# 数据挖掘：概念与技术

武永亮

# 大数据方向的宣讲图

# 数据挖掘工作缺口

# 生活中的数据挖掘

- 垃圾邮件的处理，邮件分类，过滤，筛选….
- 商品推荐，相关度推荐，猜你喜欢….
- 匹配，游戏匹配…..
- 预测，天气预报，股市，房价…

算法：可以完成任何事

# 生活中的数据挖掘

- 写出你的性别，姓名（可用昵称）
- 写出你期望的未来一半的身高，体重，大数据成绩区间.

# 常见混淆概念

- 数据挖掘、机器学习、模式识别…

# 教材-作者

- **http://www.cs.illinois.edu/homes/hanj/**
- **The book will be covered in two courses at CS, UIUC：伊利诺伊大学，厄巴纳-尚佩恩(University of Illinois at Urbana-Champaign)**
  - **CS412: Introduction to data warehousing and data mining Coverage (Chapters 1-7 of This Book)**
  - **CS512: Data mining: Principles and algorithms (Chapters 8-11 of This Book)**

*Jiawei Han*

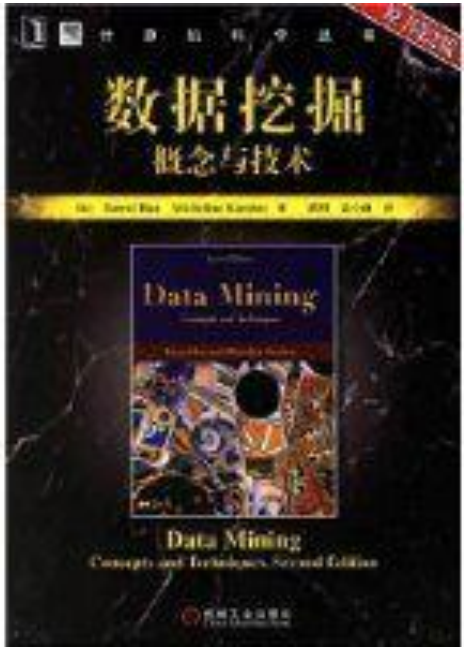
Professor, Department of Computer Science
Univ. of Illinois at Urbana-Champaign
Rm 2132, Siebel Center for Computer Science
201 N. Goodwin Avenue
Urbana, IL 61801, USA
E-mail: hanj[at]cs.uiuc.edu

Ph.D. (1985), Computer Science, Univ. Wisconsin-Madison

Data Mining and Databas

Data Mining Research Group
(Data Mining Group Summary Report: Spr
Database and Information Systems Researc
(UIUC Academic Calendar)

Fax: (217) 265-6494
Web: www.cs.uiuc.edu/~hanj



# 课程信息

- 数据挖掘的（前7章的内容），
  - 第1章 引言
  - 第2章 数据预处理
  - 第3章 数据仓库与OLAP技术概述
  - 第4章 数据立方体计算与数据泛化
  - 第5章 挖掘频繁模式、关联和相关
  - 第6章 分类和预测
  - 第7章 聚类分析
- 导论课程（从数据库角度出发）
- 相关涉及：数据仓库、数据库系统、统计学与机器学习的概念和技术

# 课时安排与考核

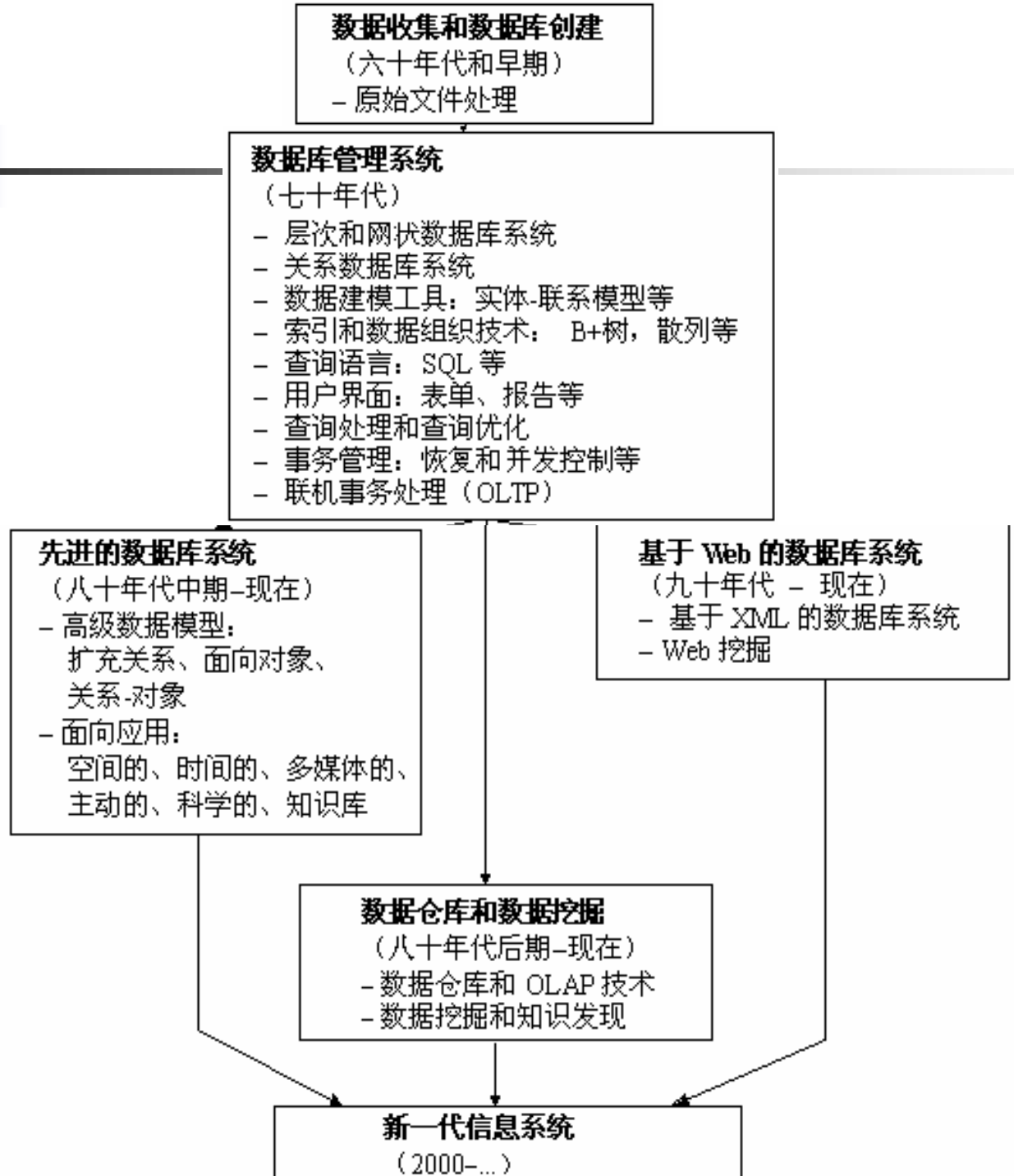
- 课时安排
  - 总学时 **18**，课次**6**半天，共**2**周
- 考核
  - 平时成绩：**6**次作业
  - 考试成绩：

# 第1章 引论

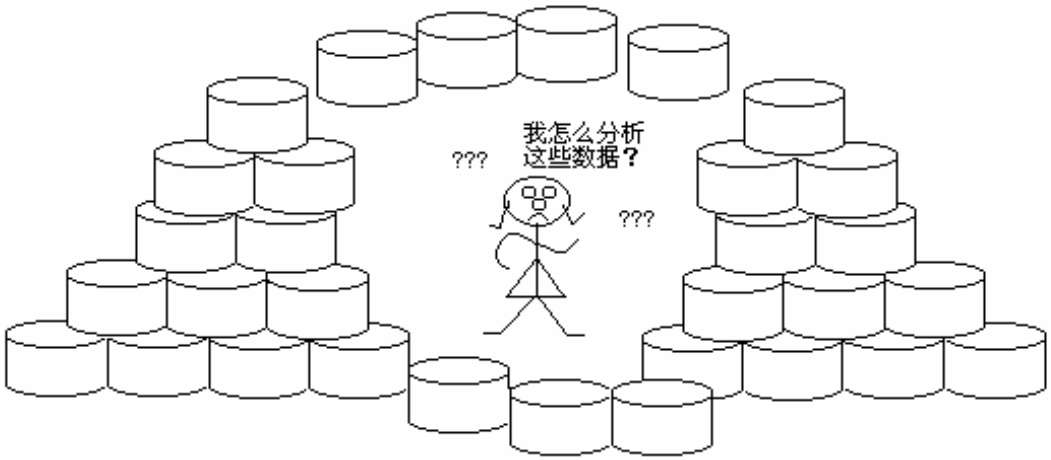
- 动机：为什么要数据挖掘?
- 什么是数据挖掘?
- 数据挖掘：在什么数据上进行?
- 数据挖掘功能
- 所有的模式都是有趣的吗?
- 数据挖掘系统分类
- 数据挖掘的主要问题

# 数据处理技术的演进

- **1960s:**
  - **数据收集, 数据库创建, IMS层次和网状 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—2000s:**
  - **数据挖掘和数据仓库, 多媒体数据库, 和 Web 数据库**


数据收集和数据库创建
（六十年代和早期）
– 原始文件处理
数据库管理系统
（七十年代）
– 层次和网状数据库系统
– 关系数据库系统
– 数据建模工具：实体-联系模型等
– 索引和数据组织技术： B+树，散列等
– 查询语言：SQL 等
– 用户界面：表单、报告等
– 查询处理和查询优化
– 事务管理：恢复和并发控制等
– 联机事务处理（OLTP）
先进的数据库系统
（八十年代中期–现在）
– 高级数据模型：
扩充关系、面向对象、
关系-对象
– 面向应用：
空间的、时间的、多媒体的、
主动的、科学的、知识库
基于 Web 的数据库系统
（九十年代 – 现在）
– 基于 XML 的数据库系统
– Web 挖掘
数据仓库和数据挖掘
（八十年代后期–现在）
– 数据仓库和 OLAP 技术
– 数据挖掘和知识发现
新一代信息系统
（2000–…）

# 动机: 需要




- 数据爆炸问题
  - 自动的数据收集工具和成熟的数据库, 数据仓库, 和其它信息存储库
    - **Business: Web, e-commerce, transactions, stocks, …**
    - **Science: Remote sensing, bioinformatics, scientific simulation, …**
    - **Society and everyone: news, digital cameras, YouTube**
- 我们正被数据淹没,但却缺乏知识
  - 数据丰富，但信息贫乏
- 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)

# 数据挖掘界简史

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.
- ACM Transactions on KDD starting in 2007

# Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

# Where 2 Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

# 什么是数据挖掘?

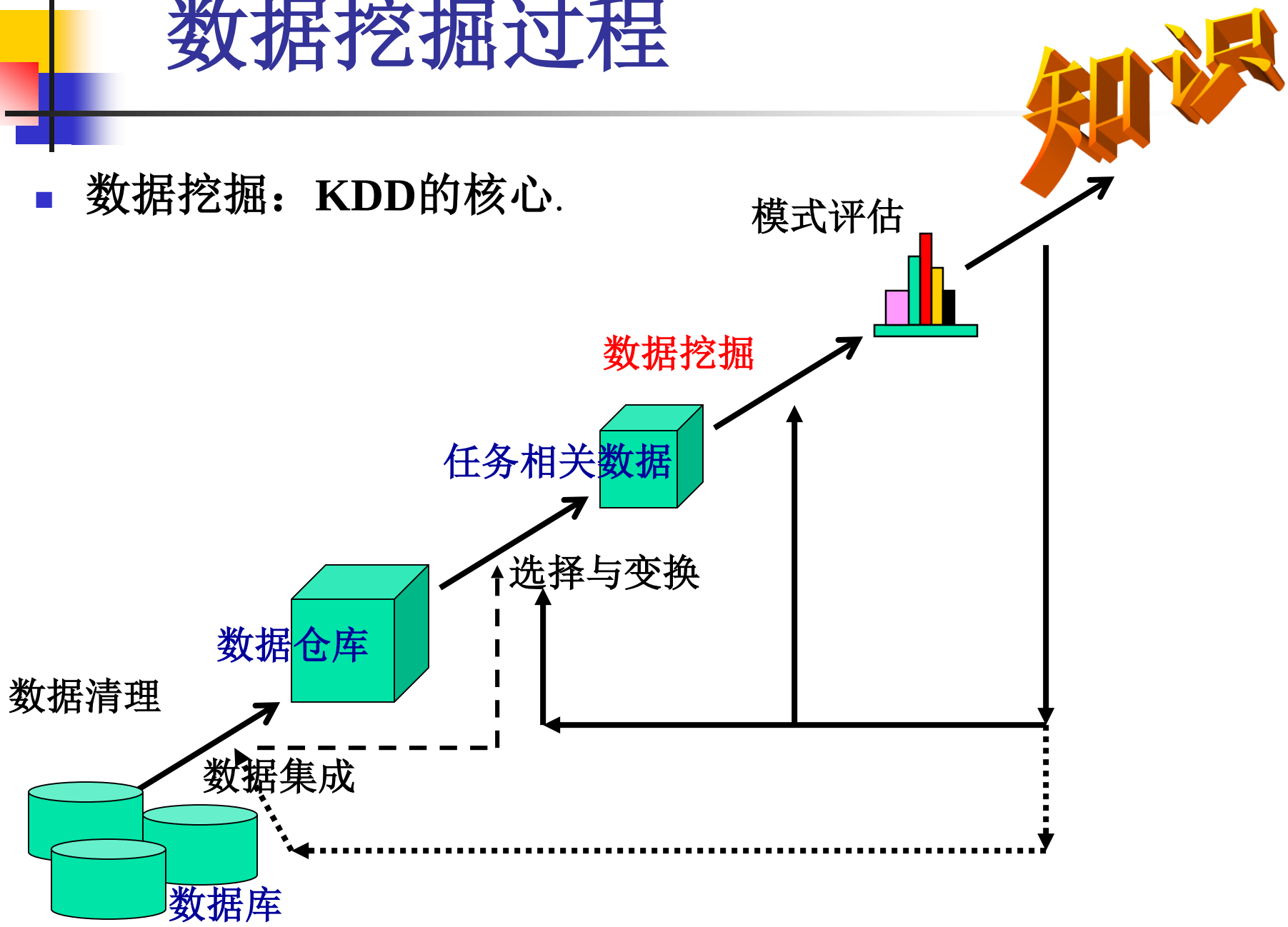
- 数据挖掘 (从数据中挖掘知识):
  - 从大型数据库中提取有趣的 (非平凡的, 蕴涵的, 先前未知的 并且是潜在有用的) 信息或模式
  - 数据挖掘: 用词不当?
- 其它叫法和“inside stories”内幕新闻 :
  - 数据库中知识发现(挖掘) (**Knowledge discovery in databases**, **KDD**), 知识提取(knowledge extraction), 数据/模式分析(data/pattern analysis), 数据考古(**data archeology**), 数据捕捞(**data dredging**), 信息收获(**information 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
    - 处理大量数据/ 有效的可伸缩的技术

# Why Not Traditional Data Analysis?

- 巨大的数据Tremendous amount of data
  - **Algorithms must be highly scalable to handle such as tera-bytes of data**
- High-dimensionality of data
  - **Micro-array may have tens of thousands of dimensions**
- High complexity of data
  - **Data streams and sensor data**
  - **Time-series data, temporal data, sequence data**
  - **Structure data, graphs, social networks and multi-linked data**
  - **Heterogeneous databases and legacy(遗产) databases**
  - **Spatial, spatiotemporal, multimedia, text and Web data**
  - **Software programs, scientific simulations**
- New and sophisticated applications

# 数据挖掘过程

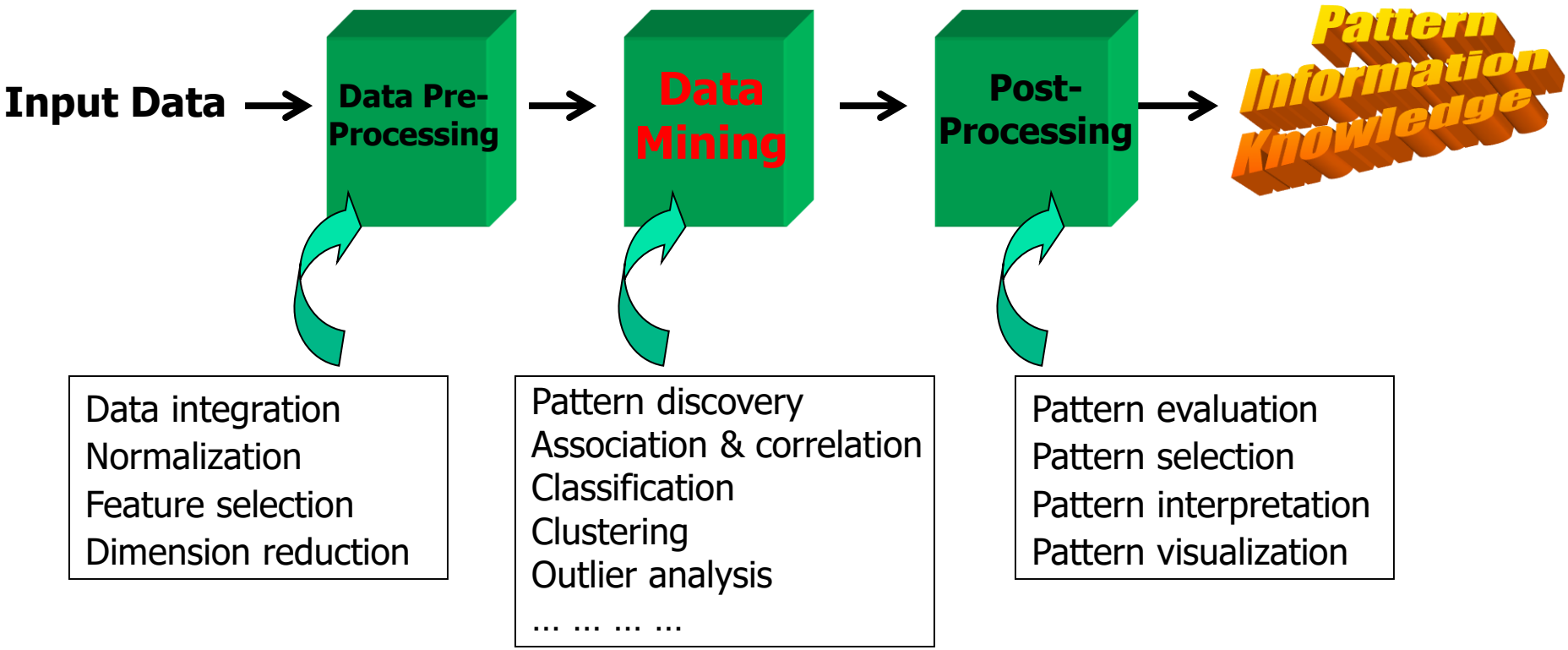
- 数据挖掘：KDD的核心.



# 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.
- 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

# KDD过程: 机器学习和统计的角度




- **This is a view from typical machine learning and statistics communities**

# 典型的数据挖掘系统结构

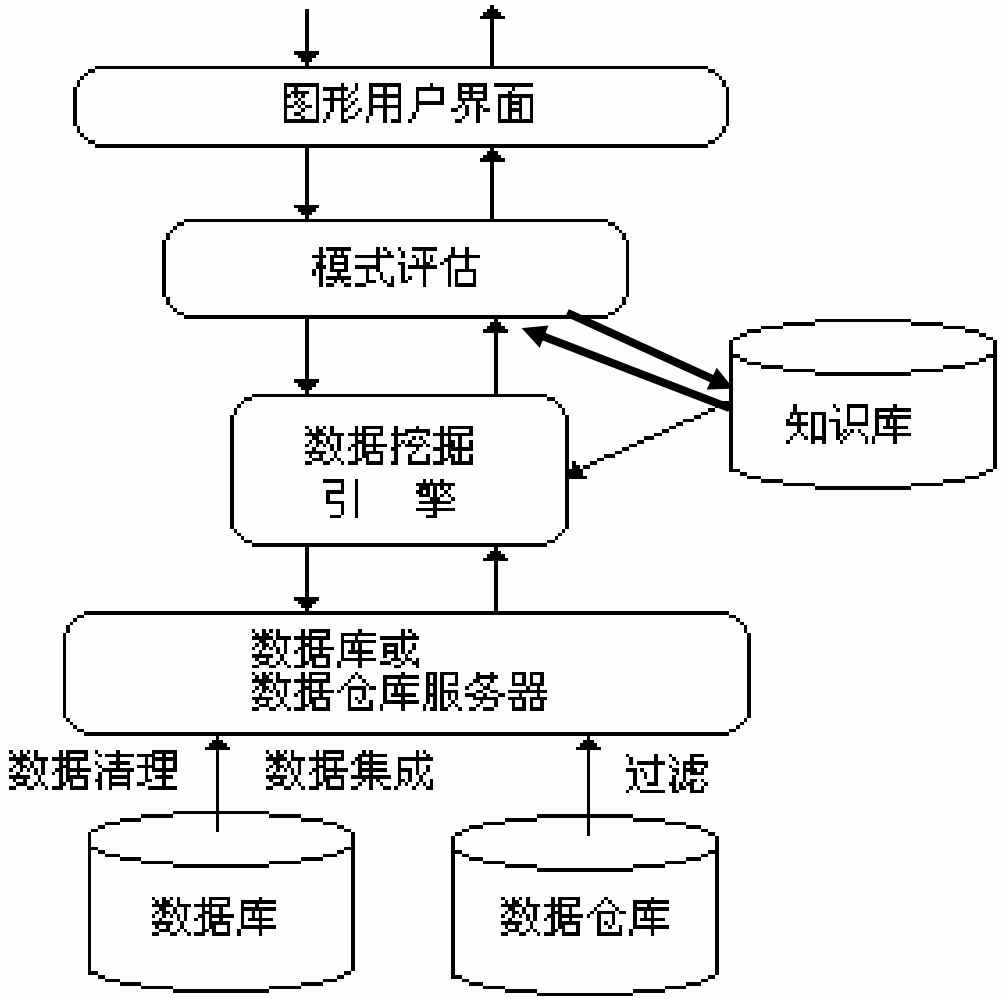
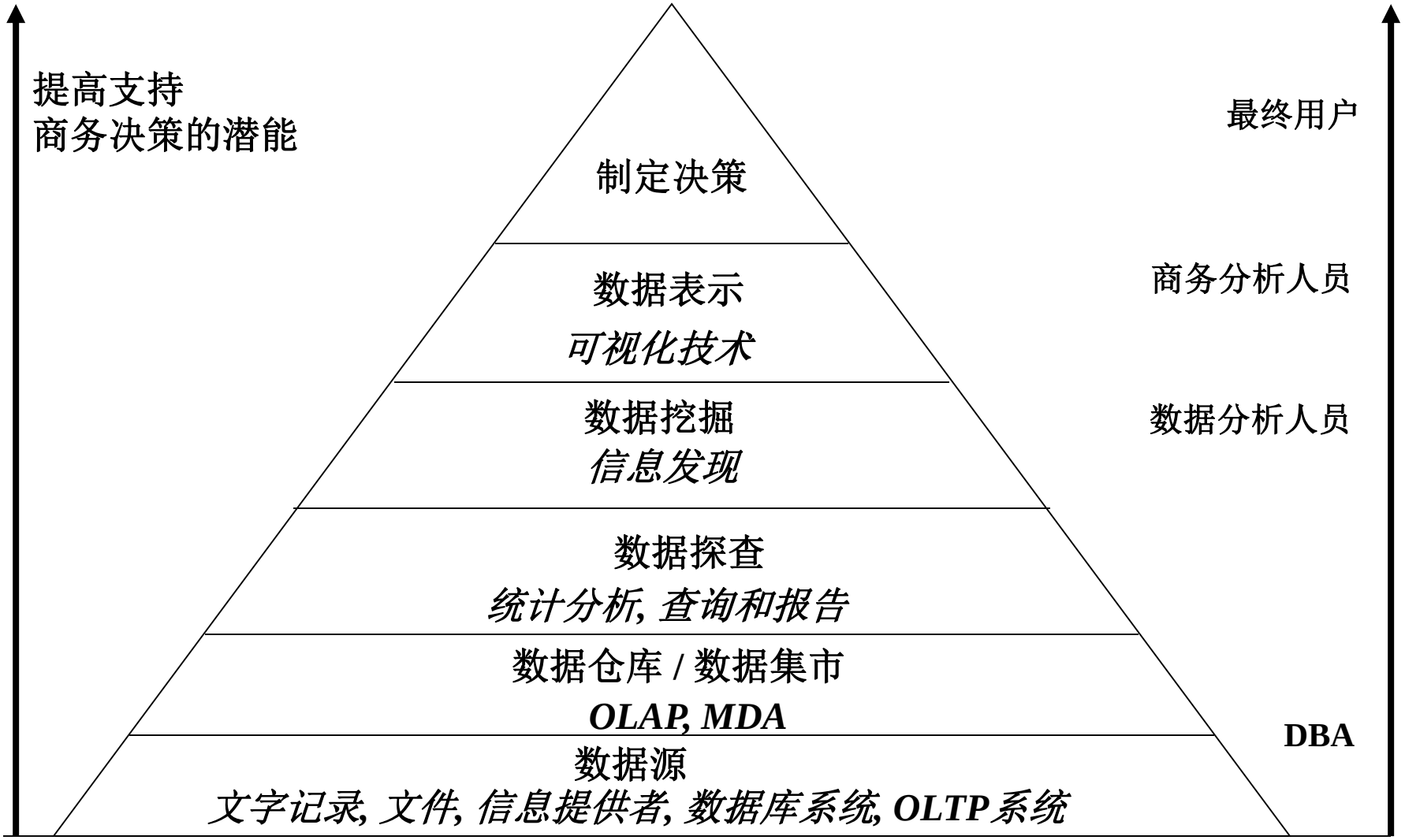



图 1.5：典型的数据挖掘系统结构

# 数据挖掘和商务智能


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

# 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.
  - 生物信息学/生物 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)

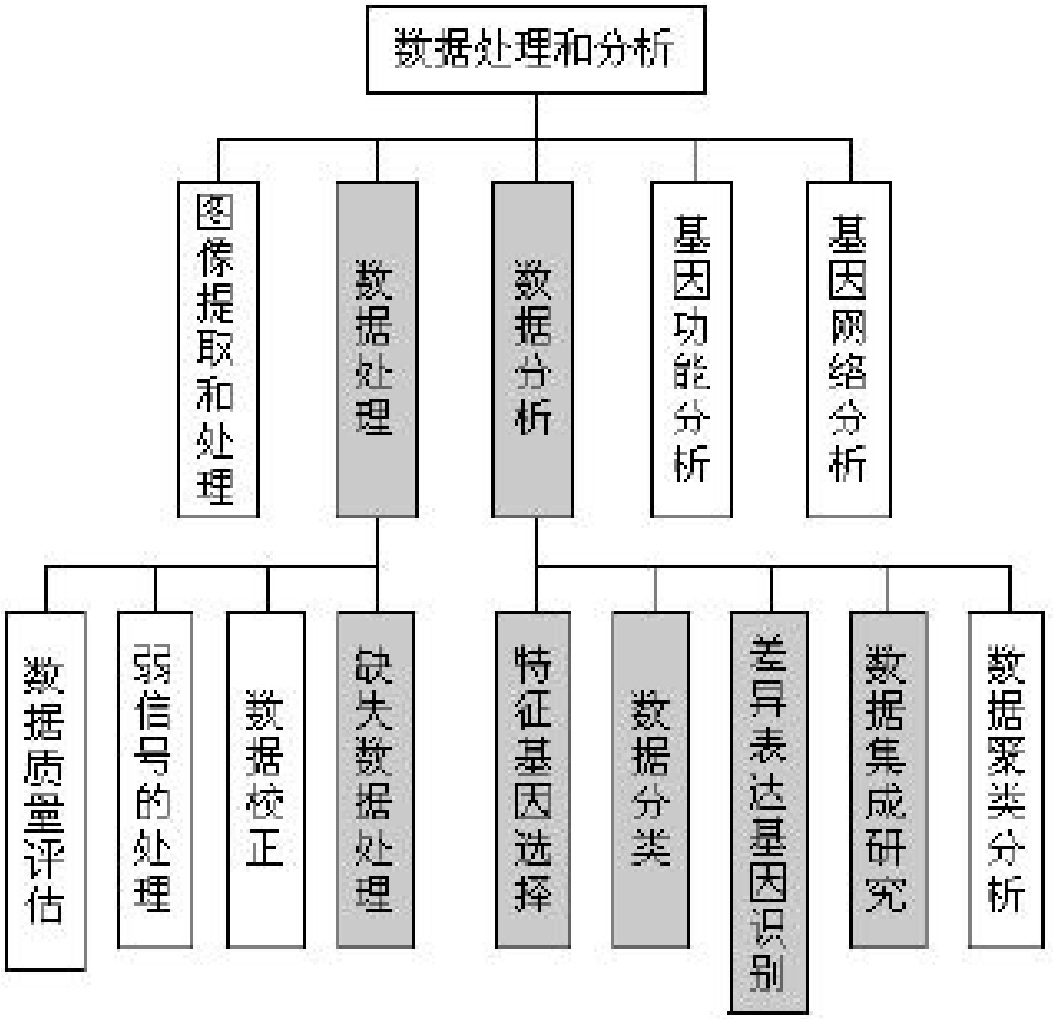
- 应用
  - 广泛用于健康照料, 零售, 信用卡服务, 电讯 (电话卡欺骗), 等.
- 方法
  - 使用历史数据建立欺骗行为模型, 使用数据挖掘帮助识别类似的实例
- 例
  - 汽车保险: 检测这样的人, 他/她假造事故骗取保险赔偿
  - 洗钱: 检测可疑的金钱交易 (US Treasury's Financial Crimes Enforcement Network)
  - 医疗保险 : 检测职业病患者, 医生和介绍人圈

# 欺骗检测和管理(2)

- 检测不适当的医疗处置
  - 澳大利亚健康保险会(Australian Health Insurance Commission) 发现许多全面的检查是请求做的，而不是实际需要的 (每年节省100万澳元).
- 检测电话欺骗
  - 电话呼叫模式: 通话距离, 通话时间, 每天或每周通话次数. 分析偏离期望的模式.
  - 英国电讯(British Telecom)识别频繁内部通话的呼叫者的离散群, 特别是移动电话, 超过数百万美元的欺骗.
- 零售
  - 分析家估计, 38%的零售业萎缩是由于不忠诚的雇员造成的.

# 生物数据分析/挖掘

- **microarray data**
- **biological seque**
- **biological netwo**
- 生物文本挖
  - 文本数据中抽取
  - 从抽取信息中**i**



# 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.**
- **Web：页面的分类、聚类、推荐/用户的访问模式**

# 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Characterization and discrimination
  - 概化, 汇总和比较数据特征, 例如, 干燥和潮湿的地区
- 频繁模式,关联,相关 **Frequent patterns, association, correlation vs. causality**
  - 频繁模式：数据中频繁出现的模式
  - 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)

- 

(a)

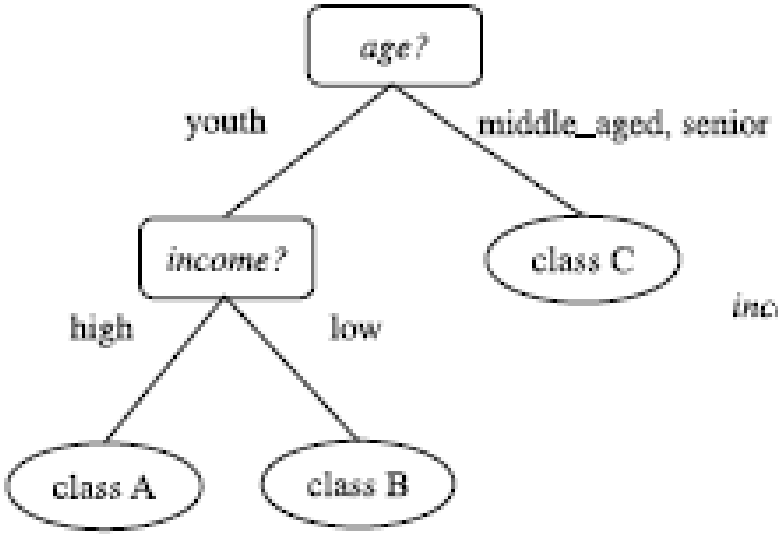
age(X, "youth") AND income(X, "high") ⟶ class(X, "A")

age(X, "youth") AND income(X, "low") ⟶ class(X, "B")

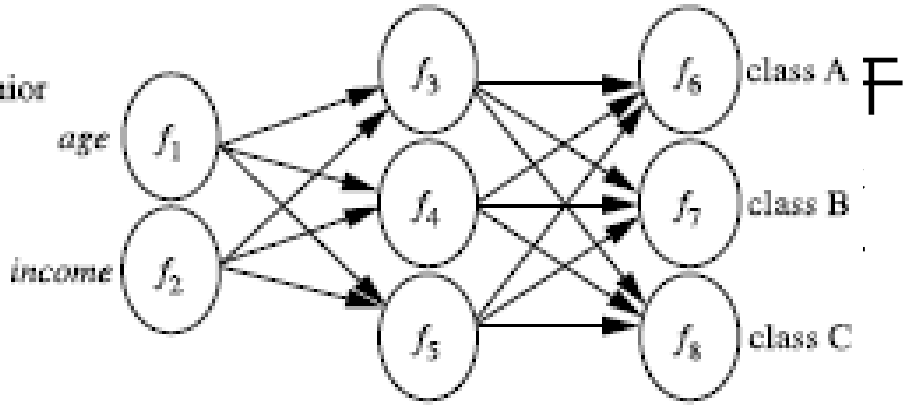
age(X, "middle_aged") ⟶ class(X, "C")

age(X, "senior") ⟶ class(X, "C")

(b)

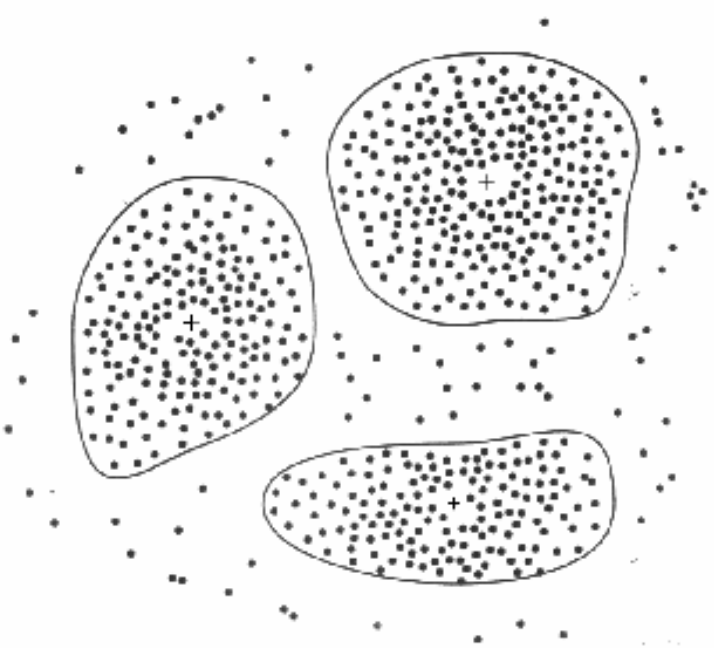



(c)



# 数据挖掘功能(3)

- **聚类分析Unsupervised learning (i.e., Class label is unknown)**
  - 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



.10 关于一个城市内顾客的 2-D 图，显示了 3 个聚类，

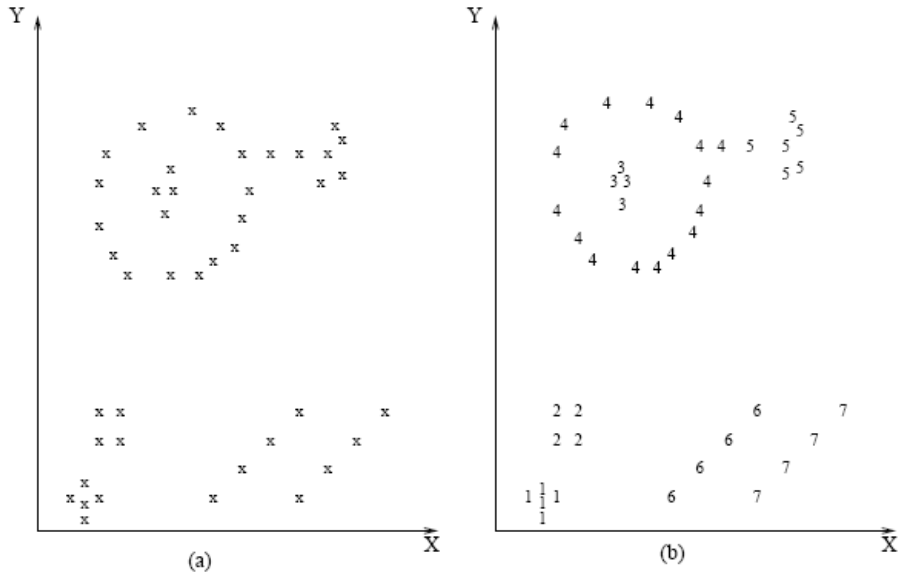



Figure 1. Data clustering.

# 数据挖掘功能(4)

- 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
    - **e.g., first buy digital camera, then buy large SD memory cards**
  - 基于相似的分析
    - **Approximate and consecutive motifs**

# 数据挖掘功能(5) -Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …


37

# Top-10 Most Popular DM Algorithms:18 Identified Candidates (I)

- Classification
  - #1. C4.5: Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann., 1993.
  - #2. CART: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, 1984.
  - #3. K Nearest Neighbours (kNN): Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. TPAMI. 18(6)
  - #4. Naive Bayes Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
- Statistical Learning
  - #5. SVM: Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
  - #6. EM: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York. Association Analysis
  - #7. Apriori: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
  - #8. FP-Tree: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.

# The 18 Identified Candidates (II)

- Link Mining
  - **#9. PageRank: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.**
  - **#10. HITS: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. SODA, 1998.**
- Clustering
  - **#11. K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.**
  - **#12. BIRCH: Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.**
- Bagging and Boosting
  - **#13. AdaBoost: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.**

# The 18 Identified Candidates (III)

- Sequential Patterns
  - #14. GSP: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. 5th International Conference on Extending Database Technology, 1996.
  - #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
- Integrated Mining
  - #16. CBA: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
- Rough Sets
  - #17. Finding reduct: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992
- Graph Mining
  - #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

# Top-10 Algorithm Finally Selected at ICDM'06

- #1: C4.5 (61 votes)
- #2: K-Means (60 votes)
- #3: SVM (58 votes)
- #4: Apriori (52 votes)
- #5: EM (48 votes)
- #6: PageRank (46 votes)
- #7: AdaBoost (45 votes)
- #7: kNN (45 votes)
- #7: Naive Bayes (45 votes)
- #10: CART (34 votes)

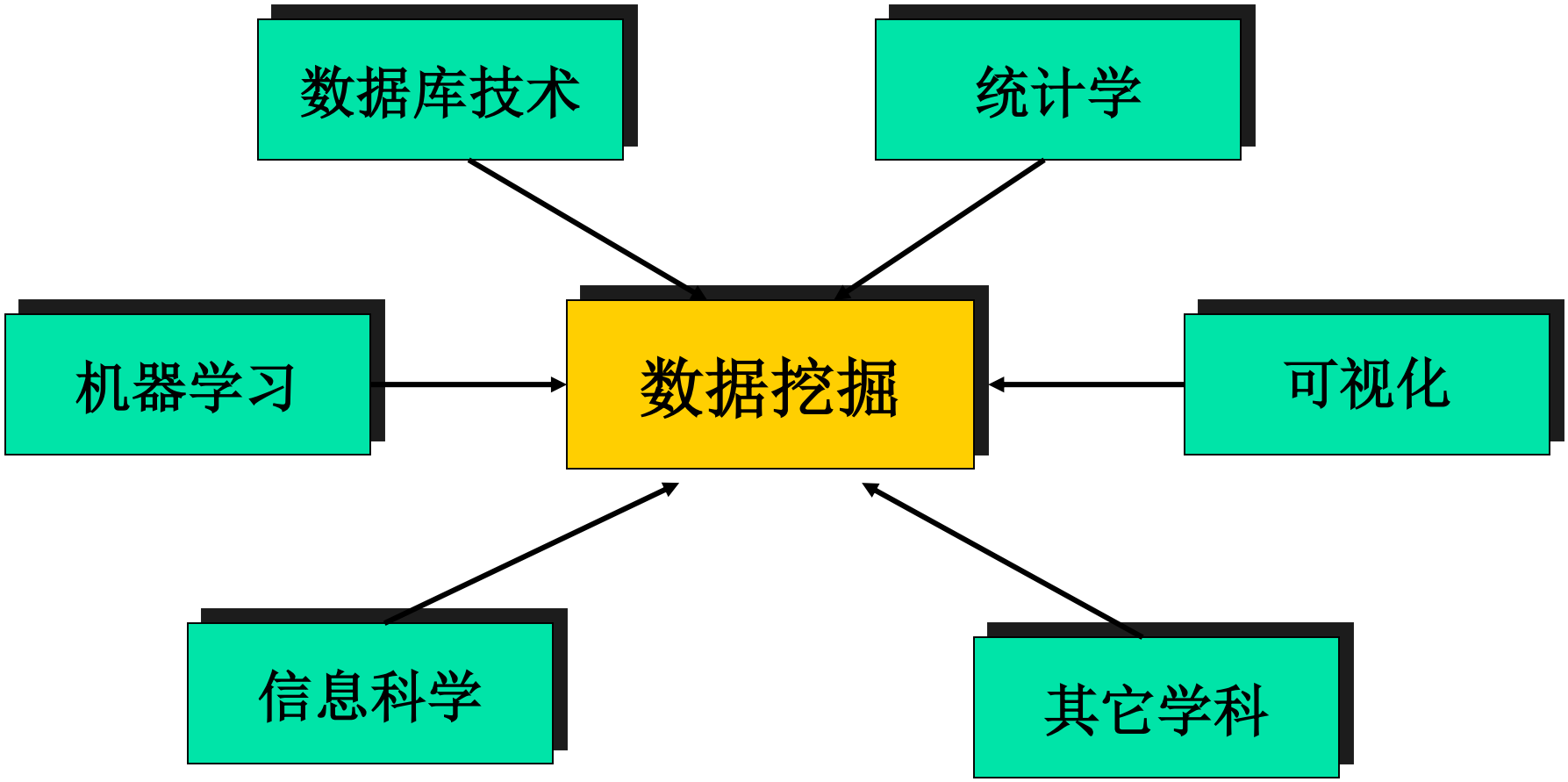
# 挖掘出的所有模式都是有趣的吗?

- 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



# 数据挖掘分类

- 一般功能
  - 描述式数据挖掘——描述数据的一般性质
  - 预测式数据挖掘——对数据进行推断，做预测
- 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

# 数据挖掘分类的多维视图

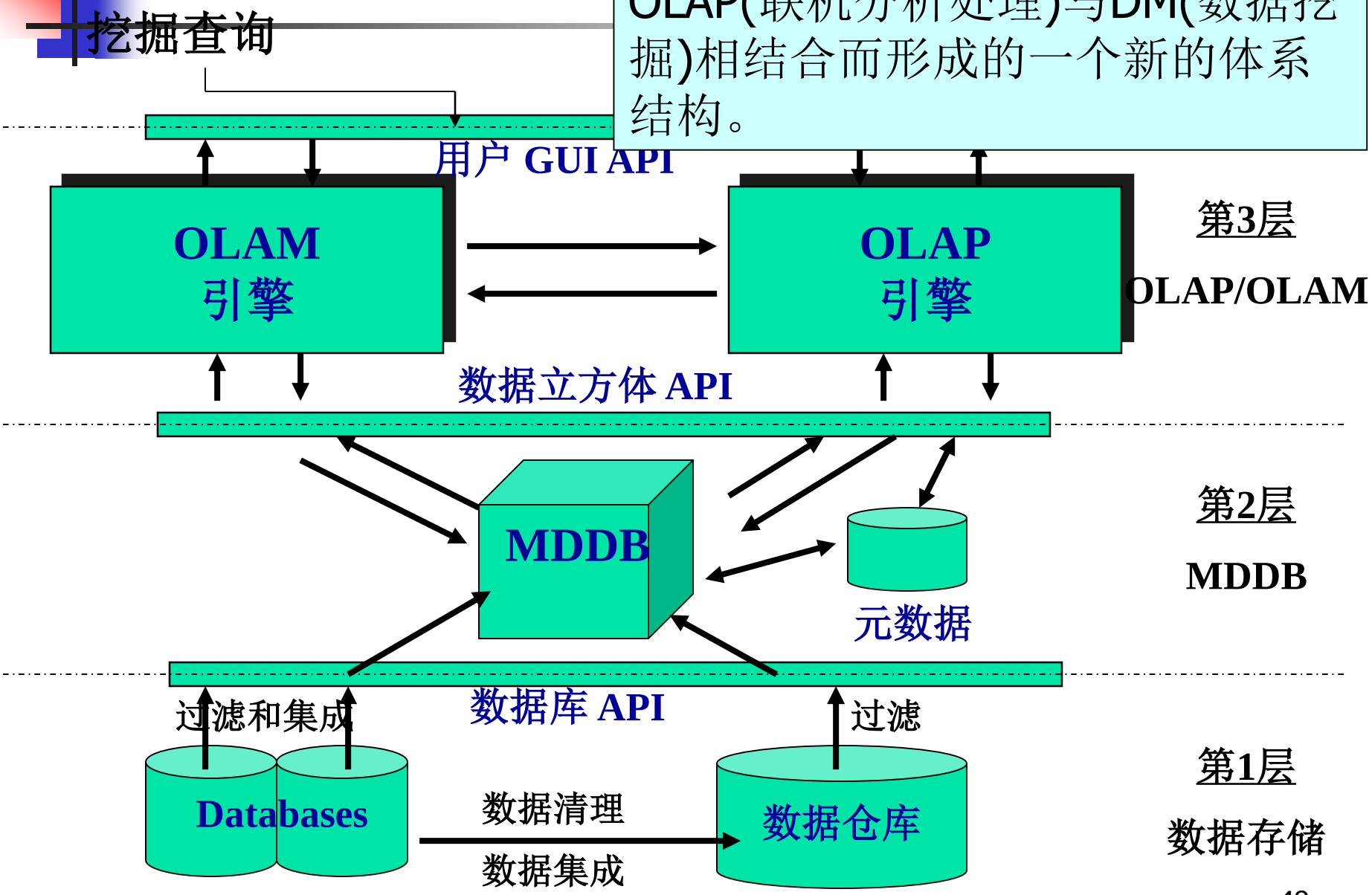
- 待挖掘的数据库
  - 关系的, 事务的, 面向对象的, 对象-关系的, 主动的, 空间的, 时间序列的, 文本的, 多媒体的, 异种的, 遗产的, **WWW**, 等.
- 所挖掘的知识
  - 特征, 区分, 关联, 分类, 聚类, 趋势, 偏离和孤立点分析, 等.
  - 多/集成的功能, 和多层次上的挖掘
- 所用技术
  - 面向数据库的, 数据仓库 **(OLAP)**, 机器学习, 统计学, 可视化, 神经网络, 等.
- 适合的应用
  - 零售, 电讯, 银行, 欺骗分析, **DNA** 挖掘, 股票市场分析, **Web** 挖掘, **Web**日志分析, 等

# OLAP挖掘: 数据挖掘与数据仓库的集成

- 数据挖掘系统, DBMS, 数据仓库系统的耦合
  - 不耦合, 松耦合, 半紧密耦合, 紧密耦合
- 联机分析挖掘
  - 挖掘与 OLAP 技术的集成
- 交互挖掘多层知识
  - 通过下钻, 上卷, 转轴, 切片, 切块等操作, 在不同的抽象层挖掘知识和模式的必要性.
- 多种挖掘功能的集成
  - 特征分类, 先聚类再关联

# OLAM 的结构

*OLAM*(数据联机分析挖掘)是OLAP(联机分析处理)与DM(数据挖掘)相结合而形成的一个新的体系结构。



# Why Data Mining Query Language?

- Automated vs. query-driven?
  - Finding all the patterns autonomously in a database?—unrealistic because the patterns could be too many but uninteresting
- Data mining should be an interactive process
  - User directs what to be mined
- Users must be provided with a set of primitives(原语,基本要素) to be used to communicate with the data mining system
- Incorporating these primitives in a data mining query language
  - More flexible user interaction
  - Foundation for design of graphical user interface
  - Standardization of data mining industry and practice

# 数据挖掘查询语言

- 通过数据挖掘查询语言，数据挖掘任务可以通过查询的形式输入到数据挖掘系统中。
- 定义数据挖掘查询语言的优势

①可以让用户透明地使用各种数据挖掘查询语句，而不管它们是怎样实现的；

②可以把数据挖掘平滑地集成到各种应用系统上，而不像当前这样，每一个数据挖掘应用系统都得从头开发设计；

③可以使数据挖掘的研究工作更有继承性，通用良好的查询语言的定义是进一步工作的基础。

# Primitives that Define a Data Mining Task

- Task-relevant data
  - Database or data warehouse name
  - Database tables or data warehouse cubes
  - Condition for data selection
  - Relevant attributes or dimensions
  - Data grouping criteria
- Type of knowledge to be mined
  - Characterization, discrimination, association, classification, prediction, clustering, outlier analysis, other data mining tasks
- Background knowledge
- Pattern interestingness measurements
- Visualization/presentation of discovered patterns

# 数据挖掘原语

用户在进行数据挖掘时，总希望能够通过使用一组数据挖掘原语来与数据挖掘系统通信，以支持有效的和有成果的知识发现。

这组原语包括：

①与任务相关的数据；

②要挖掘的知识类型；

③用于挖掘过程的背景知识：概念分层；

④评估模式的兴趣度度量和阈值；

⑤可视化发现模式的期望表示。

DMQL 就是基于对这些原语的说明而设计出来的一种有效的数据挖掘查询语言。该语言采用类似于 SQL 的语法，因此它易于和关系查询语言 SQL 集成在一起。

# Primitive 3: Background Knowledge

- A typical kind of background knowledge: Concept hierarchies
- Schema hierarchy
  - E.g., street < city < province_or_state < country
- Set-grouping hierarchy
  - E.g., {20-39} = young, {40-59} = middle_aged
- Operation-derived hierarchy
  - email address: hagonzal@cs.uiuc.edu

    login-name < department < university < country
- Rule-based hierarchy
  - low_profit_margin (X) <= price(X, $P_1$) and cost (X, $P_2$) and ($P_1$ - $P_2$) < $50

# Primitive 4: Pattern Interestingness Measure

- **Simplicity**

  **e.g., (association) rule length, (decision) tree size**

- **Certainty**

  **e.g., confidence, P(A|B) = #(A and B)/ #(B), classification reliability or accuracy, certainty factor, rule strength, rule quality, discriminating weight, etc.**

- **Utility**

  **potential usefulness, e.g., support (association), noise threshold (description)**

- **Novelty**

  **not previously known, surprising (used to remove redundant rules, e.g., Illinois vs. Champaign rule implication support ratio)**

# Primitive 5: Presentation of Discovered Patterns

- Different backgrounds/usages may require different forms of representation
  - E.g., rules, tables, crosstabs, pie/bar chart, etc.
- Concept hierarchy is also important
  - Discovered knowledge might be more understandable when represented at high level of abstraction
  - Interactive drill up/down, pivoting, slicing and dicing provide different perspectives to data
- Different kinds of knowledge require different representation: association, classification, clustering, etc.

# An Example Query in DMQL

**Example 1.11 Mining classification rules.** Suppose, as a marketing manager of *AllElectronics*, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than $40,000, and who have bought more than $1,000 worth of items, each of which is priced at no less than $100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL[3] as follows, where each line of the query has been enumerated to aid in our discussion.

```
use database AllElectronics_db
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age
mine classification as promising_customers
in relevance to C.age, C.income, I.type, I.place_made, T.branch
from customer C, item I, transaction T
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
        and C.income ≥ 40,000 and I.price ≥ 100
group by T.cust_ID
having sum(I.price) ≥ 1,000
display as rules
```

# 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

# 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.

# 谢谢大家!