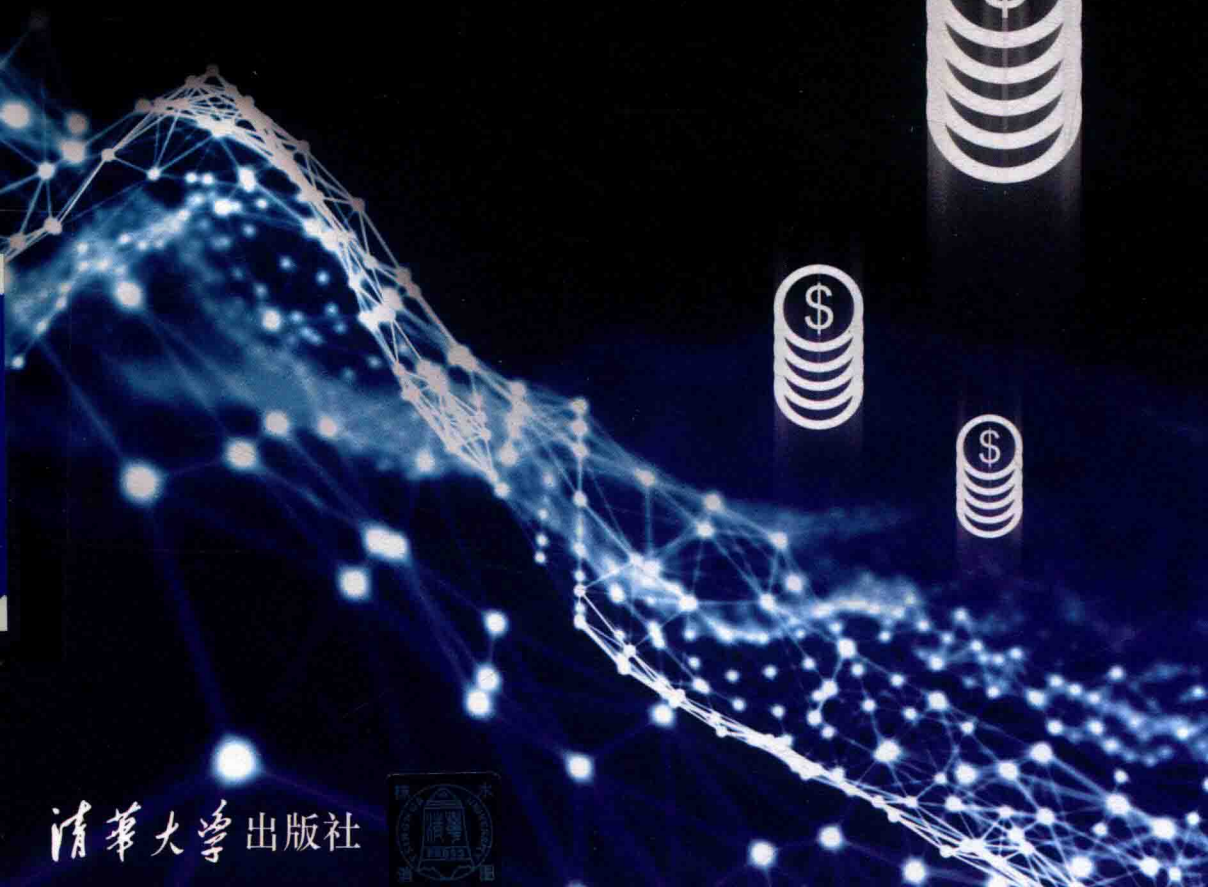
大数据金融

BIG DATA FINANCE

刘晓星 著

万物皆互联　无处不数据　大数据金融时代已悄然来临



清華大學出版社

BIG DATA FINANCE

大数据金融

刘晓星 著

清華大學出版社
北京

内 容 简 介

本书基于作者近几年来在大数据金融领域的独特观点和系列研究成果，着重介绍了大数据的提出与演化及大数据思维，并从大数据与金融融合、大数据金融的商业模式、大数据金融机构与产品创新、大数据金融服务平台创新、大数据金融算法、大数据金融生态环境建设、Fintech 与大数据金融等多个方面对大数据金融进行了深入研究和展望。

本书适合从事金融科技、数据挖掘、大数据金融等领域的研究人员以及金融机构和政府相关管理决策部门的从业人员阅读使用，同时也适合高等院校经济、金融、统计、管理等专业领域的教师、研究生阅读参考。

图书在版编目(CIP)数据

大数据金融 / 刘晓星著. — 北京：清华大学出版社，2018
ISBN 978-7-302-51611-8

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①金融－数据管理－研究 Ⅳ. ①F830.41

中国版本图书馆 CIP 数据核字(2018)第 247137 号

责任编辑：陆浥晨
封面设计：李召霞
版式设计：方加青
责任校对：宋玉莲
责任印制：董 瑾

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：三河市龙大印装有限公司
经 销：全国新华书店
开 本：170mm×240mm 印 张：24.25 字 数：433 千字
版 次：2018 年 11 月第 1 版 印 次：2018 年 11 月第 1 次印刷
定 价：79.00 元

产品编号：073204-01

前言

万物皆互联，无处不数据。21世纪的今天，随着移动互联网、云计算、物联网、区块链、人工智能、量子计算等大数据技术发展的日新月异，人类已经步入“大数据时代”，2012年被公认为我国大数据发展的历史元年，数据的大规模生产、分享和应用已成为现实。在“大数据时代”，大数据作为一种新型资产，与资本、劳动、技术、土地等生产要素一起推动着经济社会向前快速发展。

大数据的发展有其特定的社会历史背景，人类测量、记录和分析世界的渴望是推动大数据发展的核心动力，是人类社会发展到一定阶段的必然产物，顺应了时代的发展潮流。当文字、图像、音频，甚至世间万物都可转变成数据，一切都可量化时，大数据就能创造出巨大的新型价值，渗透并服务于人类生产生活的方方面面。至于大数据的重要性，赵国栋等在《大数据时代的历史机遇：产业变革与数据科学》中指出，“大数据时代公司的价值，与其数据资产的规模、活

性成正比，与其解释、运用数据的能力成正比”，甚至认为“缺乏数据资源，无以谈产业；缺乏数据思维，无以言未来”。为了更好地理解和融入大数据时代，突破传统思维模式的桎梏，形成适应大数据时代的思维方式是一种必然选择。大数据改变着我们理解世界的方式，促使我们由热衷于寻找事务间的因果关系转向寻找相关关系，从微观和宏观两个层面深入了解事物的本质。通过探求“是什么”而不是“为什么”，我们正以全新的视角更好地理解和审视这个世界。

大数据与金融的深度融合是大数据时代的一个重要发展趋势。移动支付、P2P 网贷产品、互联网基金、众筹、智能、投顾、互联网保险等互联网金融新业态的蓬勃发展，一方面挑战着传统金融机构的垄断地位，另一方面重新塑造了一种“以客户为中心、满足客户消费体验”的新型金融服务模式。大数据挖掘和分析技术的不断更新发展使海量非结构化金融数据的有效利用成为可能，通过对金融数据的多维实时分析和挖掘，可以为互联网金融机构提供客户的全方位信息，包括客户的消费习惯、资产负债、流动性状态、信用变化等，为其准确预测客户行为奠定了数据基础。这些历史性变革有助于金融机构加快业务和产品创新，实现精准营销和加强风险管控，促使企业数据资产向战略资产转化。虽然目前金融业还处于大数据应用的初级阶段，但我们有理由深信在不远的未来，大数据将成为开展金融业务的基础性资源和非常关键的金融构成要素。随处可见、随时可得的移动金融客户端和个性化金融产品正在向我们传递一个重要信息——大数据金融时代已悄然来临。大数据金融正在改变我们的生活方式，推动着金融产业变革和商业模式创新，创造出更多的数据金融价值。

本书共 11 章，主要研究内容如下。

第一章是大数据的提出与演化。从大数据概念提出的背景出发，按时间顺序梳理大数据概念的演变过程，阐释了大数据的主要发展阶段，以及国内外大数据的发展现状。从理论技术层面和实践层面对当前的大数据概念进行理论界定。分析了大数据带来产业变化的关键节点、六大趋势、五大颠覆领域、三大关键行业以及数据交易变革，揭示了数据市场的未来发展趋势，大数据将带来变革时代的力量以及未来可能面临的潜在挑战。

第二章是大数据思维。主要论述了大数据思维方式对传统思维惯性的冲击，具体分析了大数据思维带来的六大转变及相关原理，分析了大数据思维方式冲击下，社会、政府、企业、个人的传统思维发生的变化，引发的全新大数据思维方式，对传统行业和国家战略未来发展的影响。

第三章是大数据与金融的融合。这两者的相互融合是行业发展的必然趋势，

主要有以下三个方面的动因：金融行业应用大数据的优势、大数据技术的日臻成熟和金融业创新发展的必然要求。本章从大数据金融的特征切入，分析了大数据金融的四个基本特征，即数字化、开放性、高生产力和科学决策，展望了大数据金融应用的重点方向及未来趋势。

第四章是大数据金融的商业模式。本章基于大数据背景探讨了商业模式创新对金融行业的运营效率和结构效率的影响。具体从企业、产业与行业三个维度分析了大数据时代金融业务商业模式的创新之处。在企业维度，大数据技术影响商业模式创新的关键因素包括组织、产品、客户、业务、财务五个方面。在产业维度，根据产业链上从事不同环节的数据资源提供商、大数据分析咨询提供商、大数据处理服务提供商和大数据解决方案提供商这四个环节，详细分析商业模式在大数据产业链层面的创新与实践应用。在行业维度，以数据为媒介整合产业链上下游，实现数据驱动的跨界模式，通过数据连接起不同行业，实现市场、企业和客户的价值关系重组。最后从企业战略、产业生态和社会运用的角度探讨大数据金融商业模式的未来发展趋势。

第五章是大数据金融机构与产品创新。大数据对金融行业的影响同样体现在传统金融机构的大数据应用和基于互联网的新型机构创建与产品开发。本章从银行业、证券业、保险业、信托业、融资租赁业和中央银行这些传统金融机构的大数据应用实践创新出发，分析了传统金融机构如何运用大数据技术把握时代潮流，开发基于互联网大数据的系列金融产品。

第六章是大数据与供应链金融。供应链金融实现了物流、资金流与数据流的融合，形成了更广阔的产业平台。本章从供应链金融发展的背景出发，分析了大数据、互联网、区块链等 Fintech 对传统供应链金融及其风险成因的影响，展望了大数据时代供应链金融发展的新趋势。

第七章是大数据金融服务平台。从数据来源、服务内容、平台目的、服务对象、定价机制等方面对大数据金融服务平台进行了理论阐释，分析了大数据金融服务平台面临来自数据质量、行业应用、监管、市场等方面的诸多风险，通过具体事例探讨了大数据金融服务平台的竞争策略、战略规划和产业链重构。

第八章是大数据金融算法。大数据体系包括数据采集与预处理、大数据存储技术和数据分析与指标构建。本章阐释了数据挖掘经典算法的理论基础与应用实践及其面临的技术监管挑战，从人与物两个维度分析了大数据算法的未来发展路径。

第九章是大数据金融生态环境建设。本章从政策、经济、技术和交易四个

方面分析了大数据金融的外部宏观环境，结合行业内部环境，探讨市场环境对大数据金融发展的影响及其传统监管体系的不足，从宏观视角分析了大数据金融生态系统的构成及其面临的挑战。

第十章是Fintech与大数据金融。Fintech作为一种新业态，通过借助大数据、人工智能、区块链等各类先进技术提升金融行业运行效率。本章介绍了Fintech发展进程中的技术变革，中美Fintech发展进程以及细分投资领域的发展。基于资金端和资产端分析了Fintech创新的内在本质。从区块链和人工智能视角分析了Fintech的技术创新与未来前景。

第十一章是案例分析。以蚂蚁金服为例，具体分析了互联网公司如何借助大数据技术抢占时代先机；以Wealthfront为例，分析了Fintech创新引领的智能投顾；以比特币为例，分析了基于区块链技术的数字货币发展。

本书写作过程中，我的研究生吴之悦、顾诚嘉、石广平、许从宝等做了大量的文献资料收集整理工作，东南大学金融系张颖老师对本书提供了诸多有益帮助，在此一并表示感谢。本书是在学习大量国内外相关文献资料基础上编写的，书后参考文献都有列出，如有“挂一漏万”之处，敬请海涵。本书的出版得到了国家自然科学基金面上项目（71673043）、东南大学经济管理学院以及清华大学出版社的大力支持，在此表示衷心感谢。尽管想努力为读者呈现一本满意的大数据金融书籍，但由于作者水平有限，加之时间仓促，书中难免有疏漏或错误之处，恳请读者多提宝贵意见，以便今后进一步修改和完善。

刘晓星

戊戌端午于南京东南大学九龙湖畔

目 | 录

第一章

大数据的提出与演化

近几十年来，随着信息技术发展的日新月异，数据分析已广泛运用于国民经济、商业实践、国家治理和社会生活等各个领域，深刻影响着人们的日常生活，“大数据”概念开始备受社会各界的广泛关注。本章从大数据概念提出的背景出发，按时间顺序梳理大数据概念的演变过程，对当前的大数据概念进行明确的界定，并指出大数据带来的时代变革。

第一节 大数据概念提出的背景

“大数据”概念的提出建立在信息技术进步的基础上，有其清晰的社会历史发展脉络，迎合着现代产业结构转型升级的需要。硬件存储性能、光纤传输带宽等基础设施的完善，互联网、云计算与物联网技术的发展，网络社交以及智能终端的普及都为“大数据”概念的提出奠定了基础，并推动“大数据”这一概念不断渗透到更多相关领域。

一、技术进步

（一）信息基础设施的完善

作为英特尔的创始人之一，Gordon Moore 于 1965 年提出了著名的“摩尔定律”。该定律阐述了计算机存储器的未来发展趋势，即每隔 18 个月，计算机存储器的性能便会提升一倍，即计算机的计算、存储能力将相对于时间周期呈指数式上升。与此同时，计算机软件系统也会随之升级，从而使计算机的信息处理和存储功能在短期内得以迅速提升，单位信息存储的成本大幅下降。当 IBM 于 1955 年推出第一款商用硬盘存储器时，其价格是 6 000 多美元 / 兆，1960 年下降到 3 600 美元 / 兆，1993 年约为 1 美元 / 兆，2000 年再降至 1 美分 / 兆，截至 2010 年则约为 0.005 美分 / 兆。而自 1977 年美国芝加哥率先投入使用光纤通信系统以来，光纤传输带宽实现迅猛增长，其信息传输能力也得到大幅跃升，甚至超越了摩尔定律下芯片性能的提升速度。信息基础设施的持续完善，包括数据存储性能不断提升、数据传输带宽的持续增加，为大数据的存储和传播提供了物质基础，使得数据信息的大规模存储、传输与分析得以实现。目前硬件存储性能与网络带宽不再成为制约大数据应用的主要因素，并且它们的高速发展将持续为大数据时代提供廉价的存储与传输服务。

（二）互联网领域的发展

人与人之间交流沟通由于互联网的出现而极大地便利了，互联网的广泛运用改善了人们的日常生活，并逐渐渗透到人们生活的方方面面。人们在互联网的海洋里徜徉时，也留下来海量数据。于是越来越多的重要数据被保存在无数个计算机上，为了保证数据存储的安全与数据传递的高效，要求计算机之间相互传递数据、互为备份的通信机制具有更高的性能标准。目前在使用互联网数据时，一般都是通过“请求”+“响应”的模式，即只有在客户端发出请求的情况下，服务器终端才会发送所需要的数据。这种数据传递模式在一定程度上保证了数据传递的安全和高效，使得人们在使用网络时的每一个搜索请求、每一个访问请求、每一个交易记录等数据信息都忠实准确地被记录在各类服务器的日志上。因而互联网的广泛普及积累了巨量的数据信息，使大数据分析过程中的数据采集成为可能，大大降低了数据采集的成本，提高了数据信息记录的真实性和可靠性。

（三）云计算技术的进步

云计算是一种基于互联网的新兴计算方式，共享的软硬件信息资源可以通过这种计算方式按需提供给计算机和其他终端应用设备。云计算服务主要是通过提供通用的在线商业应用来实现的，云计算技术改变了以往数据分散保存在每个独立的计算机中的状况，改变了数据的存储与访问方式，为大数据的集中管理和分布式访问提供了必要的场所和分享的渠道，也为数据分析、数据挖掘奠定了坚实基础。因此从某种程度上可以说，云计算是大数据诞生的前提和必要条件，没有云计算，就缺少了集中采集和存储数据的重要基础。总之，云计算为大数据提供了存储空间和访问渠道，大数据则是云计算的灵魂和升级的必然方向。近年来，以大型互联网公司、银行、电信运营商、政府部门等为代表，各市场主体都越来越关注数据的价值，纷纷出资兴建自己的“数据中心”。其中绝大部分银行、电信、互联网公司都实现了全国级的数据库建设工作，为“大数据”应用的诞生提供了必备的储存空间和访问渠道，进一步推动了大数据时代的早日来临。

（四）物联网、网络社交及智能终端的普及

基于传感器技术的物联网迅速发展，能够持续集中收集海量数据，这成为大数据的重要来源之一。其实在我们的日常生活中，传感器的运用无处不在，

它既可以是遍布大街小巷的摄像头，将实时路况及时传达；也可以是智能手机终端的重力感应器、加速度感应器、距离感应器、电子罗盘、摄像头等各类传感器，通过数据回馈分析，实现电子导航、健康指标监测等功能，提升用户体验。如果说，物联网技术的发展改变了物与物、人与物之间的关系，使得互联网的概念延伸到实物中，那么社交网络的兴起则重新定义了人与人之间交往的方式，将实际生活中的人际关系投射到互联网空间中，大大拓展了互联网的内涵。从社交网络的信息中可以了解人们的喜好、偏爱、消费习惯等信息，还能够利用网民的关系链来传播这些信息，从而构成了以个人为枢纽的数据集合，从而提供真实有效的数据。智能终端的普及拉近了互联网与日常生活的距离，也使得物联网技术与社交网络进一步融入人们的生活中，不断产生各种类型的数据，构成了大数据的重要来源。自 2010 年第二季度开始，智能手机和平板电脑的出货量就已经超越了传统台式电脑，智能手机和平板电脑凭借其便捷性迅速占领市场，并日益渗透到日常生活、商业办公、统计调查、政府治理等各个方面，成为大数据的重要来源渠道。

二、产业升级

从哲学意义上说，世界处于永续变动之中，万事万物在其运动过程中都产生了大量的数据信息。近年来，随着互联网、云计算、物联网等信息技术的飞速发展，各行各业的产业结构不断升级，这无时无刻不在产生海量的数据，形成大数据雏形。目前，我国经济本质上仍处于传统经济阶段，缺乏具有国际竞争力的现代产业，产业结构升级已经迫在眉睫，这无疑为大数据的滋生提供了肥沃的土壤。

当前互联网的普及、信息技术的进步以及电子化时代的到来，人们以更快捷、更容易、更廉价的方式获取和存储数据，使得数据及信息量以指数方式增长。据粗略估计，一个中等规模企业每天要产生 100 MB 以上的商业数据。而电信、银行、大型零售业随着产业结构的不断调整和升级，每天产生的数据量都可以用 TB 来计算（数据的最小计量单位是字节，具体换算标准为 1 KB=1 024 B；1 MB=1 024 KB；1 GB=1 024 MB；1 TB=1 024 GB；1 PB=1 024 TB；1 EB=1 024 PB；1 ZB=1 024 EB；1 YB=1 024 ZB；1 DB=1 024 YB；1 NB=1 024 DB）。《至顶网年度技术报告》的数据统计结果显示，2013 年中国产生的数据总量超过 0.8ZB，是 2012 年数据总量的 2 倍，相当于 2009 年全球的数据总量。而且预

计到2020年，中国产生的数据总量将超过8.5 ZB，是2013年的10倍。产业结构升级所带来的数据越来越多，剧增的数据背后隐藏着许多重要的信息，人们希望对其进行更高层次的分析，以便更好地利用这些数据。现有的数据库系统虽然拥有高效地完成数据的输入、统计、查询等功能，却不能发现数据中的关系与规则，不能在现有数据的基础上来推断今后的发展趋势。大数据技术背后隐藏的知识手段的不足，使得“数据爆炸但知识匮乏”这一现象浮现出来。自此人们纷纷提出“学会选择、提炼、舍弃信息”，并思考怎样才能不被海量的信息所淹没，怎样才能及时发现有用的知识、提高信息利用效率？如何从浩瀚如烟海的资料中选择性地收集有价值的信息？这为数据分析带来了一些挑战：第一是信息过量，难以消化；第二是信息真假难以辨别；第三是信息安全难以保证；第四是信息形式不一致，难以统一处理 。为应对这些挑战，计算机数据仓库处理技术随之走向成熟，从数据中发现知识及其核心技术——大数据技术便应运而生，并得以蓬勃发展，显示出越来越强大的生命力。

三、社会历史

1998年，《科学》杂志刊登的一篇名为《大数据的处理程序》的文章中第一次明确使用了大数据（big data）一词。2008年9月*Nature*杂志刊登了名为“Big Data”的专题，“大数据”概念开始受到广泛关注，大数据的产生和发展有其特定的社会历史发展脉络。其实大数据存在的历史非常悠久，“大数据”概念的提出标志着人们已经开始意识到大数据的客观存在，而且已经感受到了大数据应用的重要性。

各种各样的海量数据构成了大数据的基石。悠久的社会历史文化为大数据的产生提供了充足的时间条件。从人类历史发展脉络来看，数据的产生与人类自身的生存、生活密切相关，也正是这种内在需求促进了数据发展为大数据。大数据分析是一种非常实用的技术，古希腊的哲学家率先让数据从实用走向抽象。哲学家们第一次抛弃经验主义的桎梏，把数据当作事物的本源，这种独特的思维模式为自然哲学的研究开辟了一条崭新的道路，也为大数据的诞生奠定了哲学历史基础。纵观数据的发展历史，数据和其他语言文字一样，都是人类文明的产物，是用于记录事物性质和互相交流的工具。从广义上看，数据可以被看作语言的一部分，但与文字语言的差别在于，数据的表达形式更简单、更加有利于交流。所以虽然不同人类文明有着不同的记数方式和数制，但随着不

同文化的相互交流融合，数据形式的高度统一超出了所有文字语言，这离不开数字简单精确的属性。回顾科学技术的发展史，科学技术的迅猛发展离不开科学数据的支撑，科学数据具有客观性、精确性、一致性和易交流性等特征。所以说，数据不仅是连接事物客观性和人类主观性的纽带，还是人类认识世界的桥梁。但从数据产生的那一刻起，人类主观因素无时无刻不在影响着数据的客观性。大量数据构成的集合形成了一种重要的研究素材，激发着科学家和哲学家们进行深入的探究，他们在研究过程中越发意识到数据的重要性，所以大数据便应运而生。

在这里，我们简要介绍一下数据科学的发展历史。

自 20 世纪中期以来，生物学领域的基因组测序技术发展迅猛，累积了海量的生物学数据，如何理解这些数据，是生物学家们面临的一种新挑战。同样的数据分析问题也存在于其他领域（如气象学、社会学等）和复杂系统的研究之中。值得注意的是，国际科学技术数据委员会（Committee on Data for Science and Technology，CODATA）于 1966 年成立，旨在提升数据的质量、可信度、可达性并加强对数据的管理，从而在世界范围内实现共享科技数据的目标。1984 年 6 月，中国科学院以国家会员的身份加入 CODATA。

基于数据的相关研究已得到学术界的广泛关注。数据科学是一门以大量观测数据、理论数据和计算机模拟数据为研究对象，通过挖掘、提取等手段寻求其内在规律的学科。1960 年，Peter Naur 首次提出“数据科学”（data science）这一术语。1996 年，在日本东京召开的分类国际联合会（the International Federation of Classification Societies，IFCS）上，第一次将数据科学用于会议题目——“数据科学，分类和相关方法”（Data Science，Classificationand Related Methods）。美国普渡大学统计学教授 William S. Cleveland 于 2001 年首次倡导将数据科学建设成一门独立的学科，他认为数据科学是统计学与数据的结合，并建立了数据科学的 6 个细分技术领域：多学科研究、数据模型和方法、数据计算、教育、工具评估、理论。

2001 年，CODATA 创办了学术刊物 *CODATA Data Science Journal*，标志着数据科学的诞生。2003 年，由中美两国学者共同创办的 *Journal of Data Science* 在哥伦比亚大学正式出版，*Journal of Data Science* 主要发表一些关于数据的研究成果，如数据的收集、分析以及建模等。

2012 年，Springer 出版集团创建了期刊 *EPJ Data Science*。该期刊的主办方认为，21 世纪出现的“数据驱动科学”是传统“假说驱动科学”研究方法的重