数据挖掘
第1章 引言

教师：王东京
学院：计算机学院
邮箱：dongjing.wang@hdu.edu.cn

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:

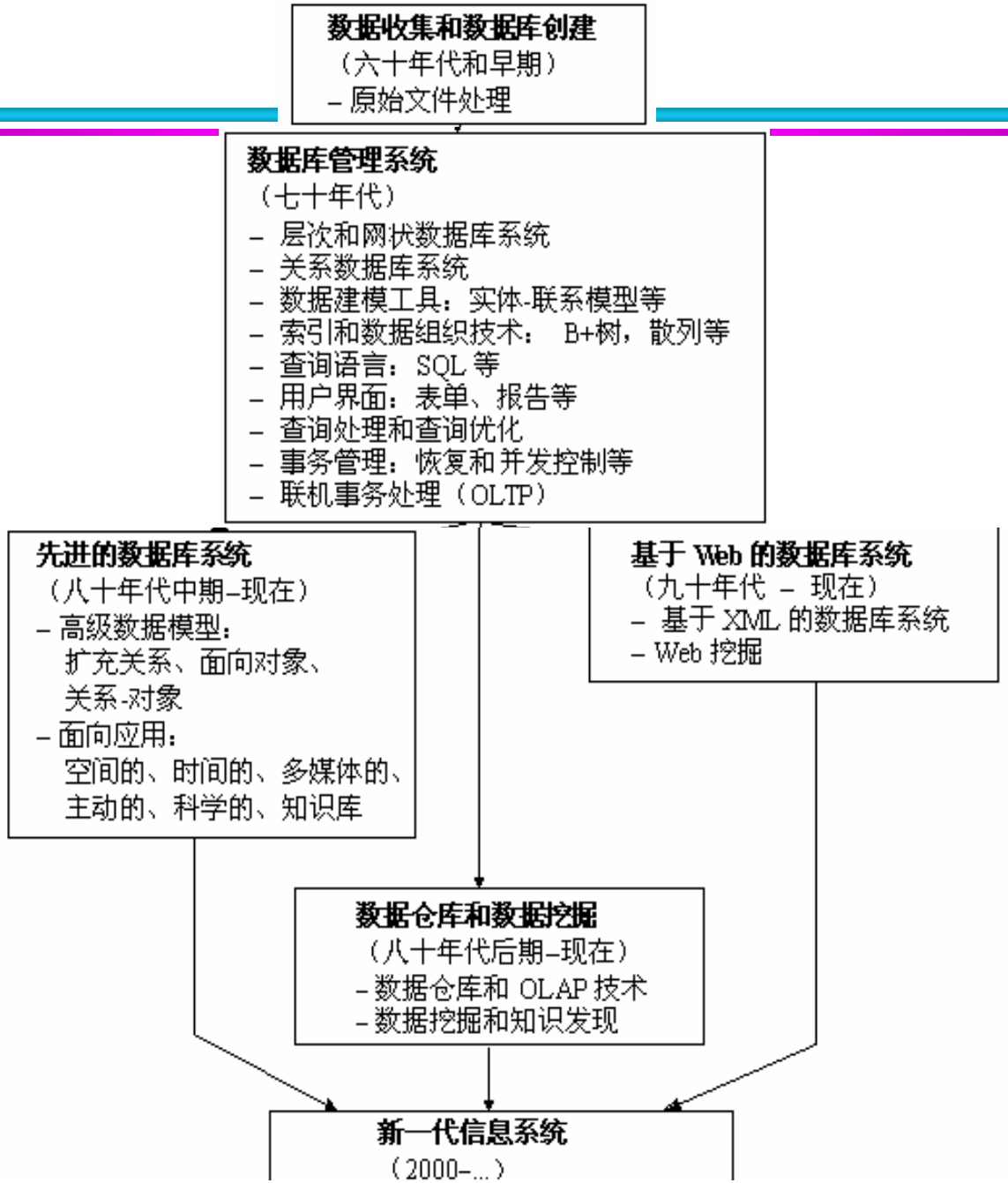
- 关系数据库模型, 关系 DBMS 实现

1980s:

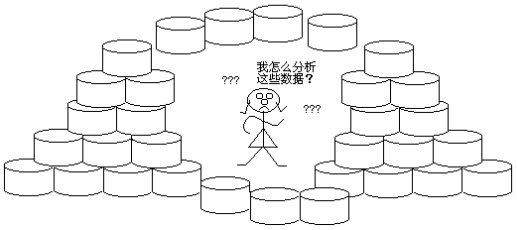
- RDBMS, 先进的数据模型 (扩充关系的, OO, 演绎的, 等.) 和面向应用 的 DBMS (空间的, 科学的, 工程的, 等.)

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
  - 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

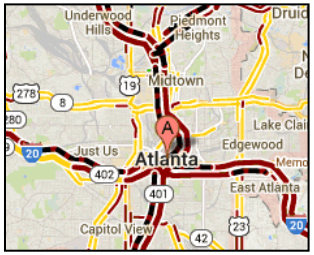
- 数据仓库与联机分析处理(OLAP)
- 从大型数据库的数据中提取有趣的知识(规则, 规律性, 模式, 限制等)

大规模数据无处不在！

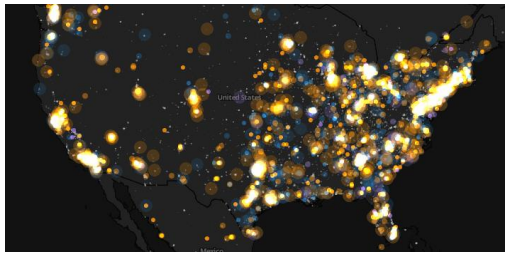
- 得益于数据生成和采集技术的发展，商业领域和科学领域产生了海量数据。
- 新口号（New mantra）
  - 随时随地收集可以获取的任何数据。
- 期望（Expectations）
  - 收集的数据将对收集的目的或未预期的目的具有价值。

网络安全

电子商务

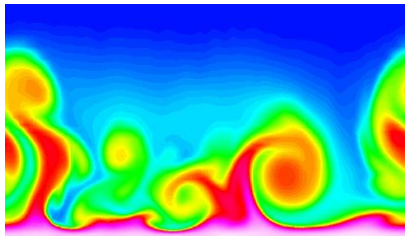


交通模式



社交网络

传感器网络



计算模拟

数据挖掘? 商业视角

商业领域已经收集并存储（warehoused）了海量数据

- 网络数据
  - Yahoo has Peta Bytes of web data
  - Facebook has billions of active users
- 购买数据，电商数据
  - Amazon handles millions of visits/day
- 银行/信用卡数据，交易数据

计算机性能越来越强，成本持续下降（cheaper and more powerful)

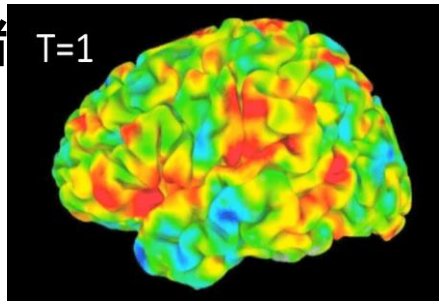
竞争压力

- Provide better, customized services for an edge (e.g. in Customer Relationship Management)

数据挖掘？科学视角

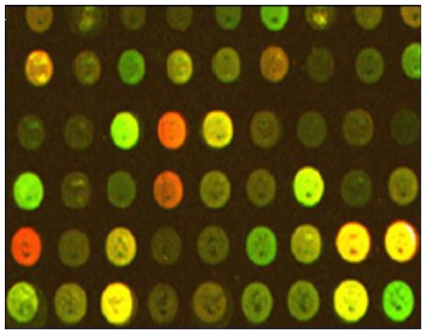
数据正在以前所未有的速度被采集和存储

- 卫星遥感数据
  - NASA EOSDIS archives over petabytes of earth science data / year
- 望远镜的扫描数据
  - Sky survey data
- 高通量（High-throughput）生物学数据
- 科学仿真
  - terabytes of data generated in a few hours

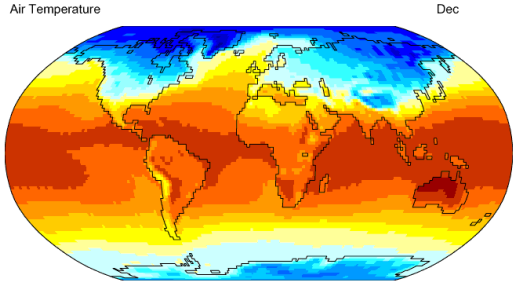


fMRI Data from Brain

Sky Survey Data



Gene Expression Data



Surface Temperature of Earth

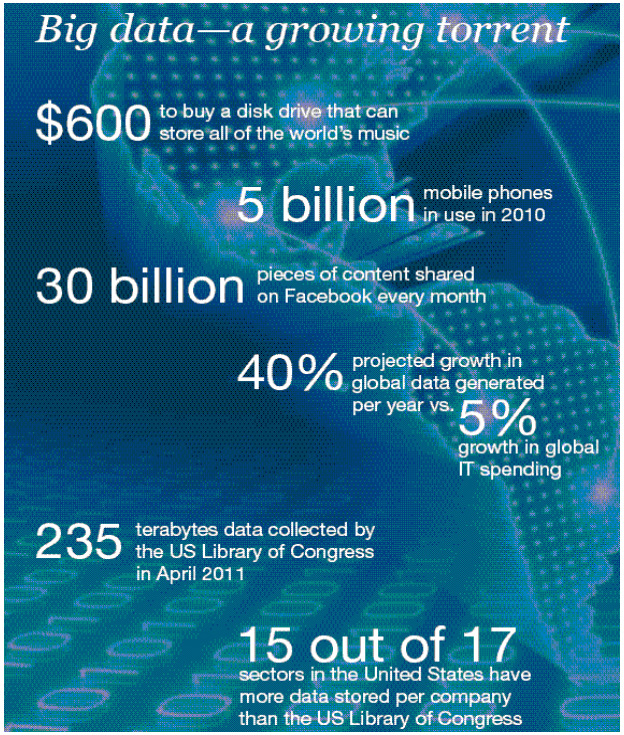
数据挖掘能够协助科学家：

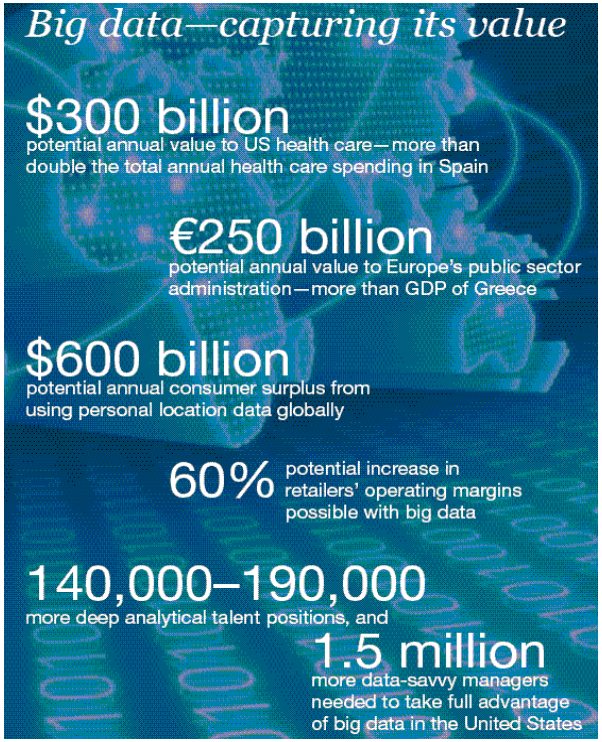
- 进行海量数据集的仿真
- 形成假说

改善各行各业生产力的绝佳机会

McKinsey Global Institute

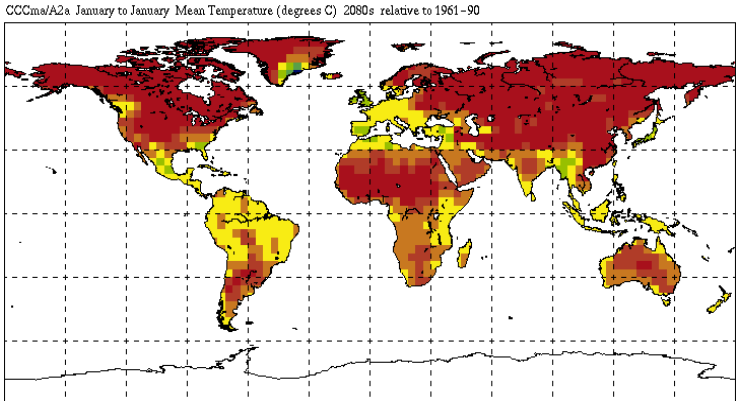
Big data: The next frontier for innovation, competition, and productivity





解决社会重大问题的良机

Improving health care and reducing costs



Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

动机与挑战

可伸缩（Scalability）

动机与挑战

可伸缩（Scalability）

高维性（High Dimensionality）

动机与挑战

可伸缩（Scalability）

高维性（High Dimensionality）

异构与复杂数据（Heterogeneous and Complex Data）

动机与挑战

可伸缩（Scalability）

高维性（High Dimensionality）

异构与复杂数据（Heterogeneous and Complex Data）

数据所有权与分布（Data Ownership and Distribution）

动机与挑战

可伸缩（Scalability）

高维性（High Dimensionality）

异构与复杂数据（Heterogeneous and Complex Data）

数据所有权与分布（Data Ownership and Distribution）

非传统分析（Non-traditional Analysis）

什么是数据挖掘

多种定义

- 从数据中轻松提取隐式（implicit）、先前未知（previously unknown）和潜在（potentially）有用的信息
- 通过自动或半自动方式对大量数据进行探索和分析，以发现有意义的模式

数据挖掘起源

来源于机器学习/人工智能，模式识别，统计分析，数据库系统等领域

传统技术可能不再适用于当前的数据：

- 大规模（Large-scale）
- 高维（High dimensional）
- 异构（Heterogeneous）
- 复杂（Complex）
- 分布式（Distributed）

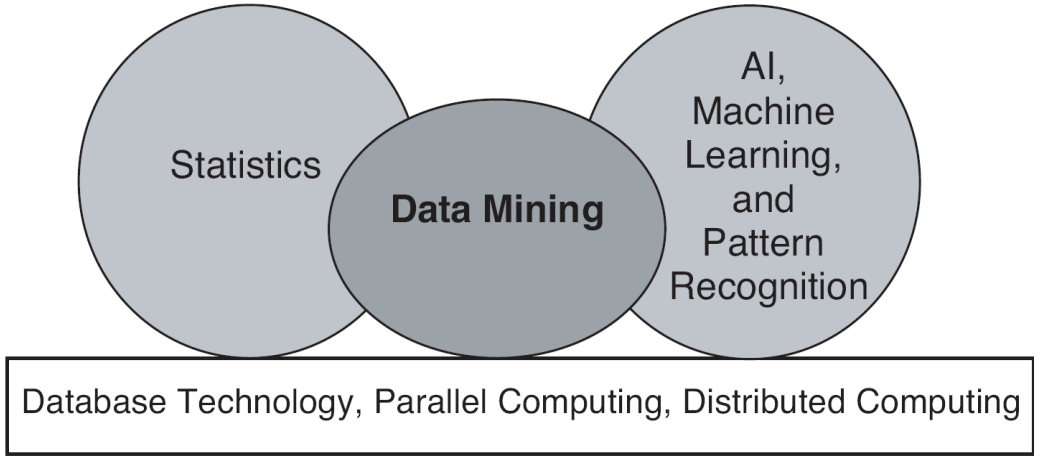
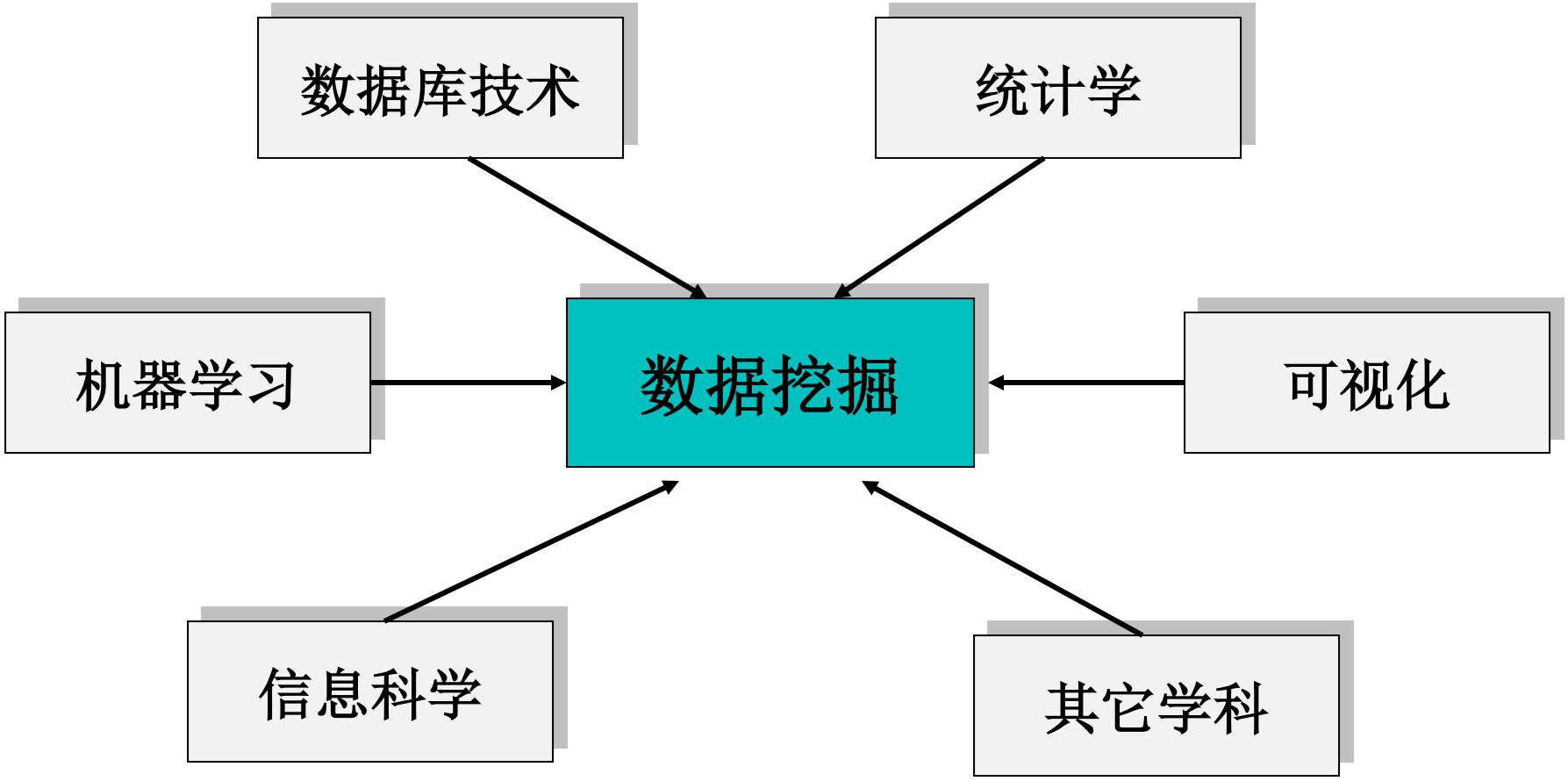


图 1-2　数据挖掘汇集了许多学科的知识

数据挖掘是当前新兴的“数据科学”和“数据驱动发现”领域的关键部分。

数据挖掘: 多学科交叉



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

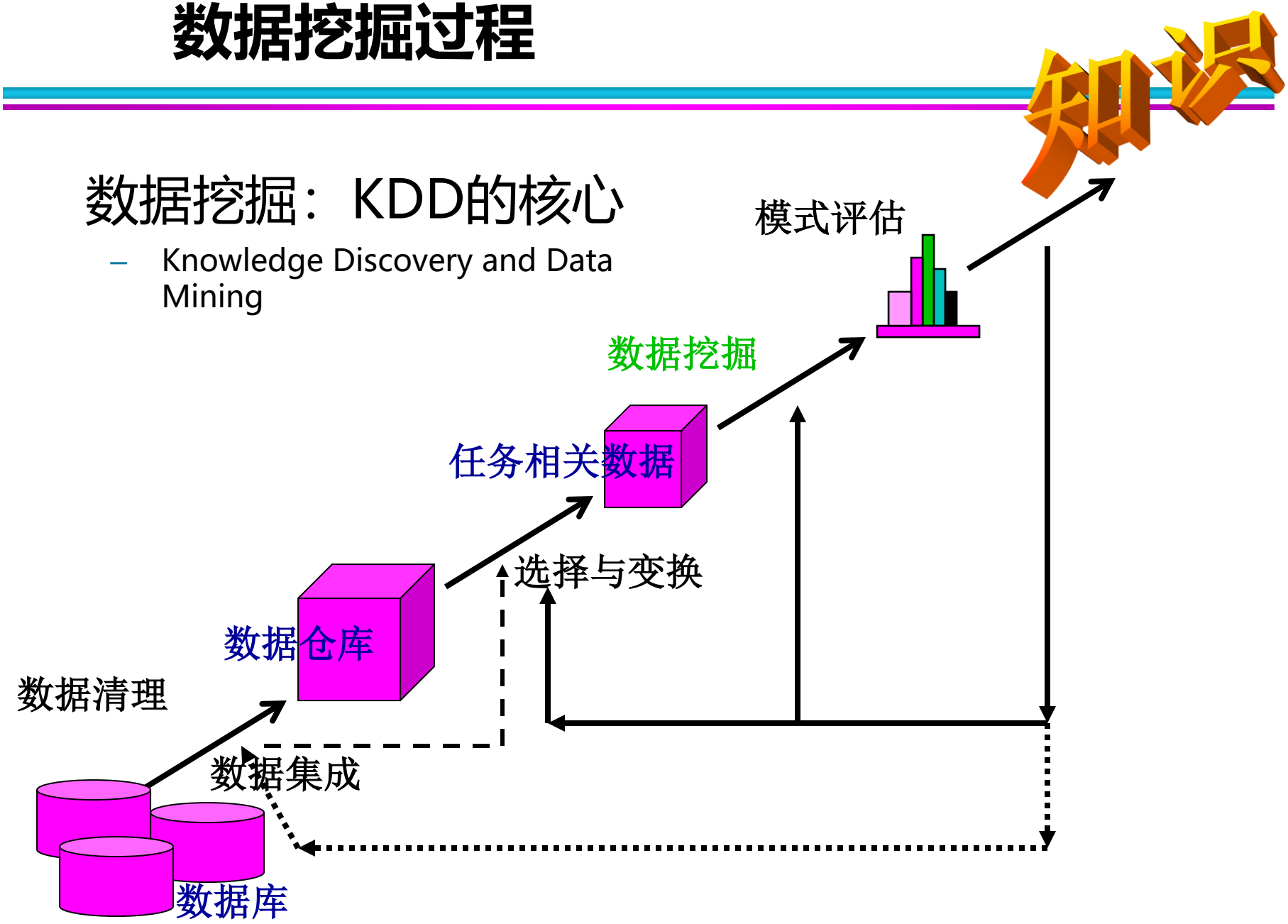
先进的数据库和信息存储

- 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

数据挖掘过程

数据挖掘：KDD的核心

- Knowledge Discovery and Data Mining



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

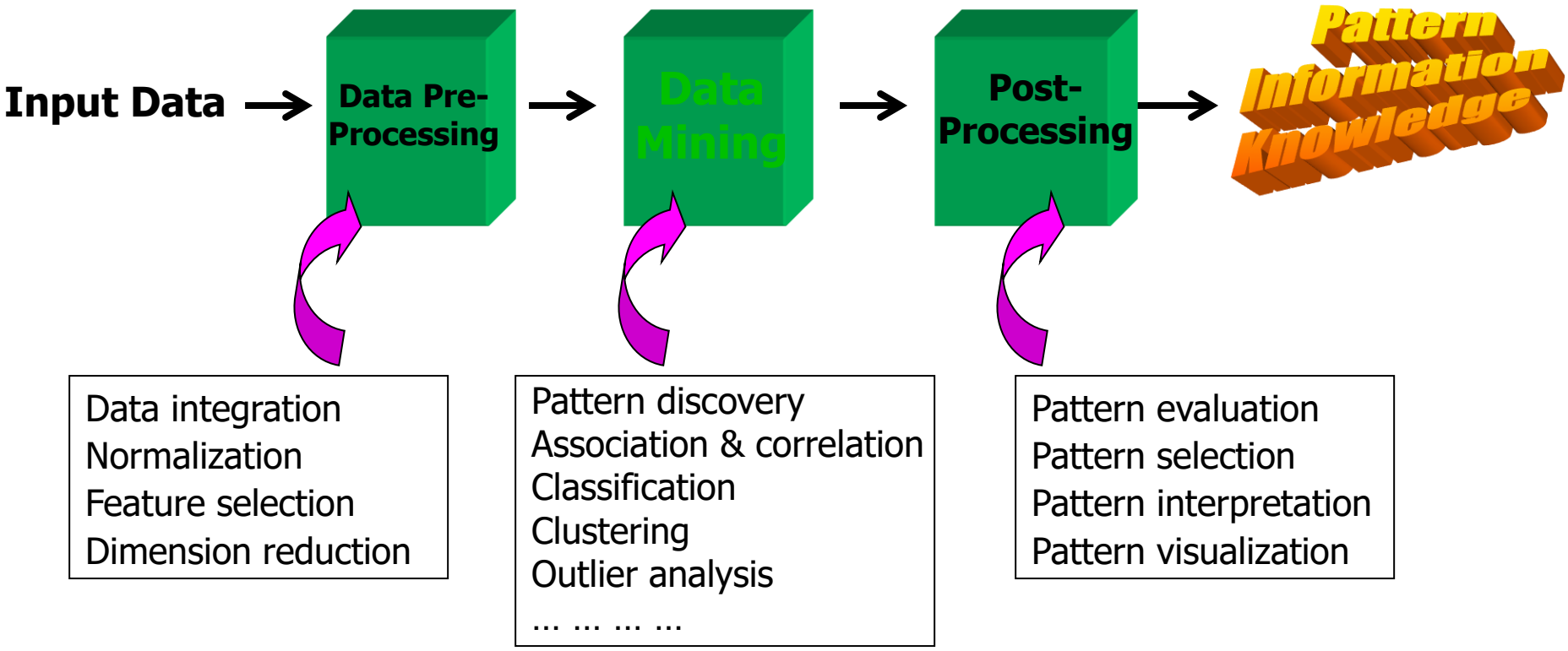
数据挖掘: 搜索有趣的模式

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

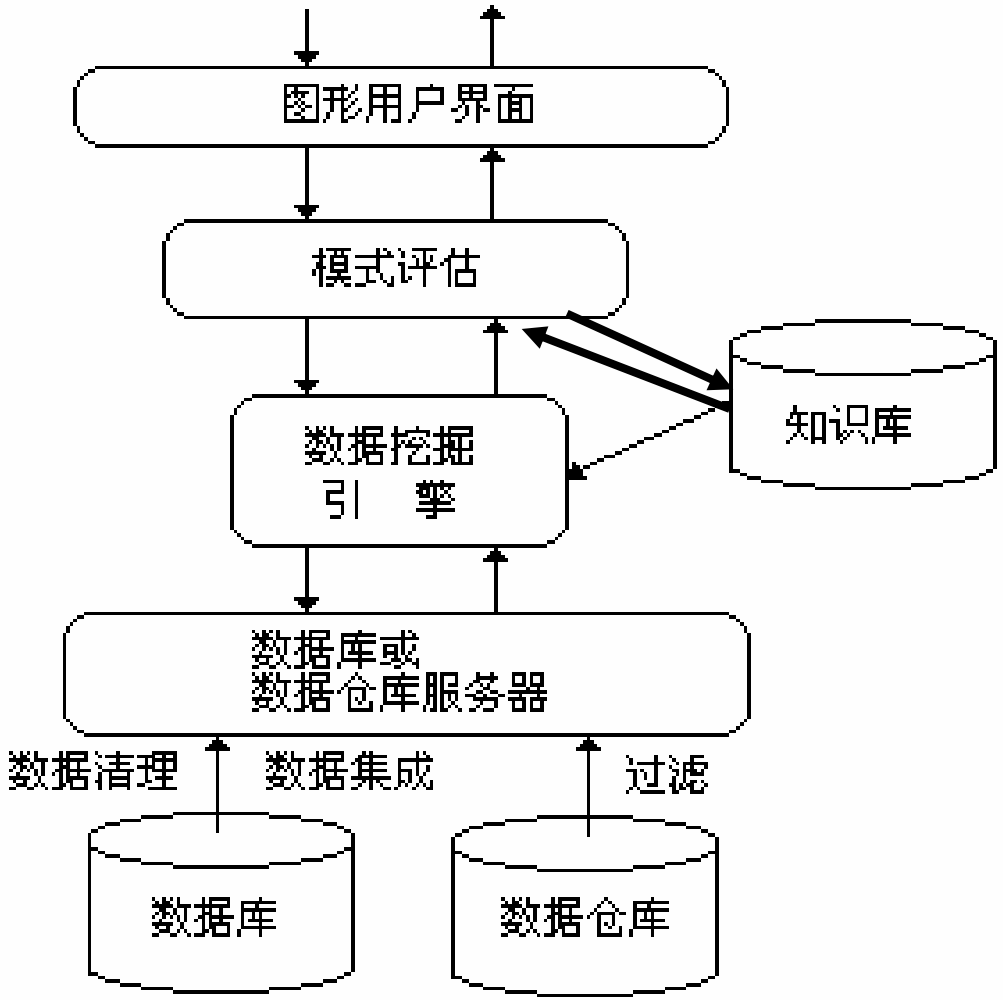
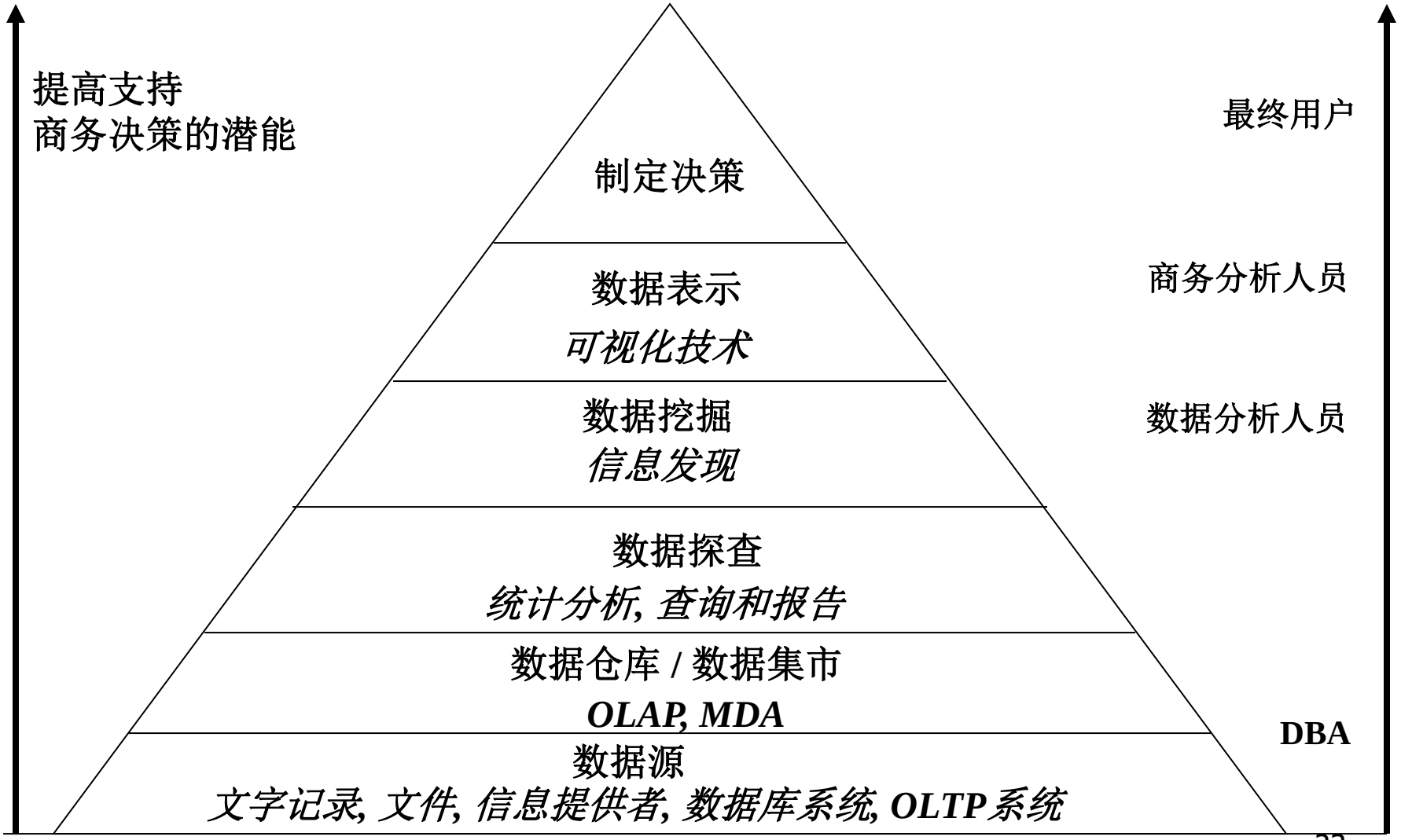


图 1.5：典型的数据挖掘系统结构

数据挖掘和商务智能


提高支持
商务决策的潜能
制定决策
最终用户
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

什么是（不是）数据挖掘

什么不是数据挖掘？

- 查找电话号码
- 在网络搜索引擎上查找关键词（信息检索）

什么是数据挖掘？

- 分析姓名和时间、地域的关系
- 将搜索引擎返回的结果进行分类/聚类
  - 苹果（水果）
  - 苹果（公司）

单选题 1分

设置

下述哪个任务不属于数据挖掘领域?

A 根据新闻的内容对其进行分类

B 根据身份证号从数据库中查找人员信息

C 根据用户的购买记录预测其是否对某件商品感兴趣

D 基于QQ的登录信息判断是否被盗号

提交

数据挖掘任务

预测任务（Prediction Tasks）

- 使用一些变量来预测其他变量的未知或将来的值。
- 自变量（independent variable）→因变量/目标变量（dependent variable）

描述任务（Description Tasks）

- 查找能够概括数据的可解释模式。
- 相关性、趋势、聚类、轨迹、异常等

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

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

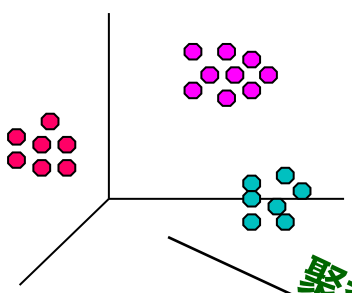
所用技术

- 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.

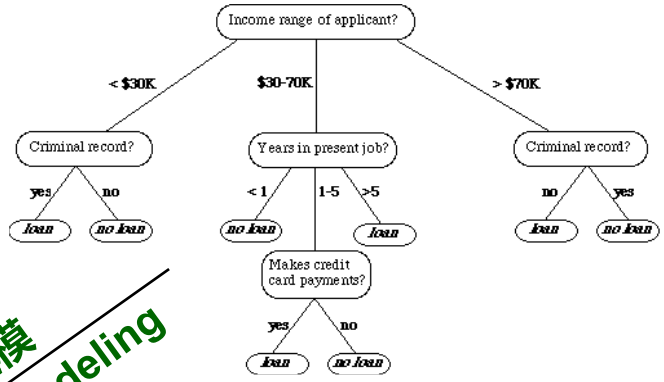
适合的应用

- 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

Data Mining Tasks ...



聚类 Clustering



预测建模 Predictive Modeling

Data

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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

关联分析 Association Rules

异常检测 Anomaly Detection

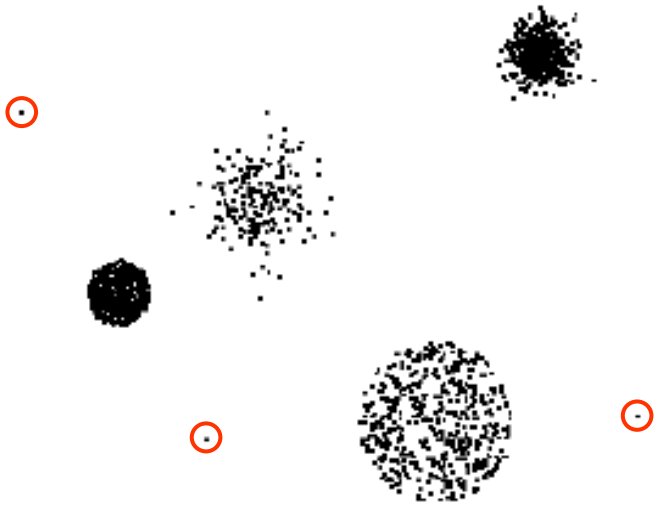
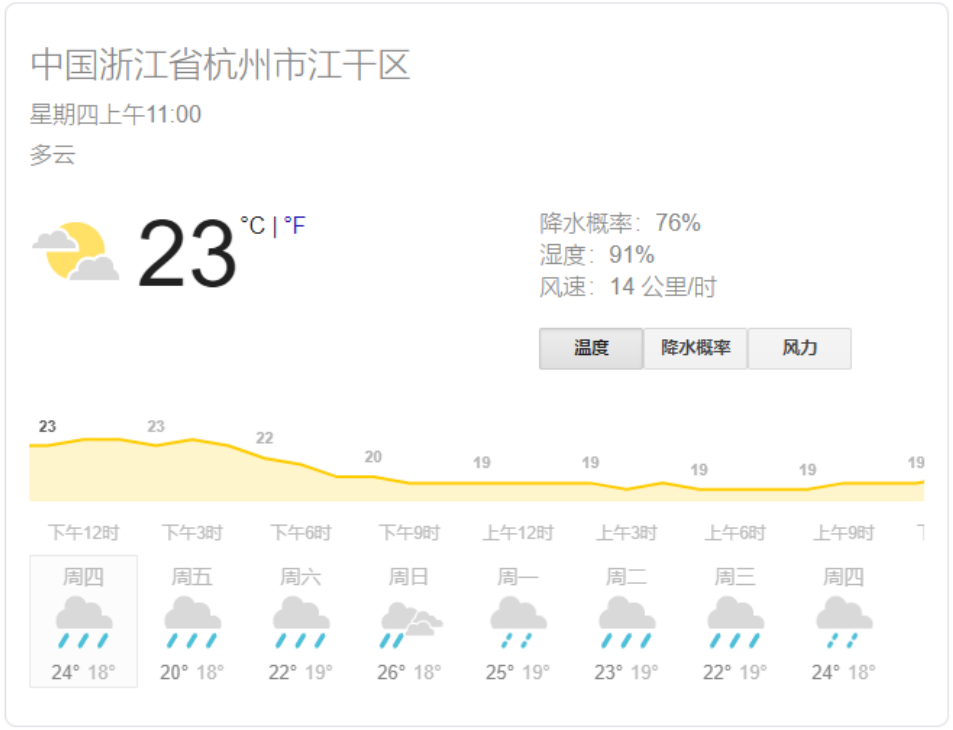


图 1-3　四种主要数据挖掘任务

预测建模：分类和回归

分类（classification）：预测离散的目标变量

回归（regression）：预测连续的目标变量



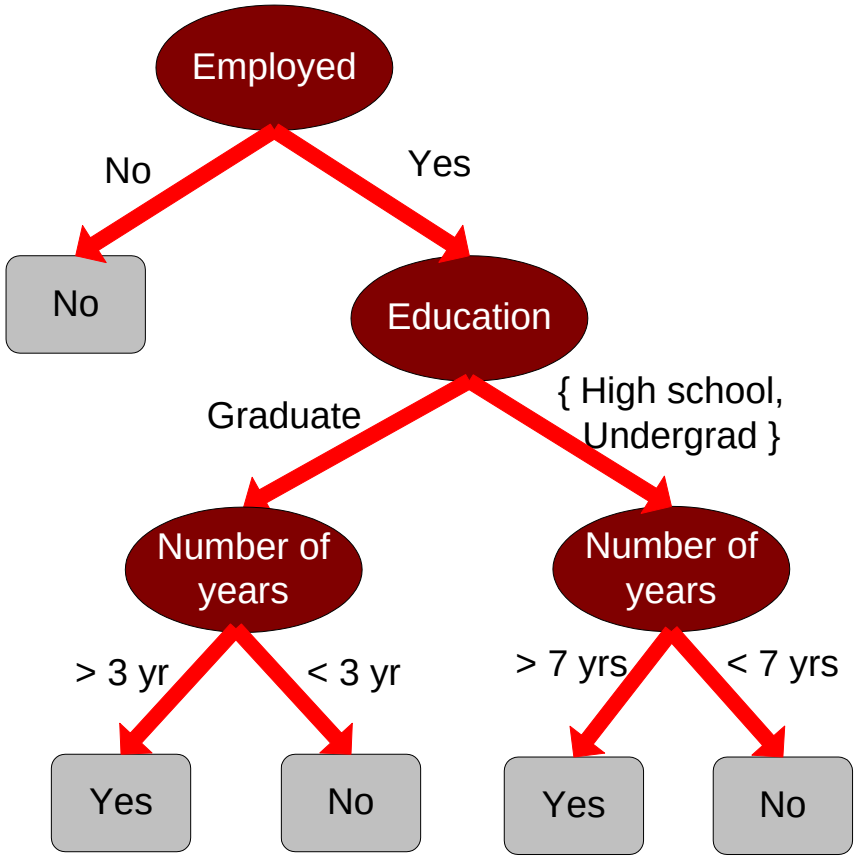
分类（classification）

找到模型：分类属性=函数（输入特征属性）

预测信用度的模型

分类属性
class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

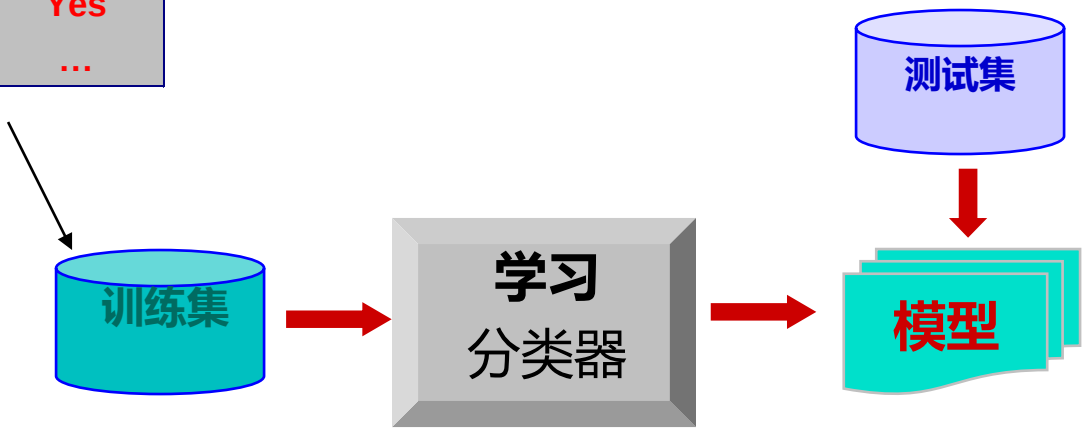


分类模型

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



分类任务：示例

将信用卡交易分类为合法或欺诈

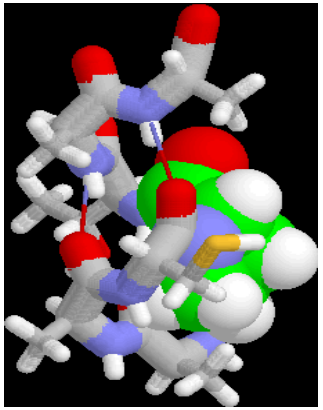
使用卫星数据对土地覆盖物（水体，市区，森林等）进行分类

将新闻报道分类为金融，天气，娱乐，体育等

识别网络空间中的入侵者

预测肿瘤细胞是良性还是恶性

将蛋白质的二级结构分为α-螺旋，β-折叠或无规卷曲



更多例子?

分类: 应用案例1

欺诈识别

- **目标**：预测信用卡交易中的欺诈案件。
- 方法：
  - 使用信用卡交易及其帐户持有人的信息作为属性。
    - 客户何时购买，购买什么，按时付款的频率等等
  - 将过去的交易标记为欺诈或公平交易。 这形成了class属性。
  - 学习能够交易类别的模型。
  - 使用此模型通过观察帐户上的信用卡交易来检测欺诈

分类：应用案例2

电话客户的客户流失预测

- **目标：**预测客户是否会流失给竞争对手
- **方法：**
  - 使用与过去和现在的每个客户进行交易的详细记录来查找属性。
    - 客户的联系人、联系频率、交流时长，他的财务状况，婚姻状况等。
  - 将客户标记为忠诚或不忠诚。
  - 获取流失模型。

From [Berry & Linoff] Data Mining Techniques, 1997

分类：应用案例3

天体测量与分类

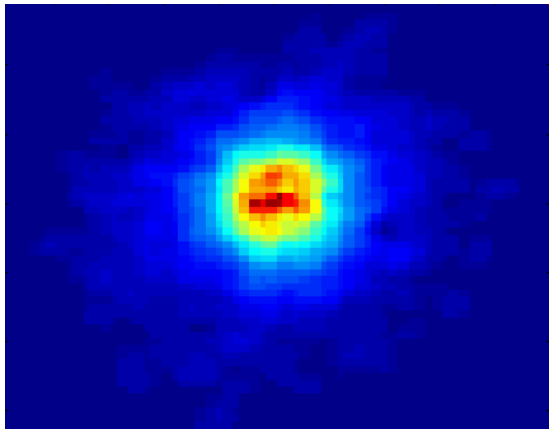
- **目标：**根据望远镜的勘测图像（来自帕洛玛天文台），预测天空物体的类别（星或星系），尤其是视觉上较暗的物体。
  - 3000 幅图像，每幅图像有 23,040 x 23,040 个像素点
- **方法:**
  - 图像分割
  - 测量图像属性(特征) – 每个对象40个属性。
  - 建模：建立特征和类别之间的关系。
  - 成功案例：可以找到16个新的高红移类星体（high red-shift quasars），其中一些难以用肉眼分辨

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

星系分类

Courtesy: http://aps.umn.edu

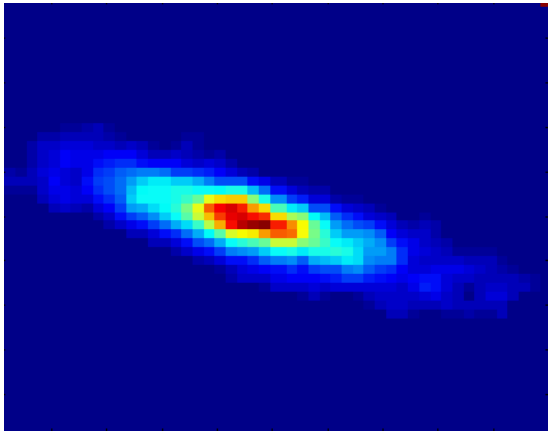
早期



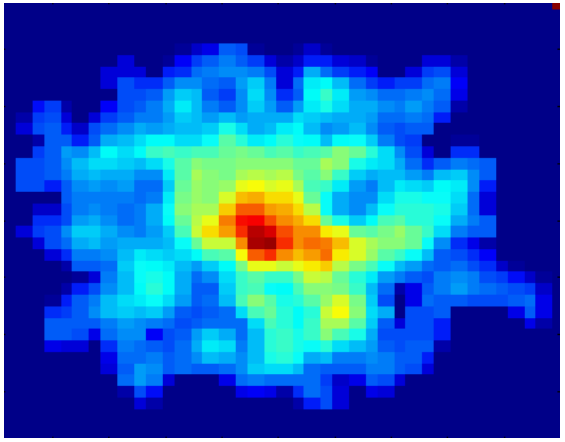
类别:
- 星系形成阶段

属性:
- 图像特征
- 光波特性等

中期



晚期



数据规模:
- 7200万颗星，2000万个星系
- 对象目录：9 GB
- 图像数据库：150 GB

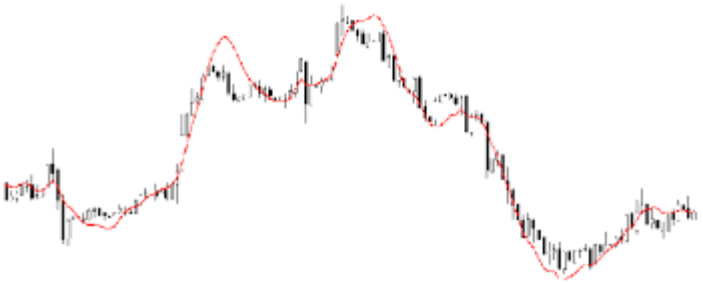
回归（Regression）

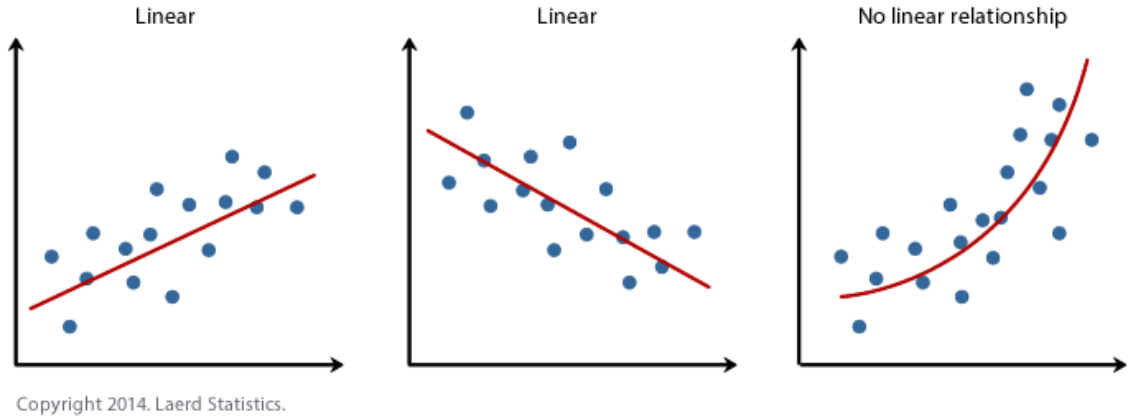
假设相关性为线性或非线性模型，则根据其他变量的值预测给定连续值变量的值。

在统计，神经网络领域进行了广泛的研究。

示例：

- 根据广告支出预测新产品的销售额。
- 根据温度，湿度，气压等预测风速
- 股市指数的时间序列预测。





单选题 1分

设置

根据历史数据预测明天的具体温度，并判断是否会下雨，分别属于（）、（）任务。

A 分类，分类

B 回归，回归

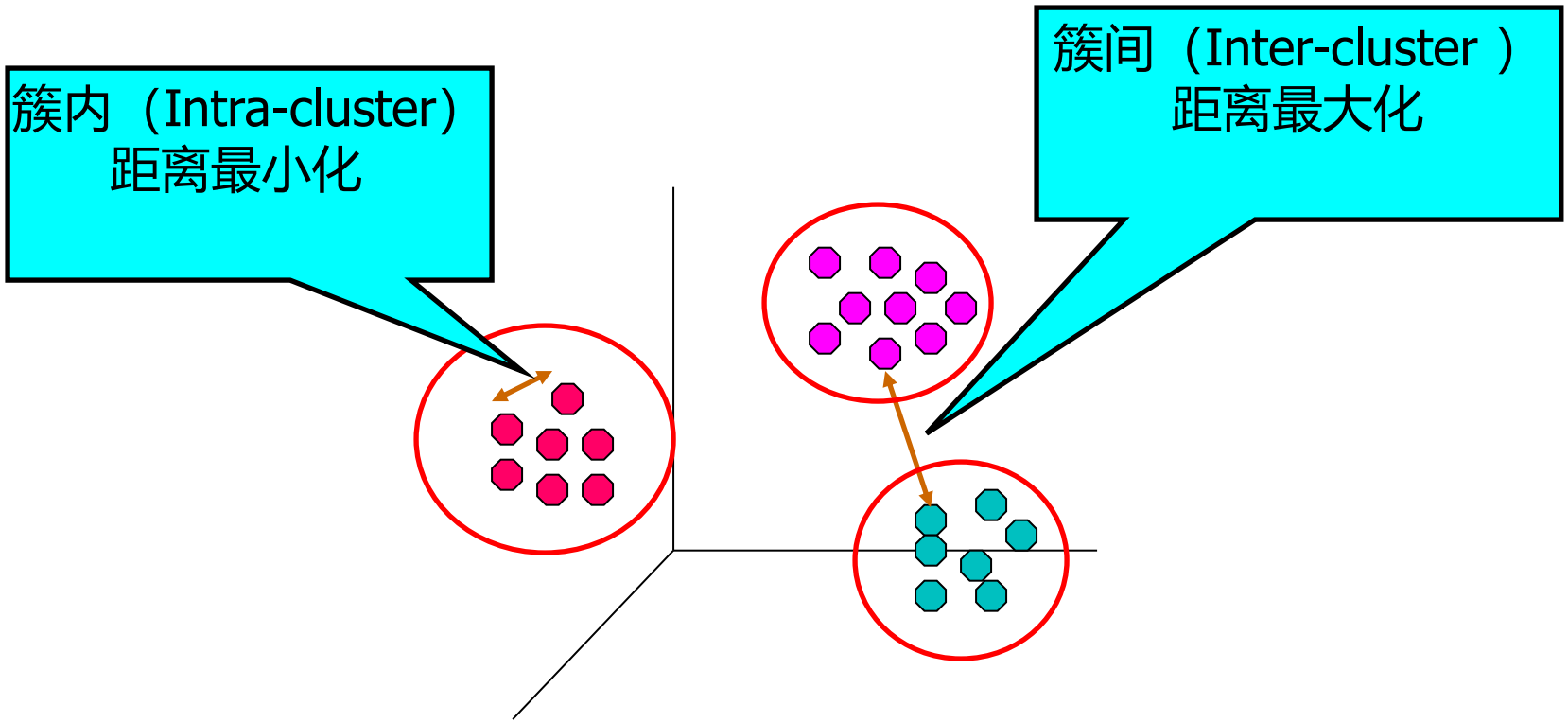
C 分类，回归

D 回归，分类

提交

聚类

查找对象组/簇（group/cluster），以使一组中的对象彼此相似（或相关），而与其他组中的对象不同（或不相关）



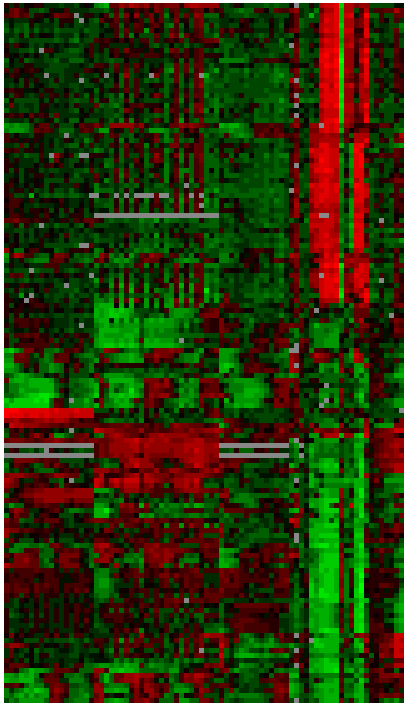
聚类分析应用

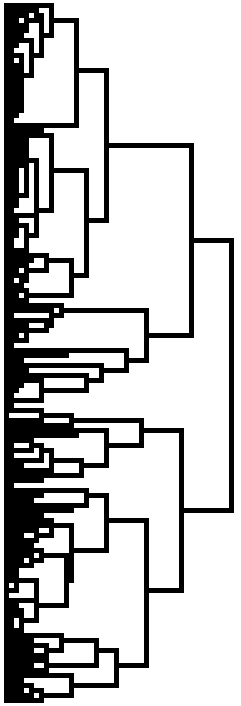
分析/理解（Understanding）

- 自定义配置文件以进行有针对性的营销
- 将相关文档分组以便浏览
- 将具有相似功能的基因和蛋白质分组
- 价格波动相似的集团股票

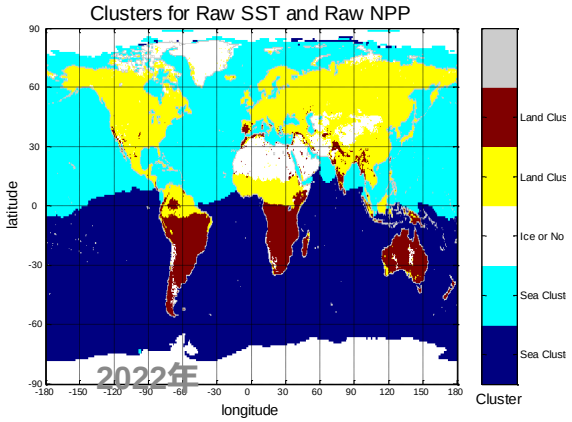
汇总/总结（Summarization）

- 减少大型数据集的大小

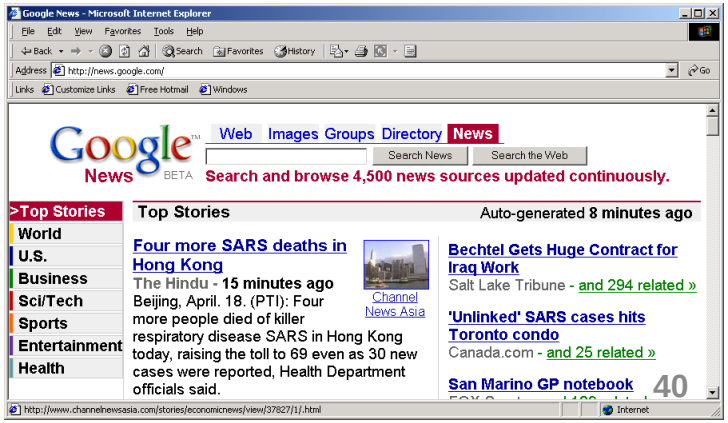




Courtesy: Michael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



聚类：应用案例1

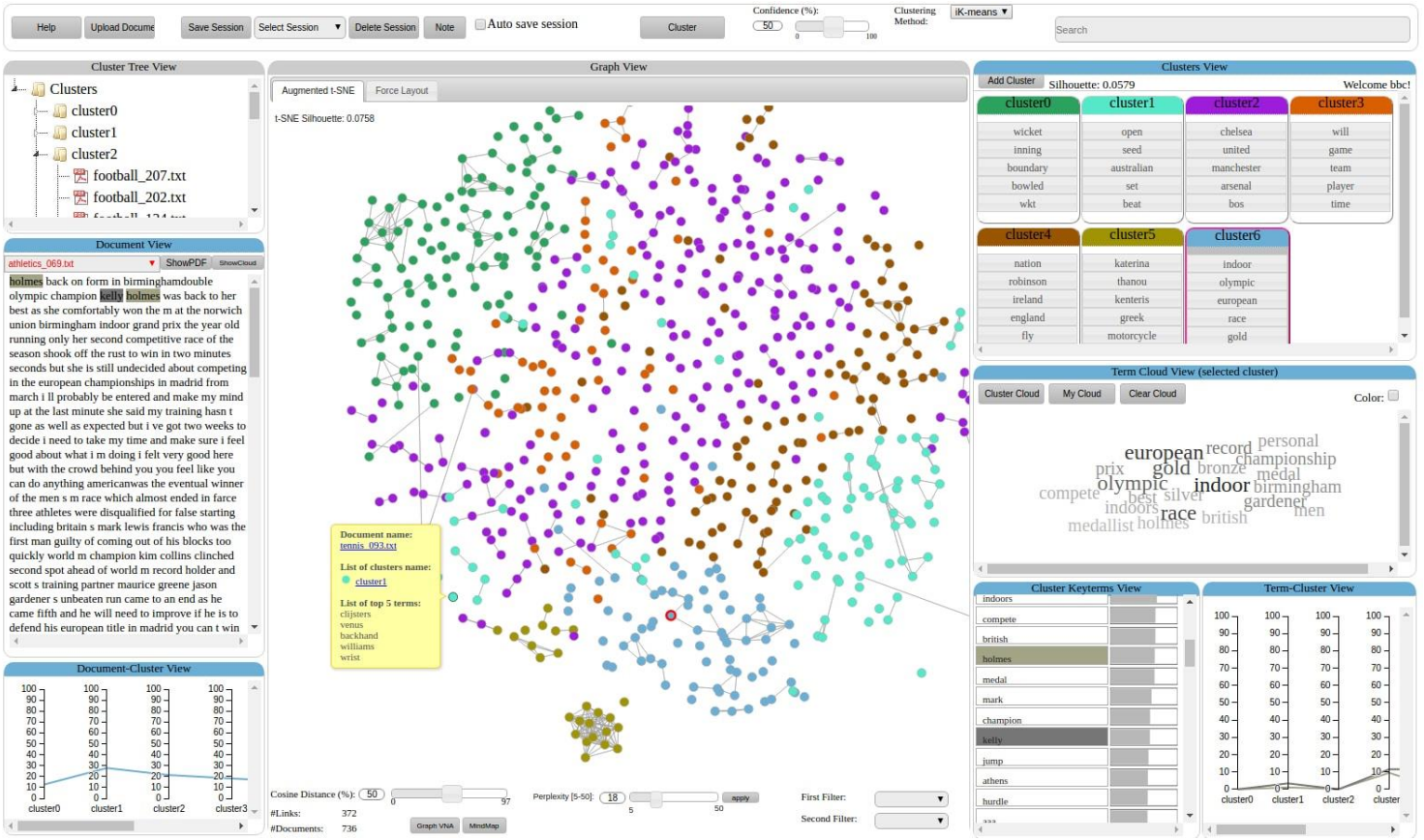
市场细分（Market Segmentation）：

- **目标：**将市场细分为不同的客户子集，可以选择合适的子集作为目标客户进行营销。
- **方法：**
  - 根据客户的地理和生活方式相关信息收集客户的不同属性。
  - 查找相似客户的集群。
  - 通过观察同一集群中的客户与不同集群中的客户的购买模式来衡量集群质量。

聚类：应用案例2

文档聚类：

- **目标：**根据文档中出现的重要术语查找彼此相似的文档组。
- **方法：**识别每个文档中经常出现的术语。根据不同术语的频率形成相似性度量。用该相似度来进行聚类。



关联规则挖掘：定义

给定一组记录，每个记录包含给定集合中的一些项目

- 生成依赖项规则：该规则将根据某些项目的出现与否来预测目标项目的出现概率。

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

关联分析：应用案例

购物篮分析

- 规则用于促销，货架管理和库存管理

电信报警诊断

- 该规则用于查找在同一时间段内经常一起发生的警报的组合

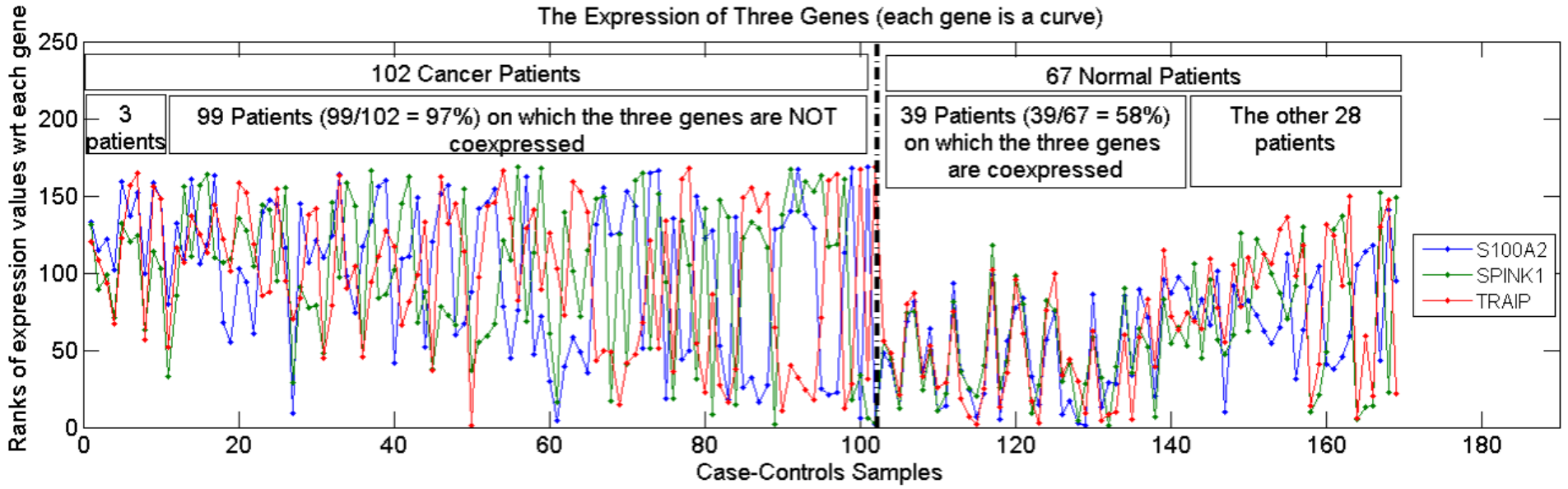
医学信息学

- 使用规则查找与某些疾病相关的患者症状和测试结果的组合

关联分析：应用案例

肺癌患者的特定基因分析

Three lung cancer datasets [Bhattacharjee et a
2001], [Stearman et al. 2005], [Su et al. 2007]



[Fang et al PSB 2010]

异常检测（anomaly detection）

识别其特征显著不同于其他数据的观测值

- 这样的观测值称为异常点（ anomaly）或离群点（ outlier）

高检测率和低误报率

- 异常检测算法的目标是发现真正的异常点，而避免错误地将正常的对象标注为异常点

信用卡欺诈检测

- 使用规则查找与某些疾病相关的患者症状和测试结果的组合

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

挖掘真实数据中规律

实际中,非常困难

(1)必须与现有的(数据)领域知识互恰

(2)对现有的(数据)领域知识体系有进展/贡献/价值

---迷雾中前行

谢谢！

数据挖掘
教师：王东京
学院：计算机学院
邮箱：dongjing.wang@hdu.edu.cn