

数据科学进展:核心理论与典型实践^{*}

朝乐门 张 晨 孙智中

摘 要 数据科学经历了近 50 年的发展与变革,对知识创新和社会进步产生了深远的影响。现阶段对数据科学的研究,除应探讨数据科学的内涵和外延之外,还需要回答一系列深层次的问题:数据科学的核心理论和代表性实践是什么,它们是如何演化而来的,仍存在的问题和挑战是什么,未来的发展趋势是什么。为此,本文从核心理论和典型实践两个维度,将数据科学的发展历程划分为萌芽期、快速发展期和逐步成熟期三个阶段并进行总结分析。现阶段数据科学的理论研究和实践应用主要存在四类问题:一是对数据科学本身的系统研究不足;二是对数据科学领域的新问题聚焦不够;三是对数据科学领域的核心问题研究不足;四是在数据科学理论的直接指导下完成的实践应用不多。对此,提出了五点对策建议:聚焦数据科学特有的新问题和主要矛盾,健全数据科学理论研究的体系,加快数据科学实践应用的产业化,推进研究方法论的突破性创新,拓展数据科学的基础理论。图 1。参考文献 66。

关键词 数据科学 核心理论 典型实践 研究现状

分类号 TP391 G203

Development of Theoretical Studies and Practical Applications in Data Science

CHAO Lemen, ZHANG Chen & SUN Zhizhong

ABSTRACT

Since Peter Naur, a Turing Award winner, coined the term of “Data Science” in 1974, data science has been developed for nearly 50 years. In addition to the previous studies of data science, there are a series of deeper questions that need to be answered: What are the core theories and representative practices of data science? How did the theories and practices develop? What are the problems and challenges in data science research? What is the future development trend of data science? Therefore, from the two dimensions of theory and practice, the development of data science is divided into three stages: the introduction period, the growth period, and the maturity period. Further, latest achievements, main challenges, and emerging trends in data science are discussed. In the introduction period (1974–2009), the theoretical research of data science was mainly conducted by a few computer scientists and statisticians. They proposed the necessity of data science from the perspective of social development as well as the criticism of traditional disciplines and discussed some basic technologies and main research topics of data science. The quality of related theoretical studies and practical applications is pretty high though the number of them is relatively low.

^{*} 本文系国家自然科学基金项目“预测性分析结果的数据故事化描述方法及关键技术”(编号:72074214)的研究成果之一。(This article is an outcome of the project “Data Storytelling Method for Predictive Analysis Results and its Key Technologies” (No. 72074214) supported by the National Natural Science Foundation of China.)

通信作者:朝乐门,Email:chaolemen@ruc.edu.cn,ORCID:0000-0001-8963-7507 (Correspondence should be addressed to CHAO Lemen, Email:chaolemen@ruc.edu.cn, ORCID:0000-0001-8963-7507)

Most of the findings in the introduction stage have deep inferences on the evolution of data science. In the growth period (2010–2014), the number of participants in data science research and industry has grown rapidly, and their research interest exceeded the discussion of the concepts or its indispensability and has turned to explore deeper topics such as the distinct methods, technologies, tools, and applications in data science. The number of related research findings and practical applications has greatly increased, but the quality is not balanced. In the maturity period (2015–), Data science is evolving into an independent discipline from statistics as well as computer science. At this stage, the main concerns on data science not only focuses on academic research but also involves more topics such as industrial application, higher education, discipline construction, and data product development. Furthermore, the knowledge system of data science is gradually formed at this stage and people have reached a consensus on some basic issues of data science.

Besides, four typical categories of challenges in the academic research or practical application of data science are discussed. First, the studies on underline topics in data science itself are ignored; second, the discussions are not focused on emerging challenges and novel theories in data science; third, some essential topics on data science are missing; fourth, the number of practical applications in data science, which were implemented by data science theory, are small. Further, the emerging trends of data science are mainly concentrated in five aspects: to focus on distinct research questions and social challenges in the big data era, to establish the data science theoretical systems, to industrialize data science applications, to propose novel methodologies and to expend the theoretical foundations.

In conclusion, the theoretical studies and practical applications of data science have to concentrate on the unique research questions of data science itself, such as solving the data challenges in the new era, solving the contradiction between new data and old knowledge, and positioning the role of data science in the scientific knowledge system, taking advantages of data science methods and tools, enhancing the problem-solving capabilities of data science, and ensuring the continues development of data science. 1 fig. 66 refs.

KEY WORDS

Data science. Core theories. Typical practices. State of the arts.

0 引言

数据科学是一门以数据尤其是大数据为研究对象,并以数据统计、机器学习、数据可视化等为理论基础,主要研究数据加工、数据管理、数据计算、数据产品开发等活动的交叉性学科^[1]。目前,数据科学已成为学术研究和产业应用的热门话题,但该领域的一些核心问题尚未得以系统研究,如数据科学是如何发展起来的,数据科学的新理论有哪些,数据科学的代表性应用是什么,现阶段数据科学发展中存在哪

些问题,数据科学的未来发展趋势是什么,等等。为此,本文综合运用文献调研、文本分析、案例分析和实证研究等方法,深入探讨数据科学领域的理论研究进展、实践应用现状以及未来发展趋势。

本文的主要结构如下:第一部分主要探讨数据科学萌芽期(1974—2009)出现的核心理论与典型实践,以及它们对数据科学产生的深远影响;第二部分主要介绍数据科学快速发展期(2010—2014)在理论和实践方面取得的标志性成果及其在数据科学领域的重要地位;第三部分主要论述数据科学在逐步成熟期(2015—)取

得的最新理论成果和实践探索,以及在数据科学专业化发展中发挥的重要作用;第四部分分析讨论了数据科学在理论研究和实践应用方面存在的问题及对策建议。

1 数据科学的萌芽

数据科学萌芽于计算机科学和统计学两个学科。在萌芽期(1974—2009),对数据科学的研究主要集中于极少数的计算机科学家和统计学家,他们主要从对未来社会的展望和对传统学科的批判角度提出了建立数据科学新学科的必要性和可行性,并探讨了数据科学的基础技术和主要研究话题。与此同时,社会对数据科学人才的需求显著增长,Python、R、Hadoop、Tableau 等新技术不断涌现,Google 流感分析等区别于传统科学认知的实践取得成功,在这些背景之下,人们开始期待数据科学这一门新学科的快速兴起。

1.1 核心理论

在萌芽期,以下核心理论对数据科学的发展具有重要推动作用。

(1)数据科学概念的提出及阐释。数据科学的诞生可追溯至数据科学这一新术语的明确定义。1974年,著名计算机科学家、图灵奖获得者 Naur 在其著作《计算机方法的简明调研》(*Concise Survey of Computer Methods*)^[2]中首次明确提出了数据科学(Data Science)的概念,并进一步阐述了术语“数据科学”和“数据学”(Datalogy)的区别——前者是处理数据(问题)的科学,而后者侧重于数据处理过程及其在教育领域中的应用。遗憾的是,至今仍有很多人在混淆和曲解数据科学和数据学这两个概念。

(2)对传统统计学的批判及在统计学领域研究数据科学的必要性的讨论。与计算机科学家首次提出数据科学的概念同等重要的是,统计学家对传统统计学的批判进一步推动了数据科学理论的诞生。1993年,贝尔实验室的

Chambers 发表题为《或优或劣的统计,未来研究的选择》(*Greater or Lesser Statistics, A Choice for Future Research*)^[3]的文章,提出从数据中学习的新研究范式对于传统统计学产生的挑战。2001年,贝尔实验室的 Cleveland 发表论文《数据科学——拓展统计学技术领域的行动计划》(*Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics*)^[4],主张将数据科学作为统计学的一个新研究方向。统计学家对传统统计学的批判及对数据科学的重视,成为数据科学理论发展的重要推动因素。目前,统计学和计算机科学已成为数据科学的两个主要理论基础学科。

(3)探索性数据分析的提出。探索性数据分析(Exploratory Data Analysis)的提出改变了以描述性分析和验证性分析为主的传统统计学,目前已成为数据科学应用尤其是数据理解的重要手段。1977年, Tukey 发表著作《探索性数据分析》(*Exploratory Data Analysis*),强调使用数据来提出假设以进行检验,探索性数据分析应该与验证性数据分析同时进行^[5]。在数据科学中,探索性分析通常为数据分析的第一步,是预测性分析、诊断性分析和规范性分析等高级数据分析的基础。

(4)CAP 理论与 BASE 原则的提出。与传统的数据问题相关研究不同,数据科学研究注重实用的现实主义。2000年, Brewer 提出 CAP 理论,指出一个分布式系统不能同时满足一致性(Consistency)、可用性(Availability)和分区容错性(Partition tolerance)这三个需求,最多只能同时实现两个^[6]。2008年, Pritchett 提出 BASE 原则,即 Basically available(基本可用)、Soft state(软状态)和 Eventually consistent(最终一致性),表示虽然不能达到分布式系统的强一致性,但可以通过合适方式达到最终一致性^[7]。CAP 理论和 BASE 原则的提出是现实主义兴起的重要表现,强调数据的存储和管理活动应重视实用性,不再盲从理论完美主义。以 CAP 理论和 BASE 原则为代表的重视实用性的现实主义思

想的兴起,是数据科学区别于传统数据问题研究的重要标志之一。

(5)云计算技术的出现及其普及应用。云计算以其经济性、虚拟化、弹性计算和按需服务等独有思想,降低了人类对数据采集、处理、计算和管理的成本,提高了人们的数据利用能力,与物联网、移动互联网、新兴科学仪器和生产制造设备一同成为推动数据科学兴起的主要力量。自2003年起,Google公司先后发表了有关GFS、MapReduce和Bigtable技术的三篇论文:《Google文件系统》(*The Google File System*)^[8]、《MapReduce:简化大型集群上的数据处理》(*MapReduce: Simplified Data Processing on Large Clusters*)^[9]、《Bigtable:用于结构化数据的分布式存储系统》(*Bigtable: A Distributed Storage System for Structured Data*)^[10],奠定了云计算在数据存储、计算和管理领域应用的基础。2009年,Oskarsson主张应将NoSQL解释为“不仅仅是SQL”(Not Only SQL, NoSQL)^[11],强调NoSQL对传统关系数据库的补充作用。目前,云计算是数据科学中进行大数据采集、处理、计算和管理的主流技术之一。

(6)有关大数据的特征和重要性的学术讨论。大数据是数据科学的主要研究对象,而大数据与传统数据的区别在于大数据是数据科学研究的立足点。2001年,Laney提出了大数据的“3V”特征——海量(Volume)、高速度(Velocity)和多样性(Variety)^[12];2006年,Humby提出“数据是新石油”(Data is the new oil)^[13]的理念,强调数据跟石油一样具有极高的价值,且都在利用之前需要被开采、提炼和加工;2017年,《经济学家》再次提到“世界上最有价值的资源不再是石油,而是数据”^[14],引发了全球对大数据的广泛关注。有关大数据的特殊性和重要性的学术讨论,是数据科学领域早期的主要研究成果。

(7)深度学习的兴起。深度学习的出现为数据科学,尤其是机器视觉和自然语言处理的学术研究和实践应用,提供了一种全新的解决方案。自2006年Hinton等提出深度信念网

络^[15]以来,深度学习在大数据处理与分析,尤其是机器视觉、图像处理和自然语言处理领域,得到了广泛利用。目前,深度学习已成为数据科学领域处理非数值型数据的关键技术。

(8)数据密集型科学发现的提出。第四范式的提出不仅反映了人类科学研究范式的变化,而且为数据科学研究提供了独有的研究范式——数据密集型科学发现。2007年,Gray将科学研究分为四种范式——实验科学、理论科学、计算科学和数据密集型科学发现,并指出人类社会正在转向第四范式——数据密集型科学发现^[16]。与传统科学不同的是,数据科学对现实世界的认识从二元世界(精神世界和物质世界)转变为三元世界(精神世界、物质世界和数据世界),通过对数据世界的研究解决现实世界的问题是数据科学的主要研究责任和目的。

1.2 典型实践

在萌芽期,对数据科学的发展具有重要意义的应用和实践包括以下内容。

(1)Python语言和R语言的相继出现及其在数据科学中的应用。与传统的面向软件开发的语言不同,Python语言和R语言能较好地支持面向数据科学的编程任务。1991年,Rossum公布了Python语言;1995年,统计学家Ihaka和Gentleman开发出R语言。之后,包括NumPy、Pandas、Statsmodel、Scikit-learn、Tidyverse和ggplot2等在内的面向数据科学的Python或R工具包的大量出现,简化和方便了人们的数据科学工作。目前,Python语言和R语言已成为数据科学领域应用最广泛的两种开源工具。

(2)Tableau及其在数据可视化中的应用。数据呈现是数据科学的主要活动之一,而数据可视化是目前最为广泛使用的数据呈现方式。2003年,Tableau在斯坦福大学诞生,其核心技术包括可视化查询语言VizQL、实时查询引擎和内存数据引擎。目前,Tableau已成为数据科学领域常用的数据可视化工具之一。

(3)数据科学家岗位的设立。数据科学家

岗位的出现为数据科学人才的职业化发展奠定了基础。2005年,美国国家科学基金委(National Science Foundation, NSF)建议“NSF应与收集管理人员和整个社区合作,采取行动推动数据科学家发展和完善其职业道路,并确保研究企业拥有足够数量的高质量数据科学家”^[17]。2008年,Patil和Hammerbacher分别在LinkedIn和Facebook成立了第一个正式的数据科学社区^[18],使得数据科学家这一岗位被广泛关注,福布斯(Forbes)将他们誉为首次提出术语“数据科学家”(Data Scientist)的人^[19]。2009年,Sadkowsky等在LinkedIn上组建了第一个数据科学家小组——The Data Scientists Group。目前,数据科学家已成为数据科学专业区别于传统数据研究学科——统计学和计算机科学的标志性概念之一。

(4) Hadoop生态系统的形成。Hadoop生态系统的形成推动了云计算技术在应用实践领域的普及,进而成为数据科学实践应用的重要手段。Hadoop 0.1.0于2006年发布,后发展成为包括HDFS、YARN、MapReduce、Pig、Hive、HBase在内的生态系统。目前,Hadoop生态系统是数据科学的一个主流平台和技术体系。

(5) 谷歌流感趋势分析工具(Google Flu Trends, GFT)的成功应用。GFT的成功应用引发了人们对大数据分析方法的热烈讨论,是社会开始对大数据问题展开科学研究的重要里程碑。2009年,Ginsberg等^[20]在Nature发文介绍GFT所采用的不同于传统统计分析的新思维,并证明GFT能够实时预测当年甲型H1N1流感病毒在美国境内的传播。GFT的成功应用以及后来的批判性研究,对于数据科学尤其是大数据分析的实践创新和认识论变化,具有重要推动作用。

(6) 数据科学应用和交流平台的相继出现。在萌芽期,出现了一些对数据科学理论研究和实践应用产生重要影响的专业网站、学术期刊和国际会议,比较有代表性的包括:KDnuggets创始人Piatetsky-Shapiro于1989年组织召开首

届数据库知识发现研讨会,并于1995年将其发展为知识发现和数据挖掘的年度大会——ACM SIGKDD会议^[21];2002年,国际科学理事会的科学和技术数据委员会(the Committee on Data for Science and Technology, CODATA)发行了第一本以“数据科学”命名的学术期刊——《数据科学学报》(The Data Science Journal);数据科学应用和交流平台的相继出现也为数据科学的快速发展奠定了基础。

2 数据科学的快速发展

与萌芽期不同,快速发展期(2010—2014)的学术研究已突破对数据科学概念及其重要性的讨论,开始探讨数据科学的方法、技术、工具和应用等核心问题,2013年涌现出一些讨论数据科学思想与应用的标志性成果:Mattmann在Nature发表题为《计算——数据科学的一种愿景》(Computing: A Vision for Data Science)的论文^[22];Dhar在《美国计算机学会通讯》(Communications of the ACM)上发表题为《数据科学与预测》(Data Science and Prediction)的论文^[23];Provost和Fawcett出版专著《面向商务的数据科学》(Data Science for Business)^[24];Mayer-Schönberger和Cukier出版专著《大数据——一场即将改变我们的生活、工作和思维的革命》(Big data: A Revolution That Will Transform How We Live, Work, and Think)^[25];Schutt和O'Neil出版专著《数据科学实践》(Doing Data Science)^[18]。

2.1 核心理论

进入快速发展期后,数据科学在其核心理论,尤其是数据科学的学科定位、大数据的新特征、数据科学家的能力要求与岗位职责、非结构化数据管理、大数据计算架构以及数据管理能力评估和数据治理等方面取得了突破性进展。

(1) Conway的数据科学维恩图。跨学科性是数据科学的公认特征之一,然而数据科学到底横跨哪些学科曾是一个有争议的话题。2010

年,Conway 提出数据科学维恩图(The Data Science Venn Diagram),认为数据科学是数学与统计学知识、领域实务知识、黑客精神与技能的交集,解决了人们对数据科学的跨学科特征的争议。目前,Conway 的数据科学维恩图被认为是数据科学的重要理论,学者们围绕该理论进行了大量的探讨。

(2)对大数据现象的定量分析及对大数据质量的重视。随着大数据应用的普及,数据科学对大数据现象的认识水平和解释能力也得到了进一步提升。2012年,McAfee等调查发现行业排名前1/3的数据驱动决策型公司的生产率平均比竞争对手高出5%,利润率高出6%^[26],用定量方法描述了大数据分析和数据驱动决策的价值。2013年,IBM在Laney定义的大数据3Vs特征的基础上新增了第四个维度,即质量维度——真实性(Veracity)^[27]。对大数据现象的定量分析以及对大数据质量的重视,增强了企业和产业对数据科学的关注程度,使数据科学成为企业管理和产业发展的新课题。

(3)围绕数据科学家进行的学术讨论。与萌芽期相比,快速发展期对数据科学家的关注焦点从实践应用领域转向理论研究领域,数据科学家的岗位职责、能力要求和团队建设成为理论研究重点。2010年,Cukier提出数据科学家是一种综合了软件程序员、统计学家和讲故事者/艺术家的技能的新型从业者^[28]。2011年,Loukides进一步明确了数据科学家的跨学科特征^[29]。2011年,Patil出版图书《数据科学团队建设》(*Building Data Science Teams*)^[30],较系统地讨论了数据科学家的能力要求及如何组建数据科学家团队等问题。2012年,数据科学中的相关思想成功地被应用于奥巴马团队的总统竞选工作,受到社会各界的广泛关注^[31];同年,Davenport和Patil提出“数据科学家将是21世纪‘最性感’的职业”^[32],数据科学家成为媒体报道的新焦点。在快速发展期,数据科学家成为数据科学领域理论研究的一个新课题。

(4)NewSQL技术的提出。由于传统关系数

据库在管理大数据时具有诸多不足之处,如不善于处理大数据读写操作,不适用于数据模型不断变化的应用场景,频繁操作的代价过大,简单处理的效率较低等。对此,数据科学研究中提出了多种补充或改进方案,比较有代表性的是NewSQL、NoSQL和关系云。2011年,Aslett率先提出NewSQL的概念,认为NewSQL是对各种新的可伸缩/高性能SQL数据库供应商的简称^[33]。通常,NewSQL不仅具有NoSQL对海量数据的存储管理能力,而且保持了传统数据库支持ACID和SQL的特性^[34]。NewSQL技术的提出为数据科学,尤其是海量非结构化数据的实时处理,提供了新的解决方案。

(5)Spark生态系统的兴起。Spark技术以及以Spark为中心的生态系统的兴起,为数据科学提供了不同于Hadoop生态系统的另一种解决方案——前者关注的是通用解决方案,而后者(每一个应用工具)关注的是单一应用场景。Spark诞生于数据科学的萌芽期,但以Spark为中心的生态系统则形成于数据科学的快速发展期。2009年,Spark作为加州大学伯克利分校AMP实验室(UC Berkeley AMPLab)的研究项目被开发,其早期生态系统由Shark SQL、Spark Streaming、MLLib、GraphX、SparkR和PySpark等组件组成。2013年诞生了Spark的商业版本——Databricks,并逐步形成以Databricks为核心的新的大数据处理生态系统。目前,Spark生态系统和Hadoop生态系统已成为数据科学的两个主流技术和平台。

(6)Lambda架构的提出。在数据科学的萌芽期,数据处理的可靠性和实时性之间存在矛盾。例如,Hadoop MapReduce的可靠性强,但实时性差,而Storm相反。为此,Marz于2011年提出一种大数据系统参考架构——Lambda架构^[35],采用分层思想较好地解决了这一矛盾。目前,Lambda架构已成为实时大数据处理系统的重要指导思想,在数据科学及其相关领域得到了广泛应用。

(7)规整数据的提出。与传统数据相关研

究不同,数据科学更加强调数据模态维度上的数据准备,而不是数据质量维度上的数据清洗。规整数据(Tidy Data)是数据模态转换的代表性理论之一。2014年,Hadley提出规整数据的思想,并给出了数据规整化的原则与过程^[36]。从本质上讲,所谓规整数据,就是可被常用数据分析软件或工具直接处理的模式形式组织的数据。目前,规整数据及数据的规整化处理已成为数据科学领域的核心理论之一。

(8)关于数据管理模型和数据治理框架的讨论。数据科学具有技术和管理的双重属性,而其管理属性主要起源于数据管理和数据治理问题的研究。2014年,CMMI研究所发布数据管理能力成熟度模型(Data Management Maturity Model, DMM模型),为企业提供了一套最佳实践标准。除了DMM模型,比较有影响力的理论还有数据管理协会(Data Management Association, DAMA)提出的数据管理知识体系指南。在数据治理框架方面,数据治理研究所(The Data Governance Institute, DGI)早在2004年就提出了数据治理框架(Data Governance Framework)。数据管理模型和数据治理框架的提出,为数据科学指导企业数据管理和数据治理实践奠定了基础,成为数据科学在企业管理中的重要应用场景。

2.2 典型实践

在应用方面,数据科学在快速发展期的标志性进展主要包括:大数据生态图谱的定期发布,资源平台Kaggle和DataCamp的成立,数据科学平台的兴起,基于Python/R的数据科学实践的探讨,大数据成为发达国家公共政策和社会治理的新焦点,关于数据科学人才培养和课程建设的讨论,以及数据科学与大数据相关标准化工作的启动。

(1)大数据生态图谱的定期发布。随着数据科学的发展步入快速期,从事数据科学产品生产的企业越来越多,亟待系统梳理数据科学领域的产业、企业和产品,而大数据生态图谱的

出现解决了这一社会需求。2012年,FirstMark资本公司(FirstMark Capital)的Turck和Zilis首次绘制了著名的大数据生态图谱(Big Data Landscape)并将其开源,有效地展示了大数据市场发展的现状与趋势。目前,大数据生态图谱已成为数据科学领域展示和观测大数据产业布局和产品发展动态的重要工具。

(2)资源平台Kaggle和DataCamp的成立。自2010年以来,数据科学及相关领域的资源平台相继出现,推动着数据科学领域的产业发展和人才培养,其中最具有代表性的是Kaggle和DataCamp。2010年,Goldbloom创立了数据科学的专业平台Kaggle,旨在通过众包方式解决数据科学的难题。2014年,Cornelissen等创立了数据科学在线学习平台——DataCamp,主要提供与数据科学、统计和机器学习相关的大量课程资源。目前,Kaggle和DataCamp已发展成为数据科学领域的两个重要的资源平台。

(3)数据科学集成开发平台和专业社区的兴起。2012年,Anaconda发布初始版本(0.8.0),并发展成为数据科学领域最流行的开发平台。2013年,Granville等创办数据科学中心(Data Science Central),为数据科学家提供交流的社区。同年,Databricks创立,它集成了Apache Spark、Delta Lake和MLflow技术,将数据工程、数据科学和数据分析聚合在一个开放、统一的平台内,使数据团队可以更快地开展协作和创新^[37]。以Anaconda为代表的集成开发平台和以数据科学中心为代表的数据科学专业社区的出现,推动了数据科学的专业化和职业化发展。

(4)基于Python/R的数据科学实践的探讨。随着Python和R在数据科学领域的广泛应用,基于Python/R的数据科学实践的专业讨论得到重视。2012年,Python Pandas包的开发者McKinney出版著作《利用Python进行数据分析:通过Pandas、NumPy和IPython进行数据整理》(*Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*),对如何基于

Python 进行数据集的控制、处理、整理、分析给出了较为系统的说明。2014 年,Zumel 等出版专著《基于 R 的实用数据科学》(*Practical Data Science with R*),较系统地介绍了如何运用 R 开展数据科学工作。目前,以 Python 和 R 为代表的开源工具已成为数据科学实践尤其是入门级数据科学实践的主流技术。

(5) 大数据成为发达国家公共政策和社会治理的新焦点。在数据科学的快速发展期,如何从大数据中获取洞察和价值成为政府治理、公共服务、社会治理和经济发展的新焦点。2012 年,美国施行“大数据研发计划”(Big Data Research and Development Initiative),计划投资两亿美元以提高政府从大量、复杂的数据集中提取知识和见解的能力^[38]。2013 年,澳大利亚出台《公共服务大数据战略》(*The Australian Public Service Big Data Strategy*)^[39],为政府部门更好地利用数据资产和保护个人隐私权提供了参考依据。2013 年,《法国政府大数据五项支持计划》^[40]发布,旨在推动数据科学家教育,优先在交通、医疗卫生等领域建设大数据示范工程,并为大数据应用提供良好的生态环境。2013 年,英国发布《把握数据机遇:英国数据能力战略》(*Seizing the Data Opportunity: A Strategy for UK Data Capability*)^[41],指出数据能力的内涵并提出具体工作部署,希望英国能在大数据分析领域领先于其他国家。发达国家在公共政策和社会治理领域重视大数据洞察和价值发现,拓展了数据科学的研究范畴,使得数据本身的治理和基于数据的治理成为数据科学重要的研究内容和实践领域。

(6) 对数据科学专业建设和课程设计的探索。数据科学实践的普及和深入对数据科学人才提出了更高的要求,数据科学专业建设和课程设计成为数据科学实践的重要课题。2012 年,Schutt 在哥伦比亚大学首次开设数据科学课程——“数据科学导论”(Introduction to Data Science);2013 年,Provost 和 Protopapas 等分别在纽约大学和哈佛大学开设“数据科学导论”课

程^[42]。从人才的专业化培养视角看,专业建设和课程设计的课程是数据科学人才培养的两个重要抓手。

(7) 数据科学与大数据标准化工作的启动。数据科学在快速发展期的另一个标志性实践项目是大数据标准化工作的开展。2014 年,ISO/IEC 成立大数据工作组 (ISO/IEC JTC1/SC22/WG9),负责制定大数据的基本标准,包括参考体系结构和术语标准,以支持 JTC1 的大数据工作。2017 年,又成立人工智能工作组 (ISO/IEC JTC1/SC42),旨在实现人工智能领域的标准化,并解散 WG9,同时将其工作计划转移至 SC42。数据科学标准化工作的启动,为数据科学的成功实践奠定了基础。

3 数据科学的逐步成熟

在逐步成熟期(2015—),数据科学逐渐从统计学和计算机学科中独立出来,成为一门新的独立学科。在此阶段,人们对数据科学的关注超越了学术研究的范畴,在科学研究、产业应用、人才培养、学科建设以及数据产品开发等方面进行了全方位的系统发展,由此使得数据科学的知识体系逐步形成,人们对数据科学的一些基本问题逐渐形成共识,数据科学领域的争议焦点从学科定位与边界、研究意义和研究责任等基本问题转向更深层次的问题,如数据科学的核心理论、数据科学中独有的新研究课题以及数据科学在其他学科中的应用等。

3.1 核心理论

在理论方面,人们开始聚焦于数据科学的核心理论、机器学习和人工智能在数据科学中的应用、数据故事化的提出及对数据可视化的补充、可解释性和可用性之间矛盾的解决、数据科学的成长曲线和魔力象限以及数据科学在其他领域中的应用研究等具体问题。

(1) 数据科学的核心理论受到关注。数据科学的理论研究从对其外围问题的关注转向对

内部核心问题的重视,这是数据科学从快速发展期转向逐步成熟期的重要标志。2017年,斯坦福大学 Donoho 教授发表论文《数据科学的五十年》(50 Years of Data Science)^[43],较为系统地回顾了数据科学的发展历程。2018年,Kelleher 等出版专著《数据科学》(Data Science)^[44],介绍了数据科学简史、数据科学与机器学习的关系、数据科学的应用、数据基础结构、数据伦理问题。2019年,Pieter 等出版了面向临床数据科学的专著《临床数据科学基础》(Fundamentals of Clinical Data Science)^[45],主要介绍了数据收集、建模和临床应用。2020年,Blum 等出版著作《数据科学基础》(Foundations of Data Science),主要介绍了数据科学的数学和算法基础^[46]。围绕数据科学自身的基础理论、关键技术、理论体系和学科发展历史进行的系统研究,为数据科学这一新学科的可持续发展奠定了基础。

(2) 机器学习和人工智能成为数据科学的理论基础。数据科学理论研究和实践应用均需要机器学习和人工智能技术的支持。2015年,Jordan 等在 Science 发表题为《机器学习:趋势、前景与展望》(Machine Learning: Trends, Perspectives, and Prospects)的论文^[47],提出机器学习是人工智能与数据科学的核心,采用数据密集型机器学习方法可以推动医疗保健、制造业、教育、金融、公安和市场营销等领域开展基于证据的决策。机器学习是人工智能的一部分,随着人工智能的进一步发展,机器感知、人机交互、认知计算、自然语言处理、智能机器人和智能应用技术等人工智能领域的核心技术也将陆续被应用于数据科学领域。

(3) 数据故事化的提出及对数据可视化的补充。数据故事化的提出为数据呈现提供了一个补充方案,这成为数据科学研究的新内容。2015年,Knaflitz 出版专著《用数据讲故事》(Storytelling with Data)^[48],从此数据故事化成为数据科学的一个新研究课题。数据可视化和数据故事化的区别在于,前者主要关注数据感知问题,后者主要解决数据认知问题,数据感知是数

据认知的前提,而数据认知是数据感知的进一步深化,数据认知高于数据感知^[49]。

(4) 可解释性和性能之间的矛盾研究成为新课题。可解释性和性能是数据科学领域的重要矛盾之一。可解释性(interpretability)是指采用人尤其是非专业人士可理解的方式表达的能力。性能(performance)是指模型预测正确率或平均预测误差^[50]。然而已有的机器学习模型中,性能较好的模型(如深度神经网络)往往是黑箱模型,其可解释性较差;具有较强可解释性的模型(如决策树)通常性能弱于黑箱模型。解决可解释性和性能之间的矛盾,是提升数据科学学科核心竞争力和社会贡献的主要挑战和重要突破口。

(5) 数据科学的成长曲线和魔力象限的提出。数据科学自身的发展水平和数据科学产品的动态评价是当今社会的一个关注焦点。为此,Gartner 公司于 2016 年首次发布研究报告《数据科学的成长曲线》(Hype Cycle for Data Science),并于 2017 年更名为《数据科学和机器学习成长曲线》(Hype Cycle for Data Science and Machine Learning)。此外,Gartner 公司于 2017 年将原有的《高级分析平台魔力象限》(Magic Quadrant for Advanced Analytics Platforms)改名为《数据科学平台魔力象限》(Magic Quadrant for Data Science Platforms),2018 年再度更名为《数据科学和机器学习平台的魔力象限》(Magic Quadrant for Data Science and Machine-Learning Platforms)。魔力象限根据实践能力和远见能力,将市场上提供数据科学和机器学习产品的企业划分为领导者、挑战者、远见者和利基参与者四类,如在 2021 年《数据科学和机器学习平台的魔力象限》中,IBM 是领导者之一,Alteryx 是挑战者^[51]。数据科学的成长曲线和魔力象限的提出,为大众了解数据科学领域的发展水平和数据科学产品的最新动态提供了较为直观的分析工具,成为该领域的标志性研究成果。

(6) 数据科学在其他领域应用研究的兴起。除了数据科学自身领域的研究,其他学科和领

域也开始关注数据科学的相关理论与应用问题,如风能、农业、新闻、材料学、化学、医疗健康、信息资源管理和供应链管理等领域均从自身专业视角研究数据科学问题,利用数据科学理论和方法,解决各自领域的实践应用问题。随着数据科学被应用到其他领域,数据科学理论出现了两个重要分支——专业数据科学和专业中的数据科学^[52],前者代表数据科学作为一门通用性的独立于具体应用领域的基础科学,后者代表不同学科领域对数据科学的差异性研究和应用,对具体领域的专业知识依赖度较高。

3.2 典型实践

在逐步成熟期,数据科学的应用成为一个新的关注点。数据科学在多个传统学科领域得到广泛应用,数据科学专业人才的培养成为新课题,数据科学领域的理论、方法、技术、工具和应用呈现出集成化发展趋势,数据的科学治理和有效管理问题成为全社会的新焦点,数据科学的应用从数据分析的简单应用转向以数据为中心的社会的治理和生产模式的变革。

(1)数据科学在农业、电力、材料、医学等多个专业领域得到实质性应用,从理论研究转向实质性应用是数据科学实践进入逐步成熟期的标志。以农业领域的应用为例,气候公司(The Climate Corporation)于2015年将其大数据产品整合为新的数字化农业平台——Climate FieldView,将数据科学应用于对农作物、田地和天气等的分析,为农民提供农作物种植决策支持^[53];Climate FieldView平台已发展成为全球知名的数字化农业平台,代表了数据科学在农业领域的最佳实践。目前,在传统专业领域,尤其是农业、电力、材料、医学等领域,数据科学的实践应用越来越多,已成为这些传统专业领域的创新动力和新的增长点。

(2)社会对数据科学家的需求趋于专业化,数据人才培养趋于体系化,表明数据科学领域进入逐步成熟期。2015年,美国白宫任命业界知名数据科学家Patil为首席数据科学家,体现

出对数据科学家职业人员的需求;同年,Baker在*Nature*发表论文《数据科学——产业诱惑》(*Data Science:Industry Allure*)^[54],使数据科学家得到各种产业的关注。基于Study Portals的数据统计显示,截至2021年12月,国外数据科学专业的本科、硕士、博士学位项目已分别达到5708、5475和280项,主要分布在美国、英国、澳大利亚、加拿大、德国等国家^[55]。中国教育在线显示,目前我国已有558所院校设立了“数据科学与大数据技术”本科专业^[56]。国内外高校都已将数据科学专业人才的培养纳入教育体系,可以应对社会对数据科学专业人才的**新需求。

(3)多种大数据技术融合应用成为一种趋势。大数据采集、计算、管理和分析技术的融合应用是数据科学实践进入逐步成熟期的另一个标志。2015年,Ryza等出版新书《Spark高级分析》(*Advanced Analytics with Spark*),详细介绍了如何利用Spark技术分析大规模数据^[57]。2016年,IBM公司创建IBM Data Science Experience(DSX)平台(现更名为Watson Studio),集成了Apache Spark、Scala、Python、R、Jupyter Notebook等多项开源技术,为数据科学家提供数据和模型全套分析工具。2017年,Piatetsky的调查显示Python语言已超越R语言成为数据分析、数据科学和机器学习领域最受欢迎的语言^[58]。多种大数据技术的融合应用改变了以单一技术为中心的数据分析实践,成为数据科学实践领域的新趋势。

(4)各国政府投入大量资金,并将大数据作为国家战略的重要部分,这是数据科学实践进入逐步成熟期的重要标志。2015年,我国国务院发布《促进大数据发展行动纲要》,提出重点开展数据理论和科学研究工作,鼓励与支持在大数据理论创新、研究方法及大数据关键技术应用等方面的探索研究。2016年,美国发布“联邦大数据研究和开发战略计划”,提出在关键领域推进大数据研发活动。随着大数据成为更多国家的战略规划的内容,数据科学实践将得到

前所未有的重视和发展。

(5) 数据安全问题成为数据科学实践的关注点。随着数据科学实践的普及,包括数据隐私及个人数据保护在内的数据安全问题受到社会各界广泛关注。2016年,欧盟通过《通用数据保护条例》(*General Data Protection Regulation, GDPR*)^[59],提出个人数据保护规章制度,成为个人数据隐私保护方面的立法先例。2016年,出于对数据隐私的担忧,Netflix取消第二届Netflix奖(The Netflix Prize)^[60],隐私保护得到算法设计和商业运营领域的重视。2018年,剑桥分析公司(Cambridge Analytica)事件^[61]曝出,隐私保护成为全球媒体讨论的热点。2018年,《加利福尼亚州消费者隐私法》(*The California Consumer Privacy Act of 2018, CCPA*)^[62]颁布,成为消费者隐私保护领域的新探索。目前,包括数据隐私在内的数据安全问题的立法、伦理、道德和偏见已成为数据科学实践领域的热门话题。

(6) 有关数据科学实践的代表性论著持续出版。在此阶段的初期,代表性著作主要集中在基于Python和R的数据科学实践。2016年,VanderPlas出版著作《Python数据科学手册:使用数据的基本工具》(*Python Data Science Handbook: Essential Tools for Working with Data*)^[63],详细介绍了基于Python的数据科学实践工具,包括IPython、Jupyter、NumPy、Pandas、Matplotlib和Scikit-Learn等。同年,Grolemund和Wickham出版著作《R数据科学》(*R for Data Science*),系统介绍了如何利用R完成数据的导入、清洗、转换、可视化和建模等工作^[64]。随着数据科学实践的不断成熟,相关著作相继出现,目前已出现多部基于Spark生态系统、Hadoop生态系统的数据科学实践著作,如2015年Chambers等出版《Spark权威指南:让大数据处理变得简单》(*Spark: The Definitive Guide: Big Data Processing Made Simple*)^[65],Grover等出版《Hadoop应用架构:设计真实世界的大数据应用》(*Hadoop Application Architectures: Designing Real-World Big Data Applications*)^[66]。

(7) 数据科学专业的人才培养和课程建设逐步进入成熟阶段。在国外,纽约大学、加州大学伯克利分校、约翰·霍普金斯大学、华盛顿大学等高校率先开展数据科学专业教育。在国内,2016年北京大学、对外经济贸易大学和中南大学首批获批开设数据科学与大数据技术专业。同年,中国人民大学朝乐门出版国内第一部系统阐述数据科学原理、方法与技术的专著——《数据科学》,之后他主讲的“数据科学导论”课程先后被评为国内数据科学领域首门国家精品在线开放课程和国家级一流本科课程,主编的教材《数据科学理论与实践》被北京市教委认定为优质教材。数据科学专业的人才培养和课程建设推动了数据科学理论研究和实践应用的可持续发展。

(8) 数据科学实践的标准化工作取得突破性进展,在大数据互操作、技术参考架构、安全与隐私、系统功能要求、数据能力评估以及核心术语等方面,相继发布了多项国际或国家标准。2015年,美国国家标准与技术研究院发布了大数据互操作框架系列标准(NBDIF),目前已发布三版;自2018年以来,ISO/IEC 20547系列标准陆续发布,主要涉及大数据架构中的应用流程、用例及需求、参考框架、安全与隐私等内容;2019年,ISO/IEC 20546发布,主要对大数据基础术语进行了标准规定。与此同时,我国在大数据术语、大数据技术参考模型、数据管理能力评估、数据质量评价、大数据存储/处理/分析和系统功能要求等方面颁布了国家标准。大数据标准工作的不断拓展和系统化发展,将进一步推动数据科学理论在实践领域的大规模应用,降低数据科学实践中的风险和成本,提升数据科学实践项目的工作效率。

4 讨论

从以上分析可知,数据科学的发展可以分为三个阶段,分别为萌芽期、快速发展期和逐步成熟期,数据科学领域已取得的主要成就体现

在两个方面。

(1) 数据科学核心理论研究取得一定成果。目前,数据科学在大数据的特征和重要性、数据科学的内涵与外延、探索性数据分析、大数据的存储/计算/管理技术、CAP 理论与 BASE 原则、深度学习、数据密集型科学发现、数据科学维恩图、数据科学家培养、规整数据、数据能力评估、大数据治理、数据可视化与故事化以及数据产品开发等各方面均取得了一定的理论创新。

(2) 数据科学的实践应用取得创新和突破。当前,数据科学的实践应用领先于理论研究,倒逼理论研究必须进行突破性创新。数据科学领域的代表性实践有:基于 Python/R 语言的数据科学编程,数据可视化工具的研发,多领域数据

科学家岗位的设置,Hadoop/Google/Spark/Oracle 大数据生态系统的建设,以谷歌流感趋势为代表的大数据实时分析,数据科学集成开发平台和专业社区的建设,基于大数据的公共政策和社会治理的开展,数据科学的专业建设和课程设计,数据科学与大数据的标准研制,大数据技术的融合应用,数据隐私和安全保护,数据科学在农业、电力、材料、医学等多个专业领域中的实质性应用,等等。

4.1 存在问题

数据科学在理论研究和实践应用方面取得的突破性进展为数据科学这一新学科领域的快速发展奠定了基础。但是,在以下四个方面仍存在问题,如图 1 所示。

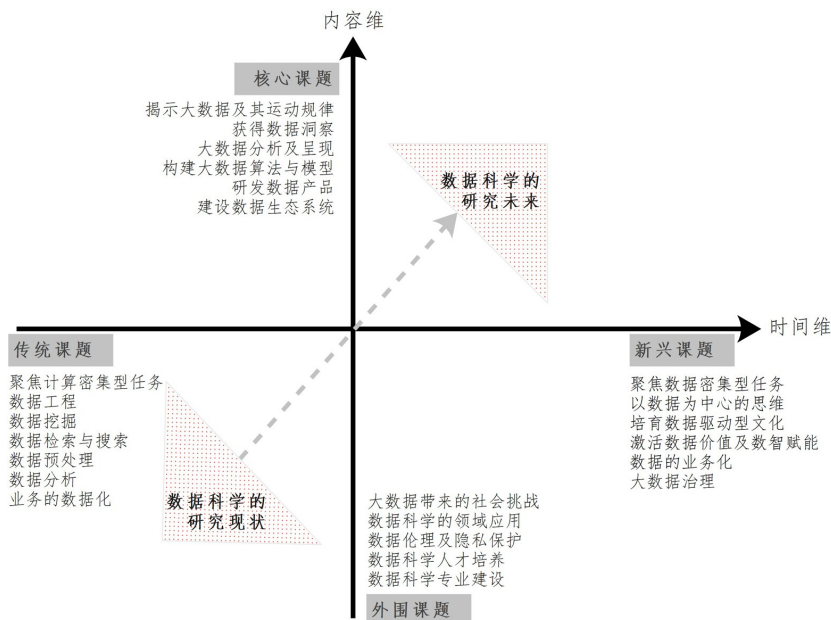


图1 数据科学的研究现状与趋势

(1) 对数据科学本身的系统研究不足。从萌芽期到逐步成熟期,数据科学研究在整体上是关注外围课题转向讨论内部核心课题,但至今仍缺乏对数据科学特有理论的系统研究,亟待针对数据科学的理论、方法、技术、工具、应用和实践开展系统的深入研究,力求在数据科

学这一新学科的研究目的、研究责任、研究内容、研究方法及理论体系等方面达成基本共识,进而推动数据科学理论与实践的持续发展。

(2) 对数据科学领域新课题的聚焦不够。在快速发展期和逐步成熟期,数据科学的多数研究仍停留在计算密集型任务、数据工程、数据

挖掘、数据检索与搜索、数据预处理、数据分析以及业务的数据化等传统课题,而缺乏对大数据时代所面临的新课题的开创性研究。与传统数据相关学科(如统计学和计算机科学)不同,数据科学需要侧重以下新课题的研究:数据密集型任务、以数据为中心的思维、培育数据驱动型文化、激活数据价值及数智赋能、数据的业务化以及大数据治理。同时,在数据科学领域亟待深入研究的新理论包括:数据密集型科学发现、大数据因果分析、数据试验设计、数据产品开发、探索型分析、规范性分析、数据加工与准备、数据可视化与故事化、可解释性算法与人工智能,以及大数据的道德、伦理和立法等。

(3) 对数据科学领域核心课题的研究不足。目前,数据科学领域的研究仍停留在外围课题,如大数据带来的社会挑战、数据科学的领域应用、数据伦理及隐私保护、数据科学人才培养以及数据科学专业建设等,而对数据科学核心课题的研究相对欠缺。未来研究需加强对数据科学本质性问题的探讨,包括:揭示大数据及其运动规律、获得数据洞察、大数据分析及呈现、构建大数据算法与模型、研发数据产品以及建设数据生态系统。

(4) 数据科学理论直接指导下完成的实践应用不多。数据科学萌芽期和快速发展期的多数实践并非是在数据科学理论的指导下完成的,而是直接模仿/移植或复制已有的成功实践。到目前为止,数据科学仍处于实践应用领先于理论研究的状态,各种开创性实践的不断发展正在倒逼数据科学理论的创新,亟待加强数据科学理论研究的深度,提升数据科学理论对实践应用的指导意义。

4.2 展望

从数据科学发展历程、存在问题及未来社会需要看,数据科学的理论研究和实践应用需要重视以下方面。

(1) 聚焦数据科学特有的新问题和主要矛盾。数据相关的研究是人类社会一直以来关注和讨论的传统问题,而并非数据科学这一新学

科诞生以后才出现的新课题。数据科学这个新学科的意义不仅在于总结人类已有的对数据的相关认识,而且更需要研究大数据,尤其是数据富足时代的新数据所带来的新问题和新的挑战。大数据时代的新问题和主要矛盾主要体现在新数据(大数据)和老知识(传统知识)之间的矛盾,而数据科学的未来研究应以解决这一矛盾为主要抓手,将研究重点放在传统知识的继承和创新,力争缩短传统知识与大数据时代的新数据之间的鸿沟。随着数据科学成熟期的到来,对数据科学的新问题的研究,尤其是围绕大数据时代到来导致的数据变化以及新数据与旧知识之间矛盾的研究,将成为数据科学研究的主要趋势。

(2) 健全数据科学理论研究的体系。在数据科学的萌芽期和快速发展期,人们对数据的研究分散在不同层次,从不同视角研究各自感兴趣和关心的问题,缺乏系统性梳理。但随着进入成熟期,数据科学的理论发展呈现出体系化趋势,数据科学已从统计学、计算机科学等基础学科中脱离出来,正在发展成为一门独立的新学科,数据科学理论的专门性系统研究将成为数据科学研究的一项重要任务。

(3) 加快数据科学实践应用的产业化。在数据科学的快速发展期,数据科学的应用往往是个别成功案例,且仅限于企业内部的业务应用,尚未实现产业化应用与普及。随着成熟期的到来,数据科学实践应用的广度和深度不断加强,越来越多的数据科学实践正成为企业生产制造及产业协同发展的新增长点,数据科学将成为以数据产品为中心的产业革命及新业态的主要推动力。

(4) 推进研究方法的突破性创新。方法层次上的突破创新是数据科学未来发展的关键所在。以数据密集型科学发现为代表的第四范式理论越来越受到数据科学领域的重视,数据科学有望在研究方法层面取得新突破。数据科学领域的新方法应具备以数据为中心、数据密集型和数据驱动三个新特征。同时,研究方法层

次变革与专业化人才培养密切相关。在我国,数据科学与大数据技术、大数据管理与应用已进入本科专业目录,而且数据科学也成为了硕博阶段人才培养的新方向。研究方法的专业化和研究视角的独特性是推动数据科学原创性发展的主要前提,也是未来研究的重点所在。

(5)拓展数据科学的基础理论。在快速发展期,数据科学主要是以统计学和机器学习为主要理论基础发展起来的。未来数据科学的基础理论需要不断拓展,以统计学为核心的理论基础需要拓展至数学与统计学,以机器学习为中心的理论基础需要拓展至人工智能,同时也应引入深度学习和区块链等新技术。数据科学的研究需要重视数据加工、数据分析、数据计算、数据管理、数据产品开发、数据伦理/道德/立法等基础理论。

5 结语

数据科学这一学科的主要使命和社会贡献不仅在于解决大数据时代到来所导致的数据

变化问题以及新数据与旧知识之间的矛盾,而且对各个领域的思维转变(从经验驱动到数据驱动)、研究方法创新(从第三范式到第四范式)、工具开发(从传统数据分析工具如 SPSS 到数据科学平台)、统计方法转变(从小样本到海量数据)做出了开创性贡献。随着数据科学的发展从快速发展期进入成熟期,人们对数据科学的关注从对外延、理念、思想认识等表层问题的讨论转向对本质、方法、技术、工具和应用等深层问题的探讨。在此过程中,数据科学的理论研究和实践应用亟待聚焦数据科学的自身问题和主要矛盾,进一步明确数据科学在人类知识体系中的合理定位,凸显其方法、技术和工具层面的价值,提升数据科学解决问题的能力,推动数据科学领域的健康发展。目前,数据科学仍是实践领先于理论的新兴学科,理论研究远远滞后于行业实践。未来,需要深入研究数据科学中的新方法、新技术和新工具,不断提炼数据科学的理论,加强理论创新,提升理论对实践应用的指导作用,推动数据科学理论与实践的协同发展。

参考文献

- [1] 朝乐门. 数据科学理论与实践[M]. 第二版. 北京:清华大学出版社,2019:9. (Chao L M. Data science theory and practice [M]. 2nd edition. Beijing:Tsinghua University Press,2019:9.)
- [2] Naur P. Concise survey of computer methods[M]. New York:Petrocelli Books,1974.
- [3] Chambers J M. Greater or lesser statistics;a choice for future research[J]. Statistics & Computing,1993,3(4):182-184.
- [4] 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.
- [5] Tukey J W. Exploratory data analysis[M]. London:Pearson,1977.
- [6] Brewer E. CAP twelve years later;how the“rules” have changed[J]. Computer,2012,45(2):23-29.
- [7] Pritchett D. BASE;an acid alternative;in partitioned databases,trading some consistency for availability can lead to dramatic improvements in scalability[J]. Queue,2008,6(3):48-55.
- [8] Ghemawat S,Gobioff H,Leung S T. The Google file system[C]//Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles. 2003:29-43.
- [9] Dean J,Ghemawat S. MapReduce:simplified data processing on large clusters [C]//Proceedings of the 6th Conference on Usenix Symposium on Operating Systems Design and Implementation. 2004,4:137-150.

- [10] Ghemawat S,Hsieh W C,Wallach D A,et al. Bigtable;a distributed storage system for structured data[C]//Proceedings of the 7th Conference on Usenix Symposium on Operating Systems Design and Implementation. 2006,7: 52-57.
- [11] Grolinger K,Higashino W A,Tiwari A,et al. Data management in cloud environments;NoSQL and NewSQL data stores[J]. Journal of Cloud Computing,2013,2(1):22.
- [12] Laney D. 3D data management;controlling data volume,velocity,and variety[J]. META Group Research Note, 2001,6(70):1.
- [13] Charles A. Tech giants may be huge,but nothing matches big data[J]. The Guardian,2013(13):1.
- [14] Economist T. The world's most valuable resource[J]. The Economist,2017,423(TN. 9039).
- [15] Hinton G E,Osindero S,Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation,2006, 18(7):1527-1554.
- [16] Hey T,Tansley S,Tolle K. The fourth paradigm:data-intensive scientific discovery[M]. Redmond:Microsoft Research,2009:177-185.
- [17] Simberloff D,Barish B C,Droegemeier K K,et. al. Long-lived digital data collections enabling research and education in the 21st century[J]. Science,2005,185(3310):1330.
- [18] O'Neil C,Schutt R. Doing data science:straight talk from the frontline[M]. Sebastopol:O'Reilly Media,Inc. 2013:8.
- [19] Forbes Inc. Tim O'Reilly:the world's 7 most powerful data scientists[EB/OL]. [2014-12-05]. <https://www.forbes.com/pictures/efej45gkjj/2-jeff-hammerbacher-chief-scientist-cloudera-and-dj-patil-entrepreneur-in-residence-greylock-ventures/?sh=6de116c22e53>.
- [20] 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.
- [21] Fayyad U M,Piatetsky-Shapiro G,Smyth P. Knowledge discovery and data mining:towards a unifying framework [C]//Simoudis E,Han J,Fayyad U M. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland,Oregon:AAAI Press,1996,82-88.
- [22] Mattmann C A. Computing;a vision for data science[J]. Nature,2013,493(7433):473-475.
- [23] Dhar V. Data science and prediction[J]. Communications of the ACM,2013,56(12):64-73.
- [24] Provost F,Fawcett T. Data science for business;what you need to know about data mining and data-analytic thinking[M]. Sebastopol:O'Reilly Media,Inc. ,2013.
- [25] Mayer-Schnberger V,Cukier K. Big data;a revolution that will transform how we live,work,and think[M]. Boston:Houghton Mifflin Harcourt,2013.
- [26] McAfee A,Brynjolfsson E,Davenport T H,et al. Big data;the management revolution[J]. Harvard Business Review,2012,90(10):60-68.
- [27] IBM. 分析:大数据在现实世界中的应用[R/OL]. [2021-12-26]. <https://ibm.com/downloads/cas/ED0JV08Q>. (IBM. Analysis;the application of big data in the real world[R/OL]. [2021-12-26]. <https://ibm.com/downloads/cas/ED0JV08Q>.)
- [28] Cukier K. Data,data everywhere;a special report on managing information[M]. Economist Newspaper,2010.
- [29] Loukides M. What is data science[M]. Sebastopol:O'Reilly Media,Inc. ,2011.
- [30] Patil D J. Building data science teams[M]. Sebastopol:O'Reilly Media,Inc. ,2011.
- [31] Hellweg E. The first big data election[J]. Harvard Business Review,2012:13.

- [32] Davenport T H, Patil D J. Data scientist[J]. Harvard Business Review, 2012, 90(5): 70-76.
- [33] Aslett M. What we talk about when we talk about NewSQL[EB/OL]. [2021-12-26]. https://blogs.451research.com/information_management/2011/04/06/what-we-talk-about-when-we-talk-about-newsq.
- [34] Raje A, Jagdale A. Sql Vs NoSql; NewSql the solution for big data[J]. IOSR Journal of Computer Engineering, 2017, 2(8): 45-51.
- [35] Warren J, Marz N. Big data: principles and best practices of scalable realtime data systems[M]. New York: Manning, 2015.
- [36] Wickham H. Tidy data[J]. Journal of Statistical Software, 2014, 59(10): 1-23.
- [37] Databricks. About databricks-the data and AI company[EB/OL]. [2021-03-26]. <https://databricks.com/company/about-us>.
- [38] Office of Science and Technology Policy Executive Office of the President. Obama administration unveils "big data" initiative; announces \$200 million in new R&D investments[EB/OL]. [2021-03-26]. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.
- [39] The Australian public service big data strategy[EB/OL]. [2021-03-26]. https://toolkit.data.gov.au/images/1/16/Big_Data_Strategy.pdf.
- [40] Filippone D. Big data; les 5 mesures du plan de soutien dévoilées[EB/OL]. [2021-03-26]. <https://www.journaldunet.com/solutions/reseau-social-d-entreprise/1125790-big-data-les-5-mesures-du-plan-de-soutien-devoilees/>.
- [41] Seizing the data opportunity: a strategy for UK data capability[EB/OL]. [2021-03-26]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/254136/bis-13-1250-strategy-for-uk-data-capability-v4.pdf.
- [42] 朝乐门. 数据科学导论的课程设计及教学改革[J]. 计算机科学, 2020, 47(7): 1-7. (Chao L M. Course design and redesign for introduction to data science[J]. Computer Science, 2020, 47(7): 1-7.)
- [43] Donoho D. 50 years of data science[J]. Journal of Computational and Graphical Statistics, 2017, 26(4): 745-766.
- [44] Kelleher J D, Tierney B. Data science[M]. Cambridge: MIT Press, 2018.
- [45] Pieter K, Michel D, Andre D. Fundamentals of clinical data science[M]. Berlin: Springer, 2019.
- [46] Blum A, Hopcroft J, Kannan R. Foundations of data science[M]. Cambridge: Cambridge University Press, 2020.
- [47] Jordan M I, Mitchell T M. Machine learning: trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.
- [48] Knafllic C N. Storytelling with data: a data visualization guide for business professionals[M]. New York: John Wiley & Sons, 2015.
- [49] 朝乐门, 张晨. 数据故事化: 从数据感知到数据认知[J]. 中国图书馆学报, 2019, 45(5): 61-78. (Chao L M, Zhang C. Data storytelling: from data perception to data cognition[J]. Journal of Library Science in China, 2019, 45(5): 61-78.)
- [50] Gosiewska A, Kozak A, Biecek P. Simpler is better: lifting interpretability-performance trade-off via automated feature engineering[J]. Decision Support Systems, 2021: 113556.
- [51] Gartner Inc. Magic quadrant for data science and machine learning platforms[EB/OL]. [2021-03-01]. <https://www.gartner.com/en/documents/3998753-magic-quadrant-for-data-science-and-machine-learning-pla>.
- [52] 朝乐门, 邢春晓, 张勇. 数据科学研究的现状与趋势[J]. 计算机科学, 2018, 45(1): 1-13. (Chao L M, Xing

- C X,Zhang Y. Data science studies;state-of-the-art and trends[J]. Computer Science,2018,45(1):1-13.)
- [53] Climate FieldView. Introducing the new climate fieldview? platform[EB/OL]. [2021-03-26]. <https://climate.com/blog/introducing-the-new-climate-fieldviewtrade-platform>.
- [54] Baker M. Data science;industry allure[J]. Nature,2015,520(7546):253-255.
- [55] Study Portals Inc. “Data science” [EB/OL]. [2021-12-27]. <https://studyportals.com/students>.
- [56] 数据科学与大数据技术[EB/OL]. [2021-12-24]. https://gkcx.eol.cn/special/1283?special_type=3&sort=1. (Data science and big data technology[EB/OL]. [2021-12-24]. https://gkcx.eol.cn/special/1283?special_type=3&sort=1.)
- [57] Ryza S,Laserson U,Owen S,et al. Advanced analytics with spark;patterns for learning from data at scale[M]. Sebastopol;O'Reilly Media,Inc.,2015.
- [58] Piatetsky G. Python overtakes R,becomes the leader in data science,machine learning platforms[EB/OL]. [2017-09-06]. https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html?imm_mid=0f5ddc&cmp=em-data-na-na-newsltr_20170906.
- [59] Goddard M. The EU general data protection regulation;European regulation that has a global impact[J]. International Journal of Market Research,2017,59(6):703-705.
- [60] Holmes D E. Big data;a very short introduction[M]. Oxford;Oxford University Press,2018;89.
- [61] Cadwalladr C,Graham-Harrison E. Revealed;50 million Facebook profiles harvested for Cambridge Analytica in major data breach[J]. The Guardian,2018,17:22.
- [62] State of California Department of Justice. California consumer privacy act (CCPA)[EB/OL]. [2021-03-26]. <https://oag.ca.gov/privacy/ccpa>.
- [63] VanderPlas J. Python data science handbook;essential tools for working with data[M]. Sebastopol;O'Reilly Media,Inc,2016.
- [64] Wickham H,Grolemund G. R for data science;import,tidy,transform,visualize,and model data[M]. Sebastopol;O'Reilly Media,Inc,2016.
- [65] Chambers B,Zaharia M. Spark;the definitive guide;big data processing made simple[M]. Sebastopol;O'Reilly Media,Inc,2018.
- [66] Grover M,Malaska T,Seidman J,et al. Hadoop application architectures;designing real-world big data applications[M]. Sebastopol;O'Reilly Media,Inc,2015.

朝乐门 数据工程与知识工程教育部重点实验室(中国人民大学)研究员,中国人民大学信息资源管理学院副教授,博士生导师。北京 100872。

张 晨 数据工程与知识工程教育部重点实验室、中国人民大学信息资源管理学院博士研究生。北京 100872。

孙智中 数据工程与知识工程教育部重点实验室、中国人民大学信息资源管理学院博士研究生。北京 100872。

(收稿日期:2021-03-26)