应用统计学系列教材 Texts in Applied Statistics

大数据分析：方法与应用

Big Data Analysis: Methods and Applications

王 星 等 编著
Wang Xing

清华大学出版社

应用统计学系列教材 Texts in Applied Statistics

大数据分析：方法与应用

Big Data Analysis: Methods and Applications

王 星 等 编著

Wang Xing

清华大学出版社

北京

内 容 简 介

本书介绍数据挖掘、统计学习和模式识别中与大数据分析相关的理论、方法及工具。理论学习的目标是使学生掌握复杂数据的分析与建模；方法学习的目标是使学生能够按照实证研究的规范和数据挖掘的步骤进行大数据研发，工具学习的目标是使学生熟练掌握一种数据分析的语言。本书内容由 10 章构成：大数据分析概述，数据挖掘流程，有指导的学习，无指导的学习，贝叶斯分类和因果学习，高维回归及变量选择，图模型，客户关系管理、社会网络分析、自然语言模型和文本挖掘。

本书可用做统计学、管理学、计算机科学等专业进行数据挖掘、机器学习、人工智能等相关课程的本科高年级、研究生教材或教学参考书。

图书在版编目(CIP)数据

大数据分析：方法与应用/王星等编著. --北京：清华大学出版社，2013
应用统计学系列教材
ISBN 978-7-302-33417-0

Ⅰ.①大…　Ⅱ.①王…　Ⅲ.①统计分析－高等学校－教材　Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字(2013)第 182786 号

责任编辑：刘　颖
封面设计：常雪影
责任校对：王淑云
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：19　　**字　　数**：459 千字
附光盘 1 张
版　　次：2013 年 9 月第 1 版　　**印　　次**：2013 年 9 月第1 次印刷
印　　数：1～3000
定　　价：39.00 元

产品编号：050949-01

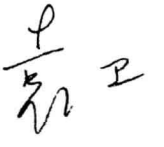
序

大数据是今天这个时代的一个符号，几乎所有的领域都在寻找着来自大数据的灵感。

今年春天，美国国家科学基金会（NSF）数学科学部组织数学、统计学相关领域专家撰写并发布了《2025 年的数学科学》报告，特别辟出一个专题来探讨大数据对数学和统计学科未来发展的驱动作用。报告中明确指出大数据给数学与统计学的发展带来了巨大的创新空间，呼吁加强方法的多样性和灵活性的研究。这些论断与 2012 年 3 月 29 日奥巴马在白宫网站上发布的《大数据研究和发展倡议》（Big Data Research and Development Initiative）遥相呼应。白宫科技政策办公室主任约翰·霍尔德伦在评论这份报告时敏锐地指出大数据的核心问题是大数据分析，他说“我们不打算强调数据本身所创造的价值。大数据的核心问题是从数据中产生新见解的能力，比如复杂关系的识别和做出越来越精准的预测。我们需要的是从数据中产生动力，获取知识和采取行动的能力”。这恰恰是本书作者想要回答的问题，作者在这本《大数据分析：方法与应用》中探讨了大数据分析的相关主题，从经典的机器学习数据挖掘方法到前沿的高维数据降维和图模型，都一一进行了深入浅出的讲解。

这是国内最早从统计学的视角阐述大数据分析问题的论著之一。书中不仅全面介绍了各种数据（噪声数据、不平衡数据、高维数据、文本数据、网络数据等）的建模方法、数据分析过程和评价方法，并将这些技术结合市场研究、知识发现、疾病诊断和舆情分析等应用领域进行了案例讨论。本书的独特之处在于以大数据分析为主线将统计学、计算机、机器学习、社会网络分析、市场营销等多学科的专业知识汇于一册，揭示了大数据研究所蕴含的跨界合作、协同创新的特质。本书为方兴未艾的大数据研究提供了及时的理论和方法支持，是一本适应科技创新发展形势，满足高层次、应用型、复合型人才培养需求的创新型教材。

总之，这不仅是一本大数据分析领域的优秀教材，而且也是一本各个领域了解大数据和大数据分析的理想参考书。



2013 年 6 月

前　言

信息技术推动了大众对数据的消费，大众对数据的消费热点经历了一个明晰的轨迹，20 世纪 80 年代是数学热，数字产生于数学模型，但数学模型对带有观测误差数据的解读能力有限，20 世纪 90 年代是信息热，信息为数字披上了外衣，然而技术的计算代价、适应能力和容错能力等还缺乏一个统一的分析标准。结果从 20 世纪 90 年代开始，统计开始成为大众消费数据的热点，这一消费的转变也将一度默默无闻、与世无争的统计学家从象牙塔带到真实世界，开始参与到从数据特点出发构建面向不同问题的统计模型的实践中来。在当今这个网络密布、数据激增的时代，统计建模为大数据分析提供了一套可扩展、可深化，并能高质高效地揭示有价值信息的方法，使透过微观数据视角洞察在“无尺度网络”中游走的人类行为成为可能。大数据分析方法已经在信用识别、垃圾过滤、过度开发、诱惑欺诈、轨迹寻踪等应用研究中显露手脚，其潜在的能量与应用前景无疑有着更为广阔的空间。

与传统的统计分析相比，大数据有着来源复杂、体量巨大、价值潜伏等特点，这使得大数据分析必然要依托计算机技术予以实现。这也逐渐演变出大数据分析的两个研究方向：第一个方向侧重于数据的处理与表示，主要强调采集、存取、加工和可视化数据的方法；第二个方向则研究数据的统计规律，侧重于对微观数据本质特征的提取和模式发现。经过多年的实践探索，业界已经越来越清晰的意识到只有在两个方向上的协同、均衡推进，才能保障大数据应用的稳健成长和可持续发展。因此，大数据分析的发展重心也逐渐由数据处理的技术向数据分析的科学倾斜，后者正是本书的焦点与重点。

相应的，我们所指的大数据分析方法主要取材于统计学习（Statistical Learning）、数据挖掘（Data Mining）和模式识别（Pattern Recognization）等领域，这些内容安排在第 3 章、第 4 章、第 5 章、第 6 章和第 7 章。第 2 章着重介绍数据挖掘流程与数据处理技术。大数据分析还是一门与应用结合很强的课程，我们精心挑选了三类典型的应用模型，内容安排在第 8 章、第 9 章和第 10 章。本书集方法与应用于一册，希望读者通过方法的学习掌握复杂数据的分析与建模；通过应用的学习能按照实证研究的规范和数据挖掘的流程开展大数据的研发。除此之外，大数据分析还有很强的实践性，为体现这一特点，我们强调了工具的作用。通过工具的学习希望读者能够熟练掌握一门数据分析语言。本书大部分方法将给出 R 软件的示范程序，R 软件是免费、开源、专业、前沿的统计分析软件，分析研究数据的功能强大，是实践和领会大数据建模的有效途径。另外，书中也使用了少量的 JMP 和 Statistica 等工具的分析结果。

本书既可用做培养应用统计专业硕士的教材，也适用于管理学、信息学、统计学等专业进行数据挖掘、机器学习、人工智能等相关领域的教学与研究。研究生或本科高年级的数据挖掘课程可通过基本原理的学习，了解不同的模型和算法的设计特点，并通过每章后面所列参考文献进行延伸阅读。

本书通过案例讲解算法，以提高读者实际解决问题的能力。书中的案例也可用做提高学生统计咨询能力的课堂训练。在习题练习中的一些题目可作为课堂案例，安排学生分组讨论，并鼓励学生演示分析思路和分享分析收获。使学生有机会诊断问题，并学会选用适当的方法和技术分析数据。通过案例教学的方式将对学生领会大数据分析方法和应用大有助益。

如上所述，本书内容由 10 章构成：大数据分析概述，数据挖掘流程，有指导的学习，无指导的学习，贝叶斯分类和因果学习，高维回归及变量选择，图模型，客户关系管理，社会网络分析，自然语言模型和文本挖掘。教学内容建议一学期 54 学时完成，其中至少应该安排 10 学时用于大数据分析项目的上机实验和讨论。

作者过去 6 年中一直在给高年级本科生和研究生讲授数据挖掘与机器学习课程，本书是作者结合多年授课的讲义与课题研究成果基础上汇编而成。全书由王星策划、统稿和校阅，其中第 1 章至第 5 章由王星主笔。贺诗源同学主要参与了第 2 章、第 6 章和第 7 章的部分编写工作，陈文同学主要参与了第 6 章和第 8 章的部分编写工作，以上两位同学还在软件实现和例题整理部分做出贡献；郑轶、李荣明、龚君泰、马璇、李沐雨对第 8 章至第 10 章做出贡献；彭非老师、张波、邱逸轩、颜娅婷、王晓航、王杰彪、陈之进和张望等同学参与了部分实验的讨论；特别感谢 SAS 软件 JMP 事业部曹建博士、周畊等在软件和相关资料方面给予的大力支持和技术解惑，他们还提供了可供学生免费试用的版本和网址（具体方法列在光盘中）；清华大学出版社责任编辑刘颖和他的同事们尽职尽责的努力，在此一并致以衷心的谢忱。写作本书是一个愉快的过程，在这个新的科研方向和应用领域上，这支由年轻人组成的团队激情澎湃、勇于探索，他们钻研探究的精神风貌为我留下诸多美好回忆，也凝聚了开拓未来前进的不竭动力。大数据分析方法和应用涉猎很广，很难一本书面面俱到，书中尚存不详不妥之处，敬请读者指正。

王 星
中国人民大学应用统计科学研究中心
中国人民大学统计学院
2013 年 7 月

目　录

第 1 章　大数据分析概述

本章内容

- □ 大数据基本概念
- □ 数据挖掘的产生与功能
- □ 数据挖掘与相关学科的关系
- □ 大数据研究方法

1.1　大数据概述

1.1.1　什么是大数据

20 世纪 90 年代后期，以信息技术、计算机和网络技术等高新技术发展为标志，人类社会迅速迈进一个崭新的数字时代。现代信息技术铺设了一条广阔的数据传输道路，将人类的感官延伸到广袤的世界中。政府和企业通过大力发展信息平台和网络建设，改善了对信息的交互、存储和管理的效率，从而提升了信息服务的水平；生物科学领域通过对分子基因数据的解读重新诠释了生物体中细胞、组织、器官的生理、病理、药理的变化过程，从而突破了人类在许多疑难杂症上的传统认识；市场研究人员通过谷歌住房搜索量的变化对住房市场趋势进行预测，已明显比不动产经济学家的预测更为准确也更有效率；手机、互联网、物联网，这些先进的信息传输平台，在生成-传播着大量数据的同时，也越来越多的改善了人们的生活。总之，政府、科学和社会等各个领域的每个细胞，都被快速发展的信息技术激活，畅游于信息海洋并获得认知效率的飞跃，沉浸于价值被认可的幸福与满足中。

精彩纷呈的数据也带来了利用数据的烦恼。日新月异的应用背后是数据量爆炸式增长带来的大数据分析的挑战，2012 年 3 月 30 日美国国家卫生研究院宣布世界最大的遗传变异研究数据集——国际千人基因组项目（截至目前数据已约达 200 TB），数据量正在由太字节（TB=10^{12}B）向拍字节 PB（=10^{15}B）、艾字节 EB（=10^{18}B）、泽字节 ZB（=10^{21}B）甚至尧字节 YB（=10^{24}B）升级，估计每两年就会增长三倍。

大数据是一个新概念，英文中至少有三种名称：大数据（big data），大尺度数据（big scale data）和大规模数据（massive data），尚未形成统一定义，维基百科、数据科学家、研究机构和 IT 业界都曾经使用过大数据的概念，一致认为大数据具有四个基本特征：数据体量巨大；价值密度低；来源广泛，特征多样；增涨速度快。业界称为 4V 特征，取自 volume, value, variety 和 velocity 四个英文单词的首字母。由此可见，大数据的核心问题是如何在种类繁多、数量庞大的数据中快速获取有价值的信息。一方面，这种信息获取能力离不开优化的复杂大规

模数据处理技术。另一方面是模式提取的程序、标准和规范。比如随着社交网络、语义 Web、云计算、生物信息网络、物联网等新兴应用的快速增长，在经济学、生物学和商务等众多领域中出现了成组数据、面板数据、空间数据、高维数据、多响应变量数据以及网络层次数据等结构复杂的数据形态，迫切需要强大的数据处理能力以实现批量信息的生产。而这种能力的一个关键问题是：对亿万个顶点级别的大规模数据进行高效分析的模型是什么？大数据不仅数据类型复杂，更重要的是数据中模式结构复杂，信噪比较低。优质数据与劣质信息的鉴别、操作便捷与垃圾信息有效过滤的平衡设计，信用危机的识别要素、稀有信息的发现、精准需求定位等问题更加突出。在数据泛滥的情况下，有价值的信息被淹没在巨大的数据海洋之中，有价值的见解和知识很难发现。而数据分析逻辑和规范的缺失必然导致垃圾信息和乱象丛生的信息环境。大数据认知在社会分析、科学发现和商业决策中的作用越来越重要。揭示数据背后的客观规律，识别信息的价值，评估信息之间的影响是合理开发数据资源和改善人类活动的重要组成部分。大数据技术已经成为科技大国的重要发展战略。数据与能源、货币一样，已成为一个国家的公共资源，金融市场上有“劣币驱逐良币”，能源开发中“并非缺乏能源，而是缺乏清洁能源”，数据的管理和再利用技术不能取代科学，在数据的结构与功能越来越复杂的客观现实面前，需要更多角度的模式探测和更可靠的模型构建，无论是运用模型生成规则还是运用结果都需要更规范的设计与分析。

系统分析方法是传统数据建模方法，在大数据分析建模设计中大有作为，然而大数据建模更为复杂，有两个鲜明的特色，首先模型不是主观设定的或普适性的，而是具体的，从数据的内部逻辑和外部关联中根据问题的需要梳理出来的。在这个过程中，基于无形数据的有形模式的探索、比较、估计、识别、确认、解释不可或缺。这在高性能计算领域的算法研究和开发中尤其迫切。在这些研究中，模型常常并非现成的，数据与模型的简单组合拼装并不总是能够切中要害。复杂问题的数据获取，大规模数据的组织、处理，模型与算法、理性决策、数据的展现方式等，都会影响到最终输出模式和结果的可用性。第二，强调建模过程中模式的变化和复杂的关系，因为数据的脉络和联系正是通过建模过程的模式发展而一一剖析出来的。数据的分布、数据的特征、数据的结构、数据的功能、数据的运动、数据在时空中的变化轨迹、数据的影响层次、不同数据变化层次之间的关系是统计科学的核心内容。总之，数据建模既不是统计理论的简单照搬，也不等同于数据的自动加工，建模的意义是更好地理解数据，增加洞见。于是，数据建模与算法技术联合，成为大数据深度认知的关键。

1.1.2 数据、信息与认知

大数据分析里的第一个问题是要明确分析的对象——即数据的概念。什么是数据呢？数据有哪些功能呢？从表象来看，数据可以理解为人类对所感兴趣的对象特性的记录，数据是用于描述事实的，它具有时间和空间属性。数据的一项重要的功能是对所立目标形成深刻理解，提供未成形概念存在的依据。其中这个未知的概念既存在于数据之中，又与数据本身有所区别，这就是新的知识。1994 年日本学者 Nonaka[17]等从人类理解与学习认知的角度给出知识的定义：知识是概念的诠释和表达，数据是揭示知识存在的模式与关系的重要素材。单一的数据记录一般并不独立形成概念，为了产生有价值的、可靠的新认知，需要将不同记录的数据进行有效的关联和组织，通过数据分析，把握体现数据共性和差异的

关键线索，从而对在数据中的信息进行有序解读，实现对稳藏于数据中的知识的线索和联系的归纳与推理。没有数据则无法形成可靠的认识。

从知识形成过程的复杂性来看，知识可以分为显性知识（explicit knowledge）和隐性知识（tacit knowledge）。与显性知识相对应的是显数据，显数据是指按照某种规律或理论通过测量能够得到的数据，用以描述观察到的现象和对概念做出量化描述。比如植物叶子的颜色、疾病的血相特征、贫困的地理分布、事件的时间发展、网民参与社交媒件的程度等。在这类问题中，显知识常用参数表示，显数据是对参数的个体、部分或整体的观测。再比如：某一课题采用中国 51 个城市的居民微观调查数据，以与政府管制相关的企业娱乐和旅游花费来度量各城市的腐败水平，定量评估腐败对中国居民幸福感的影响。这类问题中使用哪些数据和哪些测量指标形成知识是预先确定的，数据的作用是客观真实地估计出整体的影响强度。除此之外，许多知识是不可直接量化获得的，其中又分为一部分可直接测量，另一部分无法直接测量，也有完全不能直接测量的问题。对于无法直接测量的知识，则需要通过模型辅助推断。用于未知概念推理建模的数据称为隐性数据，隐性数据的主要作用是揭示隐性知识成立的可靠依据。比如：区分两类植物的关键要素、用于疾病诊断的基本症候、贫困的成因、两个异性成年人经过交往是否能够组建家庭等问题。这类问题的特点是概念构成因素多样化、内外影响机制不确定等，常常涉及不同因素或群体之间的相互影响作用关系的发现和关系变化规律的揭示。例如，北京市北三环西向东每日早晚高峰期间桥面拥堵状况的智能预测就是一个典型的难于直接测得的问题，这个问题的关键是交通状态自动识别模型，用于建立模型的数据可以有几种选择，比如固定交通监视器的速度数据和车载 GPS 传递的车速数据，这些数据可以帮助建立速度预测模型。更进一步还需要考虑偶发拥堵和常规拥堵的区别，这两类又分别与相关路段的故障车辆数、周边教育机构的分布及天气情况有关系，这显然是一个复杂的建模问题，涉及很多变量和复杂的数据类型。再比如，新近的一项科学研究指出，科学家成功研究发现“贪食基因”，该基因的存在能够导致人即使在饱腹状态下也能吃更多的食物。科学家指出，通过抑制该基因可以有效地治疗人体肥胖现象，支持这个结论的理想数据是一组参加试验的肥胖人群食物摄入数据，以及服用抑制“贪食基因”药物前后体重变化的动态跟踪试验数据，这个实验设计比较复杂，成功的关键是如何实现双盲（double blind）设计，通过尝试有效的分销管理却有可能获得支持研究且质量不错的观察数据。再比如消费行为研究中指出：消费水平较高的人主要关注投资，消费水平较低的人关注储蓄，消费水平对于存款的影响构成了公允投资定价的法则，而这一理论到底在多大范围适用，还需要数据进一步验证，有人通过网上银行直接关联两部分数据，总结出理论成立的人群特征。在社会学研究中，观察指出人们预期在有往来的两个人之间建立恒定的友谊关系，而不会在一人对另一人的单向关系中存在友谊，这个理论在实际中如何求证？这个用于形成社会组织方向关系的认知如何衡量？

总而言之，许多问题的回答需要在显性数据的基础上形成稳定的隐性数据。今天许多存储于数据库中的大数据主要实现了事实的描述性功能，但其分析潜力没有得到深度开发。复杂的问题中，无论是已知概念的统计描述还是未知概念的统计推断常常同时被需要，显性数据和隐性数据都是不可或缺的。值得注意的是，以上侧重于从知识形成的复杂性上将数据分成显性数据和隐性数据，是一种逻辑上的区分而不必事先截然分开。比如年龄是贫困人口的重要特征表示，也可以是贫困成因分析中的一个重要变量。另一方面，降雨量在

形成地区气候概念中是一个重要的数据，但对决定某篮球俱乐部是否盈利则作用微乎其微，不能用作显性数据。显性数据由于测量上的问题，常常需要增加辅助数据进行模型推断，隐性数据所构建的概念往往也需要描述性数据给予必要的解释。有的数据兼具两种知识发现功能，不仅可以反映概念特性，而且也蕴含着不同群体的特征规律。例如，心脏病患者的饮酒习惯，既是区分心脏病患者和其他患者的识别变量，也可作为医疗诊断的识别观测变量。大量待分析的规律隐藏于数据之下，必须经过科学的辨识和分析方能得以提炼，成为有别于原始数据可判别是非和预测的可靠依据。

数据不仅在认知过程中的功能不同，在对认知的理解上也有不同，这就需要对知识进行解释的数据，一般将其称为数据的语义。语义是对数据符号的解释，数据的含义就是语义。对于信息集成领域来说，数据往往是通过模式来组织，数据的访问也是通过作用于模式来获得的，这时语义就是指模式元素的含义，例如类、属性、约束等。与语义相近的另外一个概念是语法，语法是模式元素的结构，定义符号之间的组织规则和结构关系。对数据进行统一的比较和分析可以产生新的语义实体，认知同样也依赖于表示数据涵义的语义。例如在学生档案中存在着这样一条数据（王丽，女，22，1990，四川绵阳，统计学院，2008），对于这条学生记录，结合其数据含义，可以产生如下信息：王丽是位四川籍大学生，1990年出生，2008年考入统计学院。这就是这组数据的语义。在信息社会里，信息被使用就产生了价值，信息的价值随着所分析的目的不同而有所不同。比如统计了与王丽入学时间和原籍都是一样的 30 名学生，那么这个数据如果对应于地震灾区重建所需的委托培养人才库，其信息的价值就不可低估。地震灾区重建人才需求特征与其他地区的人才特征的区别则是语法分析中的核心内容。一般而言，单个信息的价值是不高的，多个信息组织在一起进行比较分析研究，可以提升信息的价值。

数据与其语义共同构成了具有时效性的、有特定含义的、有逻辑的数据，这就是信息。如果说数据是客观事物的一种符号，那么信息则可以认为是以有意义的形式加以排列和处理的数据。例如，政府通过个人贷款购买住房的数据，可以得出某城市当年贷款购买住房的总账记录，经格式化后，总账记录以贷款购房表的形式呈现出来，就可以认为是一个信息。一般而言，信息是结构化的数据，数据则不必是结构化的，非结构化或半结构化的数据可以通过结构化程序变得易于处理和分析。数据和信息既有联系又有区别。数据是信息的素材，又反过来表达了信息，信息为数据形成知识提供架构，数据是信息的内容。信息则只有通过数据的形式表示出来才能被人们用于比较、理解和接受。尽管数据和信息在概念上有所不同，从数据和信息的分析和利用来看，二者并不具有严格的区分，在不影响到理解的情况下，数据分析和信息分析很多情况下都被认为是一个概念，其共同的功能是形成有效的知识。数据本身并不自动生成认知，而是需要背景、模式和架构的分析，已经形成的认知也需要新的数据被不断地验证和调优。虽然在许多场合中，对个案的单一观测也被称为知识，但需要区分的是：仅仅收集这些观测本身并不必然形成知识，它提供的是用于加工知识的素材——数据。在我国台湾，人们甚至使用“资料”表示数据，以示其作为知识加工素材的基础作用。数据是知识的载体，数据挖掘是从复杂数据中产生认知的方法、原则和过程。

由于信息与通过结构化数据所定义的有意义的主题紧密相连，随着这些主题的时间效用失效后，信息本身的价值往往也会随之衰减，只有人们通过对信息按照新的主旨进行重

整、归纳，比较和演绎，使其有价值的部分沉淀下来，并与已存在的人类知识体系相结合，这部分有价值的信息才会经历重生，实现价值的飞跃。例如，某地，某年 6 月 30 日，最高气温为 37 摄氏度。当年 12 月 5 日最高气温为 3 摄氏度。这些信息一般会在时效性消失后，失去被直接使用的价值。但当人们收集几年甚至几百年的气温变化信息进行归纳和对比时，就会发现此地每年 7 月气温会比较高，12 月气温比较低，于是总结出全年的气温变化规律。虽然季节概念形成的时间已无从考证，但新的知识以数据的形式再次被记录。在这个例子中，作为短期预报的信息具有直接价值，其间接价值是可以辅助人们做长期的规律分析。例如，50 年内强地震前后气温的变化规律与正常相比是怎样的，显然研究这一课题需要气温数据和对数据的相应的分析和新的处理。再比如：大学生本科成绩信息的直接价值是可作为评估学生专业水平的依据，但不容易对学生本科毕业后的就业情况做出预测。学生如果不能正常就业，从学生的学业成绩中追查出一些原因也是有可能的，因为成绩和学生就业中的专业要求存在一定的关联，这也体现了数据的间接价值。

计算机普及以前，人们关注的信息问题是客观事物的特征记录，数据特征的采集、创建、检测、简约、合成、编码、存储、发布、检索、提取、重建、概念、判断、问题解决和服务等，当时通过特征观察形成认知的过程主要是人脑的主观思维和手动完成的。信息时代，计算机采集、检测、提取、更新等技术的发展扩展了数据的传播范围，其中一项突出的贡献是丰富了数据的存储格式，其结果是数据具有了多种表现形式。常见的如文件、报告、资料、数字、音频、语言、图形、视频、Web 页面等。形式多样的大规模的数据不仅激发了人们开展富有创造性的数据分析实践活动，而且推动了从数据中发现新价值规律这一科学认知过程的设计和实践。这些活动包括对数据收集、分类、概括、组织、分析和解释的工具、算法和建模的研究。

今天，数据的另一个挑战是能够被结构化的数据是非常有限的，传统的关系数据库管理的结构化数据仅占数据信息总量的 15%。有统计显示，全世界结构化数据增长率每年大概是 32%，而非结构化数据增长则是 63%。截至 2012 年，非结构化数据将占互联网整个数据量的 85%以上，难以直接通过数据库进行有效的管理。用于形成智能的大数据，往往是非结构化数据，应用范围从企业信息化、媒体出版到垂直搜索、数字图书馆、电子商务等各个领域。未来的数据分析技术将向来源的异构化、应用的标准化、建模的流程化、表达的精炼化方向发展，并在面向对象、跨媒体数据、并行计算、分布式文件系统、异构数据的结合等领域展开更为深入的研究。

1.1.3 数据管理与数据库

从有文字记载开始，人类对自然和社会认识的进程就开始加快。认识提速的关键一步是对数据实施管理，即科学地组织和存储数据。20 世纪 30 年代，随着大工业生产和数据计算的需要，数据管理逐渐发展起来成为按照需要加工数据的一种技术。数据管理的核心问题是对数据实现分类、组织、编码、存储、检索和维护等任务。利用信息技术管理数据是近半个世纪以来的新鲜事物，数据管理技术经历了人工管理、文件管理和数据管理三个阶段：

20 世纪 50 年代以前称为早期的数据管理技术阶段，计算机的主要作用是科学计算。当时存储数据的工具只有纸带、卡片、磁带，没有磁盘等直接存储器，对数据的管理方式是人工管理，人工管理阶段的主要数据处理特点是：数据没有保存、应用程序独立、数据不

共享和数据处理不独立等特点。很多业务管理是传统的纸质文件管理方式，这些文件不会长期保存，当有课题时将数据输入，用完就撤走。每一个不同的业务问题有相应的数据格式、应用程序、逻辑关系和存取方法。由于不同的文件具有独立的语义和逻辑，所以无法相互利用和互相参照。因此程序和程序之间，数据和数据之间都存有大量的冗余。

20世纪50年代至60年代，数据存储技术取得巨大进步，硬件方面有磁盘和硒鼓等直接存储设备，软件方面出现了文件系统作为统一的数据管理系统，实现了对数据的系统性联机实时处理。用操作系统管理数据具有数据可以长期储存、数据的文件系统统一管理、数据共享性差、数据冗余等特点。这个时期虽然有了统一的文件管理系统负责管理数据，由于文件是面向应用的，于是即便两个相同的文件针对不同的应用，也需要重复存储，各自管理，这造成了很大的数据冗余，不利于数据的一致性，数据版本更新和维护困难。这个时期文件管理的数据管理技术不对数据和信息的意义进行新的创造。

20世纪70年代以后，随着计算机参与社会管理的进程加快，为编制和维护系统软件和应用程序，所需要的成本相对增加，在处理方式上出现了对更多联机实时处理的需要和分布式管理的需要。于是专门负责数据管理的系统逐渐从文件系统中独立出来，形成数据库管理系统。数据库管理系统管理的数据具有高度的结构化、数据独立性高、冗余度低、数据由数据库管理系统统一管理和控制等特点，这些特点对于提高生产速度、增强准确性和降低成本方面起到了关键作用，有效地提高了生产力。数据库管理系统开始成为改善服务、共享信息和提高质量的支持平台，20世纪90年代，对数据库系统的要求已不再局限于快速、准确、低成本地处理数据，而是希望其在缩短空间和时间，增强系统的记忆性，联系组织、客户、供应商，促进业务流程优化等四方面有所作为，这四方面的要求是传统的数据库和联机事务性处理所不能企及的。为满足企业包括资源计划（ERP）、客户关系管理（CRM）、门户网站（EWP）和信息门户平台（EIP）以及内部网（intranet）在内的应用系统的一致性，底层的统一物理存储独立出来，其作用就是对企业数据实施统一调配和管理。传统的基于业务的系统虽然在了解业务流程中取得极大成功，但在辅助决策时却产生了极大的困难。传统的数据库管理与辅助决策不相适应性主要体现在以下四个方面：

1. 数据处理效率和质量

传统的数据库系统多用于事务性处理问题，如MIS（Management Information System）和OLTP（Online Transaction Processing），主要的特点是支持大量、简单、可重复使用的例行短事务处理，如插入、查询、修改和更新记录等服务，这些操作频率高，处理时间短，分时使用系统资源。在分析处理中，用户对系统和数据库的要求有新的需要，分析的特点是按主题编制、访问大量数据和处理复杂查询的长事务为主，遍历数据造成大量系统资源被消耗。

2. 数据访问和数据集成

在商业层面进行决策分析时，需要全面集成的数据，这些数据不仅包含企业内部各个部门的相关数据，而且也包含企业外部的甚至企业之间的情报数据。决策者所需的数据也不再局限于本部门、本企业，而是分布异构的多渠道数据源，如商业领域竞争对手的Web数据库、文件系统、HTML等非数据库系统等。现实中很多数据真实的存在状态是分散而非集成的，缺乏面向新主题的统一编码，数据的格式不统一。如果将这些集成问题交给决策

系统程序解决，将极大地增加决策分析系统的负担，造成系统执行时间过长，极大降低系统的性能。为实现不同来源、格式、特性的数据在逻辑和物理上有机地集成，已经有一些成熟的集成模型。例如，联邦数据库系统、中间件模式和数据仓库模型，其技术核心是解决数据源语义的统一管理，以实现高效的统一访问。

3. 数据操作和数据分析

对数据库的操作方式上，业务处理系统的关键问题是确保数据一致性和功能稳定性，于是其主要支持多事务并行处理，加锁和日志并行控制和恢复机制，而在数据访问操作方面提供开放的权限是有限的，而数据挖掘人员则往往需要运用各种工具对数据的整体进行多种形式的统一操作，并希望将数据结果以商务智能的方式表达出来。于是，数据仓库与业务数据库分离是目前数据挖掘设计中的通常做法。联机分析处理强调与决策者的交互、快速响应以及多维可视化界面。但其分析是浅层的，传统所提供的标准化的报表方式业务处理提炼信息的内涵，在形式上和内容上很难满足决策管理的需要。

4. 数据的时限

一般情况下，数据库中只存储短期数据，不同数据的保存期限很不一样，即使一些历史数据被保存下来，但也经常被束之高阁，未能得到充分利用。而对于决策而言，决策环境是动态的，历史数据非常重要，许多分析结果有赖于大量宝贵的历史数据，存储历史数据，对历史数据进行有效说明的元数据都是决策数据所需要的基本条件。

1.1.4 数据仓库

今天大部分的数据库都是围绕着单一业务功能而展开的，综合分析能力较弱。基于复杂数据的知识发现，常常需要一个有结构的体系。其中至少包括四个基本的结构：系统、环境、步骤和主题。知识需要面向主题的表示，支持主题表示的新的数据结构就是数据仓库，美国著名信息工程专家 William H. Inmon 于 20 世纪 90 年代初在其著作《Building the Data Warehouse》提出了数据仓库概念的一个表述，认为数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于决策支持的知识管理。比如淘宝的商业数据库主要围绕着支付业务展开，但每一笔交易带来的对库存和分销的管理则需要借助数据仓库的周密计算进行规划和安排。数据仓库系统汇聚了淘宝几乎所有的商业数据，这些记录包括用户的访问路径、交易过程的海量数据。通过数据仓库的清洗、整理、过滤、排序等技术手段，这些海量数据能够产生具有商业价值的业务信息，并生成反映最新市场现状的统计分析数据报表。淘宝数据仓库将用户行为模式与最新的交易结合，为用户提供精准的个性化服务。用户使用数据仓库进行决策时所关心的重点内容构成了一些分析主题，如收入、客户、销售渠道等；数据仓库内的信息不是像业务支撑系统那样是按照业务功能进行组织的，而是根据分析主题进行组织的，称为面向主题的数据库。数据仓库中的集成性指信息不是从各个业务系统中简单抽取出来的，而是经过一系列加工、整理和汇总的过程，因此数据仓库中的信息是关于整个企业一致的全局信息。随时间变化体现在数据仓库内的信息并不只是反映一个组织当前的状态，而是记录了从过去某一时点到当前各个阶段的信息。通过这些信息，可以对企业的发展历程和未来趋势做出定量分析和预测。