

自然科学与人文科学大数据

——第六届中德前沿探索圆桌会议综述*



郭华东¹ 陈润生² 徐志伟³ 孙建军⁴ 毕军⁴ 王力哲¹ 骆健俊² 沈华伟³ 顾东晓⁴ 梁栋¹
沈文庆⁵ 张旭⁵ Hans Wolfgang Spiess⁶ Thomas Lengauer⁷

- 1 中国科学院遥感与数字地球研究所 北京 100094
- 2 中国科学院生物物理研究所 北京 100101
- 3 中国科学院计算技术研究所 北京 100190
- 4 南京大学 南京 210023
- 5 中国科学院上海分院 上海 200031
- 6 Max Planck Institute for Polymer Research Mainz 55128
- 7 Max Planck Institute for Informatics Saarbrücken 66123

摘要 大数据是知识经济时代的战略高地，是国家和全球的新型战略资源。作为思维的革命性创新，大数据为科学研究带来了新的方法论。第六届中德前沿探索圆桌会议以“自然科学与人文科学大数据”为主题，在“生物医药大数据”、“物理、化学与地球科学领域大数据”、“人文与社会科学领域大数据”和“大数据处理技术与方法”4个领域进行研讨，总结了大数据对于科学发现的重要作用、意义以及面临的重大问题，形成了关于发展科学大数据研究的相关建议。

关键词 大数据，科学大数据，生命科学，地球科学，人文科学，社会科学，计算机技术，中德前沿探索圆桌会议

DOI 10.16418/j.issn.1000-3045.2016.06.014

新一轮信息技术革命与人类社会活动交汇融合，引发了数据爆炸式增长，数据类型繁多且复杂，已经超越了传统数据管理系统和处理模式的能力范围，“大数据”概念也应运而生。2014年4月，国际数据公司（IDC）发布的第7份数字宇宙研究报告中指出，全球数据量将以超过每两年翻一番的速度持续增长，2013年全球被创建和被复制的数据总量已达4.4 ZB（Zettabyte，泽字节，1 ZB=10²¹ B），预计到2020年将增至44 ZB^[1]。我国

*资助项目：中科院规划与战略研究专项

修改稿收到日期：2016年4月29日

拥有的全球数据量比例预计也将由 2012 年的 13% 提升至 21%^[2]。大数据已对全球生产、流通、分配与消费模式产生重要影响,正在改变人们生产生活方式、经济运行机制和国家治理模式。大数据作为知识经济时代的一项战略使能技术,是各国的一种新型战略资源。不久的将来,围绕大数据引起的竞争不仅将决定国际信息产业格局,还将深刻影响经济发展、国家安全、科技进步和综合竞争力^[3]。

大数据为分析和推理方法的创新提供了一个全新的、极富前景的路径,同时也为自然科学与人文社会科学的研究提供了新的契机。科学大数据作为大数据的分支体系已成为继实验、理论和计算模式之后的数据密集型科研范式的典型代表,正在从模型驱动模式向数据驱动模式进行转化,带来了科研方法论的创新。科学大数据由各学科产生或收集的规模巨大且多源异构的数据组成,例如生命科学中的基因组数据、地球科学中的观测和模拟数据、化学和材料科学中的测量数据以及数字化的人文历史数据。这些数据亟需在全球科技界实现共享,以实现其价值的充分利用。同时,如何保证数据的可持续性使用也是当前面临的一个严峻挑战。随着数据产生变得日益便捷,数据分析开始成为瓶颈。众所周知,大数据中充斥着偏差和噪声。从大数据中析取知识涉及统计分析和机器学习等技术,然而从数据中得到的往往只是关联关系而非因果关系。对因果关系的探究超出了统计学的能力范畴,至今没有系统化的解决方案。此外,如何让基于统计方法的预测看上去更合理,也是一项重大挑战。

基于以上背景,以“自然科学与人文科学大数据”为主题的第六届中德前沿探索圆桌会议于 2015 年 11 月 19—21 日在中科院上海交叉学科研究中心召开。40 余位中外学者围绕会议主题,秉承前沿领域、交叉学科、自由探索的宗旨进行了深入的探讨和前瞻。会议共设 4 个议题,分别为“生物医药大数据”“物理、化学与地球科学领域大数据”“人文与社会科学领域大数据”和“大数据处理技术与方法”,共 21 位专家作了会议报

告。在与会专家积极探讨交流以及中德青年科学家小组的努力工作下,会议达成初步共识,认为:大数据作为改变人类生活及理解世界的新方式,正驱动着科学研究范式的转化,推动着科学发展;应科学地认知大数据对于科学发现的重要作用、意义以及面临的重大问题;在建立科学大数据中心方面进行交流和合作;组建科学大数据工作组开展大数据热点问题的研究;注重大数据青年科学家的培养等。

1 大数据在不同学科领域的发展现状及挑战

大数据的特征在于:(1)海量数据;(2)数据以高度动态的方式持续产生;(3)数据的高度异质性;(4)数据质量存在噪声、不完整和偏见方面的严重问题。这些特征在各科学领域都普遍存在,而在各科学领域相对于大数据研究的需求却又有很大的不同。

1.1 生物医药大数据发展现状及挑战

20 世纪 90 年代初国际上开始人类基因组计划研究,从此开启了人类认识自身遗传密码的划时代的航程。随着人类基因组图谱工作的完成,人类基因组的数据变得更加完善与准确。以近年来增长最快的数据,人类的单核苷酸多态性(SNP)数据为例,它代表着不同人种以及正常人和某些病人基因组中碱基的差异,已有 100 135 281 个人类非冗余并被确认的 SNP 位点被数据库收录。这表明人的基因组中平均每几十个碱基就有 1 个碱基差异。但在已知 SNP 中,仅有不到 1% 的 SNP 造成蛋白的变化。GenBank 中的 dbEST 数据库收录了大约 870 多万条代表着人类基因表达小片段的表达序列标签(EST)序列,覆盖了人类基因的 95%,冗余度已远超过 10。随着对基因组数据的不断挖掘,科学家发现了一些重要事实:DNA 上编码蛋白质的区域,也就是基因,只占人类基因组的一小部分,不会超过整个基因组的 3%,其余占人类基因组 97% 左右的“非编码 DNA”序列仍不大清楚其功能,但却蕴涵着生物体复杂性的信息、具有重要的生物学功能,且与人类

疾病相关,迄今为止,我们对这些非编码序列以及相关的非编码基因和非编码 RNA 的功能只有很少的了解^[4]。

《人类基因组计划》的完成和深入发展为生命科学积累了大量的数据和资料,这将有可能从更深层次上了解人体生长、发育、正常生理活动,同时也可能了解各种疾病的病因,并提出防治途径。

现今,已经存在着包含不同种类组学,如基因组、转录组、蛋白质组、代谢组、表观遗传组等大数据的多个大型国际共享平台。获取组学数据的方法与技术已日渐成熟,关键是数据挖掘。与组学数据的海量特征相比,组学数据的复杂特征则更具有挑战性。组学数据复杂性的本质是源于生物体的结构和功能以及生命活动过程本身的多样性和复杂性。为此必须使用信息科学领域正在发展的解析大数据内涵的一系列理论、方法与技术,必须将当前国际上两大前沿领域“组学”与“大数据”融合。临床上,组学大数据的挖掘可得到大量不同人以及正常人与病人之间在分子水平的差异,关键问题是这些差异中哪些是与疾病直接相关的、相关的程度如何?只有找到了这种联系,才能得到表征特定疾病的分子标记,才能发现药物设计的分子靶标,才能实现转化,将组学分析获取的知识用于临床。因此,生物大数据在医药领域应用的前提是建立代表分子水平差异的基因型与代表疾病特征的表现型之间的桥梁。为此,需要发展生物信息学、系统生物学,包括生物网络研究的大量理论、方法与技术,建立并完善基因型与表型的关联。

1.2 地球大数据发展现状及挑战

伴随着对地观测技术的不断发展,在空间观测、地球物理、地球化学、地质勘探和地面传感器网络等领域产生着庞大的数据,其具有海量、多源、异构、多时态、多尺度、高维度、高复杂性、非平稳和非结构化等特性,为实现地球科学领域的知识发现提供了有利支撑^[5]。以全球变化研究和数字地球为例,全球变化研究对地球系统化、综合化观测的需求带动了对地观测技术的高速发展,全球已建立准实时、全天候的地

球数据获取能力,形成了高空间、高时间、高光谱分辨率的天空地一体化对地观测系统,作为面向全球可持续发展的多学科挑战性的关键问题,全球变化研究主要包括全球变化过程的监测、全球变化的模拟分析、全球变化响应策略研究等,而这些研究都依赖于地球大数据,如长时间序列多时空尺度的对地观测数据,精确的、连续的地面台站观测和试验数据,基于有科学依据的理论推测与估算数据等。因此地球大数据可为全球变化研究发展提供新的解决思路。数字地球作为多学科交叉的研究领域,其目标是呈现一个基于海量、多类型、多源、多分辨率、多时空尺度的虚拟地球,不仅涵盖大气、地理、地质、环境、生态、资源等地球科学各个学科的数据,也与信息科学、空间科学、人文社会科学密切相关,具有地球大数据的主要特征。数字地球的发展高度依赖地球大数据,从而实现对地球系统进行描述、分析、模拟和预测^[6]。

地球大数据为地球科学带来了新的动力,但在传输、存储、处理、分析、管理、共享和知识发现等方面也带来了巨大的技术挑战。为应对这些挑战,科学家们正致力于研发面向地球大数据的计算平台、算法和软件系统等,如基于高性能平台系统、大规模存储技术、全流程自动化处理技术、高效化计算技术、数据共享与服务系统等。虽然这些技术带来一些革新,但大数据技术引入地球科学领域的时间尚短,且地球大数据与互联网大数据的行业特点具有明显差异,还存在一系列关键技术亟需攻克,如大规模多元数据集成与挖掘技术,大规模并发任务、数据、算法的多层次混合并行计算技术,数据、网络、计算多资源动态协同处理技术等。另一值得关注的方面是地球大数据的密集型科学发现。地球大数据的知识发现,不仅仅是信息提取,还有挖掘隐含的、非显见的模式、规律和知识。针对地球大数据规模庞大、维度超高但信息密度低的问题,科学家正探索通过人工智能方法简化数据量与数据维度,使大数据变小后再进行后续研究。此外,数据的极大丰富使得知识发

现由“模型驱动”逐渐转变为“数据驱动”成为可能。但是,高效挖掘地球大数据所蕴藏知识仍处于起步阶段,亟需发展面向地球大数据的知识发现创新理论与方法,如适应地球大数据的认知模型、面向全体数据的数据挖掘与知识发现方法等^[3]。

1.3 人文与社会科学大数据发展现状及挑战

在人文和社会科学领域,大数据也正在成为热门话题,它为人文社会科学研究与发展带来了新的历史性机遇与挑战。当前,人文社会科学领域产生了大量的数据,如文化遗产大数据、金融大数据、商业大数据、网络舆情大数据、医疗与健康大数据等,数据的规模和信息的完整性都是以往无法比拟的。政府、工业界、高校和研究机构越来越多的数据对社会开放,极大降低了数据的获取成本,同时数据充裕带来了研究机遇的质变,以往不可研究、不能研究的问题在大数据环境下成为可能。党的十八届五中全会提出实施国家大数据战略和推进数据资源开放共享,为人文社会科学研究打开了“另一扇窗子”^[7]。

在大数据环境下,人们不仅关心数据建模、分析、管理、复用和建立大数据基础设施,还关心如何构造和利用基于数据的、开放协同的研究与创新模式^[8,9]。当前,在人文社会科学研究领域,以“人文计算”、复杂网络分析、大规模数据分析为特征的研究方法逐渐被采纳,涌现出了越来越多基于现实数据分析的量化研究成果,人文社会科学的“科学性”显著增强^[8]。不仅如此,人文社会科学研究中大数据分析方法的使用,还提高了人文社会科学研究者研究能力,开启了人文社会科学研究的新局面。网络舆情管理、互联网金融、宏观经济分析、图书情报知识服务、历史文献管理、电子商务、新闻与数字出版、旅游管理、健康管理与养老服务等许多人文社会科学领域大数据研究成果不断涌现^[10-12],所关注的内容不仅包括针对人文社会科学特定领域和问题情景下的大数据建模与处理方法,还包括大数据资源管理与利用方法,以及大数据环境下的信息共享服务、安全、隐私保护等。例如:

Wlodarczak等人^[13]基于社交大数据进行观点挖掘与情感分析, Kim与Jeong等人^[14]采用基于观点的大数据挖掘进行股票涨跌预测。

人文社会科学领域大数据研究在面临着巨大机遇的同时,也存在一系列现实问题,不仅大数据分析的“注重关联,不关注因果”、“过拟合”等问题在人文社会科学研究领域同样存在,且已有研究成果总体上偏重于大数据应用分析,针对人文社会科学特定问题情境的大数据理论和建模方法研究和创新不足^[15-17]。此外,人文社会科学大数据研究目前还面临4方面的问题。

(1) 科研资料总量的快速增加和数据质量问题给人文社会科学研究带来了巨大挑战。当前人文社会科学研究者各自研究领域都面临大量数据资料的处理问题^[8],研究范式的转变也使得人文社会科学研究越来越依赖高质量的数据,迫切需要构建人文社会科学数据的质量保障机制,以及研究新的计算机处理模式和分析方法以支持人文社会领域科学家对知识的获取、标注、比较、取样、阐释与表现。

(2) 资料数字化带来的挑战。资料数字化改变了传统人文社会科学的资料类型,数字资源的采集、加工和处理对高水平研究成果的获得作用日益显著^[8]。以“大数据”为代表的数字资源在数据粒度、碎片化、结构多元化、信息质量等方面具有更高的复杂度,对资料的汇集、保存和综合利用更加依赖计算机的辅助,人文社会科学家进行数据处理分析也越来越需要依赖信息技术手段,迫切需要开发可用于人文社会科学大数据采集、清洗、分析处理和可视化的工具和方法。传统人文社会科学学者对信息处理分析工具与技巧的缺失将影响该领域高水平研究成果的产出。

(3) 数据出版和共享方面的挑战。缺乏能够应用于大数据研究实践成果和学术著作快速出版的开放工具和平台,也是一个重要挑战。目前亟需可用于不同学科、不同制度下的数据出版(有适当标准和授信)和数据共享的集成化平台,以及多数据集集成化出版。

(4) 大数据资源管理、知识产权、安全与隐私方面的挑战。大数据运用不仅带来了更多问题的解决方法,也带来了数据资源管理、公民知识产品、数据安全与用户隐私等方面的一系列问题,这在人文社会科学领域显得尤为突出。大数据资源管理的公共政策,大数据资源与产业的深度融合,以及大数据商业价值的挖掘与知识产权、数据安全和用户隐私保护之间关系的研究方兴未艾,尚待取得突破性的进展,值得进一步的探索。

1.4 大数据处理技术与方法发展现状及挑战

大数据在数据规模、数据增速、数据类型、数据质量、数据价值等方面的特性给大数据处理技术与方法提出了新的科学技术挑战^[9]。主要体现在5个方面:

(1) 数据存储管理方面。数据产生过程和数据分析过程的分离,使得传统面向数据查询需求的关系数据库不再适用,亟需面向数据分析需求的大规模数据仓库和NoSQL数据库^[18]; (2) 数据分析方法方面。数据的产生和获取过程不再有严格的控制,相关性分析代替因果性分析逐渐成为数据分析的主要方式,问题驱动的研究方式逐渐被数据驱动的研究方式所代替^[19]; (3) 模型和算法方面。半结构化和非结构化数据的处理需求成为主流,传统基于特征工程(feature engineering)的方法逐渐被基于特征学习(feature learning)的方法超越并取代^[20]; (4) 计算体系结构方面。新型存储器件和计算器件(例如GPU等)不断涌现,使得通用处理器和单一体系结构逐渐过渡为专用处理器和异构体系结构^[21]; (5) 计算和服务方面。对于计算资源的高可靠性和高易用性的需求日增,以互联网为媒介的云计算模式和数据中心逐渐成为大数据处理的新型模式^[22]。

近几年,大数据分析处理技术和方法有了长足的发展。Hadoop分布式文件系统、Map-Reduce和Spark分布式计算框架、衔接高性能计算和大数据的DataMPI、云计算技术、深度学习技术等新技术深刻影响和改变着大数据的分析处理。一方面,计算能力和计算模式的变革为大数据分析处理提供了高易用性、高可靠性和低熵的

计算资源;另一方面,人类社会活动的信息化和数字化程度达到了空前的水平,日益丰富的大数据构成了人、机、物三元世界的详实数字记录,形成了前所未有的数据资源。计算资源和数据资源的结合,为人工神经网络的复兴和深度学习技术的发展提供了前所未有的契机,共同催生了人工智能新的春天。无论是图形图像处理和自然语言理解等基础研究方面,还是无人驾驶和智能机器人等具体应用方面,以深度学习技术和大数据分析引擎为代表的大数据分析和处理技术都带来了质的进步,产生了深远影响。相应地,为深度学习设计和开发的新型计算框架和专用计算芯片近年来也取得了很大的进步。另外,各类体现互联网思维的“互联网+”应用,也在推动着大数据分析和处理技术的进步,以“众包”为代表的群智计算在很多应用场景(例如借助互联网进行的众包光学字符识别系统reCAPTCHA^[23])中发挥了重要作用,解决了传统计算模式无法或难以解决的问题,是大数据分析和处理技术的一个新方向。

2 科学大数据发展建议

2.1 生物医药大数据

在生命科学领域,获取组学数据的方法与技术已日渐成熟,关键是数据挖掘。对占人类基因组97%左右的非编码序列信息的积累与挖掘也已引起国际上的广泛关注,预示着这一领域将取得突破。如何从海量复杂的组学数据中获取生命活动的知识已成为了基因组及相关研究的关键。当前的困难主要包括计算量大、样本量小、有效事件频率低、存在共同与特异的变化等。今后发展的目标包括需要增大计算资源与样本数目,发展与完善统计、分析、建模等方法,并构建动态的、双色(含蛋白质及RNA)的复杂网络。当下,最为活跃的研究热点包括整合分析来源成分复杂的数据,在确保病人隐私不受侵犯的前提下,更有效地整合来自生物学与临床医学的数据以用于诊断、治疗等方面的研究。更长远的研究目标,则是基于数据的进一步演绎,如,阐明基因型与

表型的关系。虽然目前已经存在着包含不同种类组学大数据的多个大型国际共享平台,为了扩大国家在生物医药大数据方面的影响力,更多数据应同时对整个科学界开放(涉及诸如病人隐私的数据除外)。比较好的做法是将数据存放在领域内已建立的全球数据存储中心。如有必要,建立国家大型计算机中心或生物医学权威数据库以方便数据的采集、处理以及共享。

2.2 地球大数据

地球大数据的诞生,使人类对地球系统的认知从传统的经验资料搜集、理论推导、局部物理过程理解和模拟转化到利用地球大数据进行信息挖掘与知识发现,从而探索地球系统中关键信息和各子系统及各生物物理变量之间的相互关联和作用。因此,建议在3个层面发展地球大数据,使其更好地服务知识发现。(1)地球大数据为地球科学,尤其是地球系统科学的研究提供了全新的方法论。基于天空地一体化的地球观测大数据,结合地球科学领域的数据挖掘与知识发现的模型、算法,发展地球大数据知识发现的理论与方法是地球科学领域亟待解决的重大科学问题。(2)地球大数据传输、存储、管理、处理、计算与共享高度依赖于大数据技术,结合互联网领域大数据技术和云计算的最新研究成果,研发面向地球大数据的平台系统、数据的高效组织与集成、算法的并行计算技术、大规模数据挖掘、资源调度与优化、信息共享与服务方法等关键技术,发展以大数据技术和云计算为核心的地球大数据处理与应用综合服务平台,是地球科学领域大数据发展的前提和基础。(3)加强地球科学领域与各相关领域的协同合作研究,推进大数据与跨学科领域大数据的交叉和融合,推动地球科学的创新发展。例如,数字地球科学作为多学科交叉的研究领域,其学科发展依赖于不同学科大数据的综合集成的解决方法。

2.3 人文与社会科学领域大数据

大数据的运用有助于形成人文社会科学研究新思维,进一步推动研究数据有序开放、跨学科深度协作,以

及人文社会科学与自然科学及工程技术学科的融合,从而开启人文社会科学研究新局面^[8]。为推动人文社会科学大数据学科发展,提出4方面建议:(1)推动构建人文社会科学大数据质量评估标准与共建共享。制定人文社会科学大数据质量评估标准和实现大数据资源共建共享是推动人文社会科学领域大数据发展的基础性工作。为此,建议制定人文社会科学大数据质量评估标准,构建人文社会科学大数据共享平台,积极开展人文社会科学大数据共建共享机制与管理方法的探索与创新,为推动大数据满足人文社会科学领域研究人员的信息服务需求提供重要保障。

(2)推动通过跨学科研究与合作开发人文社会科学大数据分析模型及公共服务平台。应鼓励国内外计算机信息科学与人文社会科学领域的学者和技术人员开展跨学科的研究与合作,突破学科壁垒,开发面向人文社会科学领域大数据分析处理需要的计算方法以及工具性软件平台,为人文社会科学大数据研究提供技术手段支持。(3)积极推动具有中国特色的人文社会科学领域大数据理论与大数据技术产业化应用实践。进一步开展面向中国经济社会发展重大现实需求、具有中国特色的人文社会科学大数据理论研究,将大数据分析方法与我国人文社会科学具体实践有机融合,促进大数据分析在我国互联网金融、网络舆情管理、数字出版、电子商务、健康管理与养老服务、物流管理、旅游管理、智慧城市与交通管理等重点领域的研究与实践;进一步推动人文社会科学大数据研究与大数据产业的融合发展。进一步开展具有中国特色的大数据资源管理公共政策,大数据资源管理领导力(即首席数据执行官),大数据商业价值,大数据知识产权、数据安全与用户隐私保护等核心问题的研究与实践。(4)加强人文社会科学领域青年学者、博士生与研究生大数据分析方法与能力培养,鼓励他们更多地参与大数据领域的国际学术交流与合作。在人文社会科学相关院系,开设大数据分析建模课程;利用科研院所、高等学校、工业界和海外的各种相关数据、平台和人才资源,对人文社会科学领域科研人员进行大数据分析处理技术培训,增强我国人文

社会科学研究人员利用大数据分析方法解决人文社会领域科学问题的能力, 大力培养人文社会科学领域青年大数据科学家和大数据分析师, 推动我国人文社会科学研究人员在国际高水平乃至顶级期刊发表更多体现中国特色的人文社会科学大数据研究成果; 为人文社会科学领域研究人员特别是青年学者参与大数据研究领域的国际学术交流合作提供更多机会, 进一步扩大我国人文社会科学领域科学家在国际学术界的影响。

2.4 大数据处理技术与方法

大数据处理技术与方法方面, 建议重点发展3个方向。(1) **深度学习技术**。深度学习技术已经在许多非结构化数据的处理方面——特别是在表达学习方面, 展现出了其强大的生命力, 但仍面临着计算代价大、模型训练慢、可解释性差等突出问题, 未来仍需探索如何在深度学习的模型方面整合人的先验知识或抽象能力, 在降低对大量训练数据的依赖性的同时提高模型的可解释性^[24]。(2) **低熵计算框架**。计算作为一种资源需要以一种低熵的方式为大数据分析处理提供服务, 即降低计算资源在使用过程中的损耗, 并提高易用性和可靠性, 这需要云计算技术、新型计算器件、数据中心网络等多个方面的技术进步。同时, 设计安全可靠、可信易用的数据共享模式, 降低数据使用过程的频繁搬迁及数据一致性约束, 也是促进低熵计算实现需要努力的方向^[25]。

(3) **数据使能的社会智能**。大数据是衔接人、机、物三元世界的纽带, 蕴含了关于人类活动和社会智能的知识^[26], 如何利用这些数据探索社会智能涌现的机理并构建数据使能的社会计算模式, 是未来大数据分析处理和人工智能的重要探索方向, 以互联网为媒介的人机互动的人计算是可能的一种尝试形式^[27], 未来期待更为柔性和易用的促使社会智能涌现的计算模型出现。

致谢: 感谢张凤、吴艳、章文峻、王东瑶、蒋芳、薛芳、滕晓龙为会议的筹备、组织所付出的辛勤劳动。

参考文献

- 1 Turner V, Gantz J F, Reinsel D, et al. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Framingham: IDC Analyze the Future, 2014.
- 2 Gantz J F, Reinsel D. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Framingham: IDC Analyze the Future, 2012.
- 3 Guo H D, Wang L Z, Liang D. Big Earth Data from space: a new engine for Earth science. Science Bulletin, 2016, 61(7): 505-513.
- 4 Consortium E P. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, 489(7414): 57-74.
- 5 何国全, 王力哲, 马燕, 等. 对地观测大数据处理: 挑战与思考. 科学通报, 2015, 60(5-6): 470-478.
- 6 Guo H D, Wang L Z, Chen F, et al. Scientific big data and Digital Earth. Chinese Science Bulletin, 2014, 59(35): 5066-5073.
- 7 孙建军. 大数据时代人文社会科学如何发展. 光明日报, 2014-07-07.
- 8 孙建军. 大数据使社科研究不再“望数兴叹”. 人民日报, 2016-02-18.
- 9 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. 中国科学院院刊, 2012, 27(6): 647-657.
- 10 冯芷艳, 郭迅华, 曾大军, 等. 大数据背景下商务管理研究若干前沿课题. 管理科学学报, 2013, 16(1): 1-9.
- 11 俞立平. 大数据与大数据经济学. 中国软科学, 2013, 2013, (7): 177-183.
- 12 McAfee A, Brynjolfsson E. Big data: the management revolution. Harvard Business Review, 2012, 90(10): 60-66, 68, 128.
- 13 Włodarczak P, Soar J, Ally M. Reality Mining in eHealth. Health Information Science, Cham: Springer International Publishing, 2015: 1-6.
- 14 Kim Y, Jeong M, Jeong S R. Using big data opinion mining to

- predict rises and falls in the stock price index. Handbook of Research on Organizational Transformations Through Big Data Analytics. Hershey: IGI Global, 2016.
- 15 Sandra G B. Social science in the era of big data. Policy & Internet. 2013, 5(2): 147-160.
 - 16 Morozov E. To Save Everything, Click Here: the folly of technological solutionism. New York: Public Affairs, 2013.
 - 17 Lazer D, Kennedy R, King G, et al. The parable of Google flu: traps in big data analysis. Science, 2014, 343: 1203-1205.
 - 18 Leavitt N. Will NoSQL databases live up to their promise? IEEE Computer, 2010, 43(2): 12-14.
 - 19 Hey T, Tansley S, Tolle K. The fourth paradigm: data-Intensive scientific discovery. Microsoft Research, 2009.
 - 20 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 38(8): 1798-1828.
 - 21 Liu D, Chen T, Liu S, et al. PuDianNao: A polyvalent machine learning accelerator // Proceedings of the 20th international conference on architectural support for programming languages and operating systems (ASPLOS 2015). New York: ACM, 2015: 369-381.
 - 22 Armbrust M, Fox A, Griffith R, et al. A view of cloud computing. Communications of the ACM, 2010, 53(4): 50-58.
 - 23 von Ahn L, Maurer B, McMillen C, et al. reCAPTCHA: human-based character recognition via web security measures. Science, 2008, 321(5895): 1465-1468.
 - 24 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444.
 - 25 Lu X, Liang F, Wang B, et al. DataMPI: Extending MPI to hadoop-like big data computing // Proceedings of the 28th IEEE international parallel and distributed processing symposium (IPDPS 2014). Phoenix: IEEE, 2014: 829-838.
 - 26 Shen H W, Barabási A L. Collective credit allocation in science. PNAS, 2014, 111(34): 12325-12330.
 - 27 Michelucci P, Dickinson J L. The power of crowds. Science, 2016, 351(6268): 32-33.

Big Data in Natural Sciences, Humanities and Social Sciences

——Review of the 6th Exploratory Round Table Conference

Guo Huadong¹ Chen Runsheng² Xu Zhiwei³ Sun Jianjun⁴ Bi Jun⁴ Wang Lizhe¹ Luo Jianjun² Shen Huawei³
Gu Dongxiao⁴ Liang Dong¹ Shen Wenqing⁵ Zhang Xu⁵ Hans Wolfgang Spiess⁶ Thomas Lengauer⁷

(1 Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China;

2 Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;

3 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;

4 Nanjing University, Nanjing 210023, China;

5 Shanghai Branch of Chinese Academy of Sciences, Shanghai 200031, China;

6 Max Planck Institute for Polymer Research, Mainz 55128, Germany;

7 Max Planck Institute for Informatics, Saarbrücken 66123, Germany)

Abstract Big data has begun to significantly influence global production, circulation, distribution, and consumption patterns. It is changing humankind's production methods, lifestyles, mechanisms of economic operation, and country governance models. It is a strategic enabling technology in the era of knowledge-driven economies, and also a new type of strategic resource for nations and the world. It offers a promising

new route for innovative methods of analysis and inference, and provides new opportunities for natural sciences, humanities and social sciences. Ubiquitous in the discussion of today's technology, the colorful and not clearly delineated term "big data" is on people's minds, regarding both its immense potential and its actual and perceived risks. The 6th Exploratory Round Table Conference (ERTC 2015) under the theme of "Big Data in the Natural Sciences and Humanities" was successfully held in Shanghai in November 2015. It was a joint project of the Chinese Academy of Sciences (CAS) and Max Planck Society (MPG), focused on topics that are only just beginning to emerge in the scientific community. Scientists from CAS and MPG met together with experts around China and the world to review the status of research and technology regarding and using big data and to discuss how it can and should be harnessed for furthering science. Big data is characterized by (1) highly accessible generation of large volumes of data which (2) are generated continuously in a highly dynamic fashion, and which feature (3) high data heterogeneity and (4) serious issues of data quality regarding noise, incompleteness, and biases. The status and requirements of big data research differ substantially among individual scientific domains. In the life sciences, the field has large, internationally shared repositories of highly diverse omics data. Current activities include bringing together biological and medical (patient) data for research on diagnosis and therapy and making patient data accessible while preserving patient privacy. In the Earth sciences, various Earth observation methods, for example, remote sensing, ground sensor networks, geophysics, geochemistry, and geological surveys, have afforded huge volumes of data, so called big Earth data. Exciting themes include global change and digital Earth science. The concept of digital Earth is a virtual representation of our planet constructed with massive, multi-resolution, multi-temporal Earth observation, and socioeconomic data of different types. This multi-disciplinary challenge relies on big data. Big data is also emerging for the humanities and social sciences. High-resolution 3D-imaging, for example, has led to the generation of large amounts of data for digital reproductions of cultural heritage artifacts that require large processing capabilities for filtering and reassembly. The key problem in social sciences is that the vast majority of data is still only available as images, texts, or websites, without appropriate metadata to enable discovery and analysis. Methodologies based on big data pose a number of challenges. (1) In order to gain trust in the data and learned predictive models, the predictions must be interpretable by a human. (2) Another challenge is the resulting loss of privacy: in some settings, complex predictive models are able to recoup partial information from different databases, and effectively deanonymize seemingly anonymous data. (3) At the infrastructure level, energy- and cost-efficient solutions are becoming a growing necessity. (4) Furthermore, the software deployed on such infrastructure must deal transparently and resiliently with the noise and heterogeneity inherent to big data. In the three-day conference, a preliminary consensus was proposed that big data, as a new way of human life and understanding the world, is driving the transformation of scientific research paradigms and promoting scientific development. It should be scientifically cognized how big data is playing a critical role for scientific discovery, what the significance is, and what major challenges are being faced. The conference also recommended establishing a Scientific Data Center in communication and cooperation, to form a scientific working group to research big data issues, and to enhance cultivation of young scientists in the realm of big data.

Keywords big data, scientific big data, life sciences, earth sciences, humanities, social sciences, computing technology, Exploratory Round Table Conference

郭华东 中科院遥感与数字地球所研究员。中科院院士、发展中国家科学院院士、国际欧亚科学院院士。现担任国际数字地球学会 (ISDE) 主席及 ISDE 中国国家委员会主席、国科联 (ICSU) 国际科技数据委员会 (CODATA) 前主席及中国国家代表、灾害风险综合研究计划 (IRDR) 科学委员会委员及 IRDR 中国委员会主席、《国际数字地球学报》主编等职。主要从事遥感科学与应用研究, 在遥感信息机理、雷达对地观测、数字地球科学等方面取得系列成果。发表论文 400 余篇, 出版专著和主编著作 16 部, 获国家和省部级科技奖励 13 项。E-mail: hdguo@radi.ac.cn

Guo Huadong Professor of Institute of Remote Sensing and Digital Earth (RADI), the Chinese Academy of Sciences (CAS), an Academician

of CAS, a Fellow of The World Academy of Sciences for the advancement of science in developing countries (TWAS), and an Academician of the International Eurasian Academy of Sciences (IEAS). He presently serves as President of the International Society for Digital Earth (ISDE), Past-President of the ICSU Committee on Data for Science and Technology (CODATA), Science Committee Member of the Integrated Research on Disaster Risk (IRDR) programme co-sponsored by ICSU, ISSC, and UNISDR, Editor-in-Chief of the *International Journal of Digital Earth*, and Chairman of the Chinese National Committee for ISDE and China Committee for IRDR. He specializes in the remote sensing science and its applications, and has conducted ground-breaking research on the information mechanisms of remote sensing, radar for Earth observation, and digital Earth science. Prof. Guo has published more than 400 papers and sixteen books, and is the principal awardee of thirteen national and CAS prizes. E-mail: hdguo@radi.ac.cn