ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 3,May 2022,pp. 471－487
DOI:10. 16337/j. 1004-9037. 2022. 03. 001

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

数据科学:从数字世界到数智世界

张清华[1,2], 高　渝[1,2], 申秋萍[1,2]

(1. 重庆邮电大学旅游多源数据感知与决策技术文化和旅游部重点实验室,重庆 400065;2. 重庆邮电大学计算智能重庆市重点实验室,重庆 400065)

摘　要: 随着大数据的持续发展,数据已经成为国家的重大战略资源,对社会影响日益明显。为了更深入地挖掘和研究大数据背后所蕴藏的基本科学问题,新的研究领域——数据科学被提出。本文从大数据的发展历程出发,介绍了数据科学的兴起和内涵;分析了大数据和数据科学的研究现状,以及数据在各行业中的应用;简述了为探索数据科学本身的内涵和规律而建设的大数据试验场;讨论了数据科学的关键问题,以及在研究数据时应具有的新思维和新观念,以推动数据科学的发展,促进现实世界向数字世界的转型,最终实现社会生活的真正智能化。

关键词: 大数据;数据科学;大数据试验场;数字世界

中图分类号: TP311.13;TP18　　**文献标志码:** A

Data Science : From Digital World to Digital Intelligent World

ZHANG Qinghua[1,2], GAO Yu[1,2], SHEN Qiuping[1,2]

(1. Key Laboratory of Tourism Multisource Data Perception and Decision of Ministry of Culture and Tourism, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: With the development of big data, data has become a major strategic resource for countries and its social impact is increasingly obvious. Thus, data science is proposed to explore and study basic scientific problems contained in big data. In this paper, the development of big data, the rise and connotation of data science are first introduced. Second, the research status of big data and data science is analyzed, and the application of data in various industries is discussed. Third, the big data proving ground that is constructed to explore laws and problems of data science is briefly described. Finally, in order to promote the development of data science, accelerate the transformation of the real world to the digital world, and realize the intelligent life, the key issues of data science and the new thinking in digital world are discussed.

Key words: big data; data science; big data proving ground; digital world

引　　言

随着互联网、移动互联网和物联网等技术的飞速发展,数字信息高速流动,人、机、物在任何时间和地点互联互通[1],源源不断的数据在万物互联中产生汇聚,并以指数形式增长。指数增长的数据充斥着

基金项目: 国家重点研发计划(2020YFC2003502);国家自然科学基金(61876201)。
收稿日期: 2022-01-30;**修订日期:** 2022-04-10

整个世界,逐渐成为重要的生产资料,“大数据”应运而生,并成为社会各界关注的焦点和讨论的热点。什么是大数据?《Science》在2011年出版的专刊中将大数据定义为“无法使用传统软件和工具在有限时间内进行采集、管理和分析的数据集合”[2]。维基百科中对大数据的定义为“涉及的数据数量巨大到无法使用现有主流软件和工具在有限且合理的时间内对其进行采集、管理和分析”。大数据研究机构Gartner对大数据的定义为“一种需要通过新的处理模式来处理的高增长率和种类繁多的巨量信息资产,从而优化处理结果使其具有更强的决策能力、更高的洞察力”[3]。实际上,对于大数据目前并没有一个统一的定义[4],而IBM提出的关于大数据的“5V”特征(Volume、Variety、Velocity、Value和Veracity)受到社会各界的广泛认可[5]。换而言之,满足数据量巨大、数据种类繁多、获取数据速度快、价值大密度较低并且能反映真实信息这5个特征的数据均可称其为大数据。随着社会对大数据认识的逐渐深入,与其相关的产业逐渐涌现,各行业的数据规模逐渐庞大,数据甚至被誉为“未来的新石油”。发展至今,大数据研究已经取得令人瞩目的成绩,数据应用到各行各业,逐渐成为其核心资产,而且数据的获取、存储和计算已不再是难题。然而,现实世界正逐步映射到数字世界,在数字世界中如何治理数据,如何有效地解释并利用数据以及如何推动智能化世界的发展成了亟待解决的问题,由此促生了一种区别于传统科学研究的新研究领域——数据科学。

1 从大数据到数据科学

大数据的发展说明了各个领域已经广泛数字化,推动了数字世界的形成。在数字世界中,大数据是研究的内容和基础,数据科学是研究大数据时新出现的理论和方法以及思维和模式。

1.1 大数据的发展

大数据作为信息技术领域的重要课题之一,从提出到现在一直受到广泛关注。1980年,著名未来学家阿尔文·托夫勒在《第三次浪潮》一书中正式提出“大数据”一词[6]。2008年,国外著名杂志《Nature》推出“Big Data”专刊,开始探讨数据量的飞速增长给各领域带来的影响[7]。2011年,麦肯锡发布研究报告“Big Data: The Next Frontier for Innovation, Competition and Productivity”指出“大数据时代已经到来”[8]。2012年,欧洲信息学与数学研究协会在出版的会刊《ERCIM News》“Big Data”专刊中讨论了数据密集型研究的创新、数据管理的技术等问题[9]。另外,各国政府相继发布大数据相关的纲领性文件,例如美国政府启动“大数据研究和发展计划”,英国发布“英国数据能力发展战略规划”,日本发布“创建最尖端IT国家宣言”,以及韩国提出“大数据中心战略”等。

中国对于大数据的研究起步较晚,但是发展速度却非常快。自2013年起,中国的大数据研究开始蓬勃发展,当年在国内召开了以“数据科学与大数据的科学原理及发展前景”为主题的香山科学会议[10];国家统计局与阿里、百度等11家企业联手,共同签署了战略合作框架协议。习近平总书记指出:“浩瀚的数据海洋就如同工业社会的石油资源,蕴含着巨大生产力和商机。谁掌握了大数据技术,谁就掌握了发展的资源和主动权”[11]。因此,2013年也被称为中国的大数据元年。2014年,大数据首次写入国家政府工作报告;2015年,国务院发布“关于促进大数据发展行动纲要”将发展大数据产业上升为国家战略,指引国内大数据发展的顶层设计和总体部署[12];2017年,习近平总书记提出“实施国家大数据战略加快建设数字中国”[13];2018年,在习近平总书记给中国国际大数据产业博览会的贺信中明确提出,要“把握好大数据发展的重要机遇,促进大数据产业健康发展,处理好数据安全、网络空间治理等方面的挑战”[14];2020年,“中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议”中明确提出要“加快数字化发展”[15];2021年,习近平总书记在向“可持续发展大数据国际研究中心成立大会暨2021年可持续发展大数据国际论坛”致贺信中提到“世界正遭受新冠肺炎疫情巨

大冲击,科技创新和大数据应用将有利于推动国际社会克服困难"[16]。

由此可见,大数据已经得到世界范围内的重视,其发展趋势势不可挡。随着大数据的飞速发展,现实世界的物理空间和人类社会空间被映射到虚拟的数字空间,形成了除现实世界外的数字世界。传统的数据处理思维和技术在新生的数字世界中已步履维艰,为了更好地挖掘和研究数字世界背后所蕴藏的科学问题,急需寻找治理数字世界的方法论和科学技术。因此,国内外学者提出了一个新的研究领域——数据科学。

1.2 数据科学的发展

数据科学一词最早出现在1974年出版的著作《Concise Survey of Computer Methods》中,书中写到"数据科学是一门基于数据处理的科学"[17],作者认为数据处理后可以和其他领域建立起联系,这种联系将为该领域的科学提供参考与借鉴。然而,数据科学研究并没有因此得到学术界的重视,经历了漫长的沉默期。直到2001年,国际杂志《International Statistical Review》上发表的"Data Science:An Action Plan for Expanding the Technical Areas of the Field of Statistics"一文中提出"数据科学是统计学的一个重要研究方向",使得统计学领域开始关注数据科学的研究[18]。2013年,Mattmann在《Nature》上发表"Computing:A Vision for Data Science",从日常研究存在的数据问题出发,讨论了数据科学存在的必要性以及数据科学的内涵,将数据科学引入计算机科学与技术领域,使得计算机科学与技术领域开始关注这一研究方向[19]。不过,数据科学正式进入大众视野,受到社会各界的广泛关注,主要是由于以下2个标志性事件的发生[20]:(1)2012年,Davenport和Patil在《Harvard Business Review》上发表的"Data Scientist: The Sexiest Job of the 21st Century"指出"数据科学家是公司竞相招聘的对象"[21];(2)2015年,Patil被聘请成为白宫首任数据科学家,这是美国白宫第一次设立数据科学家岗位。

1.3 数据科学的内涵

数字化、网络化、智能化是联结物理世界、人类社会和数字世界所构成的三元世界的载体[22]。其中,数字化正从计算机化向社会全面数据化发展,数据逐渐成为一类新的科学范式、一项新的高新技术以及一种新的决策方式,进而衍生出研究数据的科学,即数据科学。徐宗本院士基于研究对象、研究方法以及研究目标3个维度,在《数据科学:它的内容、方法、意义与发展》一书中将数据科学定义为"数据科学是有关数据价值链实现的基础理论与方法学,运用建模、分析、计算和学习杂糅的方法研究从数据到信息、从信息到知识、从知识到决策的转换,并实现对现实世界的认知与操控"[23]。一门科学的内涵应该既包括方法论和本体论的内容,还包括其学科发展的内容。因此,接下来将从这3个角度讨论数据科学的内涵。

从方法论的角度来讲,数据科学是大数据时代促成的一种新的科学研究范式。从古至今,人类的科学研究经历了经验科学、理论科学和计算科学3种范式[24],图灵奖得主Jim Gray认为现在进入了第四范式"数据科学",即数据密集型科学研究[25]。在基于前3种范式的科学研究中,人们解决问题的方法基本可以总结为:通过反复地观察自然或者做模拟实验得到一定量的实验数据,再分析这些数据得到结论,称之为定理或知识;之后遇到问题时,便可以通过被前人验证过的知识来解决问题。前3种范式的不同之处在于所研究的知识难度的深入以及研究工具的进步,而思维模式都是"从数据中获取知识,运用知识解决问题"。与前3种科学范式所认为的"知识就是力量"不同,第四范式认为"数据也是一种力量",其基本思想是数据驱动科学发现,即把数据看作现实世界在数字世界的映射,通过利用和分析数据可以揭示现实世界所蕴含的科学规律。在数据范式思维模式下,减少对精确模型与假设的依赖,通过数据挖掘出来的知识可能是人类无法理解但是机器能理解并且客观存在的知识,使得过去不能解决

的问题得到解决[26]。

从本体论的角度来讲,数据科学是“用科学的方法来研究数据”的一门新科学[27]。在数字世界中,除了可以反映现实世界中的科学规律,其本身是否也具有类似现实世界的一般性规律?既然现实世界客观存在共性规律,如能量守恒定律、牛顿定律等,那么反映现实世界的数字世界也可能具备某些特有的一般性规律[28]。数据是现实世界在数字世界中的符号化表示,是数字世界的主要构成元素。通过研究数据的历史和进化、形成和发展、类型和属性,获取其本身蕴含的规律和价值,进一步揭示数字世界的内在机理,也是数据科学研究的更基本的问题。

从学科地位的角度来讲,数学科学是一门“理工交叉、文理交融”的学科[23]。其主体构成为数学与统计学、计算机科学与人工智能学科以及各专业领域科学,其中数学与统计学为数据科学提供了研究的理论基础,计算机科学与人工智能学科为数据科学提供了研究的工具和方法,各专业领域知识为数据科学提供了研究的经验与实践应用场景[29]。换言之,数据科学是一套基于大数据时代出现的新理论、新技术、新方法、新模型、新工具和新应用来研究新挑战、新机会、新思维和新模式的知识体系[30]。数据科学生成的多源性、内涵的交叉性以及知识的多学科性搭建起沟通不同学科的桥梁,构建起自身学科体系。

总而言之,数据科学的出现不是一时兴起,也不是昙花一现,而是技术发展,尤其是计算技术、存储技术和网络技术发展的必然产物,也是技术变革的必然趋势。一方面,随着大数据产业的蓬勃发展,大量无法用传统知识解释的结果涌现,需要从理论上对其进行解释、提炼和归纳。另一方面,在大数据时代出现的新理论、新技术、新方法、新模型以及新工具已经走在了传统信息科学的前面,实践倒逼理论的发展完善,需要将其归纳整理成系统的科学理论。

2 国内外研究现状

大数据与数据科学休戚相关,大数据是数据科学研究的基础和对象,数据科学就像大数据的“灵魂”,看不见、摸不着,但却是实现数据价值的关键。基于大数据的理论、技术和应用都取得了重要的突破,本节将介绍大数据和数据科学的研究现状,以及数据与各行业的融合情况。

2.1 大数据研究现状

不同的领域具有的数据体量、数据类型以及产生数据的速度都不尽相同,因此对其数据具体的处理方法也有不同的选择,但是归根结底,对其数据的基本处理流程大同小异。孟小峰教授从数据抽取和集成到最终结果展示,归纳出了大数据的基本流程[31]。本文在此基础上将其整理为数据采集、数据预处理、数据分析和数据解释4个步骤,如图1所示。

(1)数据采集。数据采集又称数据获取,是大数据处理流程中最基础的一步。一般指通过各类与互联网结合的软硬件产品获得的各种结构化、半结构化及非结构化的大规模数据。如在使用谷歌、百度等搜索引擎,或者微信、微博、QQ等社交网络时,网络人机交互过程中产生的半结构化数据;在互联网的基础上,利用射频识别(Radio frequency identification,RFID)标签和读写器、各类传感器、M2M终端等边缘硬件获得的物联网数据[32];来自于企业内部ERP系统的行业数据;各种POS终端、多媒体终端的数据等。

(2)数据预处理。数据采集的数据源大都是多源异构的,采集得到的数据质量通常比较差,大多存在着数据缺失、不一致、冗余或者有噪声等问题。如果直接对其进行数据分析,会使得数据分析难度大,分析结果质量低,结果往往不够理想,达不到预期效果。因此,为了方便数据分析,提高分析数据的质量,需要进行数据预处理。如图2所示,数据清洗[33]、数据集成、数据变换和数据归约[34]是目前数据预处理较为常见的4种方法。

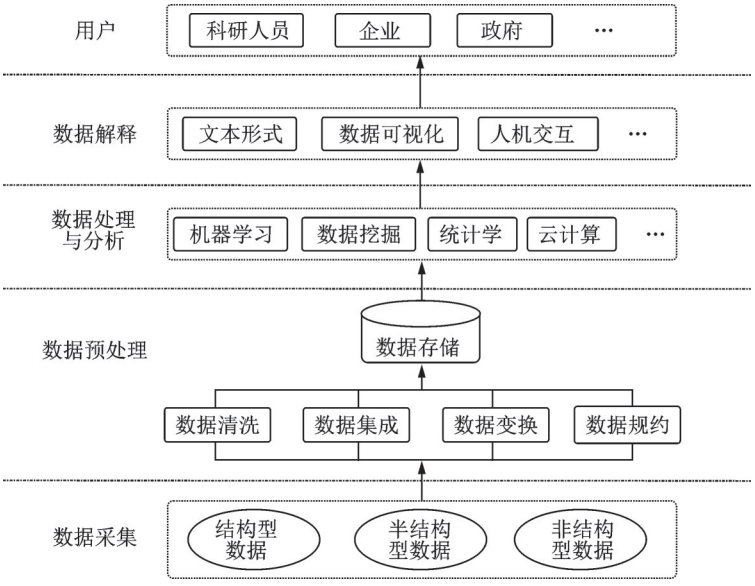

图1 大数据一般处理流程

Fig.1 General process flow of big data

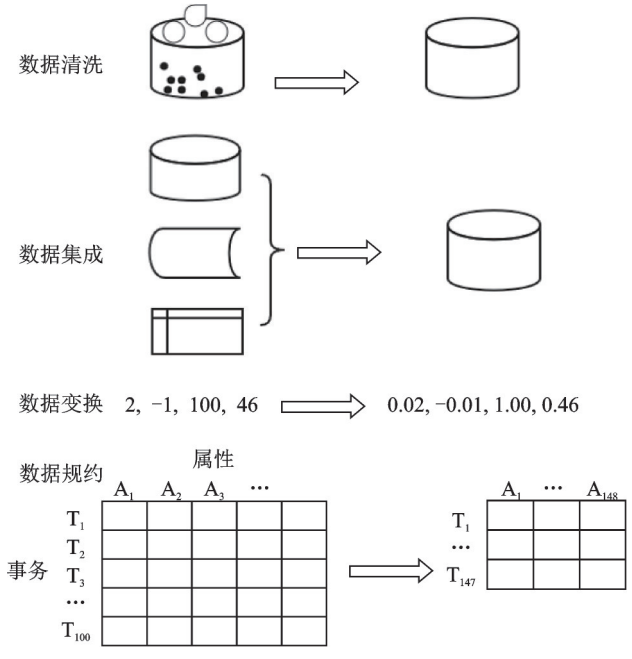

图2 数据预处理的形式

Fig.2 Form of data preprocessing

① 数据清洗,即清除数据集合中的不一致,平滑数据集合中的噪声,改善数据集合中的不完整性等。简而言之就是去除数据中的噪声和无关数据,处理遗漏数据。

② 数据集成,即将互相关联的分布式异构数据在逻辑上或者物理上集中在一起,为用户提供更全面的数据。

③ 数据变换,即通过标准化、离散化等方法让数据变得更一致,更适合分析。

④ 数据归约,即降低数据维度或者减少数据量,简而言之就是缩小数据集规模。

(3)数据分析。大数据处理流程中最直接产生价值的部分就是数据分析,这一步也是处理流程中最核心的部分[35]。因为通过数据分析可以挖掘出数据中蕴含的价值,揭示出隐藏的规律和结果,进一步可以辅助人们进行更为科学和智能化的决策[5]。经过上一步数据预处理后的数据,即为数据分析的原始数据,再根据用户对数据的应用需求对其进行进一步的处理与分析。大数据分析的核心在于如何对数据进行有效的表达、解释和学习。传统的数据分析方法比较依赖于数据的表达,由于表达能力有限,获得的学习效果不尽人意,如基于数学领域的统计分析。随着人工技术的发展,其相关方法为大数据分析提供了更多的可选择性,包括机器学习[36]、智能计算[37-38]以及知识与推理[39]等。这些方法并不都是独立存在的,它们之间互相交叉应用。

此外,云计算是目前在大数据分析领域应用比较广泛的方法,它也是大数据分析处理技术与应用的核心原理和基础平台。实际上,云计算可以根据实际需求,通过网络随时随地访问存储、计算等云端资源,是一种大规模的分布式计算模型,基础设施即服务(Infrastructure as a service,IaaS)、平台即服务(Platform as a service,PaaS)和软件即服务(Software as a Service,SaaS)三个层次组成其体系架构[40]。早在2006年,Google和亚马逊等公司就提出了云计算的构想。另外,Intel和IBM等国外著名互联网公司也都是云计算的忠实开发者和使用者。国内各大互联网公司近年也相继推出各自的云计算平台,如阿里云、百度BAE平台、腾讯云、华为云等。目前,使用较为广泛的云计算技术包括以批处理技术为核心的Hadoop,以高实时性的流处理技术为核心的Storm、Samza,同时拥有流批一体混合处理的Spark、Flink,以及以图处理技术为核心的GraphX等[41],其适用场景如表1所示。

表1 云计算技术适用场景

Table1 Applicable scenarios of cloud computing technology

云计算技术	适用场景
Hadoop	适用于处理对时间要求低的大规模数据集,成本低且灵活。
Storm	适用于对延迟需求高的纯流处理工作,可进行增量计算,实时且高效。
Samza	适用于来自不同团队的多流处理过程,可简化流处理工作,实现低延迟性能。
Spark	适用于多样化工作负载处理任务,批处理速度高于Hadoop,同时Spark Streaming还可用于延迟需求不高的流处理任务。
Flink	适用于有极高流处理需求和少量批处理需求的任务,延迟低且吞吐率高。
GraphX	适用于图数据的处理与计算。

(4)数据解释。对于用户来说,数据的分析处理过程往往不是他们最关心的,数据分析结果的解释与展示才是他们可以直接获取并使用的内容。因此,数据解释环节在数据处理流程中也是不可或缺的部分。如果不能充分且恰当地对数据分析结果进行解释与展示,那么用户可能会产生困扰,甚至被解释得不合理的分析结果误导。传统的数据解释方法大多是以文本的形式进行展示,然而在面对海量数据时,文本形式不能准确直观地解释大数据分析结果之间的关系。因此,可视化技术被引入了大数据领域,数据可视化既是一种分析方法,也是一种解释手段。现代的数据可视化技术是指借助图形化方法,将数据转换为图形图像在屏幕上显示出来,使得数据分析结果更形象[42]。另外,让人直接和机器对话进行解释的人机交互也是正在发展的一种数据解释方法。

总体来说,采集来自不同数据源的结构化数据、半结构化数据和非结构化数据,并将其预处理为统一标准的数据格式,然后再选择合适的数据分析方法进一步对其处理,并将分析结果利用可视化等技术解释并展现给用户,就是大数据处理的一般流程。

2.2 数据科学的研究现状

“用数据的方法研究科学”和“用科学的方法研究数据”是数据科学研究的两个主要角度[43]。其中,用数据的方法研究科学主要在天体信息学、生物信息学等领域应用,如著名的开普勒第三定律“行星绕太阳运行的周期的平方和行星离太阳的平均距离的立方成正比”便是基于观测到的数据归纳总结得到的,开普勒本人也并不理解其内涵;用科学的方法研究数据主要在统计学、机器学习[44]和数据挖掘[45]等领域应用,主要研究的是处理数据的技术和探索数据本身存在的共性。数据科学的出现不仅有利于研究对于海量数据的处理利用,还有利于融合不同学科领域的数据研究,解决各领域中传统知识无法解释新兴数据的矛盾。因此,近10年数据科学吸引了大量学者对其进行研究。

William认为数据科学扩大了统计分析的技术领域,提出了数据科学的6个技术工作领域,并主张为每个领域的研究分配专门的资源[46]。Grady等提出大数据分析的过程模型,在数据分析和机器学习系统开发的生命周期中实现敏捷性,以最大限度地减少实现理想任务结果所需的时间,使从数据中产生价值和花费时间之间达到最佳点[47]。Parmiggiani等将数据科学和跨学科专业的石油天然气领域相结合,不仅考虑如何分析数据,同时考虑数据的全面性以及数据未来的潜在用途[48]。Lise从整合算法和统计原理、社会科学理论和基本人文主义的角度,思考了如何理解数据科学中涉及的道德和社会问题[49]。Juan等对数据科学和人工智能在自然计算和人工计算之间的相互作用进行了总结,并分析和讨论了其应用趋势[50]。Deepak等设计和应用数据科学相关技术进行假新闻检测,以应对假新闻的威胁[51]。

叶鹰和马费成研究了数据科学与信息科学的关联,揭示了两者之间3个“三位一体”的基本原理,即数据-信息-知识、计算技术-数学方法-专业知识、人-技术-数据[52]。王仁武基于Python进行数据科学相关实践,从敏捷式角度对大数据进行开发和应用,并进行可视化展示[53]。朝乐门系统地研究了数据科学的理论、技术、实践以及人才培养,从数据科学的科学内涵、学科地位及知识体系出发,分析了数据科学的研究特点,探讨了数据科学中的争议和挑战,并提出数据科学的发展趋势[54-56]。李扬等从数据科学的起源、基础技能、分析方法和应用等方面展开讨论,建立完整的知识体系和逻辑[57]。徐宗本等从数据科学的产生背景出发,综合性论述了其科学概念与内涵、研究意义与方法、发展趋势与规律、与其他学科的关联与区别、核心问题与研究方向以及人才培养方案等多方面内容[23]。

发展至今,数据科学存在着亟待解决的重大科学技术问题。在重大科学问题方面,探索数据空间的结构与特性、建立大数据统计学、革新存储计算技术和夯实人工智能技术是值得关注与挑战的4大科学任务;在核心技术方面,物联网、大数据互操作、大数据安全、大数据存储、分布式协同计算、新型数据库、大数据基础算法、数据智能、区块链、大数据可视化与交互式分析是应该重点突破的10大技术方向[23]。

(1)探索数据空间的结构与特性。数据是构成数据空间的元素,数据空间本应是数据科学最基本的研究对象,作为研究者理应对数据空间的特征、结构、特性等有所了解。然而,现今数据科学研究大都将其作为知识发现的工具,而并非把数据空间作为最主要的研究对象。为了进一步探索数据空间的结构与特性,从数据的角度来看,可研究数据的复杂性和不确定性,以及有关数据的度量、演化与利用;从数据空间的角度来看,可赋予数据空间某种数学结构,如代数结构、拓扑结构等,使其成为数学上的空间,从而在数学意义下可以将其按照某些特定规律去运算,或使用某些特定工具去分析[23,58]。

(2)建立大数据统计学。传统统计学通常是“先问题,后数据”的模式,即根据问题需要,先通过抽样调查获取结构化数据,再对数据进行建模与分析获得结论,最后检验结论。而在大数据时代,远远超出传统记录与存储能力的半结构化和非结构化数据推动统计学向数据科学变革,衍生出“先数据,后问题”的新模式。大数据的出现,给统计学带来了挑战,建立适用于大数据分析与利用的统计学新理论和

新方法,是数据科学目前迫切需要解决的问题[23,58]。

(3)革新存储计算技术。大数据时代下,数据不再是有限、固定、不可扩充的,也不再存储在某单独的设备上,而是以“流”的方式实时给出,存储在计算外设的磁盘、不同机器或边缘端的分布式环境、甚至多处理器和共享RAM的环境中。基于传统计算理论的算法在大数据环境下失效,革新大数据存储计算技术,设计出具有低复杂性的大数据计算基础算法是数据科学当下面临的核心挑战[58]。

(4)夯实人工智能技术。作为新一代信息技术的代表,人工智能技术已经成为数据科学研究的核心工具与方法之一。然而,人工智能技术本身也仅仅是突破了从“不可用”到“可用”的技术拐点,然而从“可用”到“用得好”还存在着诸多技术瓶颈,需要夯实理论基础研究,发展技术创新与变革,探索理论禁区和未知领域[58]。

综上所述,关于数据科学的研究目前大都聚焦在其技术革新方向,而对于数据空间结构与特性间的探索、数据本身共性和规律的研究均较少,是以后值得重点关注和研究的方向。

2.3 数据在各行业中的应用

在大数据时代,几乎所有的行业都能看到大数据的身影,整体呈现从热点行业领域逐渐向传统行业渗透的趋势。大数据应用是将大量的原始数据汇集在一起,通过分析数据中潜在的规律,预测事物的发展趋势,帮助企业做出正确的决策,从而提高各个行业的运行效率,取得更大的收益。哪个行业能率先从大数据中发现其暗藏的宝藏,挖掘出“金矿”,哪个行业就能够抢占先机成为领先者。目前,已经与大数据开始融合的行业有很多,本节将主要列举以下5个大数据应用较为广泛的领域,如图3所示。

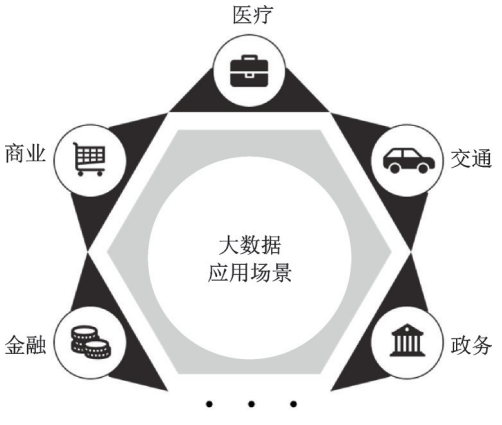

图3 大数据应用场景

Fig.3 Application scenarios of big data

(1)金融领域。金融业作为数据最密集的行业之一,在大数据时代中已然占有一席之地。在传统的银行、保险、证券行业中,可以通过获取、分析更多维度、更深层次的数据,让原来不可担保的信贷可以担保,不可保险的风险可以保险,不可预测的证券行情可以预测。另外,在信息时代特有的互联网金融行业中,大数据打破了传统金融数据的孤岛形态,使不同维度的数据相互融合,从传统的静态数据变成了可以相互融合的动态数据,数据之间的整合能力越来越强,产生更深度的联系[59]。

(2)商业领域。商业领域的数据体量巨大、集中度高、种类多,依托商业大数据分析,企业可以针对性地进行产品设计、计划生产、资源配置等,有利于精细化生产,从而提高生产效率,优化资源配置。另外,商业数据还可以记录客户的购买习惯,预测客户消费习惯、消费行为的相关性、消费趋势、流行趋势等,从而将客户可能会购买的东西精准推送给客户,既卖出了产品,又提高了客户体验。

(3)政务领域。随着大数据的出现,行政思维模式由传统的经验治理转向科学治理。目前,大数据政务应用已经逐渐获得世界各国政府的重视,中国政府也不例外。在《国务院关于印发促进大数据发展行动纲要的通知》[12]中提到,“大数据成为提升政府治理能力的新途径”,要“打造精准治理、多方协作的社会治理新模式”。基于政务大数据,一方面可以帮助政府了解城市经济发展情况、各产业发展情况、居民消费支出情况等,依据分析结果,可以提高政府宏观调控的科学性、预见性和有效性;另一方面,可以实现政务服务一号认证(身份认证号)、一窗申请(政务服务大厅)、一网办事(联网办事),大大地简化了办事手续,提升民众的幸福感[60]。

(4)医疗领域。随着医疗行业和计算机技术结合越来越紧密以及医疗信息系统的不断发展,大量的病历报告、医疗方案、药物信息被存储在数据库中,如果对这些数据进行收集整理和分析,将会给医生和患者带来很大的帮助。对于医生来说,依托医疗大数据,可以积累和分析病例档案、治疗方案,建立疾病诊断模型,从而帮助医生进行疾病诊断,并向医生推荐治疗方案;对于患者来说,借助医疗大数据,基于移动互联网和物联网,可以进行疾病自查、远程医疗等,在家也能看病,让医疗无处不在。总体来说,医疗大数据的运用从医学研究、电子病历管理、临床决策、疾病诊断以及患者参与等多个方面推动了医疗模式的转变,充分尊重了患者的个体化特征与需求,协调并整合了不同专业的医疗服务,保证了医疗服务的连续性和可及性,提高了医疗质量[61]。此外,大数据在支撑流行病病毒溯源、诊断监测和研判排查疫情的过程中也具有不可替代的作用,如COVID-19流行病防疫中出现数字接触追踪技术,即利用大数据对患者进行追踪,确定其活动场所和密切接触者等,以帮助防止疾病蔓延[62,63],以及寻求病毒传播和社会经济活动之间的平衡[64]。

(5)交通领域。交通数据资源丰富、类型繁多,且具有实时性的特征,基于交通大数据,交通运行管理完善与优化、面向车辆和出行者的智能化服务、交通应急和安全保障等方面都得到发展。通过整合分析航班、火车等公共交通工具的信息,从社会角度来看可以提高基础设施的利用效率,降低其运行成本,提高道路运输能力,减少交通事故的发生;从个人角度来看可以提供出行路径规划,实时交通情况,航班铁路动态信息服务,使出行更便捷。

除了上述提到的5个行业外,数据在其他行业的应用也非常广泛,如工业、农业和物流业等领域。从宏观的角度,大数据和各行业融合的思路可以大致归结为[59]:(1)加强企业内部的部门联系,提高管理效率;(2)以人为本,从客户的角度出发进行个性化内容和服务定制;(3)促进行业创新,发掘新需求,进行产品和服务的创新,从而降低成本,提高回报率。但是,大数据的应用还存在着受制条件,如数据质量、法律法规、社会伦理等,其实际效果也还需要时间的检验。

3 数据科学的基础设施

数据科学是支撑大数据时代发展的基础学科,要探索数据世界、治理数据世界,就必须要发展数据科学。每一个科学都需要探索,在探索过程中都需要做试验或者实验,试验是探索,实验是验证。研究数据科学,探索其内在规律需要一个大数据基础设施,统筹大数据处理的整个流程,让大数据处理更便捷、更易操作、更贴近用户,使得数据更具有生命力和价值。邬江兴院士提出的“大数据试验场”便是一个类似于基础设施的概念。邬江兴院士认为,计算技术、存储技术、网络通信技术的进步速度,如何跟上数据增长的速率是亟待解决的问题[65]。正是因为现有技术不能解决问题,故要发展新的技术、新的理论,这些理论和技术要通过试验来证明其可行性。因此,建立大数据试验场来研究数据科学的基本理论和方法势在必行。

3.1 大数据试验场的内涵

邬江兴院士表示,现阶段中国大数据技术大部分是利用国外开发的开源软件,而此次提出建设大数据科学基础设施是中国原创,是全球范围内首次提出[66]。大数据基础设施[67]区别于传统的交通、建筑和水利等硬件基础设施,也区别于数据中心、网络通讯等传统信息和通信技术基础设施,主要用于支撑大数据、区块链、云计算和物联网等新一代信息技术落地应用的底层架构和人才资源,包括“物本”和“人本”2个层面。其中,“物本”整合了云计算、边缘计算、安全多方计算和知识图谱等前沿技术;“人本”则是指具备数据思维与技能的人才以及相应的教育标准和体系。大数据基础设施的重要组成部分之一便是大数据试验场,它以支撑科学研究、技术创新、产业创新和创新创业为目标,通过科学研究引导技术创新,从而提升产业发展质量[65]。

大数据试验场是一种为了探索数据科学内在规律，解决大数据技术问题，设计出的面向数据科学的理论试验和技术研发验证环境。它拥有大规模数据存储能力和海量数据管理分析能力，服务于大数据研究与开发、科技与产业创新以及数据科学人才培养，并且面向全球开放运行。为了更直观地理解大数据试验场，以矿场来类比大数据试验场，如图4所示。海量的大数据就如同深不可测的矿场，一开始只挖掘比较表面和浅层的资源，然而浅层矿总有枯竭的一天，继续挖掘深层矿时，便会遇到区别于浅层矿的科学问题，因此也就需要专业人才学习并研究新的知识，创造新的采矿手段、新的挖掘技术以及新的工艺工具。由此可知，大数据试验场解决的便是挖深层矿的理论问题、工程技术问题、装备技术问题和人才培养问题。

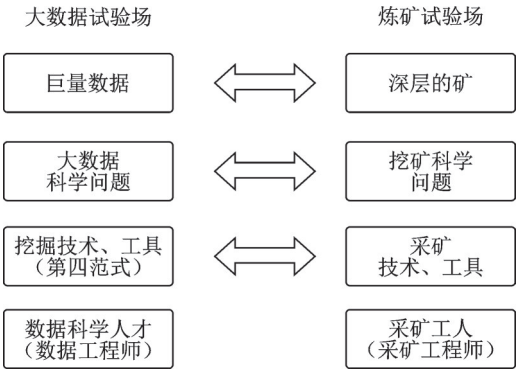

图4 大数据试验场的类比

Fig.4 Analogy of big data proving ground

3.2 大数据试验场研发现状

2016年12月2日，以解决超大规模数据的科学与应用、大数据的科技与产业创新、政策决策推演等问题为出发点，以针对新型计算、存储、网络以及液冷技术进行深入研究和市场化引导为目的，复旦大学和上海交通大学联合牵头，与29家高等院校、研究所和企事业单位在上海共同成立了大数据试验场联盟，并在复旦大学张江校区成立国家大数据试验场[68]。2019年8月2日，广东合一新材料研究院有限公司承担的大数据试验场中心平台之“液冷型大数据试验场”通过验收。液冷型大数据试验场针对大数据网络与存储需求，利用芯片级精准喷淋液冷技术以解决散热要求的难题，开发出集装箱式模块化数据中心以实现快速灵活部署，计算服务集群的功率密度和空间密度均得到提高，满足了各项功能和性能指标要求[69]。

2020年，重庆邮电大学开始筹建大数据智能计算省部共建重点实验室，为此正在建设大数据试验场算力平台，其目的在于探索数据科学本身的内涵和规律，引导产业的创新和行业的发展。在一些知名企业的指导下，构建了千万元级的算力，长期目标是打造示范性算力基础设施，构建大数据试验场基础设施，在高校中打造算力平台的典范，为重庆市实施大数据智能化发展做出贡献。

4 探索数字世界

从电子计算机发明的那一天起，人类数字化生存的帷幕就已经拉开，人类逐渐从现实世界走进数字世界，在两个世界维度自由穿梭、协同发展。互联网的出现让机器的互联互通成为可能；移动通信与互联网的结合使得数据传输从固定终端转移到移动终端，让信息共享变得更加及时高效；物联网通过传感器，使人与人之间、人与物之间、物与物之间构建起万物互联的数据世界，让现实世界精确映射到数字世界成为可能。自此，除现实世界的物理空间和人类社会空间以外，第三空间被构造出来，即虚拟的数字空间。数字世界是现实世界的基本映射，其基本要素是数据。如图5所示，淘宝、京东等购物软件就类似于现实世界中的贸易市场和商场，谷歌地图、百度地图等地图软件勾勒出现实世界的地表地貌及道路交通系统，美团、飞猪、携程等生活软件便映射出现实世界的吃住行，微信、微博等社交软件也在一定程度上反映出人类社会的社交关系。由此证明，要探索数字世界、治理数字世界，就必须发展数据科学。

4.1 数据科学的关键问题

由于科学研究的发展和外部环境的推动,科学研究范式本身也随之发生变化[24]。几千年前的第一范式“实验科学”,科学家主要通过反复观察,描述和记录自然现象,如钻木取火等;进入19世纪,科学家发现由于实验条件的限制,对自然现象无法精确理解,于是开始简化实验模型,以理论研究为主,通过脑力思考和人力计算对现实中的一般规律进行概括,如经典力学中的牛顿定律、物理学中的相对论等,这一研究模式被称为第二范式“理论科学”;随着20世纪中期计算机的出现,科学家开始利用计算机解决复杂问题中的大量计算问题,以及模拟仿真自然界中的复杂现象,如模拟伤害范围过高、伤害程度过大的核试验等,这一研究模式被称为第三范式“计算科学”;21世纪,互联网的蓬勃发展使得巨量数据源源不断产生,科学家认为数据世界就如同现实世界,本身应该也蕴藏着规律和价值,因此提出了区别于传统科学研究的第四范式“数据科学”。科学研究的4种范式总结概括如表2所示。

图5 现实世界和数字世界

Fig.5 Real world and digital world

表2 4种科学研究范式

Table 2 Four scientific research paradigms

科学范式	时间	内涵	举例
第一范式(实验科学)	几千年前	观察、描述和记录自然现象	钻木取火,比萨斜塔实验等
第二范式(理论科学)	几百年前	理论总结和概括自然界一般规律	牛顿定律,相对论等
第三范式(计算科学)	几十年前	使用计算机模拟仿真复杂现象	模拟核试验,天气预报等
第四范式(数据科学)	现今	基于数据的理论、实验和模拟	研究天文学,研制新药等

需要注意的是,4种范式并非是依次替代的关系,不是所有的问题都适合以数据科学或者其他某一种范式的思维模式解决。经验科学的理论来源是理论科学,即在现有理论的基础上进行实验;理论科学的实验过程是经验科学,即通过反复实验得到正确理论,两者相辅相成、互相推进。由于并非所有的问题都可以通过人工实验的方式解决,计算科学便被提出,用来对经验科学和理论科学进行补充和优化;而数据科学则用于处理经验科学和计算科学中出现的大数据问题,进一步完善前3种科学范式。

数据科学以数据为研究对象,其特征可概括为以下3个方面:(1)不在意数据的杂乱,但看重数据有足够的量;(2)不要求数据精准,但强调面面俱到,不一定涵盖所有的数据,但各个方面都要有代表性数据;(3)不刻意追求因果关系,但重视规律总结,这意味着不局限于追求因果关系,更多在于追求关联关系。因此,研究数据科学,本文认为可以从以下3个问题入手。

(1)数据聚合效应。数据科学研究中的数据往往来自不同的领域,具有较大的差异性,将这些来源不同、类型不同的数据在一定准则下自动聚集、自动融合、自主分析,可以挖掘更多有价值的信息,为质变提供量变基础。数据聚合有两种效应:一种是数据叠加,即数据简单地叠加变成更大的数据,从而挖掘出小数据中挖掘不到的知识,类似于现实世界中的物理变化;另一种是数据融合,即数据按照一定的规律重新结合成新的数据,数据的量不一定增加,但是所蕴含的信息已经不同于之前,类似于现实世界

中的化学变化。然而,不管是数据叠加还是数据融合,都可以实现“1+1≫2”的效果。

(2)数据成像原理。大数据之所以有用,是因为数据累积到了一定数量,到大数据临界点时可以发生质变,通过数据挖掘其背后的规律,进而还原“真相”,即还原数字世界中事物本身存在而人类可能无法事先知晓的客观规律。大数据用户画像便是数据成像的一个例子。先收集各种类型数据,包括网络行为数据、用户内容偏好数据和交易数据等,当数据足够大、足够有代表性、覆盖够全面时,便可以对用户的行为进行建模,抽象出用户的基本属性、行为特征和兴趣爱好等标签,使得用户的形象越来越完整和立体,从而不断地逼近现实中人的特征。

(3)数据态势感知。大数据通过聚合分析,发掘其背后的规律,还原真相后,主要用于预测分析,即采用态势感知、关联分析等方法对数据进行计算,挖掘数据之间的内在关联,不仅能还原真相,更要预测未来。可以尝试通过关联分析进行行为分析与预测,或者通过多粒度随机抽样进行层次化统计预测。如果数据态势感知问题得到解决,并应用到地震预警、流行病预估和慢性病预判上,将在推进社会进步方面取得重大突破。

在研究过程中,数据安全与隐私问题不容忽视。在当前数据即资源的形势下,数据逐渐成为各国博弈的资本,其安全与隐私问题值得高度重视,需要对数据的各方面采取有效的保护措施与手段,预防数据泄露、数据篡改等情况的发生。

4.2 数字世界思维

大数据发展前期的主要任务是收集数据,现在已逐渐向数据治理、数据驱动的方向转换,从而推动着数“字”世界向数“智”世界的转换。数字世界是现实世界的基本映射,这个映射空间目前还不是孪生,但是随着数据科学的不断发展,现实世界被越来越精确地映射到数字世界中,数字世界成为现实世界的孪生镜像将成为必然。从数字世界,到用数据治理世界(数治世界),从而实现数字时代的智能世界(数智世界),便是从“字”到“治”,最终实现“智”的过程,三者关系如图6所示。本文从思维和观念两个角度来进行转变。

图6 从数字世界到数智世界

Fig.6 From digital world to intelligent world

4.2.1 转换研究思维

数据科学的研究主题大致分为核心问题和周边问题[21]。其中,核心问题指的是数据科学的基础理论,即数据科学自身具有的理论、方法、模型和技术等;周边问题指的是与数据科学相关的其他现有科学研究,如统计学、机器学习、云计算、物联网和大数据应用等。现有文献表明,目前数据科学的研究主要以周边问题为热点,对于核心问题的研究相对较少,因此可以更多关注数据科学的基础理论研究。与研究数据科学的周边问题相比,研究数据科学的核心问题应当具有以下3种思维。

(1)跳转思维。未来构建从物理世界到数字世界的双射是必然趋势,现实世界与数字世界是双向互通、自由跳转的,可以通过映射现实世界来构建数字世界,分析数字世界来治理现实世界。因此,具有从宏观到微观自由切换的跳转思维有助于研究数据科学的基础理论。

(2)熵减思维。熵是衡量事物混乱程度的一个指标,在具有爆发式的数据增长,以及数据异构、多源等问题的大数据时代下,熵增也越来越快。基于负熵理论和熵增理论,在数据治理的过程中找到数据背后隐含的规则,使数据达到从无序到有序的辩证统一,从而实现熵减,是一个值得借鉴的思维模式。

(3)算法思维。在数字世界中,软件可定义一切,一切皆可计算,如何基于更多的数据设计出简单高效的算法将成为重要挑战,数据工程师将成为时代的新宠。因此,抽象现实问题并对其进行编码或设计程序解决的算法思维也是不可或缺的。

4.2.2 转变传统观念

在进行科学研究工作时,大多容易陷入改进现有理论和方法的局限里,缺乏重新审视现有方法和结论的勇气。然而在面对数字世界这个新兴的事物时,要跳出传统的研究观念,充分发挥创新思维,敢于怀疑,敢于想象,敢于探索。

(1)敢于怀疑。对于科学而言,证伪和证实同样重要,正是由于科学在不断地怀疑自己,才有今天这样的发展,如果盲目相信已有的科学成果,那么科学的发展将会停滞不前。如年轻的伽利略敢于怀疑著名思想家亚里士多德提出的"物体从高处坠落,重的下落快,轻的下落慢",并登上比萨斜塔在众目睽睽之下进行实验,用事实推翻了亚里士多德的观点,揭开了自由落体运动研究的序幕。对于现有基于传统的"数据→知识→问题"的思维模式而得到的理论,要敢于怀疑其在大数据中的可行性,从数据科学"数据→问题"的角度对现有理论进行分析和验证,换个角度可能会得到意想不到的结果,如图7所示。

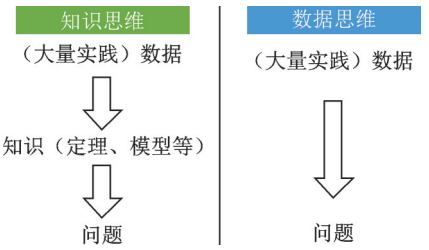

图7 解决问题的不同思维方式

Fig.7 Different ways of thinking to solve problems

(2)敢于想象。现有的科学和知识是有限的,而想象力是无限的,只有敢于想象,才能推动科学的进步,知识的发展。敢想才能实现,正是因为拥有想象力,各个时代才会出现前所未有的新事物,如电气时代的发电机、电话、飞机,信息时代的电子计算机、原子弹、人造卫星等。生活在大数据时代的我们,应当抓住数字世界研究的机遇期,充分发挥想象力,相信数字世界存在更多的类似摩尔定律的一般性规律,并通过研究论证其真实性。

(3)敢于探索。人类发展至今,对科学的探索从未停下,如牛顿被苹果砸中,探索出万有引力;爱迪生探索做灯丝的材料,发明出电灯;居里夫人反复探索3年多,发现放射性元素镭等。现今,对于数字世界的研究才刚刚开始,还有很多问题值得去探索,如果一味地做补丁式研究,则很难做出开拓性创新。在数字世界的探索中,不能仅仅"站在巨人的肩膀上"对现有方法进行改进,更应该勇于创新,发挥主观能动性,实现更多从0到1的原始创新。

总之,对于正在形成的数字世界,其研究才刚刚开始,处处都需要去挖掘和探索。面对浩瀚的数字世界,要抓住大数据带来的机遇,灵活运用数据"治"理数"字"世界,发展数"智"世界,推进世界智能化进程。

5 结束语

大数据作为一项潜在价值巨大的资产,自从其问世以来,其相关技术及应用研究一直是科学界的关注重点和研究热点。与大数据类似,数据科学的周边问题也是科学界的重点研究方向,而数据科学的核心问题作为研究难点则一直没有取得较大突破。本文从大数据谈到数据科学,将数据科学目前的相关问题进行归纳总结,并提出建立从数"字"世界(世界数字化),到数"治"世界(数据治理、数据挖掘时代),再到数"智"世界(智能化世界)的研究思维模式,以期促进大数据和数据科学相关研究的发展,加快各行业中信息公开、权威科学的公共数据库的建设进程。

参考文献:

[1] LOHR S. The age of big data[N]. New York Times, 2012-02-11.

[2] BUXTON B, GOLDSTON D, DOCTOROW C, et al. Big data: Science in the petabyte era[J].Nature, 2008, 455(7209): 8-9.

[3] 涂新莉,刘波,林伟伟.大数据研究综述[J].计算机应用研究,2014,31(6): 1612-1616.

TU Xinli, LIU Bo, LIN Weiwei. Survey of big data[J]. Application Research of Computers, 2014, 31(6): 1612-1616.

[4] 李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J].中国科学院院刊,2012,27(6): 647-657.
LI Guojie, CHENG Xueqi. Research status and scientific thinking of big data[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657.

[5] 程学旗,靳小龙,王元卓,等.大数据系统和分析技术综述[J].软件学报,2014,25(9): 1889-1908.
CHENG Xueqi, JIN Xiaolong, WANG Yuanzhuo, et al. Survey on big data system and analytic technology[J]. Journal of Software, 2014, 25(9): 1889-1908.

[6] [美]托夫勒,著.第三次浪潮[M].黄明坚,译.北京:中信出版社,2006.
TOFFLER A. The third wave[M]. HUANG Mingjian, translate. Beijing: China Citic Press, 2006.

[7] Nature. Big data[EB/OL].(2008-09-03)[2022-01-10]. http://www.nature.com/news/specials/bigdata/index.html.

[8] MANYIKA J, CHUI M, BROWN B, et al. Big data: The next frontier for innovation, competition, and productivity[EB/OL]. [2011-05-01][2022-01-10]. https://www. mckinsey. com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.

[9] ERCIIM News. Big data—Introduction to the special theme[EB/OL].(2012-04-03)[2022-01-03]. https://ercim-news.ercim.eu/en89/special/big-data-introduction-to-the-special-theme.

[10] 徐宗本,张维,刘雷,等.数据科学与大数据的科学原理及发展前景——香山科学会议第462次学术讨论会专家发言摘登[J].科技促进发展,2014(1): 66-75.
XU Zongben, ZHANG Wei, LIU Lei, et al. Data science and the scientific principles and prospects of big data—Excerpts from the 462nd Symposium of Xiangshan Scientific Conference[J]. Science & Technology for Development, 2014(1): 66-75.

[11] 中共中央文献研究室.习近平关于科技创新论述摘编[M].北京:中央文献出版社,2016.

[12] 国务院.国务院关于印发促进大数据发展行动纲要的通知[EB/OL].(2015-09-05)[2022-01-10]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

[13] 新华网.习近平:实施国家大数据战略加快建设数字中国[EB/OL].(2017-12-10)[2022-01-10]. https://www.ccps.gov.cn/xytt/201812/t20181212_123952.html.

[14] 新华社.习近平向2018中国国际大数据产业博览会致贺信[EB/OL].(2018-05-26)[2022-01-10]. http://www.gov.cn/xinwen/2018-05/26/content_5293886.htm

[15] 新华社.中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议[EB/OL].(2020-11-03)[2022-01-10]. https://www.ndrc.gov.cn/fggz/fgdj/zydj/202011/t20201130_1251646.html?code=&state=123.

[16] 人民日报.习近平向可持续发展大数据国际研究中心成立大会暨2021年可持续发展大数据国际论坛致贺信[EB/OL].(2021-09-07)[2022-01-10]. http://politics.people.com.cn/n1/2021/0907/c1024-32219363.html.

[17] NAURU P. Concise survey of computer methods[M]. New York: Petrocelli Books, 1974.

[18] CLEVELAND W S. Data science: An action plan for expanding the technical areas of the field of statistics[J]. International Statistical Review, 2010, 69(1): 21-26.

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

[20] 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 Lemen, XING Chunxiao, ZHANG Yong. Data science studies: State-of-art and trends[J]. Computer Science, 2018, 45(1):1-13.

[21] DAVENPORT T H, PATIL D J. Data scientist: The sexiest job of the 21st century[EB/OL].(2012-10-01)[2022-01-10]. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

[22] 徐宗本."数字化、网络化、智能化"新一代信息技术的聚焦点[J].科学中国人,2019(7): 36-37.
XU Zongben. The focus of the new generation information technology of "digitalization, networking and intelligence"[J]. Scientific Chinese, 2019(7): 36-37.

[23] 徐宗本,唐年胜,程学旗.数据科学:它的内容、方法、意义与发展[M].北京:科学出版社,2021.
XU Zongben, TANG Niansheng, CHENG Xueqi. Data science: Its essence, method, role and development[M]. Beijing: Science Press, 2021.

[24] KUHN T S. The structure of scientific revolutions[M].New York: Continuum, 1999.

[25] TONY H, STEWART T, KRISTIN T,等,著.第四范式:数据密集型科学发现[M].潘教峰,等,译.北京:科学出版社,2012.
TONY H, STEWART T, KRISTIN T, et al. The fourth paradigm: Data-intensive scientific discovery[M]. PAN Jiaofeng, et al, translate. Beijing: Science Press, 2012.

[26] 徐宗本.用好大数据须有大智慧——准确把握、科学应对大数据带来的机遇和挑战[J].中国科技奖励,2016(4): 27-29.
XU Zongben. Big data needs big wisdom—Accurately grasping and scientifically dealing with the opportunities and challenges brought by big data[J]. China Awards for Science and Technology, 2016(4): 27-29.

[27] 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.

[28] 程学旗,梅宏,赵伟,等.数据科学与计算智能:内涵,范式与机遇[J].中国科学院院刊,2020,35(12): 1470-1480.
CHENG Xueqi, MEI Hong, ZHAO Wei, et al. Data science and computational intelligence: Concept, paradigm and opportunities[J]. Bulletin of Chinese Academy of Sciences, 2020, 35(12): 1470-1480.

[29] 朝乐门.数据科学理论与实践[M].北京:清华大学出版社,2017.
CHAO Lemen. Data science theory and practice[M]. Beijing: Tsinghua University Press, 2017.

[30] 朝乐门,卢小宾.数据科学及其对信息科学的影响[J].情报学报,2017,36(8): 761-771.
CHAO Lemen, LU Xiaobin. Data science and its implications on information science[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(8): 761-771.

[31] 孟小峰,慈祥.大数据管理: 概念、技术与挑战[J].计算机研究与发展,2013,50(1): 146-169.
MENG Xiaofeng, CI Xing. Big data management: Concepts, techniques and challenges[J]. Journal of Computer Research and Development, 2013, 50(1): 146-169.

[32] ZHANG Y F, THORBURN P, XIANG W, et al. SSIM—A deep learning approach for recovering missing time series sensor data[J]. IEEE Internet of Things Journal, 2019, 6(4): 6618-6628.

[33] LIN W C, TSAI C F, ZHONG J R. Deep learning for missing value imputation of continuous data and the effect of data discretization[J]. Knowledge-Based Systems, 2022, 239: 108079.

[34] 代少飞,刘文波,王郑毅,等.基于双迭代聚能量字典学习的数据压缩算法[J].数据采集与处理, 2021, 36(6): 1147-1156.
DAI Shaofei, LIU Wenbo, WANG Zhengyi, et al. Data compression algorithm based on dual-iteration concentrated dictionary learning[J]. Journal of Data Acquisition and Processing, 2021, 36(6): 1147-1156.

[35] 刘智慧,张泉灵.大数据技术研究综述[J].浙江大学学报(工学版),2014,48(6): 957-972.
LIU Zhihui, ZHANG Quanling. Research overview of big data technology[J]. Journal of Zhejiang University(Engineering Science), 2014, 48(6): 957-972.

[36] SAMAR B, NADEESHA P. A data-centric review of deep transfer learning with applications to text data[J]. Information Sciences, 2022, 585: 498-528.

[37] LIU K Y, LI T R, YANG X B, et al. Granular cabin: An efficient solution to neighborhood learning in big data[J]. Information Sciences, 2022, 583: 189-201.

[38] 李金海,王飞,吴伟志,等.基于粒计算的多粒度数据分析方法综述[J].数据采集与处理, 2021, 36(3): 418-435.
LI Jinhai, WANG Fei, WU Weizhi, et al. Review of multi-granularity data analysis methods based on granular computing[J]. Journal of Data Acquisition and Processing, 2021, 36(3): 418-435.

[39] LIU H, ZHOU S W, CHEN C F, et al. Dynamic knowledge graph reasoning based on deep reinforcement learning[J]. Knowledge-Based Systems, 2022, 241: 108235.

[40] 罗军舟,金嘉晖,宋爱波,等.云计算:体系架构与关键技术[J].通信学报,2011,32(7): 3-21.
LUO Junzhou, JIN Jiahui, SONG Aibo, et al. Cloud computing: Architecture and key technologies[J]. Journal on Communications, 2011, 32(7): 3-21.

[41] DE ASSUNCAO M D, VEITH A D S, BUYYA R. Distributed data stream processing and edge computing: A survey on resource elasticity and future directions[J]. Journal of Network and Computer Applications, 2018, 103(2): 1-17.

[42] GRMPING U. Data visualization: Charts, maps, and interactive graphics[J]. Journal of Statistical Software, 2018, 98(1): 1-6.

[43] 杨旭.数据科学导论[M].北京:北京理工大学出版社,2014.
YANG Xu. Introduction to data science[M]. Beijing: Beijing Institute of Technology Press, 2014.

[44] 崔建伟,赵哲,杜小勇.支撑机器学习的数据管理技术综述[J].软件学报,2021,32(3): 604-621.
CUI Jianwei, ZHAO Zhe, DU Xiaoyong. Survey on data management technology for machine learning[J]. Journal of Software, 2021, 32(3): 604-621.

[45] 周宇,曹英楠,王永超.面向大数据的数据处理与分析算法综述[J].南京航空航天大学学报,2021,53(5): 664-676.
ZHOU Yu, CAO Yingnan, WANG Yongchao. Overview of data processing and analysis algorithms for big data[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(5): 664-676.

[46] CLEVELAND W S. Data science: An action plan for expanding the technical areas of the field of statistics[J]. International Statistical Review, 2010, 69(1): 21-26.

[47] GRADY N W, PAYNE J A, PARKER H. Agile big data analytics: AnalyticsOps for data science[C]//Proceedings of 2017 IEEE International Conference on Big Data (Big Data). [S.l.]: IEEE, 2017: 2331-2339.

[48] PARMIGGIANI E, STERLIE T, ALMKLOV P. In the backrooms of data science[J]. Journal of the Association for Information Systems, 2021, 23(1): 139-164.

[49] LISE G. Responsible data science[C]//Proceedings of 2019 IEEE International Conference on Big Data (Big Data). [S.l.] IEEE, 2019: 10.

[50] JUAN M G, JAVIER R, ANDRÉS O, et al. Artificial intelligence within the interplay between natural and artificial Computation: Advances in data science, trends and applications[J]. Neurocomputing, 2020, 410(2): 237-270.

[51] DEEPAK P, TANMOYC, CHENG L, et al. Data science for fake news-surveys and perspectives[M]. [S.l.]: Springer, 2021.

[52] 叶鹰,马费成.数据科学兴起及其与信息科学的关联[J].情报学报,2015,34(6): 575-580.
YE Ying, MA Feicheng. Data science: Its emergence and linking with information science[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 575-580.

[53] 王仁武.Python与数据科学[M].上海:华东师范大学出版社,2016.
WANG Renwu. Python and data science[M]. Shanghai: East China Normal University Press, 2016.

[54] 朝乐门,肖纪文,王解东.数据科学家:岗位职责、能力要求与人才培养[J].中国图书馆学报,2021,47(3): 100-112.
CHAO Lemen, XIAO Jiwen, WANG Xiedong. Typical responsibilities: Key qualifications and higher education for data scientist[J]. Journal of Library Science in China, 2021, 47(3): 100-112.

[55] 朝乐门,张晨,孙智中.数据科学进展:核心理论与典型实践[EB/OL].(2022-01-31)[2022-02-10].http://kns.cnki.net/kcms/detail/11.2746.G2.20211228.1820.002.html.
CHAO Lemen, ZHANG Chen, SUN Zhizhong. Development theoretical studies and practical applications in data science[EB/OL].(2022-01-31)[2022-01-31]. http://kns.cnki.net/kcms/detail/11.2746.G2.20211228.1820.002.html.

[56] 朝乐门,王锐.数据科学平台:特征、技术及趋势[J].计算机科学,2021,48(8): 1-12.
CHAO Lemen, WANG Rui. Data science platform: Features, technologies and trends[J]. Computer Science, 2021, 48(8): 1-12.

[57] 李扬,李舰.数据科学概论[M].北京:中国人民出版社,2021.
LI Yang, LI Jian. Data science[M]. Beijing: China Renmin University Press, 2021.

[58] 徐宗本.人工智能的10个重大数理基础问题[J].中国科学:信息科学,2021,51(12): 1967-1978.
XU Zongben. Ten fundamental problems for artificial intelligence: Mathematical and physical aspects[J]. Scientia Sinica (Informationis),2021,51(12): 1967-1978.

[59] TalkingData.数据科学实战指南[M].北京:电子工业出版社, 2019.
TalkingData. Practical guide to data science[M]. Beijing: Publishing House of Electronics Industry, 2019.

[60] 李军,乔立民,王加强,等.智慧政务框架下大数据共享的实现与应用研究[J].电子政务,2019(2): 34-44.
LI Jun, QIAO Limin, WANG Jiaqiang, et al. Research on the implementation and application of big data sharing under the framework of intelligent government affairs[J]. E-Government, 2019(2): 34-44.

[61] 姚琴.面向医疗大数据处理的医疗云关键技术研究[D].杭州:浙江大学,2015.
YAO Qin. Key technologies study on medical cloud for big medical data processing[D]. Hangzhou: Zhejiang University, 2015.

[62] WYMANT C, FERRETTI L, TSALLIS D, et al. The epidemiological impact of the NHS COVID-19 APP[J]. Nature, 2021, 594: 408-412.

[63] THOMPSON B, BAKER N, MAXMEN A. Coronapod: Google-backed database could help answer big COVID questions [N]. Nature, 2021-02-26.

[64] MA K C, LIPSITCH M. Big data and simple models used to track the spread of COVID-19 in cities[J]. Nature, 2020, 589: 26-28.
[65] 新华网.邬江兴院士:建设大数据试验场正当其时[EB/OL].(2022-01-31)[2022-02-10]. https://baijiahao.baidu.com/s?id=1633009144407053227&wfr=spider&for=pc.
[66] 上游新闻.中国工程院院士邬江兴:大数据就像挖矿 要让中国有挖"深层矿"能力[EB/OL].(2019-05-11)[2022-02-10]. https://www.cqcb.com/personage/2019-05-11/1613759_pc.html.
[67] 郭华东.科学大数据——国家大数据战略的基石[J].中国科学院院刊,2018,33(8): 768-773.
GUO Huadong. Scientific big data—A footstone of national strategy for big data[J]. Bulletin of Chinese Academy of Sciences, 2018,33(8): 768-773.
[68] 上海市数据科学重点实验室.大数据试验场联盟正式成立[EB/OL].(2016-12-02)[2022-02-10]. https://datascience.fudan.edu.cn/4d/71/c13525a150897/page.htm.
[69] 广东合一.我司国家"液冷型大数据试验场"项目顺利通过验收[EB/OL].(2019-08-05)[2022-02-10]. http://www.he-one.com/case/showimg.php?lang=cn&id=153.

作者简介:

张清华(1974-),**通信作者**,男,教授,博士生导师,研究方向:粗糙集、模糊集、粒计算、不确定性信息处理等,E-mail: zhangqh@cqupt.edu.cn。

高渝(1997-),女,硕士研究生,研究方向:模糊集、数据挖掘、不确定性信息处理等。

申秋萍(1997-),女,硕士研究生,研究方向:粗糙集、机器学习、数据挖掘、不确定性信息处理等。

(编辑:刘彦东)