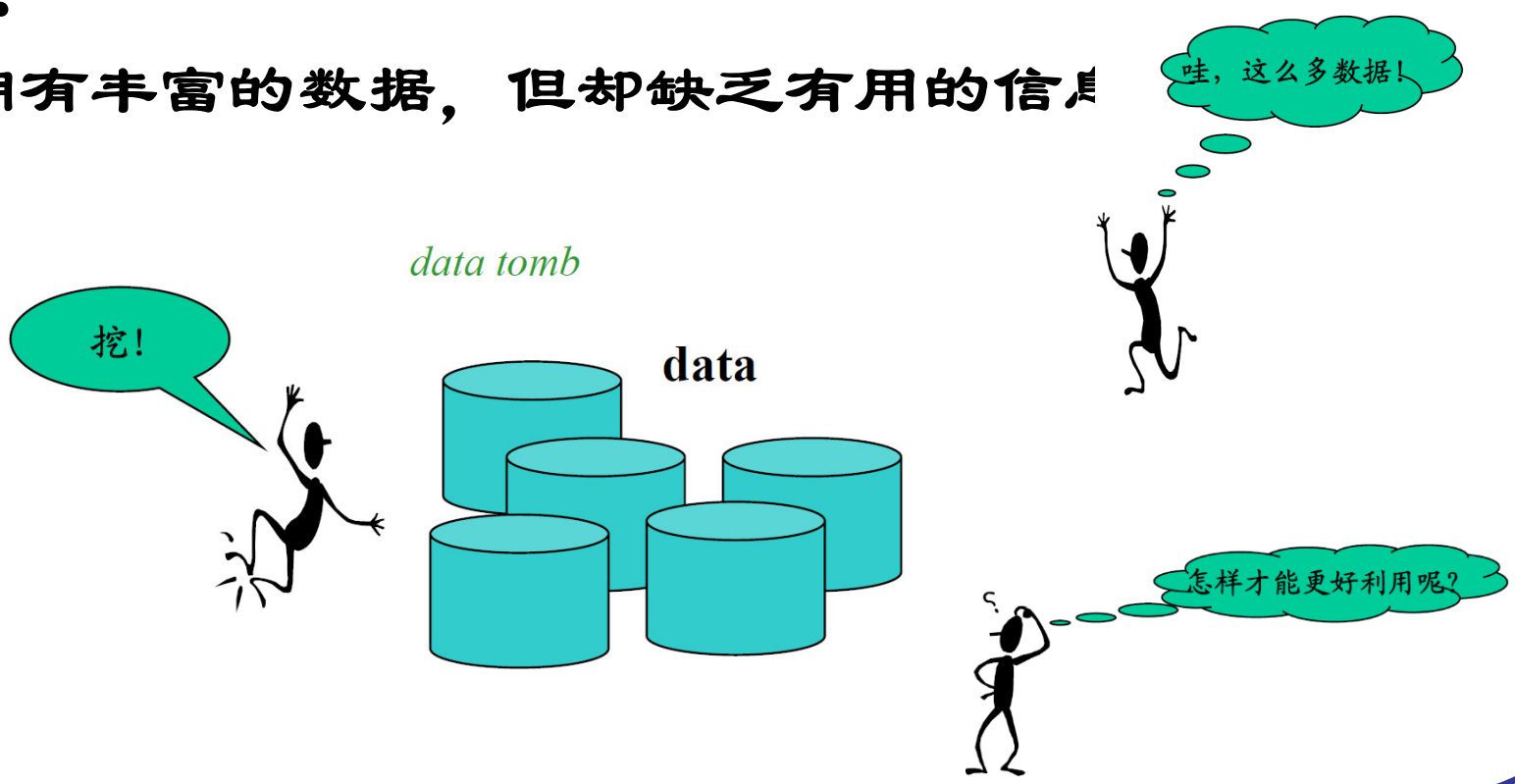
第十章 数据挖掘

- 数据挖掘的定义和发展历史
- 数据仓库和数据挖掘的OLAP技术
- 数据预处理

数据挖掘发展动力

- IT行业的快速发展产生的数据爆炸问题
  - 高性能计算机
  - 大容量存储媒介
  - 强大的数据采集设备
  - 简单的数据生成设备
  - ……
- 我们拥有丰富的数据，但却缺乏有用的信息



数据挖掘的发展历史

- 1960s及以前，文件系统
- 1970s，层次数据库和网状数据库
- 1980s早期，关系数据模型，关系数据库管理系统(RDBMS)的实现
- 1980s晚期，各种高级数据库系统(扩展的关系数据库,面向对象数据库等.）。面向应用的数据库系统（spatial数据库，时序数据库，多媒体数据库等等)
- 1990s，数据挖掘，数据仓库，多媒体数据库和网络数据库
- 2000s，流数据管理和挖掘；基于各种应用的数据挖掘；XML数据库和整合的信息系统

数据挖掘的定义

- **数据挖掘** （从数据中发现知识）

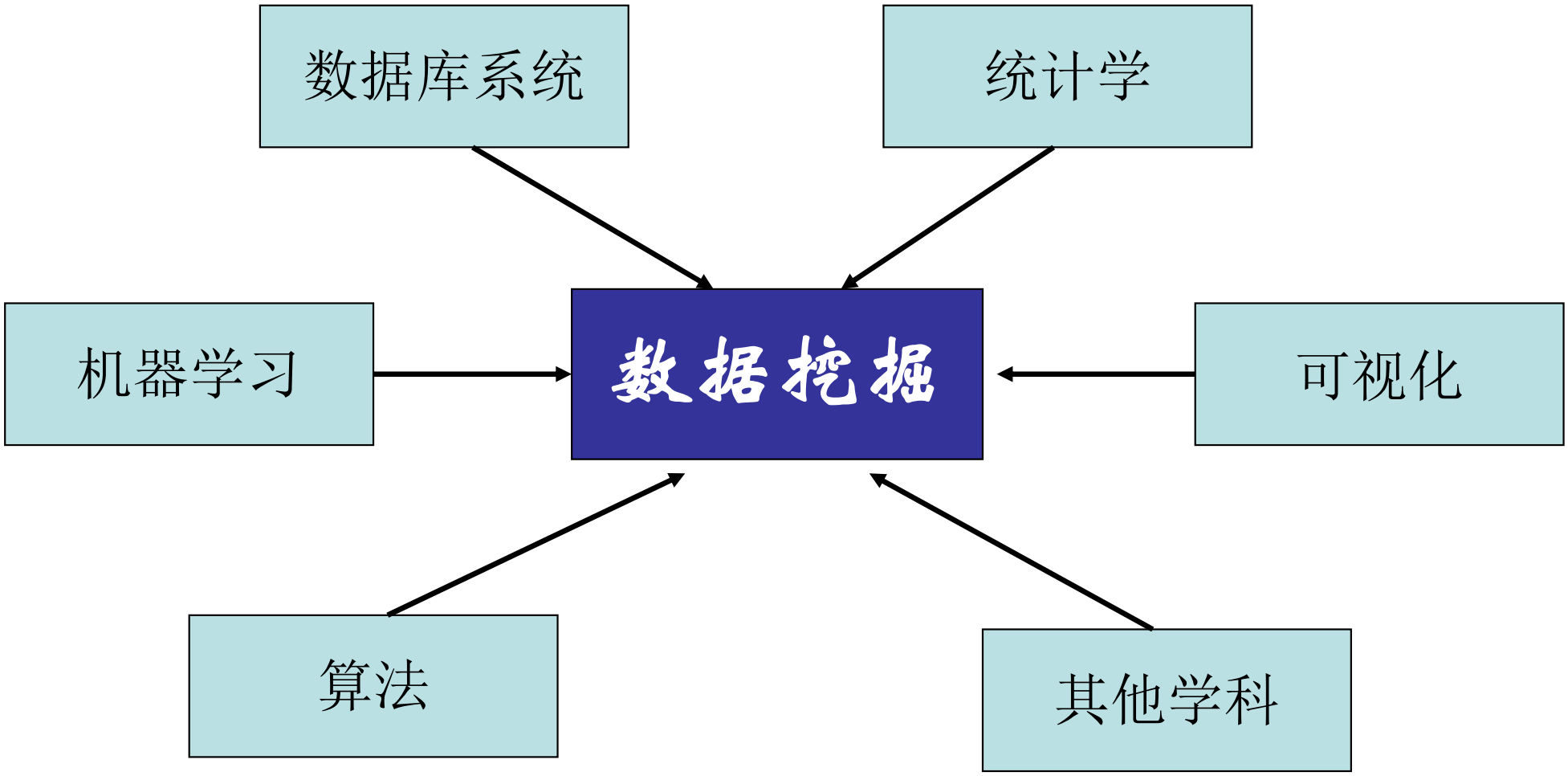
  从大量的数据中挖掘哪些令人感兴趣的、有用的、隐含的、先前未知的和可能有用的模式或知识。

  挖掘的不仅仅是数据（所以“数据挖掘”并非一个精确的用词）

- 数据挖掘的替换词

  数据库中的知识挖掘（KDD）、知识提炼、 数据/模式分析、数据考古、数据捕捞、信息收获等等。

数据挖掘-多种技术的融合



数据挖掘的应用

- 数据分析和决策支持

  市场分析和管理：

  目标市场，客户关系管理（CRM），市场占有量分析，交叉销售，市场分割

  风险分析和管理：

  风险预测，客户保持，保险业的改良，质量控制，竞争孤立点分析

  欺骗检测和异常模式的监测

- 其他的应用

  文本挖掘（新闻组，电子邮件，文档）和WEB挖掘

  流数据挖掘

  DNA 和生物数据分析

数据挖掘的应用

——市场分析和管理

- 数据从那里来?

  信用卡交易，会员卡，商家的优惠卷，消费者投诉电话，公众生活方式研究

- 目标市场

  构建一系列的“客户群模型”，这些顾客具有相同特征：兴趣爱好，收入水平，消费习惯等；确定顾客的购买模式。

- 交叉市场分析

  货物销售之间的相互联系和相关性，以及基于这种联系上的预测

- 顾客分析

  哪类顾客购买那种商品（聚类分析或分类预测）

- 客户需求分析

  确定适合不同顾客的最佳商品；预测何种因素能够吸引新顾客

- 提供概要信息

  多维度的综合报告；统计概要信息（数据的集中趋势和变化）

数据挖掘的应用

——公司分析和风险管理

- 财务计划

  现金流转分析和预测；

  交叉区域分析和时间序列分析（财务资金比率，趋势分析等等）

- 资源计划

  总结和比较资源和花费

- 竞争

  对竞争者和市场趋势的监控

  将顾客按等级分组和基于等级的定价过程

  将定价策略应用于竞争更激烈的市场中

数据挖掘的应用

——欺诈行为检测和异常模式的发现

方法：对欺骗行为进行聚类和建模，并进行孤立点分析

应用：卫生保健、零售业、信用卡服务、电信等

- 汽车保险：相撞事件的分析
- 洗钱：发现可疑的货币交易行为
- 电信：电话呼叫欺骗行为

电话呼叫模型：呼叫目的地，持续时间，日或周呼叫次数. 分析该模型发现与期待标准的偏差

- 零售产业

分析师估计有38%的零售额下降是由于雇员的不诚实行为造成的

- 反恐怖主义

数据挖掘的其他应用

- **体育竞赛**

  美国NBA的29个球队中，有25个球队使用了IBM 分析机构的数据挖掘工具，通过分析每个对手的数据（盖帽、助攻、犯规等数据）来获得比赛时的对抗优势。

- **天文学**

  JPL实验室和Palomar天文台就曾经在数据挖掘工具的帮助下发现了22颗新的恒星。

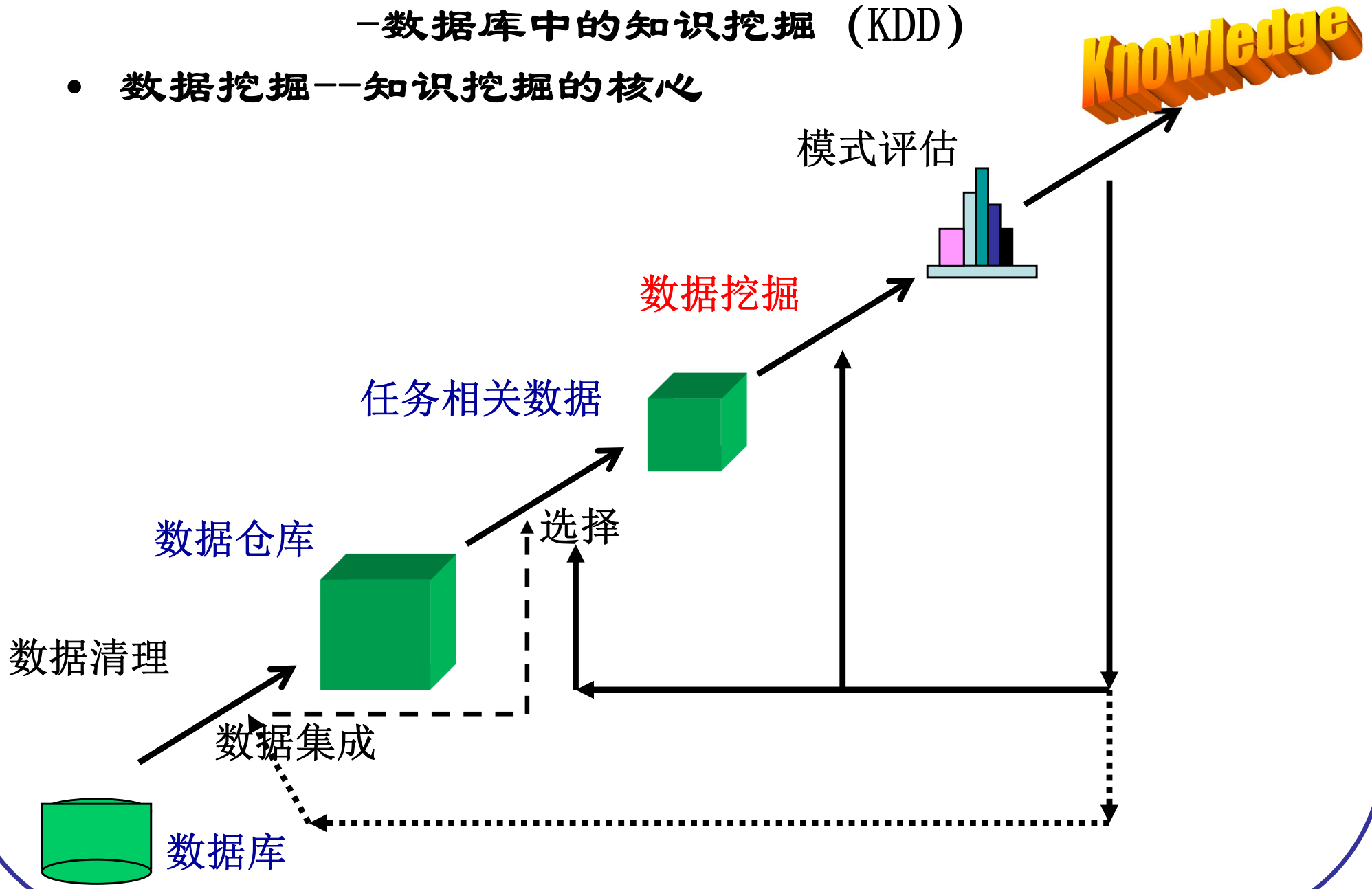
- **网上冲浪**

  通过将数据挖掘算法应用于网络访问日志，从与市场相关的网页中发现消费者的偏爱和行为，分析网络行销的有效性，改善网络站点组织。这就是新兴的WEB挖掘研究。

数据挖掘

—数据库中的知识挖掘（KDD）

- 数据挖掘--知识挖掘的核心



知识挖掘的步骤

- 了解应用领域

  了解相关的知识和应用的目标

- 创建目标数据集：选择数据
- 数据清理和预处理：（这个可能要占全过程60%的工作量）
- 数据缩减和变换

  找到有用的特征，维数缩减/变量缩减，不变量的表示。

- 选择数据挖掘的功能

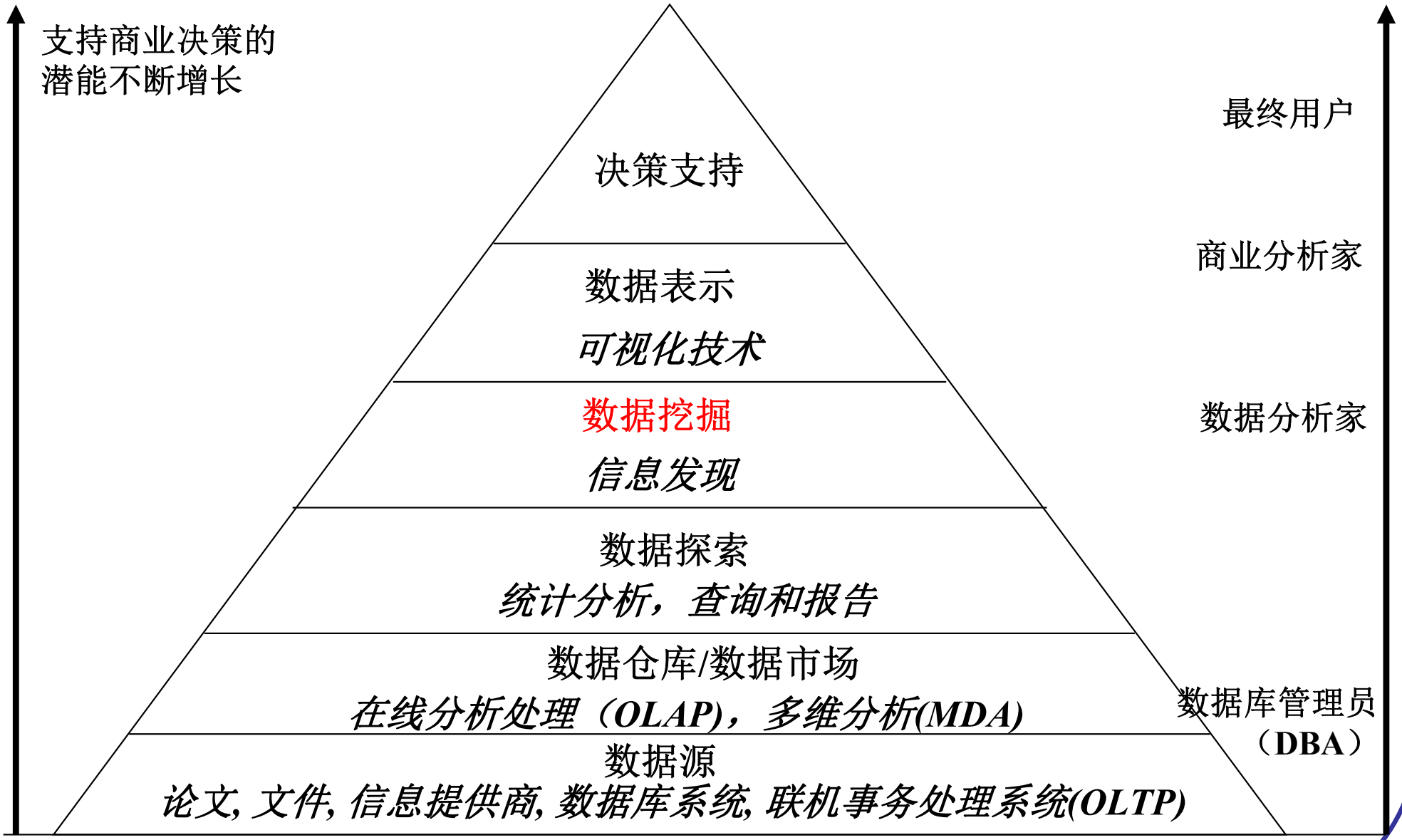
  数据总结，分类模型数据挖掘，回归分析，关联规则挖掘，聚类分析等.

- 选择挖掘算法
- 数据挖掘：寻找感兴趣的模式
- 模式评估和知识表示

  可视化，转换，消除冗余模式等等

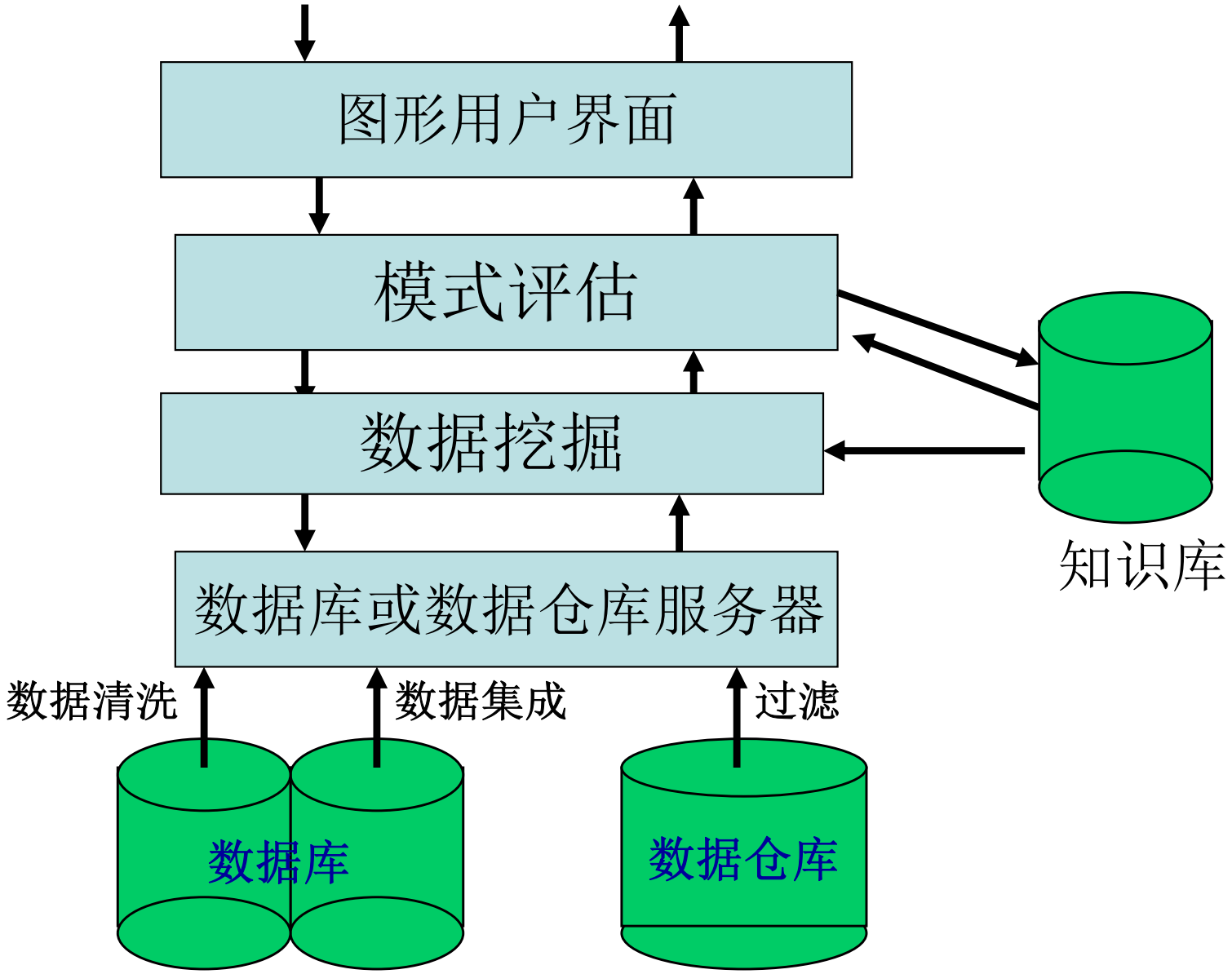
- 运用发现的知识

数据挖掘和商业智能


支持商业决策的
潜能不断增长
最终用户
决策支持
商业分析家
数据表示
可视化技术
数据分析家
数据挖掘
信息发现
数据探索
统计分析，查询和报告
数据仓库/数据市场
在线分析处理（OLAP），多维分析(MDA)
数据库管理员
（DBA）
数据源
论文, 文件, 信息提供商, 数据库系统, 联机事务处理系统(OLTP)

体系结构

—典型的数据挖掘系统



何种数据可以进行数据挖掘

- 关系数据库
- 数据仓库
- 事务数据库
- 高级数据库系统和信息库
  - 空间数据库
  - 时间数据库和时间序列数据库
  - 流数据
  - 多媒体数据库
  - 面向对象数据库和对象-关系数据库
  - 异种数据库和遗产(legacy)数据库
  - 文本数据库和万维网(WWW)

高级数据库系统和信息库

•空间数据库是指在关系型数据库（DBMS）内部对地理信息进行物理存储。空间数据库中存储的海量数据包括对象的空间拓扑特征、非空间属性特征以及对象在时间上的状态变化。

常见的空间数据库数据类型

地理信息系统(GIS)；遥感图像数据；医学图像数据。

数据挖掘技术的应用：通过空间分类和空间趋势分析，引入机器学习算法，对有用模式进行智能检索

•时间数据库和时间序列数据库都存放与时间有关的数据。时间数据库通常存放包含时间相关属性的数据。时间序列数据库存放随时间变化的值序列。对时间数据库和时间序列数据库的数据挖掘，可以通过研究事物发生发展的过程，有助于揭示事物发展的本质规律，可以发现数据对象的演变特征或对象变化趋势。

高级数据库系统和信息库

- **流数据**与传统的数据库技术中的静态数据不同，**流数据是连续的、有序的、变化的、快速的、大量的数据输入的数据**。

主要应用场合：网络监控、网页点击流、股票市场、流媒体…等等。与传统数据库技术相比，流数据在存储、查询、访问、实时性的要求等方面都有很大区别。

- **多媒体数据库**实现用计算机管理庞大复杂的多媒体数据，主要包括包括图形(graphics)、图像(image)、声音(audio)、视频(video)等等，现代数据库技术一般将这些多媒体数据以二进制大对象的形式进行存储。

对于多媒体数据库的数据挖掘，需要将存储和检索技术相结合。目前的主要方法包括构造多媒体数据立方体、多媒体数据库的多特征提取和基于相似性的模式匹配。

高级数据库系统和信息库

- **面向对象数据库和对象-关系数据库**的数据挖掘会涉及一些新的技术，比如处理复杂对象结构、复杂数据类型、类和子类层次结构、构造继承以及方法和过程等等。
- **异构数据库和历史(legacy)数据库**中，对于异构数据库系统，实现数据共享应当达到两点：一是实现数据库转换；二是实现数据的透明访问。WEB SERVICE技术的出现有利于历史数据库数据的重新利用。
- **文本数据库和万维网(WWW)**：文本数据库存储的是对对象的文字性描述。文本数据库的分类：结构类型（大部分的文本资料和网页）、半结构类型（XML数据）和结构类型（图书馆数据）；万维网(WWW)可以被看成最大的文本数据库。

数据挖掘的主要方法（1）

- 概念/类描述：特性化和区分

归纳，总结和对比数据的特性。比如：对每个月来网站购物超过5000元的顾客的描述：40－50岁，有正常职业，信用程度良好。

- 关联分析

发现数据之间的关联规则，这些规则展示属性－值频繁的在给定的数据中所一起出现的条件。

广泛的用于购物篮或事务数据分析。

数据挖掘的主要方法（2）

- 分类和预测

通过构造模型（或函数）用来描述和区别类或概念，用来预测类型标志未知的对象类。

比如：按气候将国家分类，按汽油消耗定额将汽车分类

导出模型的表示：判定树、分类规则、神经网络

可以用来预报某些未知的或丢失的数字值

- 聚类分析

将类似的数据归类到一起，形成一个新的类别进行分析。

最大化类内的相似性和最小化类间的相似性

数据挖掘的主要方法（3）

- 孤立点分析

孤立点：一些与数据的一般行为或模型不一致的孤立数据

通常孤立点被作为“噪音”或异常被丢弃，但在欺骗检测中却可以通过对罕见事件进行孤立点分析而得到结论。

- 趋势和演变分析

描述行为随时间变化的对象的发展规律或趋势

1）趋势和偏差：　回归分析

2）序列模式匹配：周期性分析

3）基于类似性的分析

- 其他定向模式或统计分析

数据挖掘系统的分类

- 一般功能

  描述性的数据挖掘

  预测性的数据挖掘

- 不同的视角，不同的分类

  根据所挖掘的数据库类型分类：关系数据库，事务数据库，流式数据，面向对象数据库，对象关系数据库，数据仓库，空间数据库，时态数据库，文本数据库，多媒体数据库，异构数据库，历史数据库，WWW。

  根据挖掘的知识类型分类：特征分析，区分，关联分析，分类聚类，孤立点分析/演变分析，偏差分析等；多种方法的集成和多层机挖掘。

  根据挖掘所用的技术分类：面向数据库的挖掘、数据仓库 、OLAP、机器学习、统计学、可视化等.

  根据数据挖掘的应用分类：金融，电信，银行，欺诈分析，DNA分析，股票市场，Web挖掘等.

OLAP 挖掘

-数据挖掘技术和数据仓库技术的集成

- 数据挖掘系统、数据库管理系统和数据仓库系统的耦合

  无耦合， 松耦合,半紧耦合， 紧耦合

- 联机分析和挖掘数据 (OLAM)

  挖掘和OLAP（联机分析处理）技术的集成

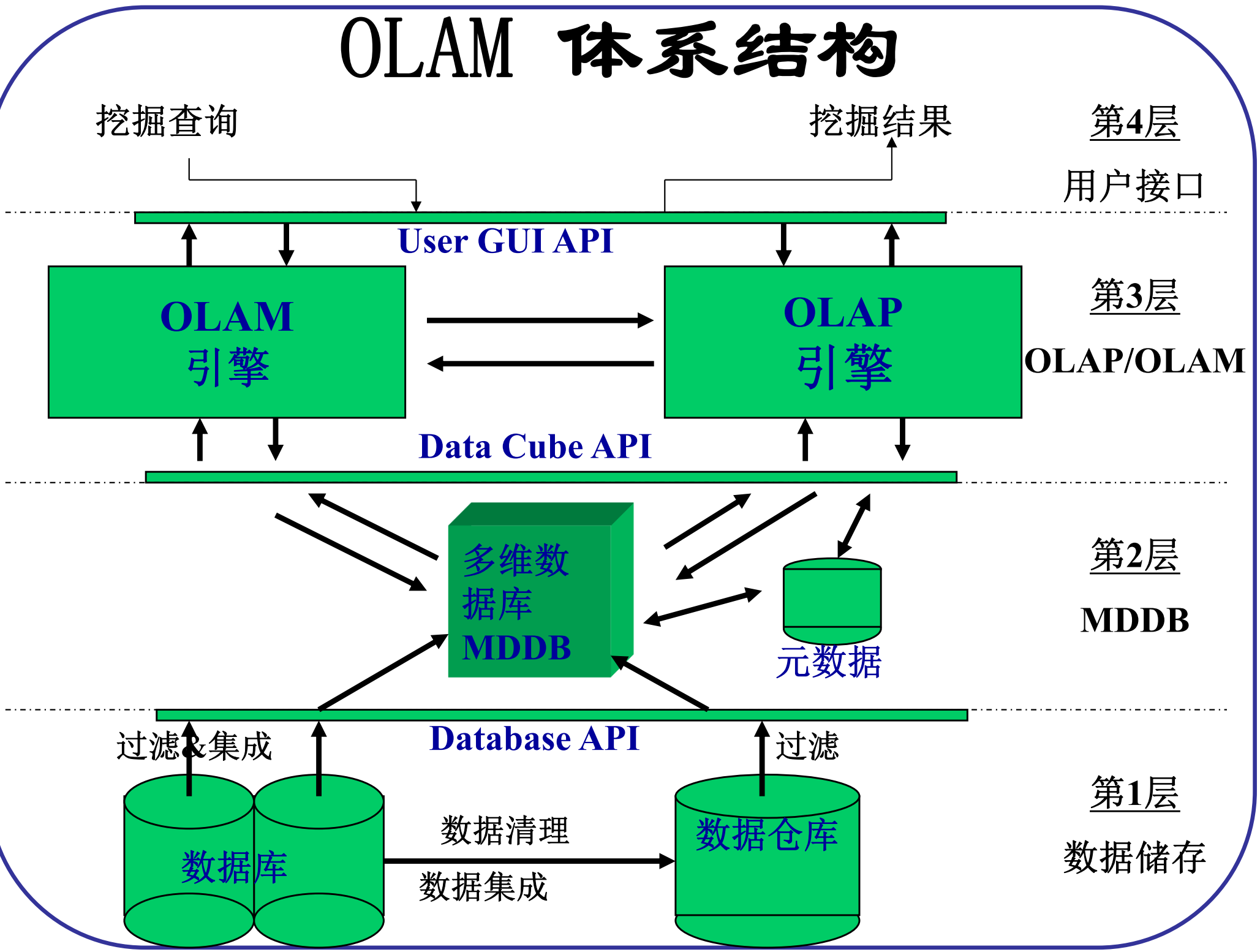
- 多个抽象层的交互知识挖掘

  在不同层次上通过交互地在数据空间和知识空间下钻、上卷和转轴来挖掘知识和模式的必要性.

- 多种挖掘功能的集成

  特性化分类， 先聚类分析后关联分析

OLAM 体系结构


挖掘查询
挖掘结果
第4层
用户接口
User GUI API
OLAM
引擎
OLAP
引擎
第3层
OLAP/OLAM
Data Cube API
多维数
据库
MDDB
元数据
第2层
MDDB
Database API
过滤&集成
过滤
数据库
数据清理
数据集成
数据仓库
第1层
数据储存

数据挖掘的主要问题（1）

- 挖掘方法

在不同的数据类型中挖掘不同类型的知识，e.g.，生物数据，流式数据，Web数据

性能：效率，有效性，和可伸缩性

模式评估：兴趣度问题

背景知识的合并

处理噪声和不完全数据

并行，分布式和增量挖掘算法

新发现知识与已有知识的集成：知识融合

数据挖掘的主要问题（2）

- 用户交互

  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

  多个抽象层的交互知识挖掘

- 应用和社会因素

  特定域的数据挖掘 & 不可视的数据挖掘

  数据安全，完整和保密的保护

数据仓库和数据挖掘的OLAP技术

- 数据仓库的定义

数据仓库是作为决策支持系统（dss）和联机分析应用数据源的结构化数据环境。数据仓库研究和解决从数据库中获取信息的问题。**数据仓库的特征在于面向主题、集成性、稳定性和时变性**。

- “数据仓库是一个面向主题的、集成的、随时间而变化的、不容易丢失的数据集合，支持管理部门的决策过程.” —比尔●恩门（数据仓库构造方面的领头设计师）

数据仓库的关键特征

- **面向主题**

围绕一些主题，如顾客、供应商、产品等；关注决策者的数据建模与分析，而不是集中于组织机构的日常操作和事务处理；排除对于决策无用的数据，提供特定主题的简明视图。

- **数据集成**

一个数据仓库是通过集成多个异种数据源来构造的。关系数据库，一般文件，联机事务处理记录；使用数据清理和数据集成技术。确保命名约定、编码结构、属性度量等的一致性，当数据被移到数据仓库时，它们要经过转化。

- **随时间变化**

数据仓库的时间范围比操作数据库系统要长的多。操作数据库系统：主要保存当前数据。数据仓库:从历史的角度提供信息（比如过去 5-10 年）

- **数据不易丢失**

尽管数据仓库中的数据来自于操作数据库，但他们却是在物理上分离保存的。

数据仓库的设计与开发

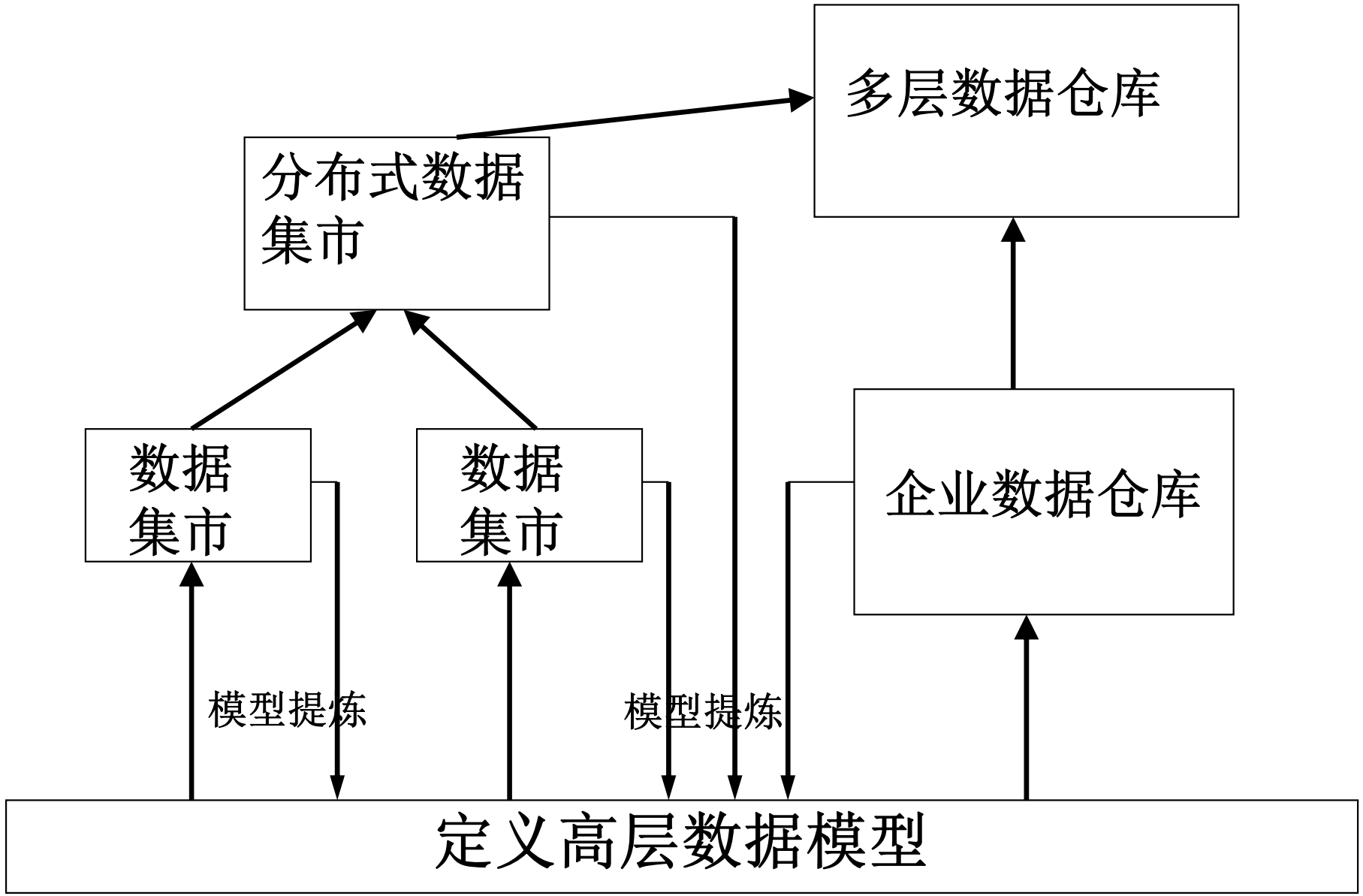
- **自顶向下开发**

一种系统的解决方法，并能最大限度地减少集成问题。但费用高，长时间开发，缺乏灵活性，因为整个组织的共同数据模型达到一致是困难的。

- **自底向上**

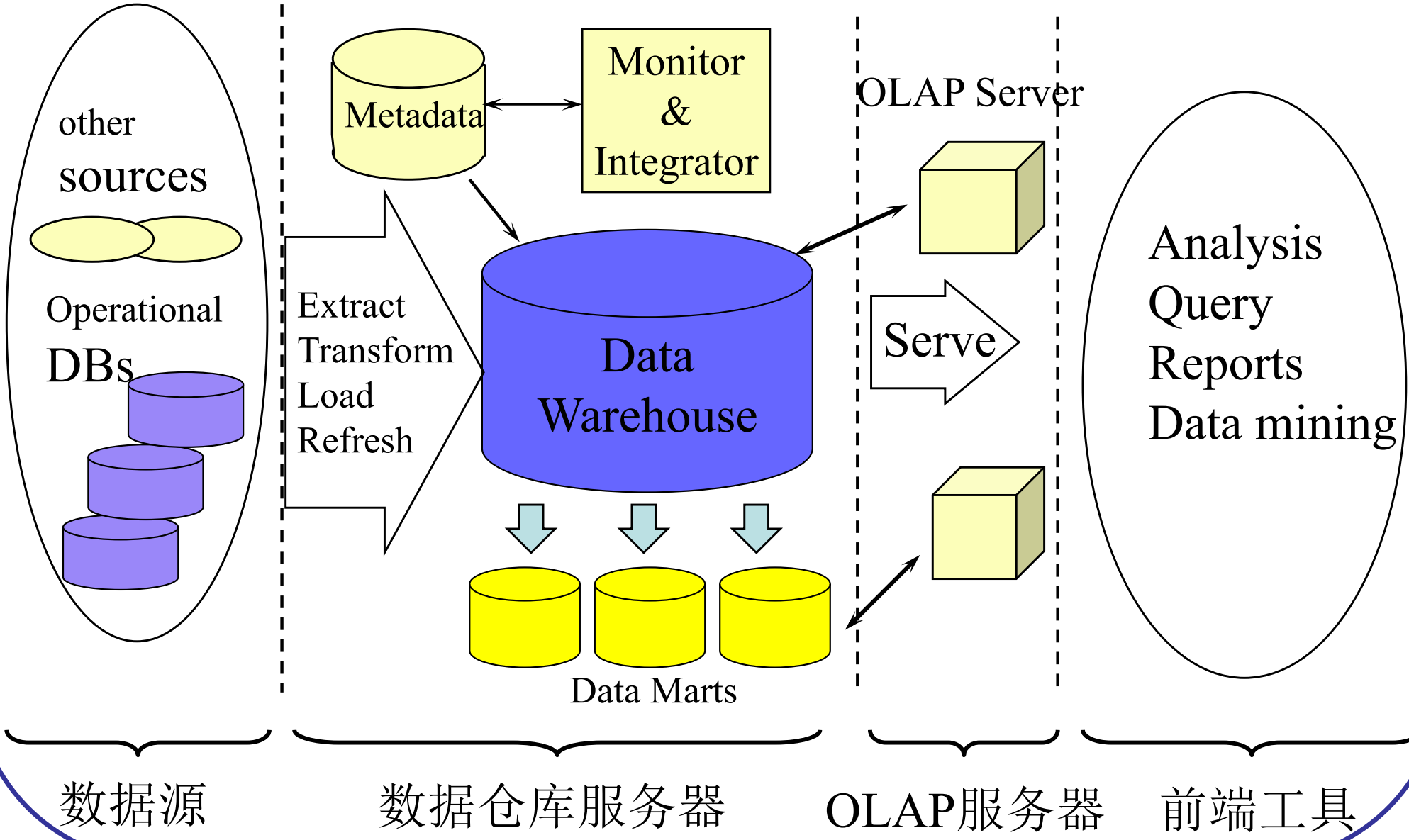
设计、开发、部署独立的数据集市方法提供了灵活性、低花费，并能快速回报投资。然后，将分散的数据集市集成，形成一个一致的企业数据仓库时，可能导致问题。

数据仓库的设计与开发



数据仓库的设计与开发

-三层数据仓库架构



数据仓库的实现

- 难点

  海量数据

  快速反应：OLAP服务器要在几秒内响应决策支持查询

- 方法

  高效的数据立方体计算技术

  高效的存取方法

  高效的查询处理技术

OLAP服务器类型

- 关系OLAP服务器(ROLAP)

  使用关系数据库或扩展的关系数据库存放并管理数据仓库的数据，而用OLAP中间件支持其余部分

  包括每个DBMS后端优化，聚集导航逻辑的实现，附加的工具和服务

  较大的可扩展性

- 多维OLAP服务器(MOLAP)

  基于数组的多维存储引擎（稀疏矩阵技术）

  能对预计算的汇总数据快速索引

- 混合OLAP服务器(HOLAP)

  结合上述两种技术，更大的使用灵活性

- 特殊的SQL服务器

  在星型和雪花模型上支持SQL查询

数据仓库后端工具和使用程序

- 用于加载和刷新它的数据

  数据提取：从多个外部的异构数据源收集数据

- 数据清理

  检测数据种的错误并作可能的订正

- 数据变换

  将数据由历史或主机的格式转化为数据仓库的格式

- 装载

  排序、汇总、合并、计算视图，检查完整性，并建立索引和分区

- 刷新

  将数据源的更新传播到数据仓库中

数据仓库的三种应用

- **信息处理**

支持查询和基本的统计分析，并使用交叉表、表、图标和图进行报表处理

- **分析处理**

对数据仓库中的数据进行多维数据分析；支持基本的OLAP操作，切块、切片、上卷、下钻、转轴等

- **数据挖掘**

从隐藏模式中发现知识；支持关联分析，构建分析性模型，分类和预测，并用可视化工具呈现挖掘的结果

从联机分析处理到联机分析挖掘

- 为什么要联机分析挖掘

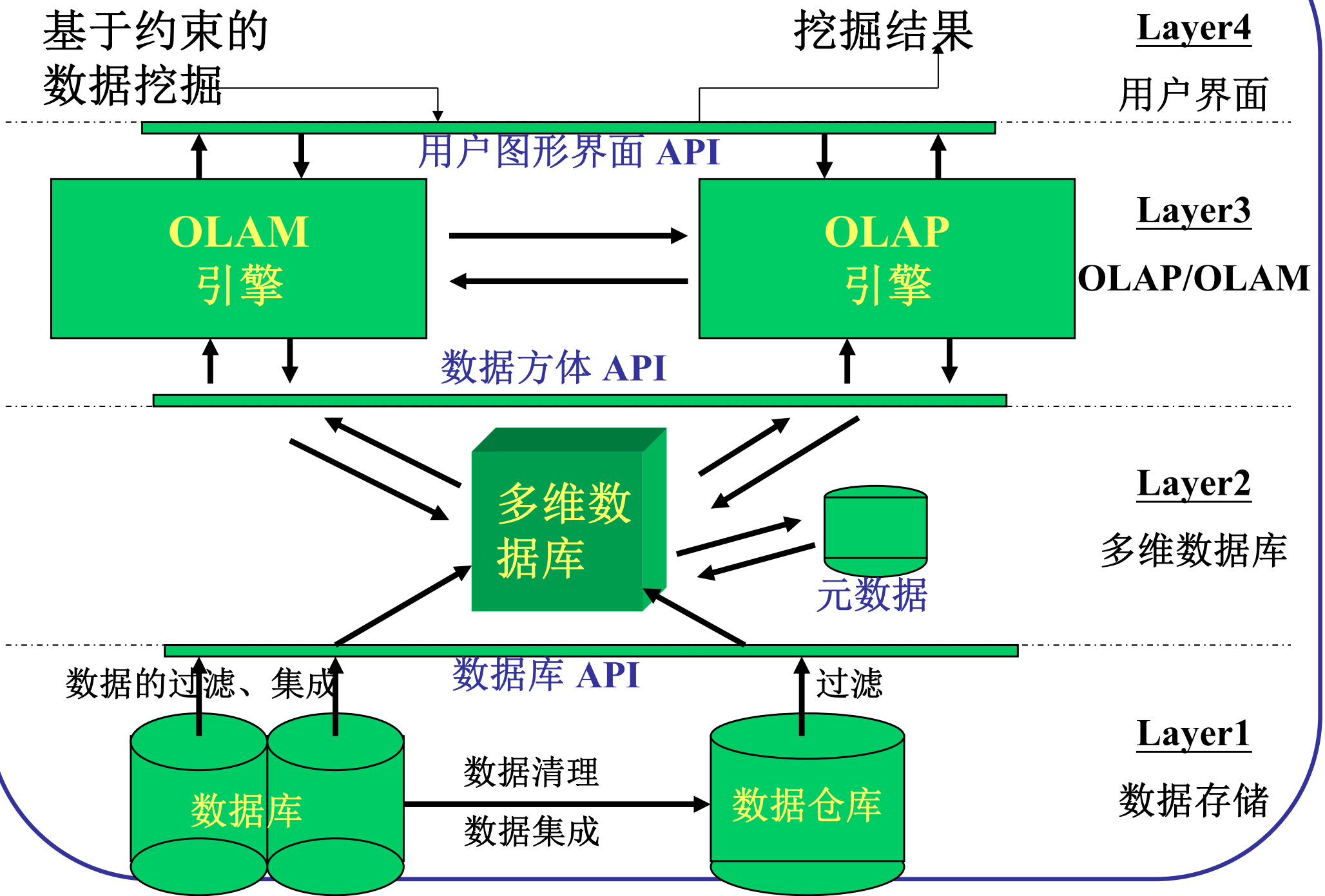
数据仓库中有高质量的数据：数据仓库中存放着整合的、一致的、清理过的数据

围绕数据仓库的信息处理结构：存取、集成、合并多个异种数据库的转换，ODBC/OLEDB连接,Web访问和访问工具等

基于OLAP的探测式数据分析：使用上卷、下钻、切片、转轴等技术进行数据挖掘

数据挖掘功能的联机选择：多种数据挖掘功能、算法和任务的整合

联机分析挖掘的体系结构



数据预处理

- 为什么要预处理数据

现实世界的数据是“肮脏的”：

1）不完整的：有些感兴趣的属性缺少属性值，或仅包含聚集数据

2）含噪声的：包含错误或者“孤立点”

3）不一致的：在编码或者命名上存在差异

没有高质量的数据，就没有高质量的挖掘结果

1）高质量的决策必须依赖高质量的数据

2）数据仓库需要对高质量的数据进行一致地集成

数据预处理

- **数据质量的多维度量**

一个广为认可的多维度量观点：

1）精确度

2）完整度

3）一致性

4）合乎时机

5）可信度

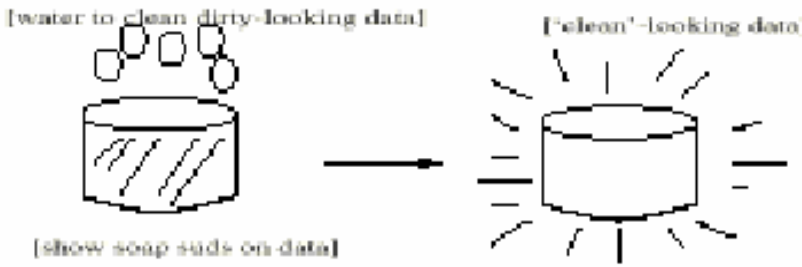
6）附加价值

7）可访问性

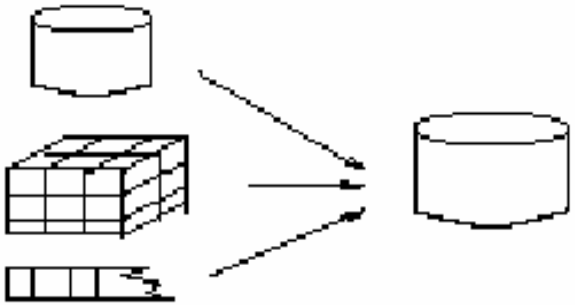
跟数据本身的含义相关的：内在的、上下文的、表象的

数据预处理的形式

数据清理



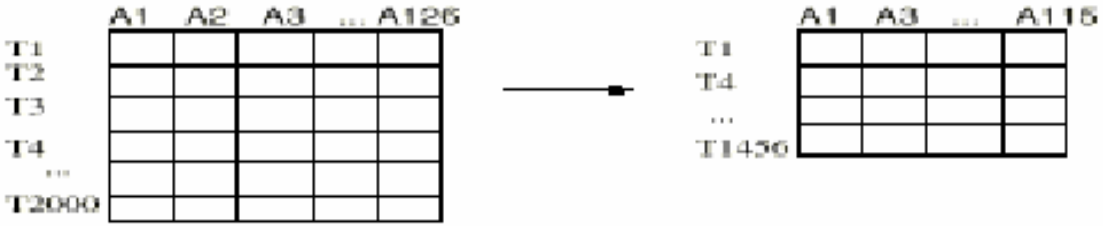
数据集成



数据变换

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

数据归约



数据预处理的主要任务

- **数据清洗**

  填写空缺的值，平滑噪声数据，识别、删除孤立点，解决不一致性

- **数据集成**

  集成多个数据库、数据立方体或文件

- **数据变换**

  规范化和聚集

- **数据归约**

  得到数据集的压缩表示，它小得多，但可以得到相同或相近的结果

- **数据离散化**

  数据归约的一部分，通过概念分层和数据的离散化来规约数据，对数字型数据特别重要

数据清洗-空缺值

- **空缺值**

1）数据并不总是完整的：例如：数据库表中，很多条记录的对应字段没有相应值，比如销售表中的顾客收入。

2）引起空缺值的原因：设备异常、与其他已有数据不一致而被删除、因为误解而没有被输入的数据、在输入时，有些数据应为得不到重视而没有被输入等。

- **处理空缺值**

1）忽略元组

2）人工填写空缺值：工作量大，可行性低

3）使用一个全局变量填充空缺值：比如使用unknown或-∞

4）使用属性的平均值填充空缺值

5）使用与给定元组属同一类的所有样本的平均值

6）**使用最可能的值填充空缺值：使用像Bayesian公式或判定树这样的基于推断的方法**

数据清洗-噪声数据

- **噪声**：一个测量变量中的随机错误或偏差

引起不正确属性值的原因：数据收集工具的问题、数据输入错误、数据传输错误、技术限制、命名规则的不一致

- **处理噪声数据**

1）分箱(binning)：

首先排序数据，并将他们分到等深的箱中；然后可以按箱的平均值平滑、按箱中值平滑、按箱的边界平滑等等

2）聚类：

监测并且去除孤立点

3）计算机和人工检查结合

计算机检测可疑数据，然后对它们进行人工判断

4）回归

通过让数据适应回归函数来平滑数据

数据集成

- **数据集成**

  将多个数据源中的数据整合到一个一致的存储中

- **处理数据集成中的冗余数据**

  集成多个数据库时，经常会出现冗余数据：同一属性在不同的数据库中会有不同的字段名；一个属性可以由另外一个表导出，如“年薪”。

  有些冗余可以被相关分析检测到。

  仔细将多个数据源中的数据集成起来，能够减少或避免结果数据中的冗余与不一致性，从而可以提高挖掘的速度和质量。

数据变换

- **数据变换**

平滑：去除数据中的噪声

聚集：汇总，数据立方体的构建

数据概化：沿概念分层向上汇总

规范化：将数据按比例缩放，使之落入一个小的特定区间：最小－最大规范化；z-score规范化；小数定标规范化

属性构造：通过现有属性构造新的属性，并添加到属性集中。

数据归约

- 数据归约

数据归约可以用来得到数据集的归约表示，它小得多，但可以产生相同的（或几乎相同的）分析结果

- 数据归约策略

数据立方体聚集、维归约、数据压缩、数值归约、离散化和概念分层产生

用于数据归约的时间不应当超过或“抵消”在归约后的数据上挖掘节省的时间。

数据归约

- 数据立方体聚集

最底层的方体对应于基本方体：基本方体对应于感兴趣的实体。

在数据立方体中存在着不同级别的汇总：数据立方体可以看成方体的格；每个较高层次的抽象将进一步减少结果数据。

数据立方体提供了对预计算的汇总数据的快速访问：使用与给定任务相关的最小方体；在可能的情况下，对于汇总数据的查询应当使用数据立方体。

数据归约

- 维归约

  通过删除不相干的属性或维减少数据量

  属性子集选择：找出最小属性集，使得数据类的概率分布尽可能的接近使用所有属性的原分布；减少出现在发现模式上的属性的数目，使得模式更易于理解。

  启发式的（探索性的）方法：逐步向前选择、逐步向后删除、向前选择和向后删除相结合和判定归纳树

数据归约

- **数据压缩**

有损压缩 VS. 无损压缩

字符串压缩：有广泛的理论基础和精妙的算法；通常是无损压缩；在解压缩前对字符串的操作非常有限

音频/视频压缩：通常是有损压缩，压缩精度可以递进选择；有时可以在不解压整体数据的情况下，重构某个片断

两种有损数据压缩的方法：小波变换和主要成分分析

数据归约

- **数值归约**

通过选择替代的、较小的数据表示形式来减少数据量

有参方法：使用一个参数模型估计数据，最后只要存储参数即可：线性回归方法：$Y=\alpha+\beta X$；多元回归：线性回归的扩充；对数线性模型：近似离散的多维数据概率分布

无参方法：直方图；聚类

数据归约

- **直方图**

一种流行的数据归约技术。

将某属性的数据划分为不相交的子集，或桶，桶中放置该值的出现频率。

桶和属性值的划分规则

1）等宽
2）等深
3）V-最优
4）MaxDiff