数据科学研究的现状与趋势

朝乐门[1,2] 邢春晓[3,4,5] 张 勇[3,4,5]

(数据工程与知识工程教育部重点实验室(中国人民大学) 北京 100872)[1]

(中国人民大学信息资源管理学院 北京 100872)[2] (清华大学计算机科学与技术系 北京 100084)[3]

(清华大学信息技术研究院 北京 100084)[4] (清华信息科学与技术国家实验室(筹) 北京 100084)[5]

摘 要 大数据时代的到来催生了一门新的学科——数据科学。首先,探讨了数据科学的内涵、发展简史、学科地位及知识体系等基本问题,并提出了专业数据科学与专业中的数据科学之间的区别与联系。其次,分析现阶段数据科学的研究特点,并分别提出了专业数据科学、专业中的数据科学及大数据生态系统中的相对热门话题。接着,探讨了数据科学研究中的10个争议及挑战:思维模式的转变(知识范式还是数据范式)、对数据的认识(主动属性还是被动属性)、对智能的认识(更好的算法还是更多的数据)、主要瓶颈(数据密集型还是计算密集型)、数据准备(数据预处理还是数据加工)、服务质量(精准度还是用户体验)、数据分析(解释性分析还是预测性分析)、算法评价(复杂度还是扩展性)、研究范式(第三范式还是第四范式)、人才培养(数据工程师还是数据科学家)。然后,提出了数据科学研究的10个发展趋势:预测模型及相关分析的重视,模型集成及元分析的兴起,数据在先、模式在后或无模式的出现,数据一致性及现实主义的回归,多副本技术及靠近数据原则的广泛应用,多样化技术及一体化应用并存,简单计算及实用主义占据主导地位,数据产品开发及数据科学的嵌入式应用,专家余及公众数据科学的兴起,数据科学家与人才培养的探讨。最后,结合文中工作,对数据科学研究者给出了几点建议和注意事项。

关键词 数据科学,大数据,数据产品开发,数据加工,数据驱动

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.01.001

Data Science Studies:State-of-the-art and Trends

CHAO Le-men[1,2] XING Chun-xiao[3,4,5] ZHANG Yong[3,4,5]

(Key Laboratory of Data Engineering and Knowledge Engineering,Renmin University of China,Beijing 100872,China)[1]

(School of Information Resource Management,Renmin University of China,Beijing 100872,China)[2]

(Department of Computer Science and Technology,Tsinghua University,Beijing 100084,China)[3]

(Reasearch Institute of Information Technology,Tsinghua University,Beijing 100084,China)[4]

(Tsinghua National Laboratory for Information Science and Technology (TNList),Beijing 100084,China)[5]

Abstract The entering big data era gives rise to a novel discipline called data science. First,the differences between domain-general data science and domain-specific data science were proposed based upon conducting an in-depth discussion on its basic concept,brief history,scientific roles and the body of knowledge. Secondly,top ten challenges faced by data science were identified via describing the debates on paradoxical topics including the shifts of thinking pattern (knowledge pattern or data pattern),perspectives on data (active or negative),implementation of intelligence(via AI or via big data),bottlenecks of data products development(computing intensive or data intensive),data preparation (data preprocessing or data wrangling),quality of services(performance of services or user experiences),data analysis (explanatory or predictive),evaluation of algorithm(by complexity or by scalability),research paradigm(third paradigm or fourth paradigm) as well as main motivations of the education(in order to cultivate data engineer or data scientist). And then,the top ten trends in data science studies were proposed:to vale predictive models and correlation analysis,to give more attention on model integration and meta-analysis,to embrace data first,model later or never paradigm,to be led by realism and ensure data consistence,to support multi-copies and data locality,the coexistence of varieties in implementation techno logies and integrated applications,to be dominated by simple computing and pragmatism,to develop data

到稿日期:2017-10-20 返修日期:2017-11-30 本文受国家自然科学基金项目(91646202,71103020),国家社会科学基金(15BTQ054,12&ZD220)资助。

朝乐门(1979—),男,副教授,博士生导师,主要研究方向为数据科学与大数据分析,E-mail:chaolemen@ruc.edu.cn(通信作者);邢春晓(1967—),男,教授,博士生导师,主要研究方向为云计算与大数据分析;张 勇(1973—),男,博士,副教授,主要研究方向为数据管理与数据分析。

products and the embedded applications of data science, to embrace the Pro-Am and metadata, and cultivate data scientist and curriculums or majors. Finally, some suggestions on how do further studies were also proposed.

Keywords Data science, Big data, Data products developement, Data wrangling, Data-driven

大数据正在改变着人们的工作、生活与思维模式[1]，进而对文化、技术和学术研究产生深远影响[2]。一方面，大数据时代给各学科领域带来了新的机遇——认识论和研究范式的转变[3]，促生了一种区别于传统科学研究中沿用至今的“知识范式”的新研究范式——“数据范式”。“数据范式”的广泛应用成为现代科学研究的一个重要转变。另一方面，大数据带来的挑战在于数据的获取、存储、计算不再是瓶颈或难题，各学科领域中的传统知识与新兴数据之间的矛盾日益突出，传统知识无法解释和有效利用新兴的大数据，进而促使传统理论与方法的革命性变化。

目前，大数据已受到各学科领域的高度关注，成为包括计算机科学和统计学在内的多个学科领域的新研究方向，表现出不同专业领域中的数据研究相互高度融合的趋势，进而即将独立出一门新兴学科——数据科学。同时，大数据研究中仍存在一些误区或曲解，如片面追求数据规模、过于强调计算架构和算法、过度依赖分析工具、忽视数据重用、混淆数据科学与大数据的概念以及全盘否定大数据等[4]。因此，现代社会需要一门新学科来系统研究大数据时代的新现象、理念、理论、方法、技术、工具和实践，即“数据科学”。

本文第 1 节探讨数据科学的内涵、发展简史、学科地位和知识体系等 4 个基本问题，并提出了数据科学的两个基本类型——专业数据科学和专业中的数据科学；第 2 节提出现阶段数据科学研究的特点——本质问题的系统研究较少，而周边问题的讨论较多，并分别分析了专业数据科学、专业中的数据科学以及大数据生态系统中的相对热门话题；第 3 节探讨数据科学研究中的 10 个争议——思维模式的转变（知识范式还是数据范式）、对数据的认识视角（主动属性还是被动属性）、对智能的认识侧重点（更好的算法还是更多的数据）、主要瓶颈（数据密集型还是计算密集型）、数据准备（数据预处理还是数据加工）、服务质量（精准度还是用户体验）、数据分析（解释性分析还是预测性分析）、算法评价（复杂度还是扩展性）、研究范式（第三范式还是第四范式）和人才培养（数据工程师还是数据科学家），并分别提出了研究挑战；第 4 节分析了数据科学研究的 10 个发展趋势——预测模型及相关分析的重视，模型集成及元分析的兴起，数据在先、模式在后或无模式的出现，数据一致性及现实主义的回归，多副本技术及靠近数据原则的应用，多样化技术及一体化应用并存，简单计算及实用主义占据主导地位，数据产品开发及数据科学的嵌入式应用，专家余及公众数据科学的兴起以及数据科学家与人才培养的探讨。最后总结全文，并对数据科学研究者提出了几点建议。

1 数据科学：大数据背后的科学

“数据科学”与“大数据”是两个既有区别又有联系的术语，可以将数据科学理解为大数据时代的一门新科学[5]，即以揭示数据时代尤其是大数据时代新的挑战、机会、思维和模式为研究目的，由大数据时代新出现的理论、方法、模型、技术、平台、工具、应用和最佳实践组成的一整套知识体系。

1.1 数据科学的内涵及兴起

1974 年，著名计算机科学家、图灵奖获得者 Peter Naur 在其著作 *Concise Survey of Computer Methods* 的前言中首次明确提出了数据科学（Data Science）的概念：“数据科学是一门基于数据处理的科学”，并提到了数据科学与数据学（Datalogy）的区别——前者是解决数据（问题）的科学（the science of dealing with data），而后者侧重于数据处理及其在教育领域中的应用（the science of data and of data processes and its place in education）[6]。

Peter Naur 首次明确提出数据科学的概念之后，数据科学研究经历了一段漫长的沉默期。直到 2001 年贝尔实验室的 Cleveland 在学术期刊 *International Statistical Review* 上发表了题为“Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics”的论文，主张数据科学是统计学的一个重要研究方向[7]，数据科学才再度受到统计学领域的关注。2013 年，Mattmann[8] 和 Dhar[9] 在 *Nature* 和 *Communications of the ACM* 上分别发表了题为“Computing: A vision for data science”和“Data science and prediction”的论文，从计算机科学与技术视角讨论了数据科学的内涵，使数据科学被纳入计算机科学与技术专业的研究范畴。然而，数据科学被更多人关注是因为后来发生了 3 个标志性事件：1）Patil D J 和 Davenport T H 于 2012 年在 *Harvard Business Review* 上发表题为“Data scientist: the sexiest job of the 21st century”[10] 的论文；2）2012 年，大数据思维首次应用于美国总统大选，奥巴马击败罗姆尼，成功连任[11]；3）美国白宫于 2015 年首次设立数据科学家的岗位，并聘请 Patil D J 作为白宫第一任首席数据科学家[12]。

Gartner 的调研及其新技术成长曲线（Gartner’s 2014 Hype Cycle for Emerging Technologies）[13] 表示，数据科学的发展于 2014 年 7 月已经接近创新与膨胀期的末端，将在 2～5 年内开始应用于生产高地期（plateau of Productivity）。同时，Gartner 的另一项研究揭示了数据科学本身的成长曲线（Hype Cycle for Data Science）[14]，如图 1 所示。从中可以看出，数据科学的各组成部分的成熟度不同：R 的成熟度最高，已广泛应用于生产活动；其次是模拟与仿真、集成学习、视频与图像分析、文本分析等，它们正在趋于成熟，即将投入实际应用；基于 Hadoop 的数据发现可能会消失；语音分析、模型管理、自然语言问答等已经度过了炒作期，正在走向实际应用；公众数据科学、模型工厂、算法市场（经济）、规范分析等正处于高速发展期。

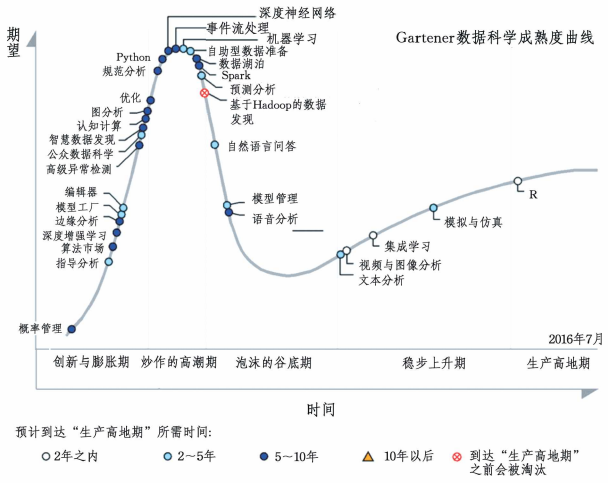

图 1　数据科学的成长曲线(2016)

Fig. 1　Growth curve of data science(2016)

1.2　数据科学的学科地位

2010 年，Drew Conway 提出了第一张揭示数据科学的学科地位的维恩图——数据科学维恩图(The Data Science Venn Diagram)(见图 2)，首次明确探讨了数据科学的学科定位问题[15]。在他看来，数据科学处于统计学、机器学习和领域知识的交叉处。后来，其他学者在此基础上提出了诸多修正或改进版本，图 3 为 Jerry Overton 于 2016 年给出的数据科学维恩图[16]。但是，后续版本对数据科学的贡献和影响远不及 Drew Convey 首次提出的数据科学维恩图。

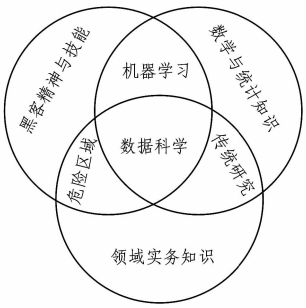

图 2　Drew Conway 的数据科学维恩图(2010)

Fig. 2　Drew Conway's venn diagram of data science(2010)

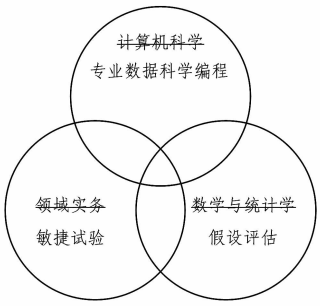

图 3　Jerry Overton 的数据科学维恩图(2016)

Fig. 3　Jerry Overton's venn diagram of data science(2016)

从 Drew Conway 的数据科学维恩图的中心部分可看出，数据科学位于统计学、机器学习和某一领域知识的交叉处，具备较为显著的交叉型学科的特点，即数据科学是一门以统计学、机器学习和领域知识为理论基础的新兴学科。同时，从图 2 的外围可看出，数据科学家需要具备数学与统计学知识、领域实战和黑客精神，说明数据科学不仅需要理论知识和实践经验，还涉及黑客精神，即数据科学具备 3 个基本要素：理论(数学与统计学)、实践(领域实务)和精神(黑客精神)。

1.3　数据科学的知识体系

从知识体系看，数据科学主要以统计学、机器学习、数据可视化以及(某一)领域知识为理论基础，主要研究内容包括数据科学基础理论、数据加工、数据计算、数据管理、数据分析和数据产品开发，如图 4 所示[17]。

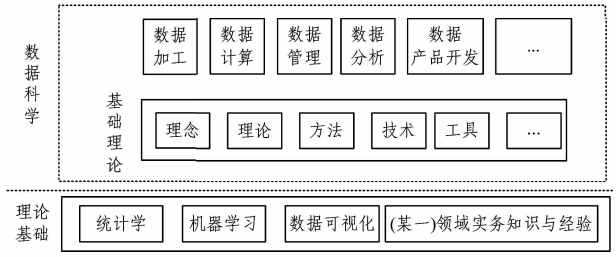

图 4　数据科学的知识体系

Fig. 4　Knowledge system of data sicence

(1)基础理论：主要包括数据科学中的新理念、理论、方法、技术及工具以及数据科学的研究目的、理论基础、研究内容、基本流程、主要原则、典型应用、人才培养、项目管理等。需要特别提醒的是，“基础理论”与“理论基础”是两个不同的概念。数据科学的“基础理论”在数据科学的研究边界之内；而其“理论基础”在数据科学的研究边界之外，是数据科学的理论依据和来源。

(2)数据加工(Data Wrangling 或 Data Munging)：数据科学中关注的新问题之一。为了提升数据质量、降低数据计算的复杂度、减少数据计算量并提升数据处理的精准度，数据科学项目需要对原始数据进行一定的加工处理工作——数据审计、数据清洗、数据变换、数据集成、数据脱敏、数据归约和数据标注等。值得一提的是，与传统数据处理不同的是，数据科学中的数据加工更强调数据处理中的增值过程，即如何将数据科学家的创造性设计、批判性思考和好奇性提问融入数据的加工活动之中。

(3)数据计算：在数据科学中，计算模式发生了根本性的变化——从集中式计算、分布式计算、网格计算等传统计算过渡至云计算。比较有代表性的是 Google 三大云计算技术(GFS、BigTable 和 MapReduce)、Hadoop MapReduce、Spark 和 YARN。计算模式的变化意味着数据科学中所关注的数据计算的主要瓶颈、主要矛盾和思维模式发生了根本性变化。

(4)数据管理：在完成“数据加工”和“数据计算”之后，还需要对数据进行管理与维护，以便进行(再次进行)“数据分析”以及数据的再利用和长久存储。在数据科学中，数据管理方法与技术也发生了重要变革——不仅包括传统关系型数据库，还出现了一些新兴数据管理技术，如 NoSQL、NewSQL 技术和关系云等。

(5)数据分析：数据科学中采用的数据分析方法具有较为明显的专业性，通常以开源工具为主，与传统数据分析有着较为显著的差异。目前，R 语言和 Python 语言已成为数据科学家应用较为普遍的数据分析工具。

(6)数据产品开发：“数据产品”在数据科学中具有特殊的

含义——基于数据开发的产品的统称。数据产品开发是数据科学的主要研究使命之一，也是数据科学与其他科学的重要区别。与传统产品开发不同的是，数据产品开发具有以数据为中心、多样性、层次性和增值性等特征。数据产品开发能力也是数据科学家的主要竞争力之源。因此，数据科学的学习目的之一是提升自己的数据产品开发能力。

1.4 专业数据科学及专业中的数据科学

数据科学是一门与领域知识和行业实践高度交融的学科。从目前的研究现状来看，数据科学可以分为两类：专业数据科学与专业中的数据科学。其中，“专业数据科学”是以独立学科的形式存在，且与其他传统学科（如计算机科学、统计学、新闻学、社会学等）并列的一门新兴科学；“专业中的数据科学”是指依存于某一专业领域中的大数据研究，其特点是与所属专业的耦合度较高，难以直接移植到另一个专业领域，如数据新闻（Data Journalism）[18]、材料数据科学（Materials Data Science）[19]、大数据金融（Big Data Finance）[20]、大数据社会、大数据伦理（Big Data Ethics）[21]和大数据教育（Big Data Education）[22]等。

专业数据科学与专业中的数据科学的联系如下：专业数据科学聚集了不同专业中的数据科学中的共性理念、理论、方法、术语与工具；相对于专业中的数据科学，专业数据科学更具共性和可移植性，并为不同专业中的数据科学研究奠定了理论基础；专业中的数据科学代表的是不同专业中对数据科学的差异性认识和区别化应用。

2 数据科学的研究热点

目前，数据科学的研究特点是对本质问题的系统研究较少，而对周边问题的讨论较多，具体可从以下 4 个方面进行分类分析。

2.1 周边问题仍为研究热点

从文献分布看，数据科学的研究主题可以分为两类：核心问题和周边问题。前者代表的是数据科学的基础理论——数据科学特有的理念、理论、方法、技术、工具、应用及代表性实践；后者代表的是数据科学的底层理论（理论基础，如统计学、机器学习等）、上层应用（应用理论，如数据新闻、大数据金融、大数据社会、大数据生态系统等）以及相关研究（如云计算、物联网、移动计算等）。文献数量和研究深度表明，现阶段的数据科学仍以周边问题的讨论为研究热点，而对数据科学核心问题的研究远远不够。对数据科学周边问题的研究主要集中在以下方面：

（1）大数据挑战及数据科学的必要性。在大数据时代，挑战和机会并存[23]。挑战不仅来自于数据量（Volume），还涉及其多个 V 特征，如种类多（Variety）、速度要求高（Velocity）和价值密度低（Value）[24-25]。因此，社会与科技的发展亟待一门新的学科——数据科学，并对大数据时代的新问题和新思路进行系统研究[26]。

（2）数据科学对统计学和计算机科学的继承与创新。一方面，数据科学作为新的研究方向，进一步拓展了统计学[27]和计算机科学与技术[28]的研究范畴；另一方面，数据科学不仅继承了统计学和计算机科学等基础理论，而且对其进行了创新与发展，逐渐成为一门独立学科[29]。

（3）新技术在数据科学中的重要地位。云计算、物联网、移动计算等新技术的兴起拓展了人们的数据获取、存储和计算能力，促使大数据时代的到来，成为数据学科诞生的必要条件。同时，数据科学中需要重点引入 Spark[30]，Hadoop[31]，NoSQL[32]等新兴技术，从而更好地面对大数据挑战。新技术的应用意味着数据科学对数据及其管理的认识发生了根本性变化——不仅开始接受数据的复杂性，而且数据管理的理念从传统的完美主义转向现实主义，“数据在先，模式在后或无模式”的数据管理范式、BASE 原则以及 CAP 理论[33]等新理念已成为数据科学的基本共识。

（4）数据科学对特定领域的影响。大数据及其背后的数据科学在特定领域的应用是近几年的热门话题，尤其在生命科学[34]、医疗保健[35]、政府治理[36]、教学教育[37]和业务管理[38]等领域的广泛应用，出现了量化自我[39]、数据新闻[40]、大数据分析学[41]等新的研究课题。

（5）数据科学领域的人才培养。与传统科学领域不同的是，数据科学领域人才培养目的是培养学生的“以数据为中心的思考能力”[42]。目前，相关研究主要涉及 4 个主题：数据科学课程的建设、相关课程的教学改革[43]、跨学科型人才培养[44]以及女性数据科学家的培养[45]。总体上看，数据科学的人才培养目的并不是培养数据工程师，而是数据科学家，尤其需要培养具有 3C 精神的数据科学家——原创性（Creative）设计、批判性（Critical）思考和好奇性（Curious）提问[46]。

2.2 专业数据科学研究中的相对热门话题

从研究视角看，数据科学的研究可以分为两类：专业数据科学和专业中的数据科学。前者代表的是将数据科学当作一门独立于传统科学的新兴学科来研究，强调的是其学科基础性；后者代表的是将数据科学当作传统学科的新研究学科和思维模式来研究，强调的是数据科学的学科交叉性。从目前的研究现状看，专业数据科学研究的热门话题如下。

（1）DIKW 模型。DIKW 模型刻画的是人类对数据的认识程度的转变过程[47]。通常认为，数据科学的研究任务是将数据转换成信息（Information）、知识（Knowledge）或（和）智慧（Wisdom）[48]。从数据到智慧的转变过程是一种从不可预知到可预知的增值过程，即数据通过还原其真实发生的背景（Context）成为信息，信息赋予其内在含义（Meaning）之后成为知识，而知识通过理解转变成智慧。

（2）数据分析学（Data Analytics）。大数据分析研究正在成为一门相对成熟的研究学科——数据分析学。需要注意的是，数据分析（Data Analysis）与数据分析学是两个不同的概念：前者强调的是数据分析活动本身，而后者更加强调数据分析中的方法、技术和工具。目前，大数据分析研究中的热门话题有两个：1）大数据分析学，尤其是大数据分析算法和工具的开发；2）面向特定领域的大数据分析，如面向物流与供应链管理[49]、网络安全[50]以及医疗健康[51]的大数据分析学。文献[52]给出了数据分析的主要类型及常见错误。

（3）数据化（Datafication）。数据化是将客观世界以及业务活动以数据的形式进行计量和记录，形成大数据，以便进行后续的开发利用。除了物联网和传感器等公认的研究课题，

量化自我（Quantified Self）[53-54]也逐渐成为数据化的热门话题。数据化是大数据时代初级阶段主要关注的问题，随着大数据的积淀，人们的研究焦点将从业务的数据化转向数据的业务化，即研究重点将放在“基于数据定义和优化业务”上。

（4）数据治理（Data Governance）。数据治理是指数据管理的管理。目前，相关研究主要集中在顶层设计[55]、实现方法[56]、参考框架[57]以及如何保证数据管理的可持续性[58]。此外，数据治理作为数据能力成熟度评估模型（Data Maturity Model）的关键过程域，重点关注的是如何通过数据治理提升组织数据管理能力的问题。DMM 中定义的关键过程域“数据治理”包括 3 个关键过程：治理管理（Governance Management）、业务术语表（Business Glossary）和元数据管理（Metadata Management）[59]。

（5）数据质量。对大数据的质量与可用性之间内在联系的讨论已成为现阶段数据科学的热点问题之一，主要研究议题集中在大数据中的质量问题会不会导致数据科学项目的根本性错误[60]以及大数据时代背景下的数据可用性的挑战及新研究问题[62]。但是，传统数据管理和数据科学对数据质量的关注点不同。传统数据管理主要从数据内容视角关注质量问题，强调的是数据是否为干净数据（Clean Data）/脏数据（Dirty Data）[63]；数据科学主要从数据形态视角关注质量问题，重视的是数据是否为整齐数据（Tidy Data）/混乱数据（Messy Data）。所谓的整齐数据是指数据的形态可以直接支持算法和数据处理的要求。例如，著名的数据科学家 Hadley Wickham 提出了整齐数据和数据整齐化处理（Data Tidying）的概念，并主张整齐数据应遵循 3 个基本原则：每个观察占且仅占一行，每个变量占且仅占一列，每一类观察单元构成一个关系表[64]。

除了上述问题之外，大数据的安全[65]、大数据环境下的个人隐私保护[66]、数据科学的项目管理及团队建设[67]、公众数据科学（Citizen Data Science）[68]等是目前在专业数据科学研究中讨论得较多的问题。

2.3　专业中的数据科学研究的相对热门话题

相对于专业数据科学，专业中的数据科学研究具有差异性和隐蔽性。差异性主要表现在各学科领域对数据科学的关注点和视角不同；隐蔽性是指专业中的数据科学研究往往间接地吸收和借鉴数据科学或类似于数据科学的思想，而并不明确采用或直接运用数据科学的规范术语。从目前的研究看，以下几个专业中的数据科学研究尤为活跃。

（1）数据新闻（Data Journalism）：新闻学领域的新研究方向之一，主要研究的是如何将大数据和数据科学的理念引入新闻领域，实现数据驱动型新闻（Data-driven Journalism）[68]。

（2）工业大数据：主要研究如何将大数据应用于工业制造领域，进而实现工业制造的创新。比较有代表性的是德国工业 4.0（Industries 4.0）、美国工业互联网（Industrial internet）和中国制造 2025（Made in China）。

（3）消费大数据：与工业大数据不同的是，消费大数据更关注产品生命周期的末端，即如何将已生产出的产品推销给更多的用户，主要包括精准营销[69]、用户画像（User Profiling）[70]以及广告推送[71]。

（4）健康大数据：主要关注大数据在健康与医疗领域的广泛应用，包括生命日志（Life Logging）[72]、医疗诊断、药物开发、卫生保健[73]等具体领域的应用。

（5）生物大数据：将大数据的理念、理论、方法、技术和工具应用于生物学领域，从而使生物学从知识范式转向数据范式[74]。

（6）社会大数据：综合运用大数据和数据科学的理论，探讨如何在大数据时代进行舆情分析、社会网络分析以及热点发现[75]。

（7）机构大数据：如何将大数据和数据科学的思想引入企业[76]、政府[77]以及公益部门[78]的日常业务、战略规划与可持续改进。

（8）智慧类应用：如何将大数据应用于智慧城市、智慧医疗、智慧养老、智慧交通、智慧教育等领域，发挥数据的驱动作用，进而实现更高的智慧。

（9）敏捷类应用：如何将大数据思维用于软件开发、项目管理以及组织管理之中，进而实现敏捷软件开发、敏捷项目管理和敏捷组织，提升其应变能力和可持续发展能力。

2.4　大数据生态系统研究中的相对热门话题

数据科学生态系统（Big Data Ecosystem）是指包括基础设施、支撑技术、工具与平台、项目管理以及其他外部影响因素在内的各种组成要素构成的完整系统。例如，大数据全景图（Big Data Landscape）[79]较为全面地展示了大数据生态系统中的主要机构及产品。现有相关研究主要从组成要素及其相互关系两个方面进行研究，相关研究中的热门话题集中在以下几个方面。

（1）基础设施：主要关注包括云计算、物联网、移动计算、社交媒体在内的基础设施对数据科学的影响及在数据科学中如何充分利用上述基础设施。

（2）支撑技术：建立在基础设施上的关键技术，现有研究主要讨论机器学习、统计学、批处理、流计算、图计算、交互计算、NoSQL、NewSQL 和关系云等支撑技术在数据科学中的应用。

（3）工具与平台：支撑技术的具体实现，目前的研究热点主要集中在 R，Python，Hadoop，Spark，MongoDB，HBase，Memcached，MongoDB，CouchDB 和 Redis 等工具与平台在数据科学中的应用。

（4）项目管理：涉及数据科学项目的范围、时间、成本、质量、风险、人力资源、沟通、采购及系统管理等 9 个方面的管理。

（5）环境因素：大数据时代对法律、政策、制度、文化、道德、伦理产生的影响与新需求。其中，大数据权属立法研究主要讨论大数据权属立法的必要性、可行性以及对策建议。从大数据的重要性的认识看，大数据不仅是一种资源，更是一种资产。大数据权属的立法已经成为大数据时代信息资源开发利用的必要条件。

3　数据科学研究的争议与挑战

在不同的学科领域，大数据时代的科学研究所面临的问题、挑战和关注点不同。从计算机科学视角看，新的数据处理

需求已经超出了现有的存储与计算能力[80]；从统计学视角看，大数据挑战在于样本的规模接近总体时，如何直接在总体上进行统计分析[1]；从机器学习角度看，训练样本集接近测试样本集时，如何用简单模型及模型集成方法实现较高的智能水平[81]；从数据分析角度看，如何从海量数据中快速洞察有价值的数据，并通过试验设计和模拟仿真实现数据到智慧的转变[82]。但是，从数据科学视角看，其研究中的常见争议及背后的研究挑战可以归纳为 10 个方面。

3.1 思维模式——知识范式还是数据范式

在传统科学研究中，由于数据的获取、存储和计算能力所限，人们往往采取知识范式（“数据→知识→问题”的范式），从数据尤其是样本数据中提炼出知识之后，用知识来解决现实问题。大数据时代的到来及数据科学的出现为人们提供了另一种研究思路，即数据范式（“数据→问题”范式），在尚未从数据中提炼出知识的前提下，用数据直接解决问题。数据范式强调的是在尚未将数据转换为知识的前提下，直接用数据解决现实世界中的问题。以机器翻译为例，传统机器翻译方法是基于自然语言理解，准确地说是基于语言学和统计学的知识进行，属于知识范式的范畴。但是，这种传统机器翻译效果一直不理想，且尚无突破性进展。然而，近几年兴起的机器翻译方法改变了传统机器翻译的思维模式，采取的是“数据范式”——直接从历史跨语言语料库中快速洞见所需结果。20 世纪 50 年代以来的 IBM 机器翻译的缓慢发展以及 2000 年以后的 Google 机器翻译的迅速兴起也反映了这种思维模式的变革。

与传统认识中的“知识就是力量”类似，在大数据时代，数据也成为一种重要力量。如何组织、挖掘和利用数据成为现代组织的核心竞争力。目前，思维模式变革的主要挑战在于如何完成以数据为中心的设计、数据驱动型决策[83]和数据密集型应用[84]。

3.2 数据的认识——主动属性还是被动属性

在传统科学研究中，数据一直被当作被动的东西，人们主要从被动属性方面来对待数据。以关系数据库为例，人们先定义关系模式，然后将数据按照关系模式的要求进行强制转换后放入数据库中，从而完成数据挖掘和分析任务。

在大数据思维模式的背后，一个根本性的变革在于人们开始意识到数据的主动属性——不再简单认为数据是一种死的、被动的东西，而更加重视数据的积极作用，从而提出了数据在先模式在后或无模式、让数据说话、数据驱动型应用、数据业务化、数据洞察和以数据为中心的思维模式等新术语。

因此，如何正确认识数据及如何充分发挥数据的主动属性成为数据科学的重要研究任务。目前，相关研究的主要挑战在于如何实现数据洞察、以数据为中心的设计、敏捷软件开发、数据驱动型决策以及智慧类应用研发。

3.3 智能的认识——更好的算法还是更多的数据

在传统学术研究中，智能主要来自于算法，尤其是复杂的算法。算法的复杂度随着智能水平的提升而提升。例如，KNN 算法是机器学习中常用的分类算法，其算法思想非常简单。人们根据不同应用场景提出多种改进或演化方案，虽然智能水平有所提高，但会提升算法的复杂度[85]。但是，数据范式表明，数据也可以直接用于解决问题，引发了一场关于“更多数据还是更好模型（More data or Better Model debate）”的讨论[86]，经过这场大讨论，人们得出了相对一致的结论——“更多数据＋简单算法＝最好的模型（more data＋simple algorithm＝the best model）”。

因此，如何设计出简单高效的算法以及算法的集成应用成为数据科学的重要挑战。目前，关于智能的实现方式的挑战在于算法设计、算法集成、维度灾难和深度学习。

3.4 研发瓶颈——数据密集型还是计算密集型

传统的软件开发与算法设计的重点是解决计算密集型的问题，计算是研究难点和瓶颈。但是，随着大规模分布式计算尤其是云计算的普及，计算不再是人们需要解决的首要瓶颈。因此，软件开发与算法设计的主要矛盾从计算转向数据，出现了数据密集型应用。在数据密集型应用中，数据是主要的关注点与瓶颈[76]。数据密集型问题的研究将进一步推动以数据为中心的研究范式。

目前，数据密集型应用的主要挑战在于副本数据技术、物化视图、计算的本地化、数据模型的多样化和数据一致性保障。

3.5 数据准备——数据预处理还是数据加工

在传统数据研究中，数据准备主要强调的是将复杂数据转换为简单数据，对脏数据进行清洗处理后得到干净数据，从而防止出现“垃圾进垃圾出”现象，主要涉及重复数据的过滤、错误数据的识别以及缺失数据的处理。可见，数据预处理主要关注的是数据的质量维度问题。但是，由于小数据到大数据之间存在质量涌现现象——个别小数据的质量问题（如缺失数据、错误数据或重复数据）不影响整个大数据的可用性，大数据处理中关注的并非是传统意义上的数据预处理，而转向另一个重要课题——数据加工。

在数据科学中，数据加工是指数据的创造性增值过程，包括两种表现形式：数据打磨（data wrangling）和数据改写（data munging）。与数据预处理不同的是，数据加工更加强调如何将数据科学家的 3C 精神融入数据处理工作之中，从而达到数据增值的目的。因此，数据加工并不仅限于技术工作的范畴，而且还涉及到艺术层面的创造，如需要采用数据柔术（Data Jujitsu）和整齐化处理（Data Tidying）的方法进行数据加工处理。

数据加工概念的提出意味着人们对数据复杂性的认识发生了重要的变革，即开始接受数据的复杂性特征，认为复杂性是数据本身的固有特征。与此同时，数据准备的关注点转向另一个重要问题，即如何发挥人的增值作用。目前，数据加工研究的主要挑战集中在以下几个方面。

（1）数据打磨或数据改写理念的提出：如何在数据科学项目中充分发挥数据科学家的作用，进而实现数据处理活动的增值效果；

（2）数据打磨或数据改写技术的实现：基于 Python、R 以及大数据技术实现数据加工的理念与方法；

（3）数据柔术：如何有艺术性地将数据转换为产品；

（4）整齐化处理：将数据转换为大数据算法和大数据技术能够直接处理的形态。

3.6 服务质量——精准度还是用户体验

查全率和查准率是传统数据研究中评价服务质量的两个核心指标。但是，当总体为未知、数据量迅速增长、数据种类不断变化和数据处理速度要求较高时，查全率和查准率的追求成为不可能。因此，在大数据环境下，更加重视用户体验，而不是查全率和查准率。在用户体验的评价中，响应速度是最为重要的指标之一。Aberdeen Group 的调查发现“页面的显示速度每延迟 1s，网站访问量就会降低 11%，从而导致营业额减少 7%，顾客满意度下降 16%”；Google 发现“响应时间每延迟 0.5s，查询数将会减少 20%”；Amazon 发现“响应时间每延迟 0.1s，营业额下降 1%”[88]。

目前，用户体验研究的主要挑战在于如何确保较快的响应速度，设计人机交互，实现服务虚拟化以及提供按需服务。

3.7 数据分析——解释性分析还是预测性分析

理论完美主义者认为，只有掌握了因果关系，才能正确认识和有效利用客观现象。传统数据分析往往是在理论完美主义的指导下完成的，试图通过对历史数据进行深度分析之后，达到深刻理解自我或解释客观现象的目的，其侧重的是因果分析，即以解释型分析为主。

在大数据环境下，数据分析的重点从因果分析转向相关分析，更加重视事物之间的相关关系[89]。然而，这种变革的背后是数据分析指导思想的根本性变化——从理论完美主义转向现实实用主义，侧重于数据分析的实用性，更加重视对未来的预测，即预测型分析。相对于解释性分析，预测性分析具有更强的时效性，可以迅速洞见事物之间的内在联系及其商业价值。

因此，数据科学的一个重要特点是预测性分析和解释性分析的分离。预测性分析主要由数据科学家完成，一般不需要领域知识；解释性分析则发生在预测性分析之后，数据科学家将预测性分析中的洞察结果转交给领域专家，由领域专家负责完成解释性分析。可见，数据科学家一般不做解释性分析，或者说，解释性分析往往超出数据科学家的能力范畴，需要由具体领域的专家完成。预测性分析和解释性分析的分离也是数据科学家和领域专家之间协同工作的主要实现方式。

大数据分析的主要挑战源自于数据的复杂性、噪声数据的分析、数据的依赖度[90]。提出面向大数据分析的新方法、技术与工具，尤其是大数据分析方法的动态演化、实时计算和弹性计算，成为相关研究中亟待解决的问题。

3.8 算法评价——复杂度还是可扩展性

复杂度，尤其是时间复杂度和空间复杂度，是传统算法的两个重要评价指标[91]，分别代表算法运行所需的时间成本和内存成本。但是，在大数据环境下，算法设计的一个重要特点是上层需求和底层数据处于动态变化之中，因此，算法应支持按需服务和数据驱动型应用。例如，Google 于 2008 年推出预测流感疫情工具——Google 流感趋势(Google Flu Trends，GFT)，通过该工具及时准确地预测了当时 H1N1 在全美范围的传播[92]，但是，2013 年 1 月的估计比实际数据高两倍，主要原因之一是缺乏算法动态性(Algorithm Dynamics)和用户使用行为习惯的变化[93]。

在大数据时代，算法的可扩展性主要代表算法的可伸缩能力。目前，相关研究的主要挑战在于低维度算法在高维数据中的应用、维度灾难、数据规约以及数据密集型应用。

3.9 研究范式——第三范式还是第四范式

图灵奖获得者 Jim Gray 曾提出，人类科学研究活动已经历过 3 种不同范式的演变过程(原始社会的“实验科学范式”、以模型和归纳为特征的“理论科学范式”和以模拟仿真为特征的“计算科学范式”)，目前正从“计算科学范式”转向“数据密集型科学发现范式(Data-intensive Scientific Discovery)”。第四范式，即“数据密集型科学发现范式”的主要特点是科学研究人员只需要从大数据中查找和挖掘所需要的信息和知识，无须直接面对所研究的物理对象。例如，在大数据时代，天文学家的研究方式发生了新的变化——其主要研究任务变为从海量数据库中发现所需的物体或现象的照片，而不再需要亲自进行太空拍照[94]。

第四范式的提出反映了人们对世界的固有认识发生了根本性的变化——从二元认识(精神世界/物理世界)转向三元认识(精神世界/数据世界/物理世界)，即在原有的“精神世界”和“物理世界”之间出现了一个新的世界——数据世界。因此，科学研究者往往直接面对的是数据世界，通过对数据世界的研究达到认识和改造物理世界的目的。对于科学研究者而言，数据世界中已积累的“历史数据”往往足以完成一项科研任务，数据科学家不需要亲自到物理世界采用问卷和访谈的方法收集数据——“调研数据”。同时，与“调研数据”相比，“历史数据”更具有客观性和可信度。目前，相关研究的主要挑战在于第三范式与第四范式的区别、第四范式的内涵、理论深入研究以及领域应用。

3.10 人才培养——数据工程师还是数据科学家

传统科学领域中，数据相关的人才培养的目标定位于数据工程师——从事数据的组织、管理、备份、恢复工作的人才。但是，在大数据时代，数据工程师无法胜任数据科学的研究任务，需要一类全新的人才——数据科学家。二者的主要区别在于：数据工程师负责的是数据的管理，而数据科学家擅长的是基于数据的管理，如基于数据的决策、产品开发、业务定义等。

目前，关于数据科学家的研究及人才培养的挑战在于正确分析岗位职责与用人需求、数据科学家的素质与能力要求、数据科学项目管理以及数据科学家的职业规划。

4 数据科学研究的发展趋势

在梳理研究热点、争议及挑战的基础上，我们需要进一步分析数据科学研究的发展趋势。从整体上讲，数据科学研究的主要发展趋势可以总结为以下几点。

(1)“思维模式的多样化和研究范式的变迁”是根本趋势。其中，思维模式的多样化主要体现在数据范式的兴起及其与传统的知识范式并存；研究范式的变迁是指科学研究范式从“计算科学范式”转向“数据密集型科学发现范式”，进而改变人们对世界的二元认识，相关研究重点将转变为通过数据世界的研究认识和改造物理世界。思维模式的多样化和研究范式的变迁对数据科学研究产生深远影响，将改变人们对数据的认识视角、开发动因和利用方式。

(2)“专业中的数据科学”是研究热点。大数据时代，各专

业领域面临的主要挑战在于如何解决新兴数据与传统知识之间的矛盾，即数据已经变了，但知识没有更新，各学科中的传统知识无法解决大数据带来的新问题。因此，大数据时代的机遇与挑战即将成为各学科领域研究的新方向，也就是说，专业中的数据科学成为相关研究的热点问题。

(3)“专业数据科学”是研究难点。“专业中的数据科学”从不同专业视角解读数据科学，存在研究兴趣点和研究发现(如理论、方法、技术、工具和典型实践等)的差异性，甚至可能出现相互重叠与冲突的现象。在这种背景下，如何将分散在不同学科领域中的共性问题及通用结论提炼成一门新的学科——“专业数据科学”，进而为各个学科领域的研究提供新的理论基础，是未来研究的难点所在。

(4)“数据生态系统的建设”是终极问题。数据学科是一门实践性极强的学科，其研究和应用均不能脱离具体领域。数据科学的研究和应用将会超出技术范畴，还涉及到发展战略、基础设施、人力资源、政策、法律与文化环境等诸多因素。因此，数据科学需要解决的终极问题是将大数据放在一个完整的生态系统之中进行认识与利用，从生态系统层次统筹和规划，避免片面认识数据问题，进而推动数据、能源和物质之间的相互转化。

4.1 预测模型及相关分析的重视

数据科学的研究责任在于预测模型，而不在于解释模型。以预测模型为中心的数据科学更偏向于实用主义，更加关注“对未来的预测能力”，而不是“对过去的解释水平”[95]。因此，数据科学的研究更加重视“现在能为未来做什么”，而不是“过去对现在的影响是什么”。

数据科学中重视预测模型而不是解释模型的另一个现实基础在于“人们往往先发现规律，后发现原因”[96]。从方法论层次看，以发现预测模型为目的的研究往往提倡的是假设演绎(Hypothetico-Deductive)研究范式[97]，先提出研究假设，然后采用试验设计和演绎分析方法论证研究假设成立与否。然而，一个好的研究假设的提出依赖于研究者尤其是数据科学家的特有素质——创造力、批判性思考和好奇心。

与解释模型不同的是，预测模型更加重视模型的简单性，而不是复杂性，主要原因有两个：1)预测模型对计算时间的要求较高，甚至需要进行实时分析，然而简单模型的计算效率往往高于复杂模型；2)经验证明，正如奥卡姆剃刀定律(Occam's razor)[98]所言，在其他条件相同的情况下，就预测而言，简单模型比复杂模型更可靠。

预测模型往往基于相关关系，而不是因果关系。通常，相关关系可以帮助我们预测未来，而因果关系有助于进一步理解和控制未来。从表面上看，预测模型依赖的是相关关系的分析，但在本质上属于一种数据驱动型的“数据范式”，与基于知识范式的解释模型有着本质性的区别。

4.2 模型集成及元分析的兴起

传统数据分析的通用做法是用一个数据模型即可解决一项数据处理任务。在这种以单一模型为基础的数据分析中，为了提升数据处理的信度和效度，需要对模型进行优化和调整，这会提升数据模型的复杂度。也就是说，传统数据分析中的数据模型有两个基本特征：单一性和复杂性。

但是，在大数据背景下，很难找到一个能够处理动态且异构数据的单一模型，因此，人们开始寻求多个模型的集成应用。与传统数据分析不同的是，大数据分析中所涉及的模型往往极其简单，即大数据分析中的数据模型也有两个基本特征：多样性和简单性。

可见，模型集成成为数据科学研究的一个新问题。通常，大数据分析采用多个较为简单的数据模型，将数据分析任务分解成分散在多个层次、多个活动中的小任务，并通过简单模型及其集成方法达到最终数据处理的目的。例如，在深度学习中，由多处理层组成的计算模型可通过多层抽象来学习数据表征[99]。

模型集成的背后是元分析的兴起。传统统计学重视基于零次或一次数据的基本分析，包括描述性统计、参数估计和假设检验。在大数据环境下，二次数据和三次数据的分析显得更为重要，数据分析工作往往在众多小模型的分析结果的基础上进行二次分析，即元分析。

4.3 数据在先、模式在后或无模式的出现

传统数据管理，尤其是关系型数据库中采用的是“模式在先、数据在后(Schema First，Data Later)”的建设模式[100]，即先定义模式，然后严格按照模式的要求存储和管理数据；当需要调整模式时，不仅需要重定义数据结构，而且还需要修改上层应用程序。然而，在大数据环境下，无法沿用“模式在先、数据在后”的建设模式，主要原因有两个：1)数据模式可能不断变化或根本不存在；2)按照预定模式进行数据的存储和处理时，容易出现信息丢失。

因此，“数据在先、模式在后或无模式(Data First，Schema Later or Never)”成为数据产品设计的主要趋势。以 NoSQL 为例，采用非常简单的键值数据模型，通过模式在后(Schema Later)或无模式(Schemaless)的方式确保数据管理系统的敏捷性。当然，模式在后或无模式也会带来新问题，如限制数据管理系统的处理能力及加大应用系统的开发难度。

在“数据在先、模式在后或无模式”的兴起背后，是信息系统建设模式的历史性变革——从先行支付(Pay-before-you-go)转向现收现付(Pay-as-you-go)的建设模式。信息系统建设中的先行支付模式的特点是根据特定时间点的需求定义信息系统，信息系统一旦开发完毕，便在一定时间内相对稳定。先行支付模式的缺点在于无法适应底层数据的复杂性和上层应用的动态变化。

4.4 数据一致性及现实主义的回归

在传统数据管理中，对数据一致性的要求接近于完美主义——强一致性，即任何时候从任何地方读出的任何数据均为正确数据。为了保证数据的一致性，在关系数据库中引入了事务、两端封锁协议和两端提交协议等方法或机制。强一致性的优点在于不仅可以保证数据质量，而且可以降低后续计算的成本。但是，强一致性不符合大数据时代的数据管理要求——高扩展性、高性能、高容错性、高伸缩性和高经济性。

因此，NoSQL 等新兴数据管理技术从根本上改变了人们对数据一致性的传统认识，主要表现在提出 CAP 理论和 BASE 原则等新兴数据管理理念，引入弱一致性、最终一致性等概念，并提供了不同的解决方案，如更新一致性、读写一致

性和会话一致性等。可见，在数据科学研究中，数据的一致性出现了多样化趋势，即根据不同应用场景，有针对性地选择具体的一致性及其实现方法。

对数据一致性的多样化认识的转变反映了人们对数据管理目标的根本转折——从完美主义回归至现实主义。以 CAP 理论[101]为例，人们对分布式系统的设计目的发生了改变，不再追求强一致性（Consistency）、可用性（Availability）和分区容错性（Partition Tolerance）3 个指标的同时最优，反而意识到了三者中的任何两个特征的保证（或争取）可能导致另一个特征的损失（或放弃）。例如，Cassadra 和 Dynamo 为了争取可用性和分区容错性而放弃了一致性。

4.5　多副本技术及靠近数据原则的应用

传统关系数据库更加看重数据冗余的负面影响——冗余数据导致的数据一致性保障成本较高。与此不同的是，数据科学中更加重视冗余数据的积极作用，即冗余数据在负载均衡、灾难恢复和完整性检验中的积极作用。同时，还通过引入多副本技术和物化视图的方法丰富冗余数据的存在形式，缩短用户请求的响应时间，确保了良好的用户体验。以 Google 搜索为例，采用缓存和照相（images）技术重复利用搜索结果。

同时，在计算和应用系统的部署上，改变传统的“数据靠近计算的原则”，反而开始采取“计算靠近数据的原则”。例如，Spark 系统提供了操作 getPreferredLocations（），支持 RDD 的本地化计算[102]；在 MapReduce 中，尽量将 Map 任务调度至存放了副本数据的机器上。可见，多副本技术和靠近数据原则均表明传统的“以计算为中心”的产品部署模式正向“以数据为中心”的产品部署模式转变。

4.6　多样化技术及一体化应用并存

传统关系数据库类产品虽多，但标准化程度较高，如均采用关系模型和 SQL 语言。但是，新兴的 NoSQL 数据库代表的不是一种特定技术，而是包括基于不同数据模型和查询接口的多种数据管理技术，如 Key-Value、Key-Document、Key-Column 以及图存储模型等。可见，在技术实现层次上，新兴技术表现出了多样化发展及高度专业化的趋势，即一项新技术专注于一个问题、一项功能或一种应用场景。例如，MapReduce，Tez，Storm，Druid 等技术的定位相对单一，分别专注于分布式批处理、Map/Reduce 过程的拆分与组合、实时处理和面向 OLAP 的列存储等较为单一功能的实现。当然，Spark，YARN 等较为通用性技术的出现也为技术层次上的高度专业化趋势提供了一种补充的解决方案。

同时，在传统数据计算/管理环境中，不同数据产品的界限是比较清楚的，所依赖的技术也是单一的，要么是关系模型，要么是层次或网状模型。但是，大数据时代的到来导致不同计算/管理技术的高度融合，进而出现一些支持多种数据计算/管理技术集成产品，甚至显现出了软硬件一体化或嵌入式应用趋势。例如，Oracle 大数据解决方案（Big Data Appliance）[103]集成了 HDFS、Oracle NoSQL、Cloudera CDH、数据仓库、内存计算和分析型应用。

可见，在数据科学研究中，一体化应用和专业化趋势并存。在产品与服务的实现层次上，一体化趋势越来越显著，一种产品的实现往往涉及多种不同技术的集成应用；在技术本身的实现层面，专业化趋势成为主流，一项新技术专注于解决相对单一的问题。

4.7　简单计算及实用主义占据主导地位

“简单”是数据科学的基本原则之一，代表着采用相对简单的技术来应对复杂的基础数据及不断变化的应用场景。与此不同的是，传统数据管理中采用的技术实现往往较为复杂。例如，传统关系数据库技术采用 Join 运算实现了多表查询等复杂操作。但是，这些复杂操作反而成为了关系数据库在提升数据管理能力方面的一个重要瓶颈。如 Join 操作要求被处理数据不能分布在不同节点，为此，NoSQL 放弃了 Join 等复杂处理操作，突出了简单计算较高的效率和较好的效果。

从复杂计算到简单计算的转变表明，人们对数据产品开发的理念从完美主义回归至实用主义。数据科学是一门实践性很强的学科，现阶段主要关注的是实用性，即解决当前社会亟待解决的实际问题，而不是复杂计算的实现。

4.8　数据产品开发及数据科学的嵌入式应用

作为数据科学的特有研究内容，数据产品开发将成为未来研究的重要课题。在数据科学中，所谓的数据产品（Data Products）并不限于“数据形态”的产品，而泛指“能够通过数据来帮助用户实现其某一个（些）目标的产品”[104]。可见，数据产品是指在数据科学项目中形成，能够被人、计算机以及其他软硬件系统消费、调用或使用，并满足他们（它们）某种需求的任何产品，包括数据集、文档、知识库、应用系统、硬件系统、服务、洞见、决策及它们的各种组合。以 Google 眼镜为例，虽然从产品形态上看其似乎是“眼镜类产品”，但从主要竞争力之源看，其确实属于“数据产品”。

数据产品开发主要关注如何将数据科学的理论融入传统产品开发实践之中，进而实现产品的更新换代和用户体验的提升。未来，数据产品开发将嵌入至传统产品的研发之中，二者的界限会越来越模糊。如何将数据科学家的创造性设计、批判性思考和好奇性提问的职业素质融入产品研发之中，从而实现传统产品的增值和核心竞争力的提升，是未来数据产品开发的难点所在。在此背景下，以数据为中心的设计思维将会成为数据产品开发的主要思维模式。同时，良好的用户体验将成为产品开发的主要评价指标之一。

数据产品开发的兴起将推动数据科学的嵌入式应用。数据科学将作为传统产品的创新点、增值点和竞争力之源，成为产品开发的必要环节，数据科学与领域呈现出了高度融合的趋势。

4.9　专家余及公众数据科学的兴起

在传统数据分析中，专家尤其是领域专家是知识的主要来源之一。例如，本体的建设需要由领域专家完成；专家系统中的知识库建立在专家的知识之上。但是，在大数据时代，专家余（ProAm）[105]成为数据处理项目的主要贡献者。与专家不同的是，专家余是指其能力在专家与业务之间的准专家型人群。近年来，众包（包括众创、众筹等）成为大数据时代的重要数据处理模式，其主要参与者均为专家余，而并非是严格意义上的专家或业余人群。例如，与传统意义上的专家编写的百科全书不同，Wikipedia 是由来自各领域的专家余共同完成的知识库。

众包的广泛应用为传统知识库建设中的数据量与形式化程度之间的矛盾提供了新的解决方案。在传统知识库建设中,要么形式化程度高,但数据量不够;要么数据量足够,但形式化程度不高。众包数据处理模式的出现使位于数据链长尾的专家余成为知识的主要贡献者和积极参与者。从协同方式看,众包中大规模协同可以分为机器协同、人机协同和人际协同3种表现形式。其中,人机协同是数据科学研究的重要课题。例如,混合智能——人与机器的互补型智能正成为人工智能的新课题。再如,语义Web技术的出现为人机协同提供了一种重要的技术支撑。

公众数据科学(Citizen Data Science)是专家余和大规模协同在数据科学领域的应用的主要表现形式之一。所谓的公众数据科学属于公众科学(Citizen Science),是指公众参与的数据科学,它与数据科学(Data Science)的区别在于参与研究者以非职业的兴趣爱好者和志愿者为主。也就是说,公众数据科学是一种基于众包和专家余的准数据科学,也是在数据科学成为一门广为接受的正式科学之前的过渡型理论。

4.10 数据科学家与人才培养的探讨

数据科学项目任务往往是富有挑战性的工作,每一项任务都是独一无二的,对工作人员的要求超出数据工程师的能力范畴,亟待一类新型人才——数据科学家来承担。从Drew Convey的数据科学维恩图[106]可以看出,数据科学具有3个基本要素,即理论(统计学与数学知识)、实践(领域实战)和精神(黑客精神)。可见,数据科学与传统科学的人才需求不同,前者不仅要求传统科学中的理论与实践,而且还需要有数据科学家的"精神"素质,即原创性设计、批判性思考和好奇心提问的能力。

因此,如何培养"理论、实践和精神为一体"的综合性人才是未来研究的重要课题。相关研究主要从以下4个层面开展:1)办学层次,即如何培养本科[107]、硕士[108]和博士[109]层次的数据科学人才。目前,国内和国外对数据科学人才培养层次的关注点不同,分别关注的是本科层次和硕士层次人才的培养,但对博士层次人才的讨论相对少。2)专业设置,即是否需要设立数据科学专业?例如,国内主要讨论的是如何建设"数据科学与大数据技术"专业。3)学科方向的选择,即如何将数据科学与传统学科相结合,确定数据科学的学科地位。4)课程改革[110],即如何完成传统课程的改革以及数据科学新课程的创造性设计。

结束语 数据科学是一门极其特殊的新兴学科,具有与其他学科不同的新特征,例如思维模式的转变(从数据范式到知识范式的转变)、对数据认识的变化(从数据的被动属性到主动属性的转移)、指导思想的变化(实用主义和现实主义的回归)、以数据产品开发为主要目的(数据成为传统产品的主要创新点)、专业数据科学与专业中的数据科学的差异性以及数据科学的三要素(不仅涉及理论和实践,而且还包括精神素质)。因此,数据科学的研究不能简单照搬传统学科的经验,应尊重其特殊使命和属性。为此,我们对数据科学研究者提出如下几点建议:

(1)正确认识数据科学。正确认识数据科学的内涵是有效学习和规范研究数据科学的前提。目前,部分学者误以为"数据科学=统计学+机器学习",过于强调统计学和机器学习,而忽略了数据科学本身。其实,统计学和机器学习是数据科学的理论基础,而并非其核心内容。数据科学具有区别于其他学科的独特的研究使命、研究视角、思维模式、做事原则和知识体系。如果脱离了这些独到之处,数据科学的学习和研究将发生方向性的误读和本质性的扭曲。

(2)突出数据的主动属性。数据科学的一个重要贡献或价值就在于它改变了人们对数据的研究方向,即从被动属性转向主动属性。一直以来,人们习惯性地把数据当作被动或死的东西,关注的是"你能对数据做什么",如模式定义、结构化处理和预处理,都试图将复杂数据转换成简单数据。但是,大数据时代更加关注数据的另一个属性——主动属性,强调的是"数据能给你带来什么",如数据驱动型应用、以数据为中心的设计、让数据说话、数据洞见等,将复杂性认为是数据的自然属性,开始接受数据的复杂性。研究方向从数据的被动属性到主动属性的转变,是学习和研究这门新学科的基本出发点。如果忽略了这一点,容易将数据科学当成数据工程来学习和研究。

(3)平衡数据科学的3个要素。与其他课程尤其是技术类课程不同的是,数据科学既包括理论和实践,又需要精神——原创性设计、批判性思考和好奇性提问的素质。因此,数据科学的学习中不仅要强调理论联系实际,而且还不能忽略对数据科学家精神的培养。积极参与数据科学相关的开源项目和竞赛类项目,是兼顾数据科学的3个基本要素的两条重要捷径。

(4)侧重培养信心和兴趣,学会跟踪数据科学的最新动态。一方面,数据科学建立在统计学和机器学习等基础理论之上,学习门槛较高,因此,培育自己对数据科学的学习信心和兴趣尤为重要;另一方面,数据科学仍属于一门快速发展的新兴学科,其理念、理论、方法、技术和工具在不断变化,要求我们必须掌握动态跟踪数据科学领域的国际顶级会议、重要学术期刊、主要研究机构、代表性人物和标志性实践的能力。

(5)重视试验设计及假设检验。试验设计是数据科学项目的重要活动之一。数据科学家应根据数据科学项目的研究目的,有创造性地提出研究假设,并设计对应的试验,最终通过这些试验达到假设检验的目的。以华盛顿大学和加州大学伯克利分校的数据科学专业人才培养方案为例,其分别开出了课程应用统计与试验设计(Applied Statistics & Experimental Design)和试验与因果分析(Experiments and Causality),以重点培养学生的试验设计和假设检验的能力。

(6)不要忽视因果分析。在大数据时代,很多人误以为"因果分析不再重要了",并把研究重点局限于相关分析。相关分析只能用于识别事物之间的关联关系,而无法指导如何优化和干预这种相关关系。因此,当相关关系发生变化或需要人为干预相关关系时,必须进一步研究其因果关系。在数据科学项目中,数据科学家的关注重点是发现各种可能的关联关系,而关联关系的产生机制和优化方法需要由领域专家完成。加州大学伯克利分校和哥伦比亚大学分别开设课程实验与因果分析(Experiments and Causality)和因果推理与数据科学(Causal Inference for Data Science),均反映了因果分

析在数据科学中的重要地位。

（7）以数据产品开发为主要抓手。数据产品开发是学习与研究数据科学的主要抓手之一。需要注意的是，数据产品不限于数据形态的产品，任何用数据来帮助目标用户实现其某一目的的产品都可以被视为数据产品。数据是未来产品的创新点和增值点。因此，向数据产品的转变是传统产品的重要发展趋势。以 Google 眼镜[111]为例，其创新源自于数据，而不在于其外观和选材，以数据为中心的产品设计才是该产品与传统眼镜类产品的根本区别。可见，数据产品开发是数据科学的最为直接且最为普遍的应用。

（8）准确定位人才培养目的。数据科学的学习和人才培养的目的是培养数据科学家而不是数据工程师。二者的区别在于，数据工程师负责的是“数据本身的管理”，而数据科学家的主要职责是“基于数据的管理”，包括基于数据的分析、决策、流程定义与再造、产品设计和服务提供等。因此，相对于数据工程师，数据科学家对人才的要求更高，不仅要有理论功底和实践经验，而且还要求有精神素质，即创造性设计、批判性思考和好奇性提问的能力。

参 考 文 献

[1] WALKER J S, NAIMI AI. Big data: A revolution that will transform how we live, work, and think[J]. Mathematics & Computer Education, 2013, 47(17): 181-183.

[2] BOYD D, CRAWFORD K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon[J]. Information, Communication & Society, 2012, 15(5): 662-679.

[3] KITCHIN R. Big data, new epistemologies and paradigm shifts [J]. Big Data & Society, 2014, 1(1): 1-12.

[4] JAGADISH H V. Big data and science: myths and reality[J]. Big Data Research, 2015, 2(2): 49-52.

[5] PROVOST F, FAWCETT T. Data science and its relationship to big data and data-driven decision making[J]. Big Data, 2013, 1(1): 51-59.

[6] NAUR P. Concise survey of computer methods[M]. Studentlitteratur AB, 1974.

[7] CLEVELAND W S. Data science: an action plan for expanding the technical areas of the field of statistics[J]. International Statistical Review, 2001, 69(1): 21-26.

[8] MATTMANN C A. Computing: A vision for data science[J]. Nature, 2013, 493(7433): 473-475.

[9] DHAR V. Data science and prediction[J]. Communications of the ACM, 2013, 56(12): 64-73.

[10] PATIL D J, DAVENPORT T H. Data scientist: the sexiest job of the 21st century[J]. Harvard Business Review, 2012, 90(10): 70-76.

[11] KITCHIN R. Big data and human geography: Opportunities, challenges and risks[J]. Dialogues in Human Geography, 2013, 3(3): 262-267.

[12] SMITH M. The White House names Dr, DJ Patil as the first US chief data scientist[OL]. https://obamawhitehouse.archives.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist.

[13] GARTNER J. Gartner's 2014 hype cycle for emerging technologies maps the journey to digital business[OL]. http://www.gartner.com/newsroom/id/2819918.

[14] GARTNER J. Hype Cycle for Data Science[OL]. https://www.gartner.com/doc/3388917/hype-cycle-data-science.

[15] SCHUTT R, O'NEIL C. Doing data science: Straight talk from the frontline[M]. O'Reilly Media, Inc., 2013: 7.

[16] OVERTON J. Going Pro in Data Science [M]. O'Reilly Media, Inc., 2016: 12.

[17] 朝乐门. 数据科学理论与实践[M]. 北京：清华大学出版社，2017：15.

[18] GRAY J, CHAMBERS L, BOUNEGRU L. The data journalism handbook: how journalists can use data to improve the news [M]. O'Reilly Media, Inc., 2012.

[19] KALIDINDI S R, DE GRAEF M. Materials data science: current status and future outlook[J]. Annual Review of Materials Research, 2015, 45: 171-193.

[20] FANG B, ZHANG P. Big Data in Finance[M]// Big Data Concepts, Theories, and Applications. Springer International Publishing, 2016: 391-412.

[21] DAVIS K. Ethics of Big Data: Balancing risk and innovation [M]. O'Reilly Media, Inc., 2012.

[22] WEST D M. Big data for education: Data mining, data analytics, and web dashboards [J]. Governance Studies at Brookings, 2012, 4: 1-10.

[23] LABRINIDIS A, JAGADISH H V. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment, 2012, 5 (12): 2032-2033.

[24] KAISLER S, ARMOUR F, ESPINOSA J A, et al. Big data: Issues and challenges moving forward[C]// 2013 46th Hawaii International Conference on System Sciences (HICSS). IEEE, 2013: 995-1004.

[25] CHEN H, CHIANG R H L, STOREY V C. Business intelligence and analytics: From big data to big impact[J]. MIS Quarterly, 2012, 36(4): 1164-1188.

[26] PROVOST F, FAWCETT T. Data science and its relationship to big data and data-driven decision making[J]. Big Data, 2013, 1(1): 51-59.

[27] CLEVELAND W S. Data science: an action plan for expanding the technical areas of the field of statistics[J]. International Statistical Review, 2001, 69(1): 21-26.

[28] MATTMANN C A. Computing: A vision for data science[J]. Nature, 2013, 493(7433): 473-475.

[29] SCHUTT R, O'NEIL C. Doing data science: Straight talk from the frontline[M]. O'Reilly Media, Inc., 2013.

[30] SHANAHAN J G, DAI L. Large scale distributed data science using apache spark[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 2323-2324.

[31] HOLMES A. Hadoop in practice[M]. Manning Publications Co., 2012.

[32] SHARMA S, SHANDILYA R, PATNAIK S, et al. Leading NoSQL models for handling Big Data: a brief review[J]. International Journal of Business Information Systems, 2016, 22(1): 1-25.

[33] SADALAGE P J, FOWLER M. NoSQL distilled: a brief guide to the emerging world of polyglot persistence[M]. Pearson Education, 2012.

[34] MARX V. Biology: The big challenges of big data[J]. Nature, 2013, 498(7453): 255-260.

[35] RAGHUPATHI W, RAGHUPATHI V. Big data analytics in healthcare: promise and potential[J]. Health Information Science and Systems, 2014, 2(1): 3.

[36] KIM G H, TRIMI S, CHUNG J H. Big-data applications in the government sector[J]. Communications of the ACM, 2014, 57(3): 78-85.

[37] DANIEL B. Big data and analytics in higher education: Opportunities and challenges[J]. British Journal of Educational Technology, 2015, 46(5): 904-920.

[38] GEORGE G, HAAS M R, PENTLAND A. Big data and management[J]. Academy of Management Journal, 2014, 57(2): 321-326.

[39] SWAN M. The quantified self: Fundamental disruption in big data science and biological discovery[J]. Big Data, 2013, 1(2): 85-99.

[40] LEWIS S C. Journalism in an Era of Big Data: Cases, concepts, and critiques[OL]. https://doi.org/10.1080/21670811.2014.976399.

[41] RAHM E. Big Data Analytics[J]. IT-Information Technology, 2016, 58(4): 155-156.

[42] BAUMER B. A data science course for undergraduates: Thinking with data[J]. The American Statistician, 2015, 69(4): 334-342.

[43] HARDIN J, HOERL R, HORTON N J, et al. Data science in statistics curricula: Preparing students to "think with data"[J]. The American Statistician, 2015, 69(4): 343-353.

[44] CASSEL L N, POSNER M, DICHEVA D, et al. Advancing data science for students of all majors[C]// Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education. ACM, 2017: 722.

[45] BERMAN F D, BOURNE P E. Let's make gender diversity in data science a priority right from the start[J]. PLoS biology, 2015, 13(7): e1002206.

[46] CHAO L. Data Science [M]. Tsinghua University Press, 2016.

[47] COOPER P. Data, information, knowledge and wisdom[J]. Anaesthesia & Intensive Care Medicine, 2014, 15(1): 44-45.

[48] ERL T, KHATTAK W, BUHLER P. Big data fundamentals: concepts, drivers & techniques[M]. Prentice Hall Press, 2016.

[49] WANG G, GUNASEKARAN A, NGAI E W T, et al. Big data analytics in logistics and supply chain management: Certain investigations for research and applications[J]. International Journal of Production Economics, 2016, 176: 98-110.

[50] CARDENAS A A, MANADHATA P K, RAJAN S P. Big data analytics for security[J]. IEEE Security & Privacy, 2013, 11(6): 74-76.

[51] RAGHUPATHI W, RAGHUPATHI V. Big data analytics in healthcare: promise and potential[J]. Health Information Science and Systems, 2014, 2(1): 3.

[52] LEEK J T, PENG R D. What is the question? Mistaking the type of question being considered is the most common error in data analysis[J]. Science, 2015, 374(6228): 1314-1315.

[53] SWAN M. The quantified self: Fundamental disruption in big data science and biological discovery[J]. Big Data, 2013, 1(2): 85-99.

[54] RUCKENSTEIN M, PANTZAR M. Beyond the quantified self: Thematic exploration of a dataistic paradigm[J]. New Media & Society, 2017, 19(3): 401-418.

[55] KHATRI V, BROWN C V. Designing data governance[J]. Communications of the ACM, 2010, 53(1): 148-152.

[56] KHATRI V, BROWN C V. Designing data governance[J]. Communications of the ACM, 2010, 53(1): 148-152.

[57] THOMAS G. The DGI data governance framework[OL]. http://www.datagovernance/the-dgi-framework.

[58] LEE S U, ZHU L, JEFFERY R. Design Choices for Data Governance in Platform Ecosystems: A Contingency Model[J]. arXiv preprint arXiv: 1706.07560, 2017.

[59] CMMI Institute. Data Management Maturity (DMM)? Model [OL]. http://cmmiinstitute.com/data-management-maturity.

[60] LIU J, LI J, LI W, et al. Rethinking big data: A review on the data quality and usage issues[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 115: 134-142.

[61] LI J Z, WANG H Z, GAO H. State-of-the-Art of Research on Big Data Usability[J]. Journal of Software, 2016, 27(7): 1605-1625. (in Chinese)
李建中, 王宏志, 高宏. 大数据可用性的研究进展[J]. 软件学报, 2016, 27(7): 1605-1625.

[62] RAHM E, DO H H. Data cleaning: Problems and current approaches[J]. IEEE Data Engineering Bulletin, 2000, 23(4): 3-13.

[63] WICKHAM H. Tidy data[J]. Journal of Statistical Software, 2014, 59(10): 1-23.

[64] LAFUENTE G. The big data security challenge[J]. Network Security, 2015, 2015(1): 12-14.

[65] PERERA C, RANJAN R, WANG L, et al. Big data privacy in the internet of things era[J]. IT Professional, 2015, 17(3): 32-39.

[66] PATIL D, NOREN A. Building Data Science Teams: The Skills, Tools and Perspectives Behind Great Data Science Groups[M]. O'Reilly, 2011.

[67] BANERJEE S. Citizen Data Science for Social Good: Case Studies and Vignettes from Recent Projects[OL]. https://www.researchgate.net/publication/283119007_Citizen_Data_Science_for_Social_Good_Case_Studies_and_Vignettes_from_Recent_Projects.

[68] PARASIE S, DAGIRAL E. Data-driven journalism and the public good: "Computer-assisted-reporters" and "programmer-journalists" in Chicago[J]. New Media & Society, 2013, 15(6): 853-871.

[69] DU D, LI A, ZHANG L. Survey on the applications of big data in Chinese real estate enterprise[J]. Procedia Computer Science, 2014, 30: 24-33.

[70] MIDDLETON S E, SHADBOLT N R, DE ROURE D C. Ontological user profiling in recommender systems[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 54-88.

[71] MARSHALL P, TODD B, RHODES M. Ultimate Guide to Google AdWords[M]. Entrepreneur Press, 2014.

[72] GURRIN C, SMEATON A F, DOHERTY A R. Lifelogging: Personal big data[J]. Foundations and Trends® in Information Retrieval, 2014, 8(1): 1-125.

[73] RAGHUPATHI W，RAGHUPATHI V. Big data analytics in healthcare：promise and potential [J]. Health Information Science and Systems，2014，2(1)：3.

[74] MARX V，Biology：The big challenges of big data[J]. Nature，2013，498(7453)：255-260.

[75] BELLO-ORGAZ G，JUNG J J，CAMACHO D. Social big data：Recent achievements and new challenges[J]. Information Fusion，2016，28：45-59.

[76] MOHANTY S，JAGADEESH M，SRIVATSA H. Big data imperatives：Enterprise 'Big Data' warehouse，'BI' implementations and analytics[M]. Apress，2013.

[77] BERTOT J C，GORHAM U，JAEGER P T，et al. Big data，open government and e-government：Issues，policies and recommendations[J]. Information Polity，2014，19(1/2)：5-16.

[78] AGGARWAL A. Opportunities and Challenges of Big Data in Public Sector[M]// Managing Big Data Integration in the Public Sector. 2015：289-301.

[79] MATT T. Big Data Landscape 2016 v18 FINAL[OL]. (2016-4-28). http：// mattturck. com/big-data-landscape-2016-v18-final.

[80] KAISLER S，ARMOUR F，ESPINOSA J A，et al. Big data：Issues and challenges moving forward[C]// 2013 46th Hawaii International Conference on System Sciences (HICSS). IEEE，2013：995-1004.

[81] AL-JARRAH，OMAR Y，et al. Efficient machine learning for big data：A review[J]. Big Data Research，2015，2(3)：87-93.

[82] BATRA S. Big data analytics and its reflections on DIKW hierarchy[J]. Review of Management，2014，4(1/2)：5.

[83] DONHOST M J，ANFARA J V A. Data-driven decision making [J]. Middle School Journal，2010，42(2)：56-63.

[84] CHEN C L P，ZHANG C Y. Data-intensive applications，challenges，techniques and technologies：A survey on Big Data[J]. Information Sciences，2014，275：314-347.

[85] VOULGARIZ Z，MAGOULAS G D. Extensions of the k nearest neighbour methods for classification problems[C]// Proc. of the 26th IASTED International Conference on Artificial Intelligence and Applications (AIA). Innsbruck，Austria，2008，13：23-28.

[86] Datawocky. More data usually beats better algorithms[OL]. (2008-03-24). http：// anand. typepad. com/datawocky/2008/03/more-data-usual. html.

[87] KLEPPMANN，MATRIN. Designing Data-Intensive Applications：The Big Ideas Behind Reliable，Scalable，and Maintainable Systems[M]. O'Reilly Media，Inc. ，2017.

[88] BREWER E. Parallelism in the Cloud[OL]. [2013-06-24]. https：// www. usenix. org/sites/default/files/conference/protected-files/brewer_hotpar13_slides. pdf.

[89] MCAFEE A，BRYNJOLFSSON E，DAVENPORT T H. Big data：the management revolution[J]. Harvard Business Review，2012，90(10)：60-68.

[90] FAN J Q，HAN F，LIU H. Challenges of big data analysis[J]. National Science Review，2014(1/2)：293-314.

[91] EDGAR，ROBERT C. MUSCLE：a multiple sequence alignment method with reduced time and space complexity[J]. BMC Bioinformatics，2004，5(1)：113.

[92] GINSBERG J，MOHEBBI M H，PATEL R S，et al. Detecting influenza epidemics using search engine query data[J]. Nature，2009，457(7232)：1012-1014.

[93] LAZER D，KENNEDY R，KING G，et al. The Parable of Google Flu：Traps in Big Data Analysis[J]. Science，2014，343(6176)：1203-1205.

[94] HEY T. The fourth paradigm：data-intensive scientific discovery [J]. Proceedings of the IEEE，2011，99(8)：1334-1337.

[95] PROVOST F，FAWCETT T. Data science and its relationship to big data and data-driven decision making[J]. Big Data，2013，1(1)：51-59.

[96] DHAR V，CHOU D. A comparison of nonlinear models for financial prediction[J]. IEEE Transactions on Neural Networks，2001，12(4)：907-921.

[97] FØLLESDAL，DAGFINN. Hermeneutics and the hypothetico-deductive method[J]. Dialectica，1979，33(3/4)：319-336.

[98] BLUMER A，EHRENFEUCHT A，HAUSSLER D，et al. Occam's razor[J]. Information Processing Letters，1987，24(6)：377-380.

[99] LECUN Y，BENGIO Y，HINTON G. Deep learning[J]. Nature，2015，521(7553)：436-444.

[100] LIU Z H，HAMMERSCHMIDT B，MCMAHON D. JSON data management：supporting schema-less development in RDBMS [C]// Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. ACM，2014：1247-1258.

[101] BREWER E. CAP twelve years later：How the "rules" have changed[J]. Computer，2012，45(2)：23-29.

[102] ZAHARIA M，CHOWDHURY M，FRANKLIN M J，et al. Spark：Cluster computing with working sets [J]. HotCloud，2010，10(10)：95.

[103] PLUNKETT，TOM，et al. Oracle Big Data Handbook[M]. McGraw-Hill Osborne Media，2013.

[104] PATIL D J. Data Jujitsu：the art of turning data into product [M]. O'Reilly Media，Inc. ，2012.

[105] LEADBEATER C，MILLER P. The Pro-Am revolution：How enthusiasts are changing our society and economy[M]. Demos，2004.

[106] CONWAY D. Data Science in the US Intelligence Community [J]. IQT Quarterly，2011，2(4)：24-27.

[107] ANDERSON P，MCGUFFEE J，UMINSKY D. Data science as an undergraduate degree[C]// Proceedings of the 45th ACM Technical Symposium on Computer Science Education. ACM，2014：705-706.

[108] MARSHALL L，ELOFF J H P. Towards an Interdisciplinary Master's Degree Programme in Big Data and Data Science：A South African Perspective [C]// Annual Conference of the Southern African Computer Lecturers' Association. Springer International Publishing，2016：131-139.

[109] SUGIMOTO C R，EKBIA H R，MATTIOLI M. The Data Gold Rush in Higher Education[M//. Big Data Is Not a Monolith. MIT Press，2016：129.

[110] ANDERSON P，BOWRING J，MCCAULEY R，et al. An undergraduate degree in data science：curriculum and a decade of implementation experience[C]// Proceedings of the 45th ACM Technical Symposium on Computer Science Education. ACM，2014：145-150.

[111] MUENSTERER O J，LACHER M，ZOELLER C，et al. Google Glass in pediatric surgery：an exploratory study [J]. International Journal of Surgery，2014，12(4)：281-289.