第3章 数据

目录

- 数据基础
- 数据特征
- 数据预处理
- 数据存储
- 数据分析

1 总览

数据

- 数据是符号的集合，是表达客观事物的未经加工的原始素材
  - 数据模型是用来描述数据表达的底层描述模型，它包含数据的定义和类型
- 数据也可看成是数据对象和其属性的集合
  - 其中属性可被看成是变量、值域、特征或特性

大数据时代

- 麦肯锡公司在2011年开始，提出“大数据时代”，“大数据”特征
  - 海量的数据规模
  - 快速的数据流转、动态的数据体系
  - 多样的数据
  - 巨大的数据价值

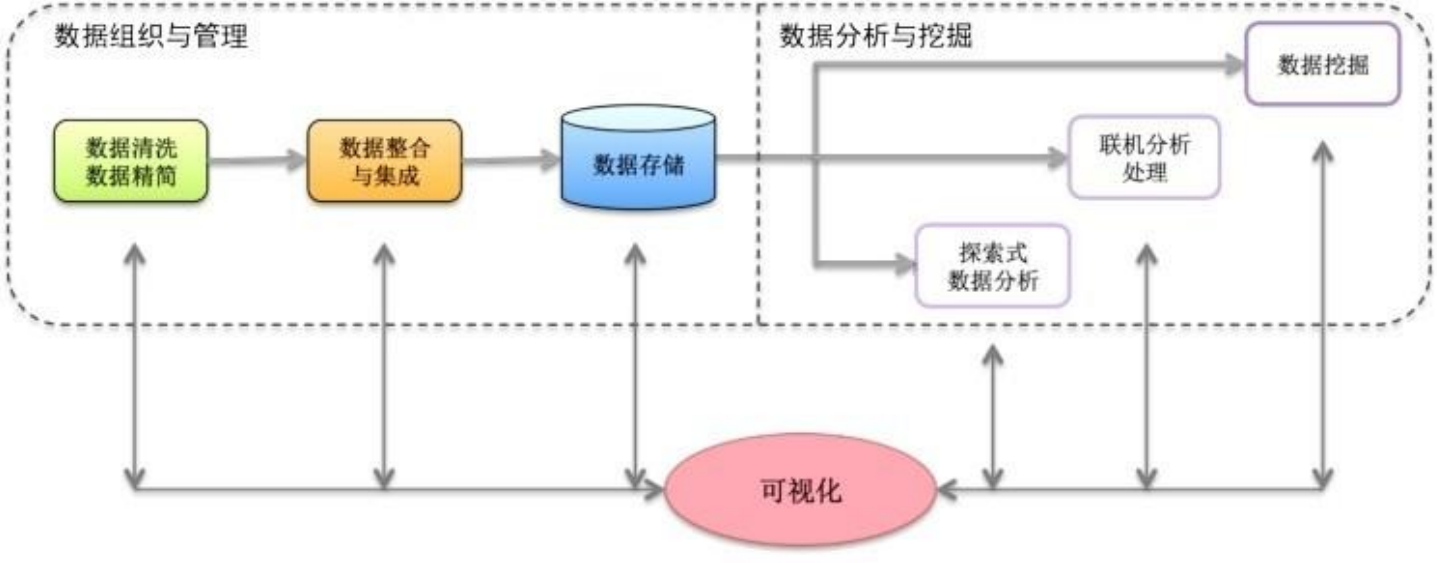
- 互联网时代，拥有大量客户数据的IT企业和金融企业，所积累的商业数据能够发挥的作用与日俱增
- 科学研究领域，传统的科学探究模式正在遭受来自大数据的强烈冲击
- 数据在政府管理、国家安全等领域的价值也越来越明显
- “数据即服务（DaaS）” 。用户可以随时随地按需求获取数据和信息

数据科学的一些基本主题

- 《第四范式：数据密集型的科学发现》
- 数据和信息的历史
- 数据、信息、知识概念和最新进展
- 数据与信息科学的学术基础
- 信息学简介
- 科学的数据生命周期
- 数据获取、保存和保护
- 数据集成
- 元数据
- 数据模型和架构
- 数据工具、基于数据的服务范式
- 数据网、网页上的数据、深层网
- 数据工作流管理
- 数据可视化
- 数据发现
- 数据和信息管理

数据科学过程

- 作为数据内涵信息的展示方法和人机交互接口，数据可视化已成为数据科学的核心要素之一



2 数据基础

数据分类

- 从关系模型的角度讲，数据可被分为实体和关系两部分
  - 实体，是被可视化的对象
  - 关系，定义了实体与其他实体之间关系的结构和模式
  - 实体关系模型，描述数据之间的结构，但不考虑基于实体、关系和属性的操作

- 数据属性可分为离散属性和连续属性
  - 离散属性的取值来自有限或可数的集合
  - 连续属性则对应于实数域

- 针对基本数据类型的交互方法主要有
  - 概括、缩放、过滤、查看细节、关联、查看历史和提取。这些基本任务构成了可视语言设计的基础

数据集

- 数据集是数据的实例。常见的数据集的表达形式有三类
- 数据记录集
  - 由一组包含固定属性值的数据元素组成
    - 数据矩阵
    - 文档向量
    - 事务型
- 图数据集
  - 由一组节点和一组连接两个节点之间的加权边组成
  - 世界航线图、万维网链接图、化学分子式等
- 有序数据集
  - 有序数据是具有某种顺序的数据集
  - 空间数据、时间数据、时空数据、顺序数据和基因测序数据等

数据相似度与密度

- 相似度(Similarity) ，衡量多个数据对象之间相似的数值，通常位于0 和1 之间
- 相异度（Dissimilarity), 其下限是0, 上限与数据集有关，可能超过1
- 邻近度是相似度和相异度的统一描述

数据相似度与密度

- 一些常用的距离和相似度定义有
  - 欧几里得距离
  - 明科夫斯基距离（欧几里得距离的推广）
  - 余弦距离
  - Jaccard 相似度

- 在基于密度的数据聚类时，需要衡量数据的密度，通常定义有三类
  - 欧几里得密度（单位区域内的点的数目）。
  - 概率密度。
  - 基于图结构的密度。

3 数据获取、清洗和预处理

数据获取

- 数据获取的手段有实验测量、计算机仿真与网络数据传输等
  - 传统的数据获取方式以文件输入/输出为主
  - 在移动互联网时代，基于网络的多源数据交换占据主流

- 数据获取的挑战：主要有数据格式变换，异构异质数据的获取协议
  - 数据的多样性导致不同的数据语义表述，这些差异来自不同的安全要求、不同的用户类型、不同的数据格式、不同的数据来源

数据获取协议

- 数据获取协议（Data Access Protocol,DAP），通用的数据获取标准。通过定义基于网络的数据获取句法，以完善数据交换机制，维护、发展和提升数据获取效率
- 理论上，数据获取协议是一个中立的、不受限于任何规则的协议，它提供跨越规则的句法的互操作性，允许规则内的语义互操作性
  - 数据获取协议以文件为基础，提供数据格式、位置和数据组织的透明度，并以纯Web化的方式与网格FTP/FTP、HTTP、SRB（Source Route Bridging，源路由网桥）、开放地理空间联盟（如WCS,WMS,WFS）、天文学（如SIAP,SSAP,STAP）等协议兼容
  - 第二代数据获取协议DAP2已提供了一个与领域无关的网络数据获取协议，业已成为NASA/ESE标准
- OPeNDAP（http://www.opendap.org）
  - 是一个研发数据获取协议的组织，它提供了一个同名的科学数据联网的简要框架，允许以本地数据格式快速地获取任意格式远程数据的机制

数据清洗

- 对于海量数据来说，未经处理的原始数据中包含大量的无效数据，这些数据在到达存储过程之前就应该被过滤掉。解决问题的方法称为数据清洗（Data Cleaning）

- 在原始数据中，常见的数据质量问题包括
  - 噪声指对真实数据的修改；离群值指与大多数数据偏离较大的数据。
  - 数值缺失，主要原因包括：信息未被记录；某些属性不适用于所有实例。处理数据缺失的方法有：删除数据对象；插值计算缺失值；在分析时忽略缺失值；用概率模型估算缺失值等。非结构化数据通常存在低质量数据项（如从网页和传感器网络获取的数据），构成了数据清洗和数据可视化的新挑战
  - 数值重复，主要来源是异构数据源的合并，可采用数据清洗方法消除

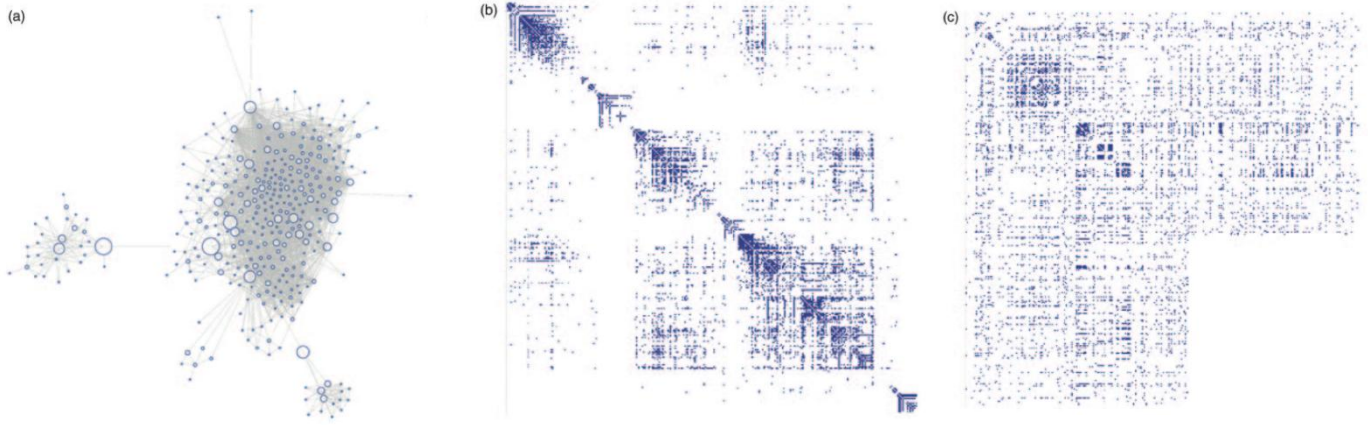
数据清洗目标

- 数据清洗最终需要达到的目标，包括有效性、准确性、可信性、一致性、完整性和时效性六个方面

目　　标	含　　义
有效性	数据是否真实合理
准确性	数据是否精确，有无误差
可信性	数据来源和收集方式是否可信
一致性	数据（格式、单位等）是否一致
完整性	数据是否有缺失
时效性	数据适用范围（相对分析任务）

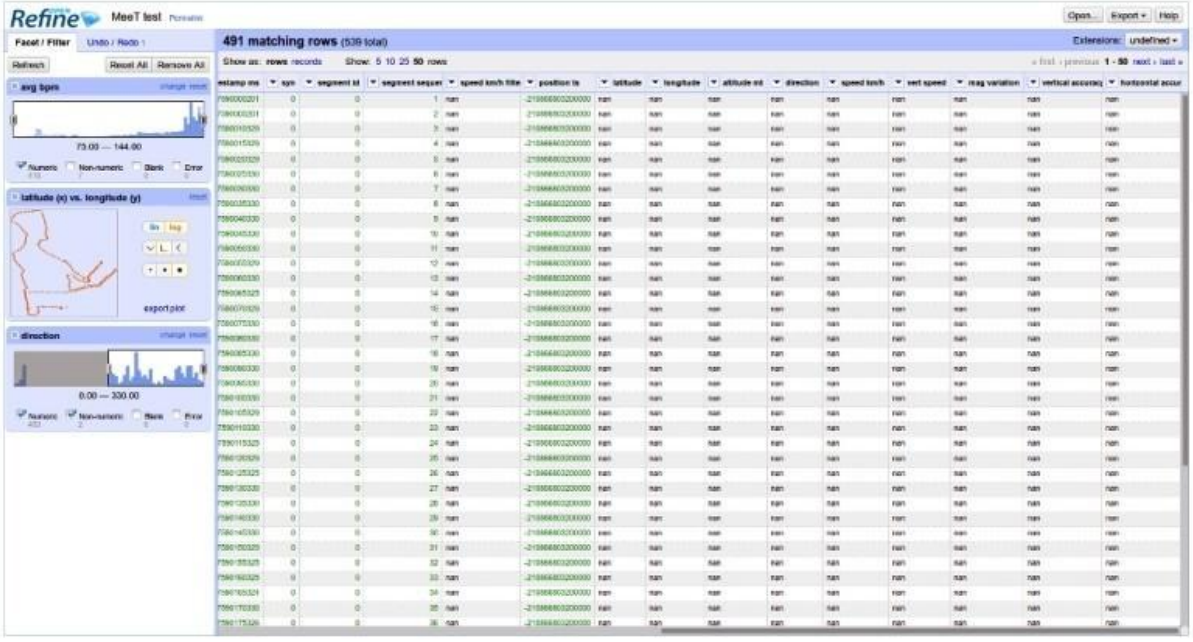
可视数据清洗

- 使用可视化工具进行数据清洗。33种脏数据类型，并且强调其中的25种在清理时需要人的交互。



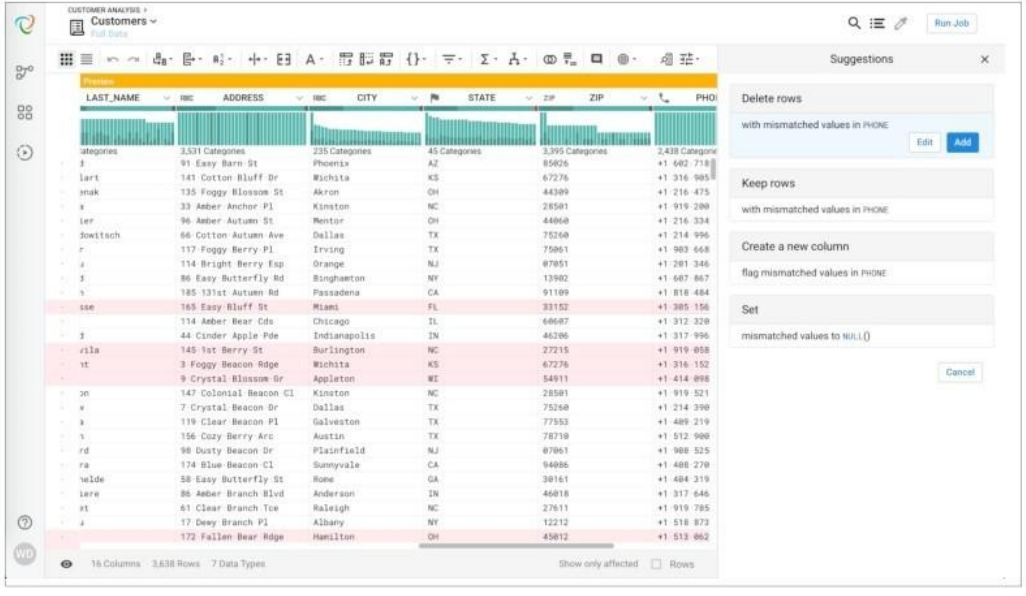
数据清洗

- OpenRefine以数据表格为主要展示方式，配以交互式和脚本式数据编辑操作，来支持用户的数据清洗任务



数据清洗

- Trifacta Wrangler系统界面。左侧使用数据表格和直方图形式展示数据原貌；右侧面板提供了一系列推荐的数据编辑和清洗操作



数据精简

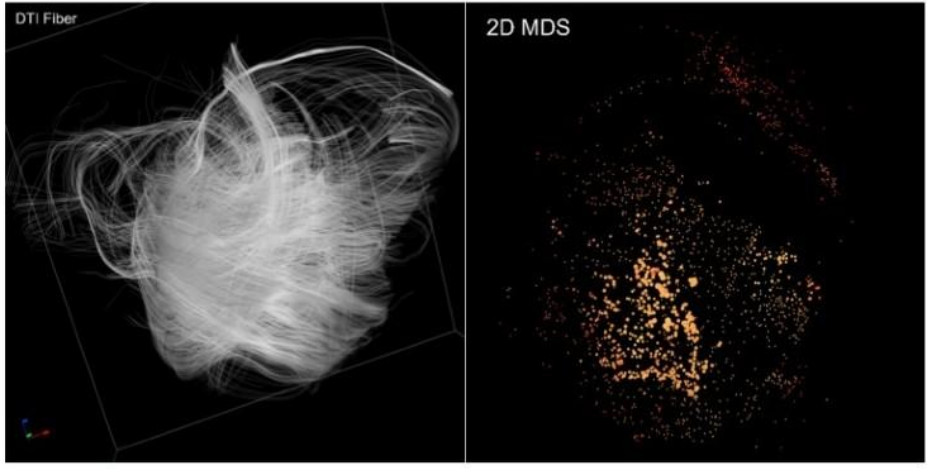
- 直接对海量高维的数据集进行可视化通常会产生杂乱无章的结果，这种现象被称为视觉混乱

- 为了能在有限的显示空间内表达比显示空间尺寸大得多的数据，需要进行数据精简

- 经典的数据精简
  - 包括，如统计分析、采样、直方图、聚类和降维
  - 也可采用各类数据特征抽取方法，如奇异值分解、局部微分算子、离散小波变换等

数据精简

- 将从弥散张量成像数据中抽取的人脑纤维看成高维空间的点（左），定义相似性，并采用多维尺度分析算法，将纤维集投影到二维空间（右），允许用户快速地选择、操纵和观察三维纤维丛的结构和分布



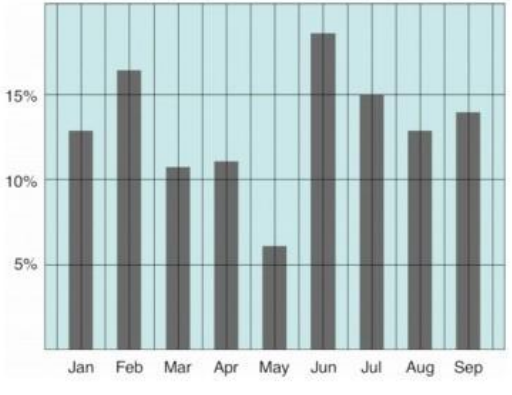
数据精简

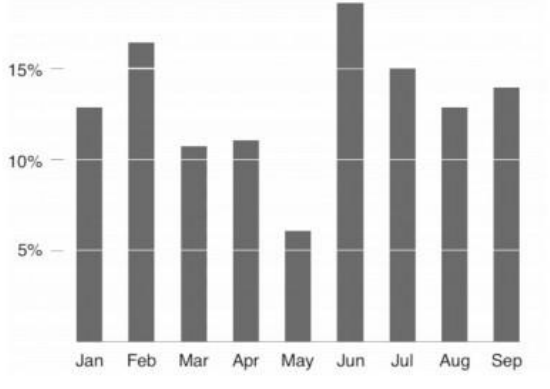
- 数据精简方法可分为两类
  - 使用质量指标优化**非视觉因素**，如时间、空间等
  - 使用质量指标优化**数据可视化**，称为可视数据精简

- 可视数据精简需要自动分析数据以便选择和衡量数据的不同特征，如关联性、布局和密度。这些量度指导和评估数据精简的过程，向用户呈现优化的可视化结果。

可视化质量指标

- 常用的可视化质量指标包括尺寸、视觉有效性和特征保留度
- 尺寸是可量化的量度，如数据点的数量，构成了其他计算的基础
- 视觉有效性用于衡量图像退化（如冲突、模糊）或可视布局的美学愉悦程度
  - 常见方法有数据密度和数据油墨等特征
    - 数据油墨比是Edward Tufte提出的一个数据可视化质量评估标准，它被定义为用于展现数据的像素数目与全部油墨像素数目的比值
- 特征保留度是评估可视化质量的核心，它衡量可视化结果在数据、可视化和认知方面正确展现数据特性的程度





其他常用的数据预处理步骤

- 通常需要对数据集进行进一步的处理操作，以符合后续数据分析步骤要求。这一类操作通常被归为数据预处理步骤。常用的预处理操作有
- 合并
  - 将两个以上的属性或对象合并为一个属性或对象
  - 合并操作的效用包括：有效简化数据；改变数据尺度（例如，从乡村起逐级合并，形成城镇、地区、州、国家等）；减少数据的方差
- 采样
  - 采样是统计学的基本方法，也是对数据进行选择的主要手段，在对数据的初步探索和最后的数据分析环节经常被采用
  - 统计学家实施采样操作的根本原因是获取或处理全部数据集的代价太高，或者时间开销无法接受
  - 最简单的随机采样可以按某种分布随机从数据集中等概率地选择数据项。

其他常用的数据预处理步骤

- 降维
  - 维度越高，数据集在高维空间的分布越稀疏，从而减弱了数据集的密度和距离的定义对于数据聚类和离群值检测等操作的影响
  - 将数据属性的维度降低，有助于解决维度灾难，减少数据处理的时间和内存消耗；可以更为有效地可视化数据；降低噪声或消除无关特征等
  - 降维是数据挖掘的核心研究内容，常规的做法有主元分析、奇异值分解、局部结构保持的LLP、ISOMAP等方法
- 特征子集选择
  - 从数据集中选择部分数据属性值可以消除冗余的特征、与任务无关的特征
  - 特征子集选择可达到降维的效果，但不破坏原始的数据属性结构
  - 特征子集选择的方法包括：暴力枚举法、特征重要性选择、压缩感知理论的稀疏表达方法等

其他常用的数据预处理步骤

- 特征生成
  - 特征生成可以在原始数据集基础上构建新的能反映数据集重要信息的属性
  - 三种常用的方法是：特征抽取、将数据应用到新空间、基于特征融合与特征变换的特征构造
- 离散化与二值化
  - 将数据集根据其分布划分为若干个子类，形成对数据集的离散表达，称为离散化
  - 将数据值映射为二值区间，是数据处理中的常见做法
  - 将数据区间映射到[0,1]区间的方法称为归一化
- 属性变换
  - 将某个属性的所有可能值一一映射到另一个空间的做法称为属性变换
    - 指数变换、取绝对值等
  - 标准化与归一化是两类特殊的属性变换，其中标准化将数据区间变换到某个统一的区间范围，归一化则变换到[0,1]区间

4 数据组织与管理

数据管理与存储

- 数据管理包括对数据进行有效的收集、存储、处理和应用的过程
- 面向应用的数据管理，它的管理对象是数据生命周期所涉及的应用过程中描述构成应用系统构件属性的元数据，包括
  - 流程、文件、数据元、代码、规则、脚本、档案、模型、指标、物理表、ETL、运行状态等
- 通常数据按照一定的组织形式和规则进行存储和处理，以实现有效的数据管理。从逻辑上看，数据组织具有一个层层相连的层次体系
  - 位、字符、数据元、记录、文件、数据库。
  - 记录是逻辑上相关的数据元组合
  - 文件是逻辑上相关的记录集合
  - 数据库是一种作为计算机系统资源共享的数据集合

文件存储

- 以文件作为数据输入和输出的形式
  - 作为一种高度灵活的数据存储形式，它允许使用者非常自由地进行数据处理
  - 弊端。如，数据可能出现冗余、不一致，数据访问烦琐，难以添加数据约束，安全性不高等问题
- 电子表单（Spreadsheet）是多功能的数据组织形式，被广泛使
  - 电子表单文件的变种，如逗号分隔值（CSV）文件格式，也已经被大量的数据交换程序支持
  - 缺少类型和元数据，因而在使用时需要预先给出对每个数据项的语义解释

结构化文件格式

- 为方便通用型数据存储和交换，数据导向型的应用程序采用标记语言格式将数据进行结构化组织
  - XML（Extensible Markup Language，可扩展标记语言），典型代表
  - 除此之外，一些科学领域使用特定的结构化文件记录数据，以满足特殊领域知识的表达高性能处理的需求
    - VOTable是一种由国际虚拟天文台联盟（IVOA）团队制定的XML数据格式，统一了记录天文星表等表列数据的格式
    - NetCDF（网络通用数据格式）是由美国大学大气研究协会针对科学数据的特点开发的面向数组型并适合于网络共享的数据的描述和编码标准，被广泛应用于大气科学、水文、海洋学、环境模拟、地球物理等诸多领域
    - HDF（层次型数据结构）是由美国国家超级计算应用中心创建、以满足不同群体的科学家对不同工程项目领域之需的多对象文件格式

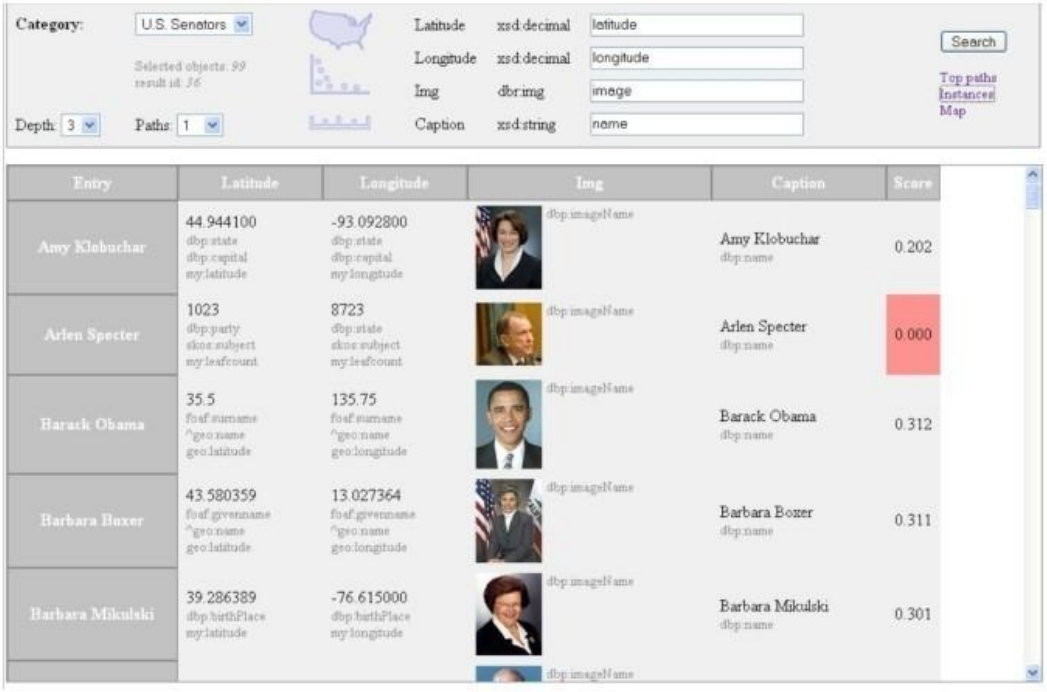
数据库

- 数据组织的高级形式是数据库
  - 即存储在计算设备内、有组织的、共享的、统一管理的数据集合

- 数据库中保存的数据结构既描述了数据间的内在联系，便于数据增加、更新与删除，也保证了数据的独立性、可靠性、安全性与完整性，提高了数据共享程度和数据管理效率
- 关系数据库模型是当前数据库系统最为常用的数据模型

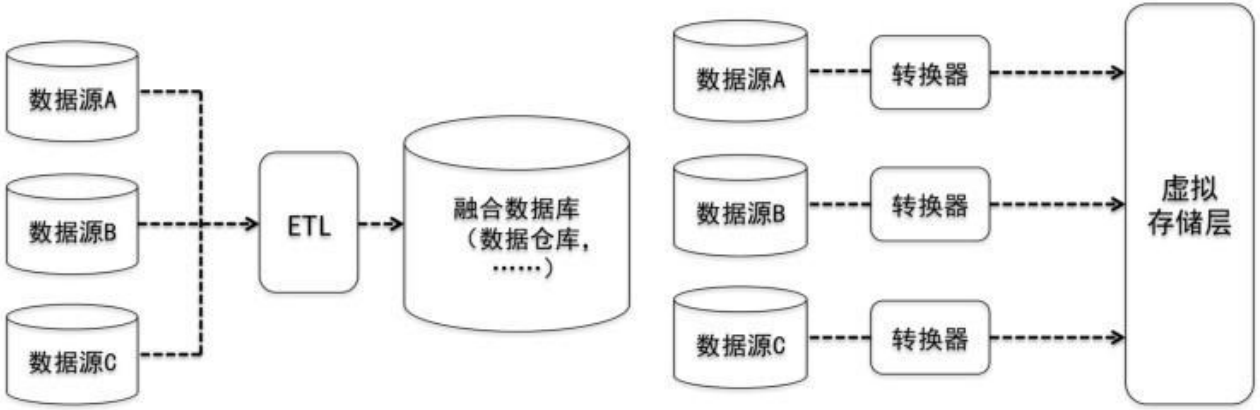
数据整合

- 数据整合指将不同数据源的数据进行采集、清洗、精简和转换后统一融合在一个数据集合中，并提供统一数据视图的数据集成方式



数据整合

- 数据整合的常用方式有以下两种
  - 物化式：查询之前，涉及的数据块被实际交换和存储到同一物理位置（如数据仓库等）。物化式数据整合需要对数据进行物理移动，即从源数据库移动到其他位置
  - 虚拟式：数据并没有从数据源中移出，而是在不同的数据源之上增加转换策略，并构建一个虚拟层，以提供统一的数据访问接口
    - 虚拟式数据整合通常使用中间件技术，在中间件提供的虚拟数据层之上定义数据映射关系。同时，虚拟层还负责将不同数据源的数据在语义上进行融合，即在查询时做到语义一致



数据整合

- 数据整合的需求来源于多个方面
  - 从数据获取的角度看，数据获取的不精确、大范围的不协调数据采集策略、商业竞争和存储空间限制等都是进行多数据源数据整合的原因
  - 在实际的数据可视化应用中，来自不同数据源的数据可能具有不同的质量，这也是面向不同质量数据的整合方法的动机
- 交互分析和可视数据要求数据整合采用**集中式、虚拟式**的模式

数据集成

- 数据集成指数据库应用中结合不同资源的数据并为用户提供数据集合的统一访问，其涵盖范围要比数据整合广

数据库

- 数据库，数据的集合，并且同时包含对数据的相关组织和操作
- 数据库管理系统，帮助维护大量数据集合的软件系统，用来满足对数据库进行存储、管理、维护以及提供查询、分析等服务的需要
- 数据库管理系统需要考虑以下几方面的因素
  - 数据库模型设计
  - 数据分析支持
  - 并发和容错
  - 速度和存储容量

数据库

- 数据库结构的基础是数据模型，它是数据描述、数据联系、数据域以及一致性约束的集合
- 现有的数据模型主要有基于对象和基于记录的逻辑模型
  - E-R模型，基于对象的逻辑模型，它根据现实世界中的实体及实体间的关系对数据进行抽象构建
  - 关系模型，基于记录的逻辑模型，广泛应用在当前各种关系型数据库系统中

关系数据库管理系统(RDBMS)

- 现代的关系数据库管理系统（RDBMS）对数据结构和数据内容提供了明确的分离，允许用户通过控制和管理的方式来访问数据，同时采用稳定的方法来处理安全性和数据一致性
  - 它通过将数据管理设计成符合原子性（Atomic）、一致性（Consistent）、隔离性（Isolated）和持久性（Durable）的事务（事务的ACID特性），确保上述数据管理要求的实现，并使分布在计算机网络不同地点的数据库（Distributed RDBMS）的并发数据访问和数据恢复得到支持

关系型数据库

- 关系模型作为一种最常见的基于记录的逻辑模型，广泛应用在当前各种关系型数据库系统中
  - 它借助于关系代数等数学概念和方法来处理数据库中的数据，由关系数据结构、关系操作集合、关系完整性约束三部分组成
  - 当前主流的关系型数据库有IBM DB2、Oracle、Microsoft SQL Server、MySQL等。标准查询语言（SQL）是关系型数据库的结构化查询语言，它提供了对关系型数据库中的表记录进行查询、操纵的功能

关系型数据库存在一些缺陷

- 关系型数据库存在一些缺陷
  - 交互式数据可视化应用通常需要将数据存储于内存，以保证足够的性能（通常需要亚秒级的响应时间）。除了一些内存数据库，普通的关系型数据库在数据量较大的情况下难以满足可视化交互的高性能要求
  - SQL支持的数据类型是存储导向而不是语义导向的。因此，对复杂关系数据进行处理和可视化时，使用者需要在数据库中添加更多的数据描述来表达记录间的语义关联，然而这样做会增加数据库设计的复杂度以及存储、查询开销
  - 关系型数据库中的事件通知通常用触发器机制实现。这种低效的通知机制难以满足数据可视化的实时性要求

NoSQL数据库

- 新的挑战
  - 在数据爆炸时代，相对于传统的结构化数据（表格记录、关系记录等）来说，新增加的大部分数据类型都是在自然和社会环境中产生的，这些数据类型并不具有结构化特征
  - 从应用目标的角度看，数据库系统同时还可以提供对所存储的数据进行分析，继而进行决策支持的功能
- NoSQL数据库被认为是不同于传统关系型数据库的数据库管理系统的总称，这种数据库能够满足对数据的高并发读写、高效存储和访问、数据库高扩展性和高可用性等需求
  - 谷歌公司内部采用特别优化的分布式存储系统BigTable
  - 亚马逊公司开发的Dynamo使用私有的键-值结构的存储系统处理Web服务等

NoSQL数据库

- “Not Only SQL”（不仅仅是SQL）
  - 面向海量数据（并且数据不需要关系模型）
  - 通常不使用表结构，并且不使用SQL进行查询

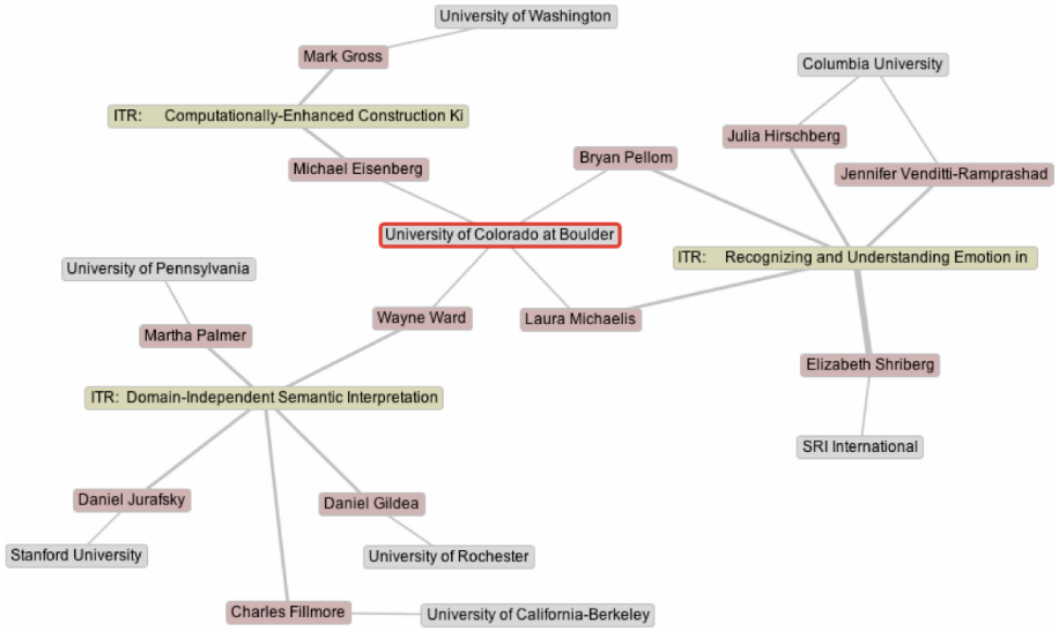
- NoSQL 数据库被认为是不同于传统关系型数据库的数据库管理系统的总称，这种数据库能够满足对数据的高并发读写、高效存储和访问、数据库高扩展性和高可用性等需求，为SNS 网站等规模大、并发数高的应用提供了符合其性能标准的解决方案
- 在数据分析方面，NoSQL 数据库能为大型数据集的在线分析提供更快、更简单、更专门的服务

NoSQL数据库实例

- 文档存储 – CouchDB
- 图结构存储 – Neo4j
- 键-值存储 – Redis（内存数据库），MongoDB（磁盘数据库）
- 表格数据 – Apache HBase (基于Hadoop)

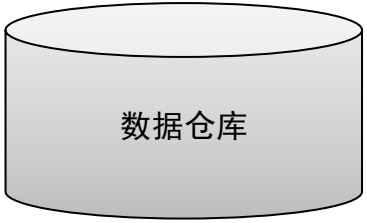
关系数据库可视化

- 现有一些数据可视化应用开始直接针对关系型数据库进行可视化，并且拥有简单的统计、分析功能



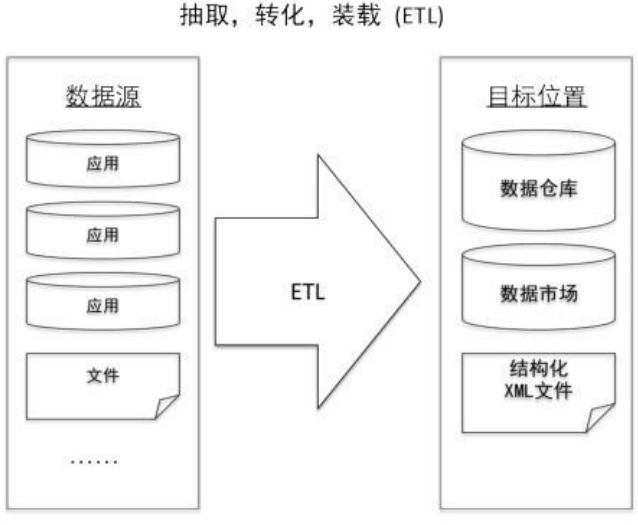
数据仓库

- 数据仓库，指“面向主题的、集成的、与时间相关的、主要用于存储的数据集合，支持管理部门的决策过程”
- 数据仓库目的，是构建面向分析的集成化数据环境，为分析人员提供决策支持
  - 数据仓库通常有特定的应用方向，并且能够集成多个异构数据源的数据
  - 同时，数据仓库中的数据还具有时变性、非易失性等特点



数据仓库流水线

- 数据仓库中的数据来源于外部，开放给外部应用，其基本架构是数据流入/流出的过程，该过程可以分为三层——源数据、数据仓库和数据应用。其流水线简称为ETL（抽取Extract、转化Transform、装载Load）
  - 抽取阶段从一个或多个数据源中抽取原始数据。
  - 转化阶段主要进行数据变换操作
    - 主要有清理、重构、标准化等
  - 装载阶段将转化过的数据按一定的存储格式进行存储



数据库和数据仓库的异同

- 数据仓库通常围绕某个应用目标、应用领域或使用者所感兴趣的内容制定，包含了一些相关的、由外部产生的数据
- 数据仓库可以不断更新和增长，这意味着数据可以被源源不断地积累起来，从而允许用户分析数据的趋势变化、模式和相互关系
- 数据仓库为复杂的决策支持查询进行了大量优化。数据仓库的不同目标和数据模型也同时引发了不同于传统数据库的技术、方法论和方法的各种研究
- 对于结构化或非结构化数据，数据仓库都能有效地进行处理，并且还能够提供两种数据的整合功能

5 数据分析与挖掘

数据分析

- 数据，用于记录事实，是实验、测量、仿真、观察、调查等的结果
- 数据分析，指组织有目的地采集数据、详细研究和概括总结数据，从中提取有用信息并形成结论的过程，其目的是从一堆杂乱无章的数据中集中、萃取和提炼出信息，探索数据对象的内在规律
- 概念上
  - 数据分析的任务分解为，定位、识别、区分、分类、聚类、分布、排列、比较、内外连接比较、关联、关系等活动
  - 基于数据可视化的分析任务包括
    - 识别、决定、可视化、比较、推理、配置和定位。基于数据的决策则可分解为确定目标、评价可供选择方案、选择目标方案、执行方案等
- 从统计应用上讲，数据分析可以被分成
  - 描述性统计分析、探索式数据分析和验证性数据分析三类

数据分析

- 数据分析从统计学中发展而来，在各行业中体现出极大的价值。具有代表性的数据分析方向有统计分析、探索式数据分析、验证性数据分析等
- 探索式数据分析主要强调从数据中寻找出之前没有发现过的特征和信息
- 验证性数据分析强调通过分析数据来验证或证伪已提出的假说
- 统计分析中的传统数据分析工具包括：排列图、因果图、分层法、调查表、散布图、直方图、控制图
  - 面向复杂关系和任务，又发展了新的分析手段，如关联图、系统图、矩阵图、计划评审技术、矩阵数据图等
  - 流行的统计分析软件如R、SPSS、SAS，支持大量的统计分析方法

数据分析与其它学科的结合

- 数据分析与自然语言处理、数值计算、认知科学、计算机视觉等结合，衍生出不同种类的分析方法和相应的分析软件
  - 科学计算领域的MATLAB
  - 机器学习领域的Weka
  - 自然语言处理领域的SPSS/Text、SAS Text Miner
  - 计算机视觉领域的OpenCV
  - 图形处理领域的Khoros、IRIS Explorer等

数据分析流程

- 从流程上看，数据分析以数据为输入，处理完毕后提炼出对数据的理解
- 在整个数据工作流中，数据分析建立在数据组织和管理基础上，通过通信机制和其他应用程序连接，并采用数据可视化方法呈现数据分析的中间结果或最终结论
- 面向大型或复杂的异构数据集，数据分析的挑战是结合数据组织和管理的特点，考虑数据可视化的交互性和操控性要求需求
  - 一方面，部分数据分析方法采取增量式的策略，但不提供给用户任何中间结果，阻碍了用户对数据分析中间结果的理解和对分析过程的干预。解决方案之一是设计标准化协议。例如，微软公司定义了用于分析XML的协议，其核心是数据挖掘模型（Data Mining Model）和可预测模型标记语言（PMML）
  - 另一方面，用户对部分数据分析结果或可视化结果可能会进行微调、定位、选取等操作，这需要数据分析方法针对细微调节快速修正

数据挖掘与联机分析处理

- 数据挖掘被认为是一种专门的数据分析方式，与传统的数据分析（如统计分析、联机分析处理）方法的本质区别
  - 数据挖掘在**没有明确假设的前提**下去挖掘知识，所得到的信息具有未知、有效和实用三个特征
  - 数据挖掘的任务往往是**预测性的**而非传统的描述性任务
  - 数据挖掘的输入可以是数据库或数据仓库，或者是其他的数据源类型，例如网页、文本、图像、视频、音频等
- 联机分析处理是面向分析决策的方法
  - 传统的数据库查询和统计分析工具负责提供数据库中的内容信息，而联机分析处理则提供基于数据的假设验证方法。这个过程是一个**演绎推理的过程**
  - 数据挖掘并不验证某个假定的模型的正确性，而是从数据中计算未知的模型，因此本质上是一个**归纳的过程**，通过构建模型对未来进行预测

数据挖掘与联机分析处理

- 数据挖掘和联机分析处理都致力于模式发现和预测，具有一定的互补性
  - 数据挖掘并不能替代传统的统计分析和探索式数据分析技术
  - 在实际应用中，需要针对不同的问题类型采用不同的方法
- 将数据可视化作为一种可视思考策略和解决方法，可以有效地提高统计分析、探索式数据分析、数据挖掘和联机分析处理的效率

探索式数据分析

- 探索式数据分析(Exploratory Data Analysis，EDA）是统计学和数据分析结合的产物
- 著名的统计学家、信息可视化先驱John Tukey将探索式数据分析定义为一种以数据可视化为主的数据分析方法，其主要目的包括
  - 洞悉数据的原理
  - 发现潜在的数据结构
  - 抽取重要变量
  - 检测离群值和异常值
  - 测试假设
  - 发展数据精简模型
  - 确定优化因子设置

探索式数据分析

- 探索式数据分析（ Exploratory Data Analysis）是一种有别于统计分析的新思路，不等同于以统计数据可视化为主的统计图形方法
  - 传统的统计分析关注模型，即估计模型的参数，从模型生成预测值
  - 大多数探索式数据分析关注数据本身，包括结构、离群值、异常值和数据导出的模型
- 从数据处理的流程上看，探索式数据分析和统计分析、贝叶斯分析也有很大不同
  - 统计分析的流程是：问题，数据，模型，分析，结论
  - 探索式数据分析的流程是：问题，数据，分析，模型，结论；贝叶斯分析的流程则是：问题，数据，模型，先验分布，分析，结论
- 探索式数据分析与数据挖掘
  - 前者将聚类和异常检测看成探索式过程
  - 而后者则关注模型的选择和参数的调节

联机分析处理

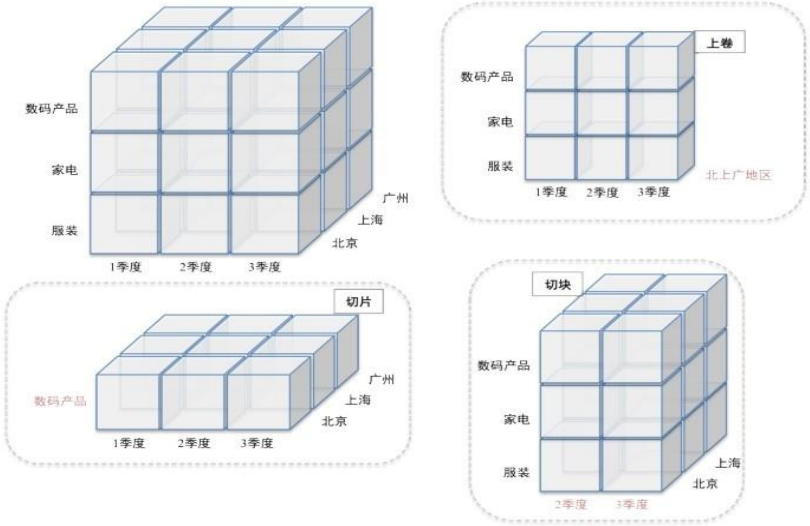
- 联机分析处理（OLAP），一种交互式探索大规模多维数据集的方法
- 关系型数据库将数据表示为表格中的行数据，而联机分析处理则关注统计学意义上的多维数组
- 将表单数据转换为多维数组
  - 首先，确定作为多维数组索引项的属性集合，以及作为多维数组数据项的属性。作为索引项的属性必须具有离散值，而对应数据项的属性通常是一个数值
  - 然后，根据确定的索引项生成多维数组表示

联机分析处理

- 联机分析处理的核心表达是多维数据模型
  - 多维数据模型又可表达为数据立方，相当于多维数组。数据立方是数据的一个容许各种聚合操作的多维表示
- 数据立方可用于记录包含数十个维度、数百万数据项的数据集，并允许在其基础上构建维度的层次结构
- 联机分析处理被广泛看成一种支持策略分析和决策制定过程的方法，与数据仓库、数据挖掘和数据可视化的目标有很强的相关性

联机分析处理

- 联机分析处理基本操作分为两类
  - 切片和切块（Slicing and Dicing）
    - 切片（Slicing）指从数据立方中选择在一个或多个维度上具有给定的属性值的数据项。切块（Dicing）指从数据立方中选择属性值位于某个给定范围的数据子集。两个操作都等价于在整个数组中选取子集
  - 汇总和钻取（Roll-up and Drill-down）
    - 属性值通常具有某种层次结构。可以通过向上汇总或向下钻取的方法获取数据在不同层次属性的数据值

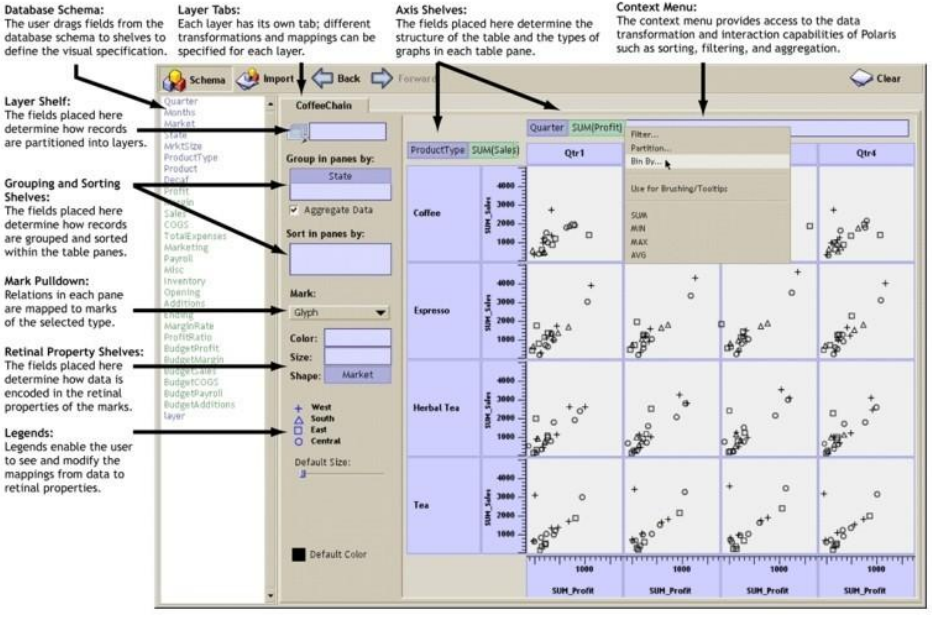


联机分析处理

- 联机分析处理是交互式统计分析的高级形式
- 面向复杂数据，联机分析处理方法的发展趋势，是融合数据可视化与数据挖掘方法，转变为数据的在线可视分析方法
  - 例如，联机分析处理将数据聚合后的结果存储在另一张维度更低的数据表单中，并对该数据表单进行排序以便呈现数据的规律。这种聚合-排序-布局的思路允许用户结合数据可视化的方法（如时序图、散点图、地图、树图和矩阵等）理解高维的数据立方表示
  - 特别地，当需要分析的数据集的维度高达数十维时，采用联机分析处理手工分析力不从心，数据可视化则可以快速地降低数据复杂度，提升分析效率和准确度

联机分析处理

- Polaris是由斯坦福大学开发的用于分析多维数据立方的可视化工具，它针对基于表格的数据进行可视化及分析，可以认为是对表格数据的一种可视化扩展



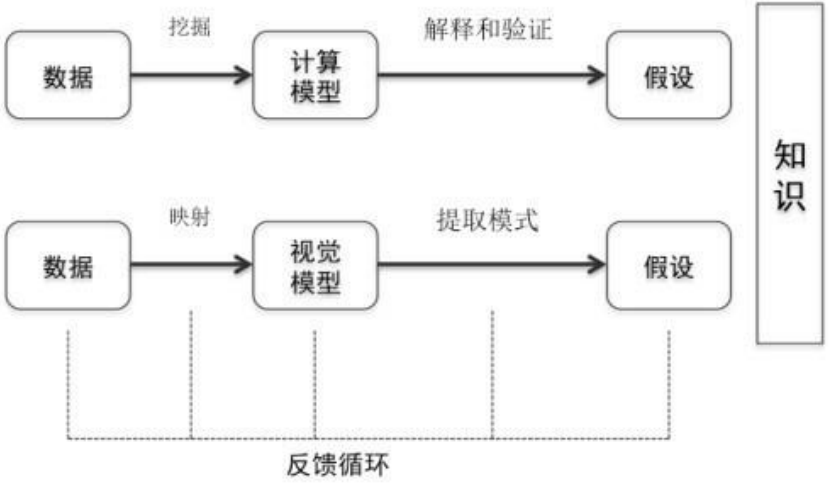
数据挖掘

- 数据挖掘，指设计特定算法，从大量的数据集中去探索发现知识或者模式的理论和方法，是知识工程学科中知识发现的关键步骤
- 面向不同的数据类型可以设计特定的数据挖掘方法
  - 如数值型数据、文本数据、关系型数据、流数据、网页数据和多媒体数据等

- 数据挖掘直观的定义是通过自动或半自动的方法探索与分析数据，从大量的、不完全的、有噪声的、模糊的、随机的数据中**提取隐含在其中的、人们事先不知道的、潜在有用的信息和知识**
  - 数据挖掘不是数据查询或网页搜索，它融合了统计、数据库、人工智能、模式识别和机器学习理论中的思路，特别关注异常数据、高维数据、异构和异地数据的处理等挑战性问题

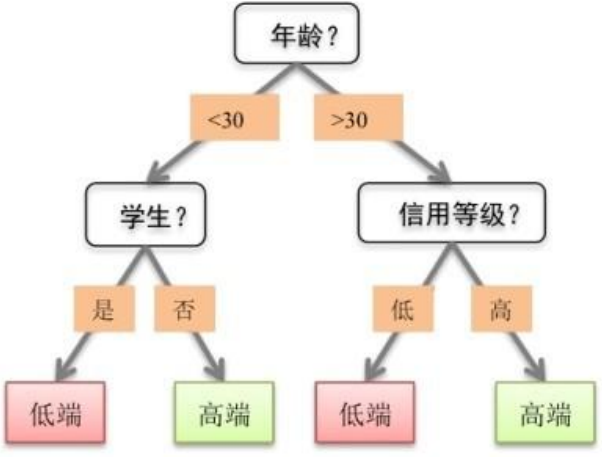
数据挖掘与数据可视化

- 基本的数据挖掘任务分为两类：
  - 基于某些变量预测其他变量的未来值，即预测性方法（例如分类、回归）
  - 以人类可解释的模式描述数据（如聚类、模式挖掘、关联规则发现）
- 直观地说，数据挖掘指从大量数据中识别有效的、新颖的、潜在有用的、最终可理解的规律和知识
- 可视化将数据以形象直观的方式展现，让用户以视觉理解的方式获取数据中蕴含的信息



数据挖掘的主要方法

- 分类（预测性方法）
  - 给定一组数据记录（训练集），每个记录包含一组标注其类别的属性。分类算法需要从训练集中获得一个关于类别和其他属性值之间关系的模型，继而在测试集上应用该模型，确定模型的精度

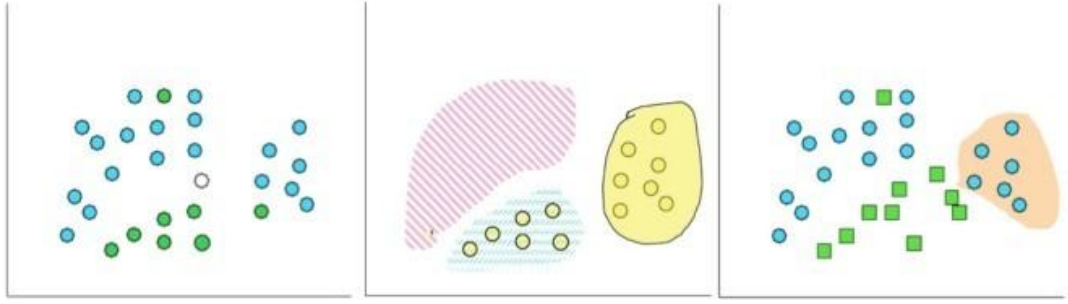


数据挖掘的主要方法

- 聚类（描述性方法）
  - 给定一组数据点以及彼此之间的相似度，将这些数据点分成多个类别，满足位于同一类的数据点彼此之间的相似度大于与其他类的数据点的相似度
  - 聚类技术的要点是，在划分对象时不仅要考虑对象之间的距离，还要求划分出的类具有某种内涵描述，从而避免传统技术的某些片面性

数据挖掘的主要方法

- 概念描述，指对某类数据对象的内涵进行描述，并概括这类对象的有关特征
- 概念描述分为特征性描述和区别性描述
  - 前者描述某类对象的共同特征；后者描述不同类对象之间的区别
  - 生成一个类的特征性描述只涉及该类对象中所有对象的共性，生成区别性描述的方法很多，如决策树方法、遗传算法等



数据挖掘的主要方法

- 关联规则挖掘（描述性方法）
  - 关联规则描述在一个数据集中一个数据与其他数据之间的相互依存性和关联性
  - 数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的属性值之间存在某种规律性，就称为关联
  - 关联可分为简单关联、时序关联、因果关联
  - 如果两个或多个数据之间存在关联关系，那么其中一个数据可通过其他数据预测。
  - 关联规则挖掘则是从事务、关系数据中的项集合对象中发现频繁模式、关联规则、相关性或因果结构

数据挖掘的主要方法

- 序列模式挖掘（描述性方法）
  - 针对具有时间或顺序上的关联性的时序数据集，序列模式挖掘就是挖掘相对时间或其他模式出现频率高的模式
  - 序列模式挖掘主要针对符号模式，而数字曲线模式属于统计时序分析中的趋势分析和预测范畴
  - 序列模式挖掘应用广泛，如交易数据库中的客户行为分析、Web访问日志分析、科学实验过程的分析、文本分析、DNA分析和自然灾害预测等
  - 时序模式是指通过时间序列搜索出的重复发生概率较高的模式

数据挖掘的主要方法

- 回归（预测性方法）
  - 回归在统计学上定义为：研究一个随机变量对另一组变量的相依关系的统计分析方法
    - 其中，线性回归是利用数理统计中的回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法
    - 此外，当自变量为非随机变量、因变量为随机变量时，分析它们的关系称为回归分析；当两者都是随机变量时，称为相关分析

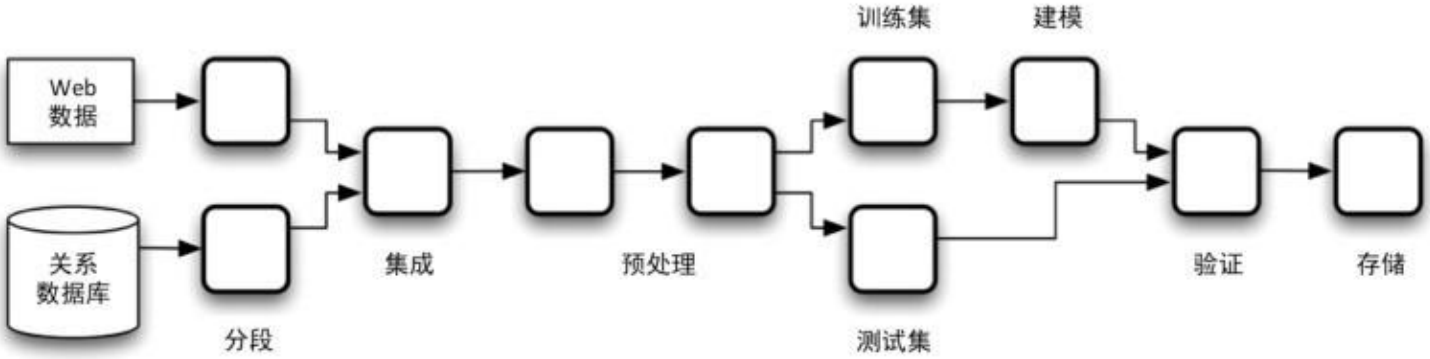
数据挖掘的主要方法

- 偏差检测（预测性方法）
  - 大型数据集中常有异常或离群值，统称为偏差
  - 偏差包含潜在的知识
    - 如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。
  - 偏差检测的基本方法是，寻找观测结果与参照值之间有意义的差别
  - 偏差预测的应用广泛，如信用卡诈骗监测、网络入侵检测等
  - 偏差检验的基本方法就是寻找观察结果与参照值之间的差别

6 数据科学与可视化

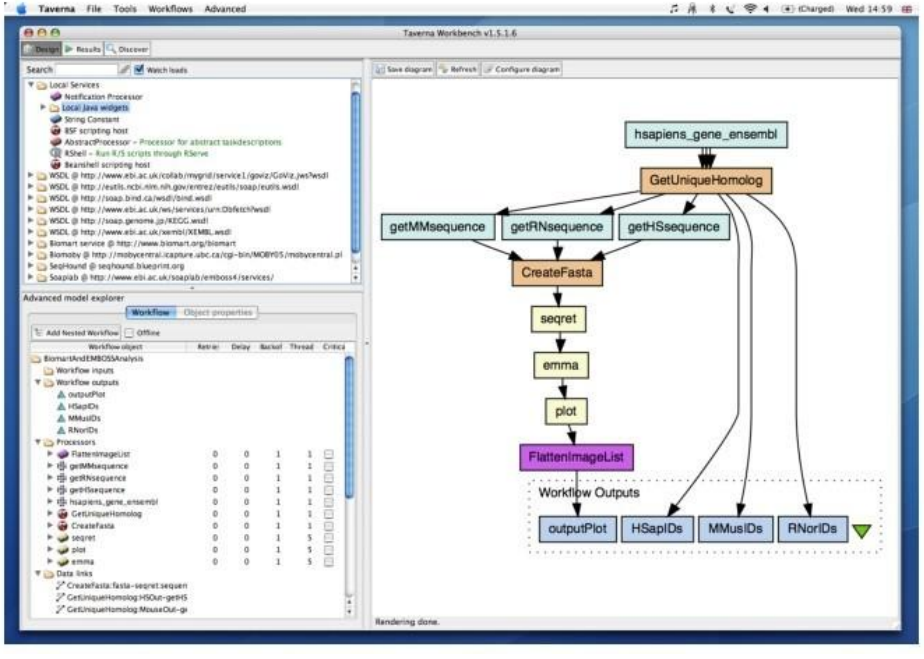
工作流

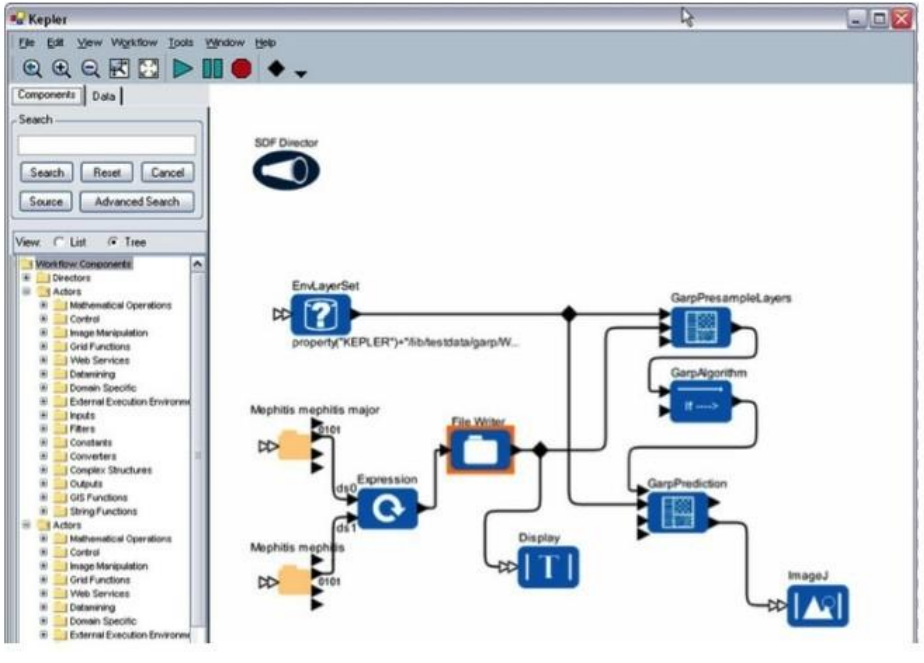
- 工作流
  - 多个用户之间按照某种预定义的规则传递文档、信息或任务的自动过程
  - 起源于生产组织和办公自动化领域，用于描述一个特定的、实际的过程步骤，在计算机应用环境下属于计算机支持的协同工作的研究范畴
  - 定义和遵循工作流有助于以标准化、自动化的方式实现某个预期的业务目标，便于协同、分享、发布和传播有效的工作模式



工作流

- 工作流常见的两种形式
  - 面向商业流程处理和商业数据处理的商业工作流
  - 面向科学研究过程控制和数据处理流程控制的科学工作流





数据工作流

- 数据工作流
  - 特指为数据处理和分析流程定义的自动过程
  - 其本质是计算业务过程的部分或整体在计算机应用环境下的自动化，与自动化工程学科密切相关

- 将工作流应用于科学研究是一个新兴的研究方向
  - 如，专注于科学工作流研究的国际研讨会（IEEE Scientific Workflow）和国际期刊（IEEE Transactions on Automation Science and Engineering）

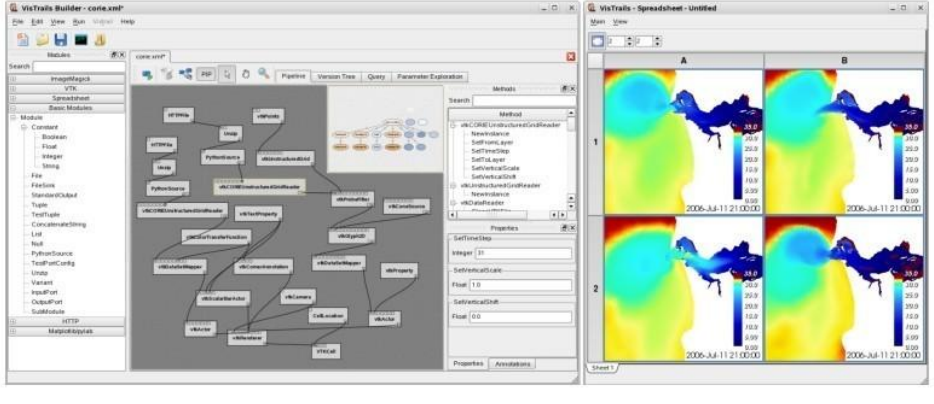
数据工作流与可视化

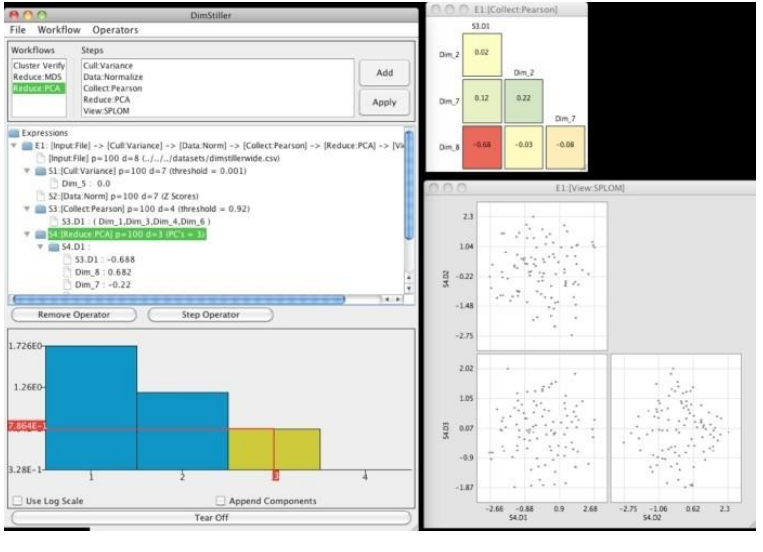
- 可视化在工作流系统中的应用非常广泛
  - 在处理复杂数据和任务时，数据的中间结果是工作流的一个环节

- 将数据可视化理念融合到数据工作流中，将带来的一些新的特点包括
  - 图形化、可视化设计流程图
  - 支持各种复杂流程
  - B/S结构
  - 表单功能强大，扩展便捷
  - 处理过程可视化管理
  - 统计、查询和报表功能

数据工作流与可视化

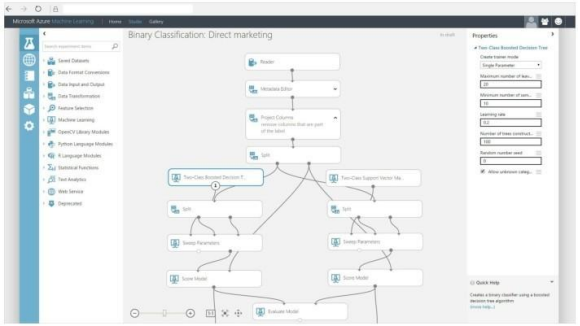
- 工作流的思想应用于数据分析和可视化的工作
  - VisTrails，一个开源的面向计算机仿真、数据浏览和可视化的科学工作流管理系统
    - VisTrails将数据操作的历史信息、数据中间状态和产生中间结果的工作流状态等记录在XML文件或关系型数据库中。简单的交互界面允许用户基于实例定制工作流，并将工作流模式应用于科学数据的探索
  - DimStiller





数据工作流与机器学习

- 以机器学习和数据挖掘为主要数据分析方法的数据科学工作流系统
  - 以数据处理和分析模块作为数据流的基本组成单元，通过拖曳等交互来构建整个数据分析流程
    - RapidMiner,Dataiku,H2O.ai等
  - 云计算平台厂商在推出机器学习功能后，也通过工作流的方式将机器学习模块进行有机结合，使用户通过网页即可定制数据处理和机器学习过程
    - 如，微软Azure Learning、阿里云DTpai和Salesforce MetaMind等



可视数据挖掘

- 可视数据挖掘
  - 数据挖掘领域，用户需要对数据挖掘结果或过程进行理解、检测和验证，可视化为辅助手段
  - 用于辅助增强数据挖掘的可理解性、可交互性，算法精度等
- 可视数据挖掘的目的
  - 用户能参与对大规模数据集进行探索和分析的过程，并在参与过程中搜索感兴趣的知识
  - 同时，呈现数据挖掘算法的输入数据和输出结果，提升数据挖掘模型的可解释性

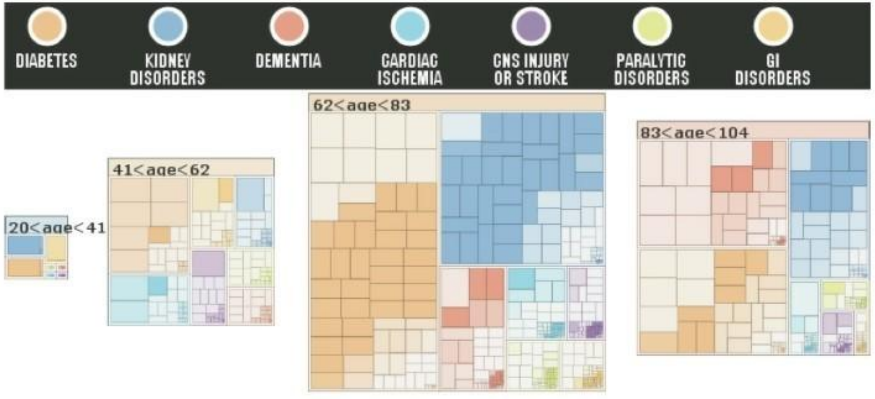
- 可视数据挖掘在一定程度上解决了将人的智慧和决策引入数据挖掘过程这一问题，使得人能够有效地观察数据挖掘算法的结果和一部分过程
- 可视数据挖掘方法主要分为两大类
  - 可视化增强的通用数据挖掘方法
  - 面向应用场景的方法

可视化增强的通用数据挖掘方法

- 可视化方法与通用数据挖掘方法的组合方式大致分为“黑盒”和“白盒”两种
  - 基于黑盒方式的可视化增强方法
    - 可视化设计者仅根据数据挖掘算法所面向的任务进行设计，并不考虑具体算法的内部机制
  - 基于白盒方式的可视化增强方法
    - 对算法过程本身进行展示，使用户能够更好地理解计算结果与输入数据、参数之间的关系

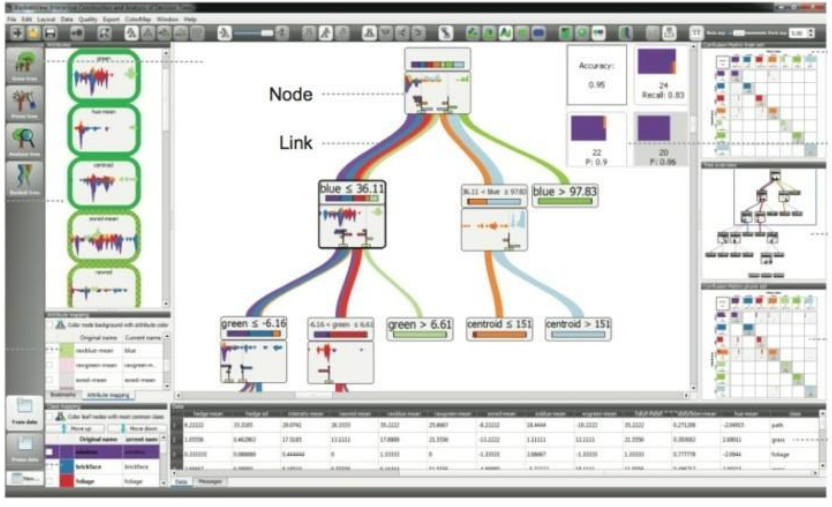
基于黑盒方式的可视化增强方法

- 仅根据数据挖掘算法所面向的任务进行设计，并不考虑具体算法的内部机制
  - 面向输入数据的可视化方法
    - 用户在使用数据挖掘方法之前，可先使用可视化方法对数据进行预探索
  - 面向算法结果的可视化方法
    - 可视化方法通常用于结果分析、模型验证等方面
  - 迭代式可视化方法
    - 帮助用户基于已有探索结果对输入参数或数据进行修改，继而生成新的模型和输出结果，并使用迭代策略引导用户对算法结果进行优化



基于白盒方式的可视化增强方法

- 对算法过程本身进行展示，使用户能够更好地理解计算结果与输入数据、参数之间的关系
- 以可解释性较强的决策树为例
  - 可视化为非数据挖掘领域用户提供了一定程度的辅助
    - 文献[Caragea2001]针对支持向量机提出了基于投影的可视化方案
    - 文献[Erhan2009]提出了一种面向图像识别的深度神经网络可视化方法，其使用可视化方式展示出不同层次神经元上的特征，帮助用户进行参数调整等工作



面向应用场景的方法

- 1.文本分析
- 2.图像分析
- 3.用户行为分析
- 4.时空数据分析
- 5.深度学习

文本分析

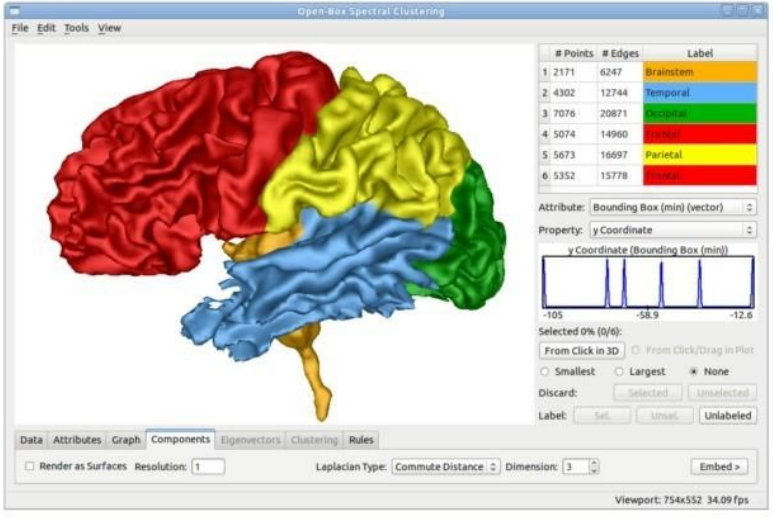
- 常见的应用类型有文本分类、聚类及异常检测等
- 1）对主题抽取结果进行可视化
  - Heimerl等人 [Heimerl2012]提出了一种交互式文本分类器训练系统，以辅助普通文本分析人员进行文本分类分析。该系统借鉴机器学习中的主动学习策略，降低了用户标记文本标签的成本，并加快了分类器的训练过程。同时，作者设计了一系列可视化视图以帮助用户理解文本集的大致词频分布和特征，以简化用户对所需文本的搜索过程
  - 文献[Chuang2012]设计了一种基于LDA（Latent Dirichlet Allocation）的可视化文本探索方法。针对有类别标签的文本数据，该方法首先抽取所有文本的主题分布，并计算不同文本类别间的主题相似度，最终使用基于散点图设计的地形视图（Landmark View）展示类别间的相似度

文本分析

- 2）对主题生成进行干预
  - UTOPIAN系统 [Choo2013]对传统LDA模型进行改进，提出了一种基于交互式非负矩阵分解的用户驱动式主题抽取方法。该方法允许用户通过该模型提供的接口交互式调整模型参数，并根据先验知识对主题生成进行干预
  - Jian Zhao等人提出的FluxFlow交互式可视分析系统 [Zhao2014]用于探测和分析社交媒体文本流中的异常信息传播。该系统利用单类条件随机场（One-Class Conditional Random Field,OCCRF）作为核心机器学习算法，用户可通过基于线索时间线的可视化方法发现当前文本流中的异常传播模式，并深入探索相关文本内容

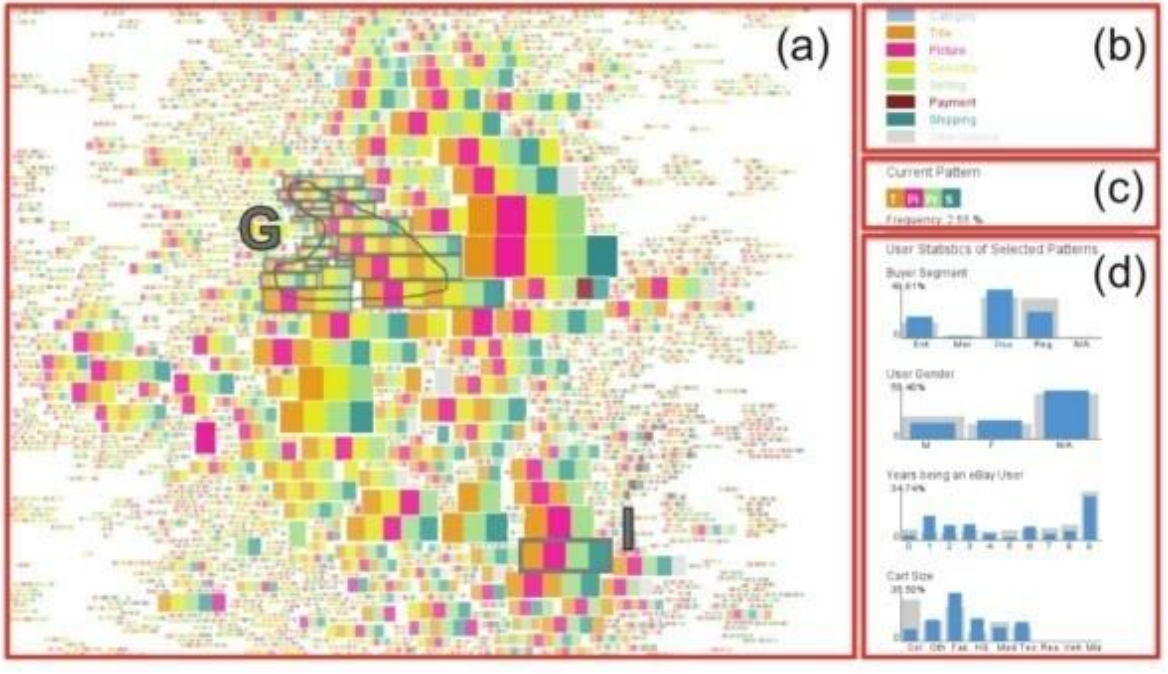
图像分析

- Hoferlin等人提出了针对视频图像数据的交互式主动学习（Inter-Active Learning）策略[Höferlin2012]。在训练图像分类器时，该策略将专家用户对图像的标记判断融入分类过程中。专家用户可通过搜索或查看分类结果的方式找出最需要人工标记的图像，标记后的训练图像能够使分类器的训练效率最大化
- 文献[Schultz2013]基于传统的谱聚类方法，提出了一种白盒式可视化增强的聚类系统，如图3.28所示。该系统面向CT、MRI等三维医学图像数据，支持用户利用专家知识，通过交互式可视化方式调整谱聚类方法所需的参数、聚类数等变量



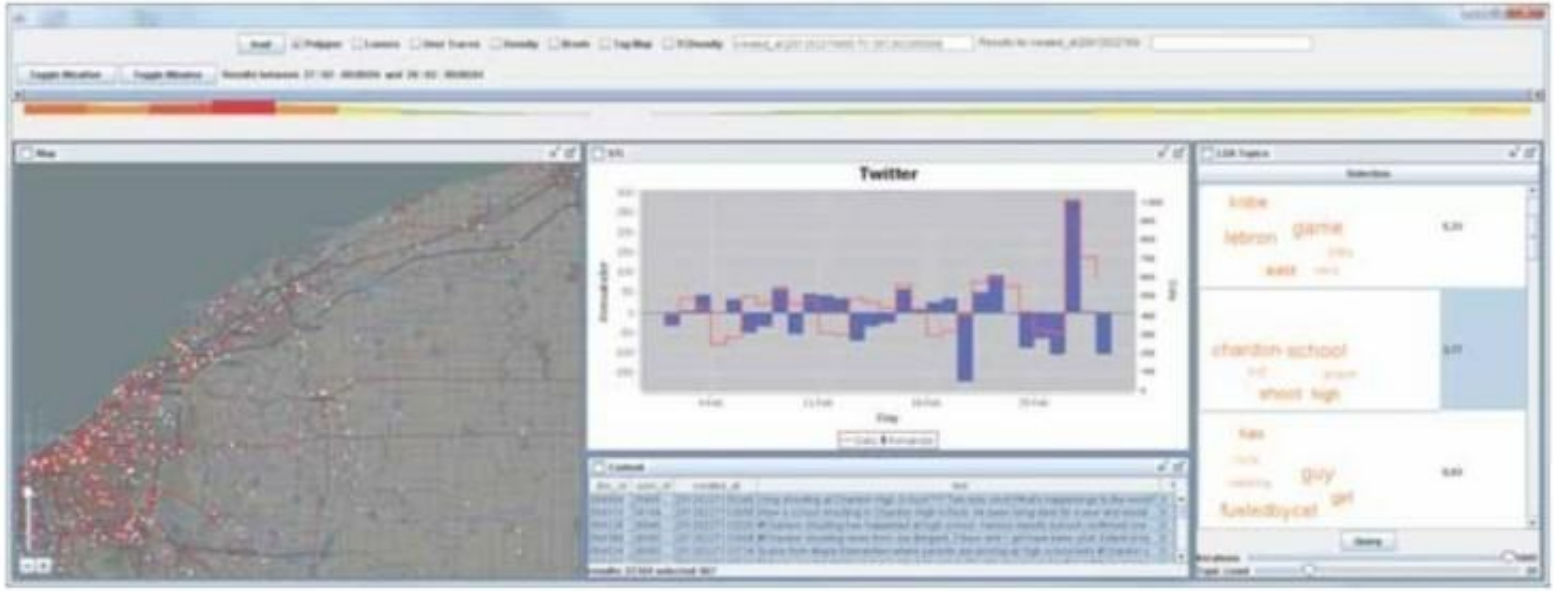
用户行为分析

- 用户在同一网站上的页面浏览顺序形成点击流。针对用户点击流进行分析，可以刻画出用户在网页间的浏览行为
  - 文献[Wei2012]使用自组织映射方法对用户点击流进行投影分析，使用色条的大小和在二维平面上的位置分布表示点击流的出现频率，以及点击流之间的相似关系。用户在投影视图上进行选中等交互操作，可以详细查看某些点击流的具体属性分布



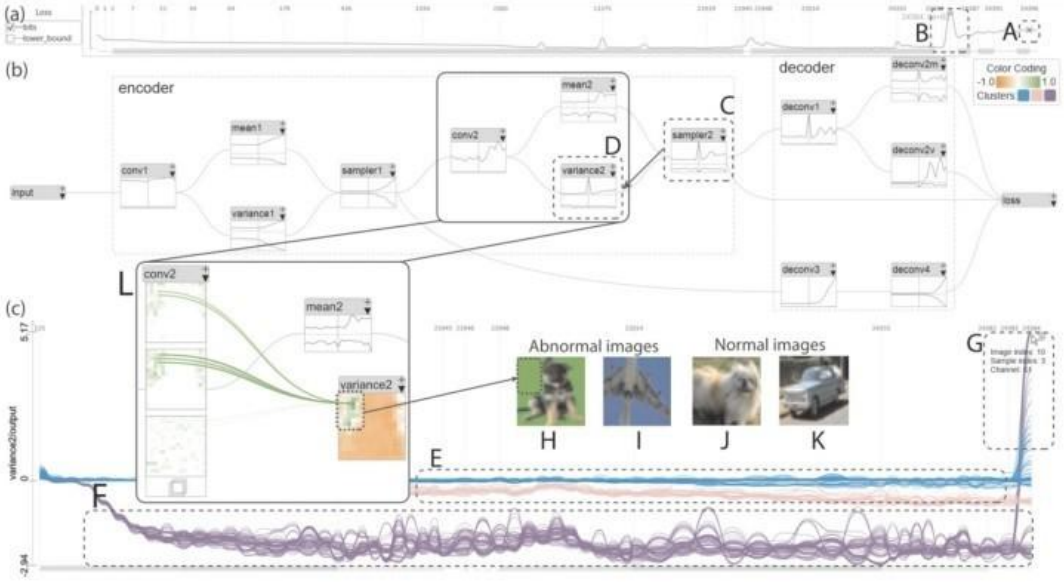
时空数据分析

- 时空数据挖掘重点关注带有时间和空间属性的数据
  - 在可视化方面，同样有很多基于时序数据和地理空间数据的方法来帮助展示数据挖掘算法的结果和算法过程
  - 针对微博或其他社交平台上带有时间戳和地点信息的文本流数据，文献[Chae2012] 基于LDA模型和STL（seasonal-trend decomposition procedure based on loess smoothing）方法分别提取某时间段内的文本主题和异常关键词，并可视化相关文本信息的地理位置和时序波动信息。用户可以通过多种交互方式渐进式探索和发现异常文本信息，并通过相关文本提取出事件



深度学习

- 研究集中在对模型所包含的信息及内涵进行解释，以及使用交互式可视化界面方便机器学习专家对模型进行调整和重新训练
  - 深度生成模型（DGM,Deep Generative Model）是近几年来深度学习方面最前沿的研究之一，其训练过程相对于其他模型较为复杂，且十分依赖使用者对深度学习算法的深入理解和实践经。DGMTracker系统[Liu2017]展示了在训练过程中模型activation在时间上的发展变化，以及哪些神经元对输出做出了贡献，或者导致训练失败



深度学习

- RNN（循环神经网络）常用于序列数据的学习和生成。RNNVis交互式可视化系统 [Ming2017]用于诊断RNN模型在训练过程中的预测性能变化。系统可适配于RNN、LSTM、GRU等模型，并以文本情感分析为主要应用场景，添加了很多文本数据分析的特殊功能
- 目前在工业实践中，深度神经网络的层数已达到数百层甚至上千层。如此多的网络层所带来的问题是模型使用者很难对整个训练过程有一个全貌，并且难以理解训练过程中参数的传播，以及训练失败的归因和改进策略分析

- 在实际使用神经网络时，如何构建网络层结构和设定参数以适配应用场景，是开发实践者常遇到的问题。DeepEyes系统 [Pezzotti2017]使用渐进式可视分析策略，帮助用户从头开始设计适合的神经网络结构，包括网络层类型、神经元个数、层次间连接方式等
- 渐进式可视分析策略的特点是为用户提供了一个渐进式迭代分析环境。用户可以从一个起始位置开始对目前的网络结构进行测试和分析，并基于测试和分析结果对已有结构进行改进，继而进入下一次分析迭代中，直至达到预定的模型构建目标。在渐进式迭代的过程中，交互式可视化界面的介入使得用户能够方便地查看和探索模型，直观理解目前模型的性能和需要改进的方向

其他应用场景

- 文献[Hossain2012]提出了结合用户反馈的聚类方法，用户可根据动物学专业知识对数据点的类别进行人为限制，以分析蝙蝠耳朵形状和声学特征
- 在高维数据分析方面，INFUSE系统 [Krause2014]借助交互式可视化方法对特征选取任务进行辅助和增强。通过展示特征在多种特征评价方法上的排名，用户可以看到特征的排名分布，以及不同的特征集合在分类算法上的测试效果，进而决定最终需要选出的特征
- 在回归分析方面，HyperMoVal系统 [Piringer2010]和文献[Muhlbacher2013]均提出了新的可视化设计，以展示回归分析中的数据分布和模型的训练结果

7 数据科学的挑战

数据科学的挑战

- 围绕数据处理的各个学科方向都开始遇到前所未有的挑战
  - 如何有效地获取数据、有效地处理数据获取的不确定性、对原始数据进行清理、分析，进而高效地完成数据存储和访问，达到去重、去粗取精的目的，是急需解决的问题
    - 同时，如何构建结构化数据与非结构化数据之间的有效关联及语义信息，存储异构、多元数据之间的关系并支持后续分析，并提供足够的性能保证，也是面临的挑战之一
  - 从大数据中获取更加有效的信息和知识，是研究重点
    - 需要考虑理论与工程算法两个方面的问题
  - 可视化作为数据科学中不可或缺的重要一环，也开始高度关注大数据带来的问题，其中包括
    - 高维数据可视化
    - 复杂、异构数据可视化
    - 针对海量数据的实时交互设计
    - 分布式协同可视化
    - 针对大数据的可视分析流程等