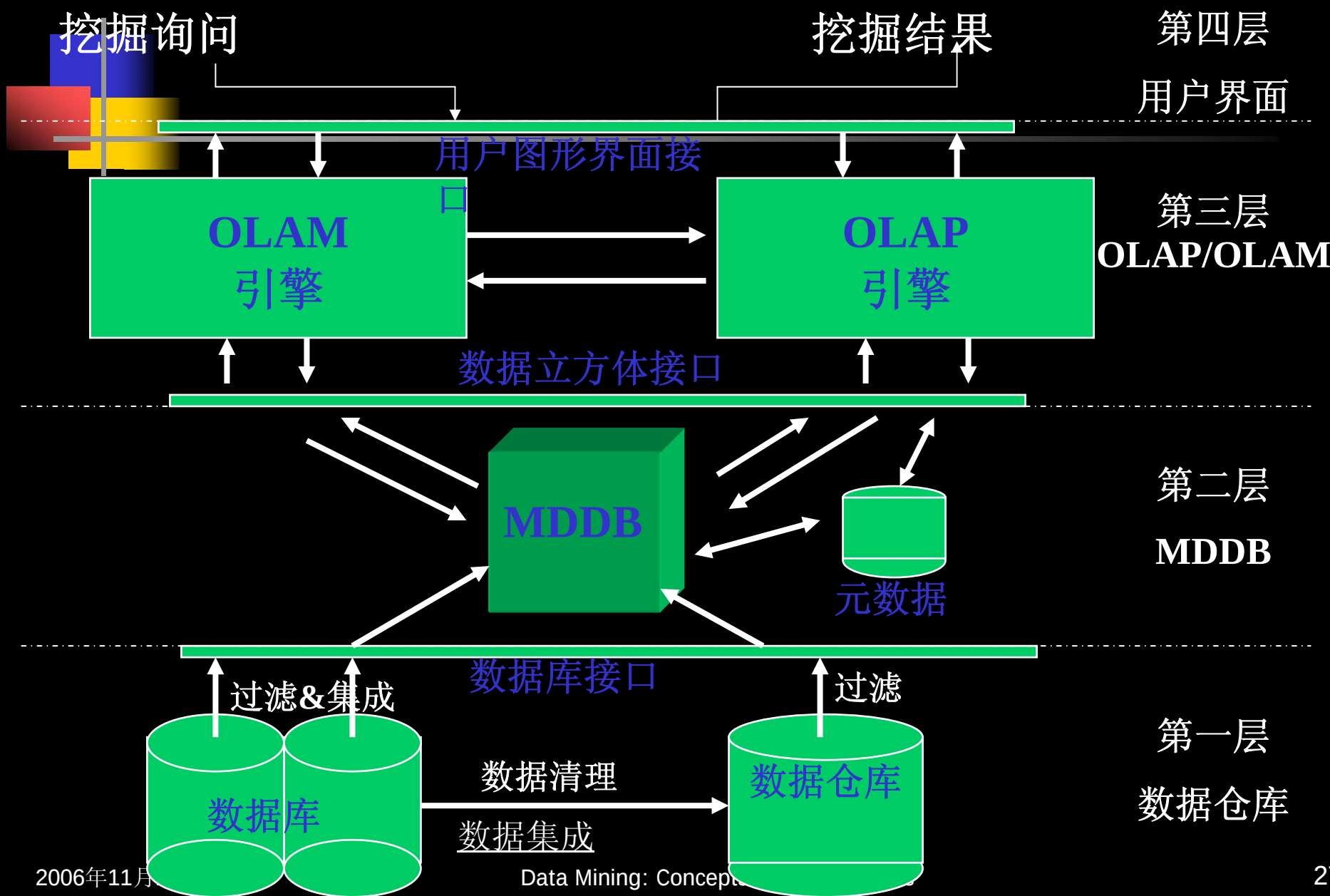
- 根据挖掘的数据库类型分类：
  - 我们可以有关系的，事务的，面向对象的，对象-关系的，空间的，时间序列的，文本的，异源的，，www的数据挖掘系统。
- 根据挖掘的知识分类：
  - 特征化，区分，关联，分类聚类，趋势分析，偏差分析，孤立点分析。
  - 多层/完整功能和多层次上的挖掘。
- 根据所用的技术分类
  - 面向数据库，数据仓库，机器学习，统计学，可视化，神经网络等。
- 根据应用分类
  - 零售、电信、银行、错误分析、DNA、股票市场、web挖掘、日志分析等等。

OLAP挖掘：数据挖掘和数据仓库结合



- 数据挖掘系统，数据库管理系统，数据仓库
  - 非耦合，疏松耦合，半紧密耦合，紧密耦合
- 联机分析数据挖掘
  - 数据挖掘和**OLAP**的结合
- 交互式挖掘多层知识
  - 通过下钻/上卷，转轴，切片/切块等，在不同的层次，挖掘知识和模式的必要性。
- 多种挖掘功能的综合
  - 特征化的分类，先聚集再关联

一个OLAM结构

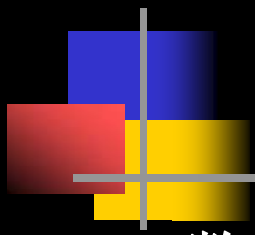

挖掘询问
挖掘结果
第四层
用户界面
用户图形界面接口
OLAM
引擎
OLAP
引擎
第三层
OLAP/OLAM
数据立方体接口
MDDB
元数据
第二层
MDDB
数据库接口
过滤&集成
过滤
数据清理
数据集成
数据库
数据仓库
第一层
数据仓库

数据挖掘的主要问题（1）

- 挖掘方法和用户交互问题
  - 在数据库上挖掘不同类型的知识
  - 多个抽象层的交互知识挖掘
  - 结合背景知识
  - 数据挖掘查询语言和特定的数据挖掘
  - 数据挖掘结果的表示和显示
  - 处理噪声和不完全数据
  - 模式评估——兴趣度问题
- 性能问题
  - 数据挖掘算法的有效性和可伸缩性
  - 并行，分布式和增量挖掘算法

数据挖掘的主要问题(2)

- 关于数据库类型的多样性问题
  - 关系和复杂的数据类型的处理
  - 有异种数据库和全球信息系统挖掘信息（**www)**
- 关于应用和社会冲击问题
  - 被发现知识的应用
    - 特殊领域的数据挖掘工具
    - 智能询问的回答
    - 过程控制和决策制造
  - 被发现数据和存在知识的综合：知识融合问题
  - 数据安全性，完整性和隐私性的保护。



小结

- 数据挖掘：从大量数据中发现有趣模式
- 随着数据库技术的广泛应用对其技术的自然演变的需求也越来越大。
- 只是发现过程包括：数据清理，数据集成，数据选择，数据变换，数据挖掘，模式评估，知识表达。
- 挖掘可以在各种各样的信息库中进行。
- 数据挖掘的功能：特征化，区分，关联，分类，聚类，孤立点分析和趋势分析。
- 数据挖掘的主要问题

文献的出处

- Data mining and KDD (SIGKDD member CDROM):
  - Conference proceedings: KDD, and others, such as PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery
- Database field (SIGMOD member CD ROM):
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.
- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics:
  - Conference proceedings: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization:
  - Conference proceedings: CHI, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

参考文献

- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
- T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.
- G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

http://www.cs.sfu.ca/~han

Thank you !!!