



# 大数据时代的 科研活动

*Scientific Discovery In Big Data Era*

CODATA中国全国委员会 编著



科学出版社

# 大数据时代的科研活动

CODATA中国全国委员会 编著

科学出版社

北 京

## 内 容 简 介

本书是由国际科技数据委员会(CODATA)中国全国委员会编著的研究报告,旨在揭示科研活动在大数据时代所表现出的新特点和在研究模式上的变化,以及这些变化对数据技术、基础设施、政策、人才等方面提出的挑战和需求,主要介绍了所需要的关键数据技术以及数据科学、数据政策的发展现状和态势,为我国政府有关部门制定相关科技发展战略和行动计划提供参考。

本书可作为政府部门、科研机构、高等院校和相关企业进行科技战略决策的参考书,也可供国内外专家、学者研究和参考。

### 图书在版编目(CIP)数据

大数据时代的科研活动 / CODATA中国全国委员会编著. —北京: 科学出版社, 2014.4

ISBN 978-7-03-040183-0

I. ①大… II. ①C… III. ①科学研究事业—研究—中国 IV. ①G322

中国版本图书馆CIP数据核字(2014)第048195号



责任编辑: 王 哲 / 责任校对: 李 影  
责任印制: 张 倩 / 封面设计: 迷底书装

科学出版社出版

北京东黄城根北街16号  
邮政编码: 100717

<http://www.sciencep.com>

新科印刷有限公司印刷

科学出版社发行 各地新华书店经销

\*

2014年4月第 一 版 开本: 787×1 092 1/16

2014年4月第一次印刷 印张: 14 1/4 插页: 2

字数: 329 000

定价: 72.00元

(如有印装质量问题, 我社负责调换)

# 序

---

步入21世纪,在信息技术持续快速发展的推动下,一个大规模生产数据的时代已经开启。人们不仅利用各种新型望远镜、卫星、传感器网络、加速器、医疗成像设备、测序仪等,通过计算、模拟等活动产生了大量的数据,而且通过使用各种智能终端和互联网应用不间断地产生着海量的数据,从而形成了“大数据”。

数据是科学发现和知识产生的源泉。大数据标志着过去不能够测量和记录的很多事物都被数据化了,这就使得科学家们有机会去更深入地探索世界,做很多以前无法进行的研究,解决很多以前难以解决的科学问题,产生很多突破性的或意料不到的科学发现。近年来国内外的一些科学研究成果也证实了这一点,例如,生物学领域和医学领域的科学家们利用大规模DNA序列数据得到了对生命王国的新认识,高能物理领域的科学家们通过在大型强子对撞机上采集海量实验数据并进行处理和分析寻找到了希格斯粒子,等等。这是大数据为科学研究带来的机遇。

与此同时,对于各学科领域来说大数据也是巨大的挑战。大数据意味着新的科研方法,以数据为中心来思考、设计和实施科学研究的第四范式(又称“数据密集型科学研究”)正在成为一种普遍范式,科学发现依赖于对海量数据的收集、管理和分析。然而,随着数据规模、复杂度和产生速度的增加,数据获取、管理、分析和应用的难度必然增大。如何保证数据的流动性和可获取性,如何有效管理大数据并保证其可被高效访问,如何集成和融合多种来源的复杂海量数据,如何快速处理大数据以及如何高效地从大数据中分析和挖掘出信息和知识等,只有解决了这些关键的问题,科研人员才能真正掌控大数据。解决这些问题不仅依赖于研发和应用新的技术,而且依赖于强大的数据基础设施、配套的科技政策的建立健全,以及科研体制的转变和新型人才的培养。

国际科技数据委员会(CODATA)中国全国委员会顺应新时代的需要,组织国内专家学者探讨大数据对科学研究的挑战与机遇问题,历时两年完成了《大数据时代的科学活动》研究报告。该研究报告反映了高能物理、天文、空间、对地观测、生物、社会计算、经济管理与金融研究等学科领域的大数据现象和科学研究方法的变化,剖析了大数据和科研方法的变化对于相关信息技术、基础设施、政策法律、人才等方面的挑战,探讨了这些方面的发展现状与态势,并提出了针对性的发展建议。

大数据已成为重要的战略资源,国家层面的科研竞争力将在很大程度上取决于

国家所拥有大数据的规模、活性以及将大数据转化为信息和知识的能力。CODATA中国全国委员会应以崇高的科学使命感和时代紧迫感,积极推动科学大数据管理和运用方面的人才培养和国际国内交流,以实干精神投身到“创新驱动发展”国家战略的落实中去。

CODATA中国全国委员会主席

丁仲礼

2014年4月

# 前 言

经过两年多的调研和研讨,编写委员会举办了4次大型专题研讨会议,针对本研究报告中的每一篇文章进行了多次推敲和修改后,《大数据时代的科研活动》终于完成。可以说,研究报告只是我们编写成员当前的认识水平。我们相信,随着大数据应用服务的深入以及数据密集型科研活动实践的不断开展,人们对这一现象和问题会有新的思考和看法。

本报告作为中国科学院国际组织人才团队及中国委员会支持计划资助的项目——“CODATA中国全国委员会国际合作与可持续发展”的成果之一,在编写的过程中得到了中国科学院国际合作局国际组织处和中国科学院信息化工作领导小组办公室信息化工作处等单位的全力支持,并得到了国际科技数据委员会(The Committee on Data for Science and Technology, CODATA)中国全国委员会各位委员的指导。参与本报告编写的成员来自中国科学院各研究所以及国内著名高校,他们在繁忙的工作之余参与报告的研究和编写工作,付出了辛勤的劳动。在此,我们一并表示感谢。

CODATA(<http://www.codata.org>)是国际科学联合会(International Council for Science, ICSU)于1966年成立的一个跨学科的科学委员会,是科技数据领域的国际权威学术机构。在其近50年的历史中,CODATA一直致力于推动科技数据在全球范围的广泛共享与应用,并不遗余力地在全球范围内倡导数据科学,以促进科技发展和造福人类。我国于20世纪80年代加入CODATA,并以中国科学院牵头,成立了CODATA中国全国委员会,委员来自于国内各研究院所、高校和相关政府部门。近年来,得益于我国科技数据共享和科研信息化等工作的深入推进,我国科学家在CODATA中的影响和作用日益加大。我们相信,未来无论对CODATA的发展还是对数据科学的发展,中国科学家都将能起到更大的作用。

期望本研究报告对科研人员认识大数据以及即将或已经到来的新型科研模式有所帮助,对科研工作的管理者和政府相关部门进行决策具有一定的参考价值。

CODATA中国全国委员会副主席

黄向阳

2013年7月

# 概 述

得益于信息技术的持续发展和广泛应用,人类的数据化能力和范围在不断扩张。进入21世纪,这种扩张让我们看到大量从宏观到微观、从自然到社会的观察、计算、传播等仪器设备和活动正在快速产生着海量、多样的数据,几乎每个学科领域都在经历着空前的数据爆炸,科学研究正在进入大数据时代。

随着大数据在各领域的广泛出现,科学研究的模式发生了变化。继实验科学、理论科学、计算科学之后出现了被称为“数据密集型科学”的第四种研究范式。这一范式强调数据作为科学方法的特征,它不仅意味着新的科研方法,而且意味着重要的思维转变。采集、存储、管理、分析和可视化数据的能力成为科学研究必须适应的新事实。

已有若干对基于大数据的数据密集型科学研究的论述见诸文献。我们有理由相信,大数据为科学研究带来了重大的机遇,基于对大数据的分析,我们能更好地理解世界,解决以前认为难以解决的或甚至认为不可能解决的很多科学问题,产生意料之外的科学发现。

但是,大数据也意味着大挑战。挑战是多方面的,只有应对好这些挑战,才能最大限度地把握机遇。

- **技术挑战:** 大数据的3V特征在从数据中创造价值的工作的各阶段都造成阻碍,关键的技术挑战包括如何更好、更高效地传输、存储、长期保存、组织、发现与访问、集成、融合、分析、挖掘和可视化这些数据。

- **基础设施挑战:** 为了探索利用大数据,数据基础设施必须是易使用、开放和可扩展的,必须支持科学数据的整个生命周期,支持数据的整合和跨学科转移,支持数据密集型科学研究。

- **政策挑战:** 在保障隐私和国家安全的前提下,最大限度地促进数据的流动性和可获取性的数据开放政策至关重要。

- **人才挑战:** 需要解决掌握大数据开发利用技术和数据密集型科学研究方法的人才不足的挑战。

一个国家的科学研究在国际上的地位,将取决于其在科学数据上的优势以及将数据转换为信息和知识的能力。美国和欧洲等发达国家和地区已经将大数据和对数据的开发利用能力提高到了国家科技战略的高度,并投入巨大的人力物力应对上述挑战。为了加快我国科技创新的步伐,建议我国政府也从国家战略的高度对此加以重视,推动建立起良性的科学数据开发利用环境。

本研究报告的“引言”部分对科研领域的大数据现象、数据密集型科学研究范式的内涵、为把握大数据所带来的机遇而需要应对的挑战、欧美等发达国家和地区的有关战略和

计划等进行了综述。

本研究报告的第一部分“以数据为核心的科研活动”中,重点选择了高能物理、天文学、空间科学、地球科学、生物学、社会计算、经济学等学科领域,分别介绍了相应学科领域的大数据现象,探讨了相应学科领域中科学研究方法的变化,以及这种变化对相关技术、基础设施、政策法律等因素中的一个或多个方面带来的挑战和(或)提出的需求。通过这些文章,我们能够更清楚地看到大数据和数据密集型科学研究是现在及未来的科学研究避免不了的问题,需要从各方面采取行动。

第二部分“大数据时代支撑科研的共性技术”中,重点从数据采集、传输、存储与处理、数据整合、数据与文献互操作、数据挖掘及可视化技术等方面,探讨技术发展的状况、出现的新技术和未来的发展趋势。

第三部分“大数据时代的数据科学”中,《数据科学发展与展望》提出了将数据作为一个“自然体”来研究的观点,以及基于此对数据科学定义和内涵、数据科学基本框架及其与自然科学和社会科学的关系等问题的认识;《数据政策的发展》介绍了国际层面和国家层面的数据开放政策,以及期刊出版界在推动数据开放方面所做的努力,并探讨了数据开放政策面临的一些挑战。

本研究报告的“展望和建议”部分,提出了5条具体的建议,期望能够为我国有关政府部门在制定科技发展战略和行动计划时提供一些参考。

建议1:研究和制定配套的科技政策。政策通过推动科研资助模式、科研成果发表和评审机制、科研人员和科研机构激励机制等的变革,促进科学数据的共享和重用,促使开放科学和开放数据的原则被广泛接受。

建议2:建设新型国家级科学数据基础设施。新型国家级科学数据基础设施由若干可互操作的国家级科学数据中心、数据档案馆、数字图书馆等组成,应能支持科学数据的整个生命周期,支持数据密集型研究和多学科研究,支持数据的跨学科转移和开放链接的数据空间,以及支持科学数据与文献的互操作。

建议3:发展大数据存储、传输、管理、分析和共享所需要的核心技术。需要重点投资的技术方向包括大数据存储技术、大规模数据传输技术、数据集成技术、工作流技术、非结构化和半结构化数据处理技术、大数据挖掘技术、大规模数据可视化技术、数据长期保存技术等。

建议4:积极培养新的专业人才和支持数据科学的发展。培养包括数据科学家、数据分析师、数据工程师、数据管理员、数据档案员等在内的专业人才,他们对于科学数据的成功管理和利用起着关键作用。支持建立数据科学研究机构、开设数据科学相关专业和课程,以促进数据科学基础理论的发展和数据技术的研发。

建议5:引导相关学科对新的科研方法进行探索与利用。通过应用示范的方式,引导相关学科科研人员对数据密集型科学研究方法进行探索和利用,加快相关学科领域科研模式的转变。在这些引导性示范中,应注意对新类型研究团体组建和运转方式的探索。

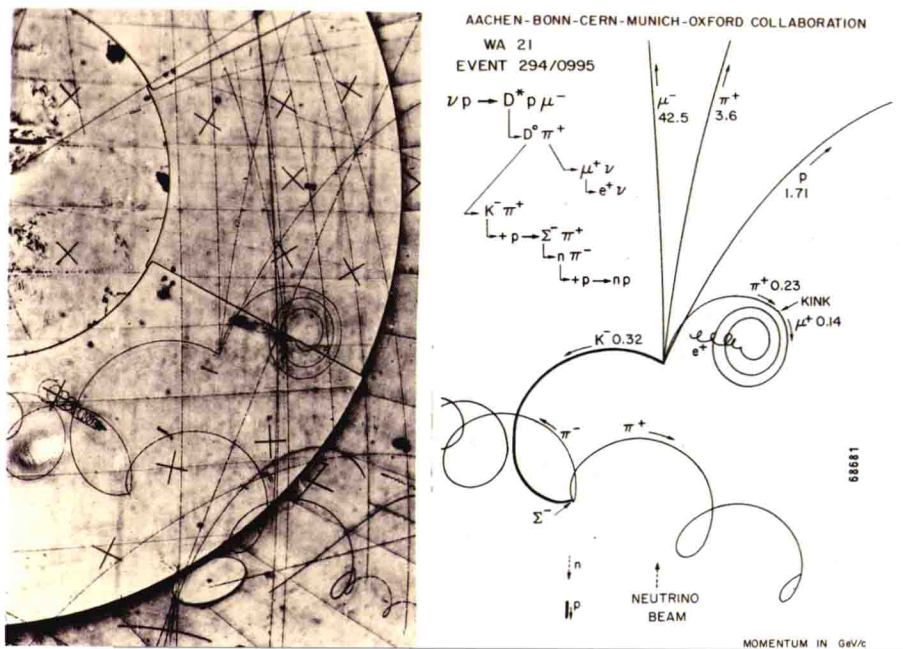


图1 20世纪80年代的实验：粒子在探测器(泡室)中的径迹

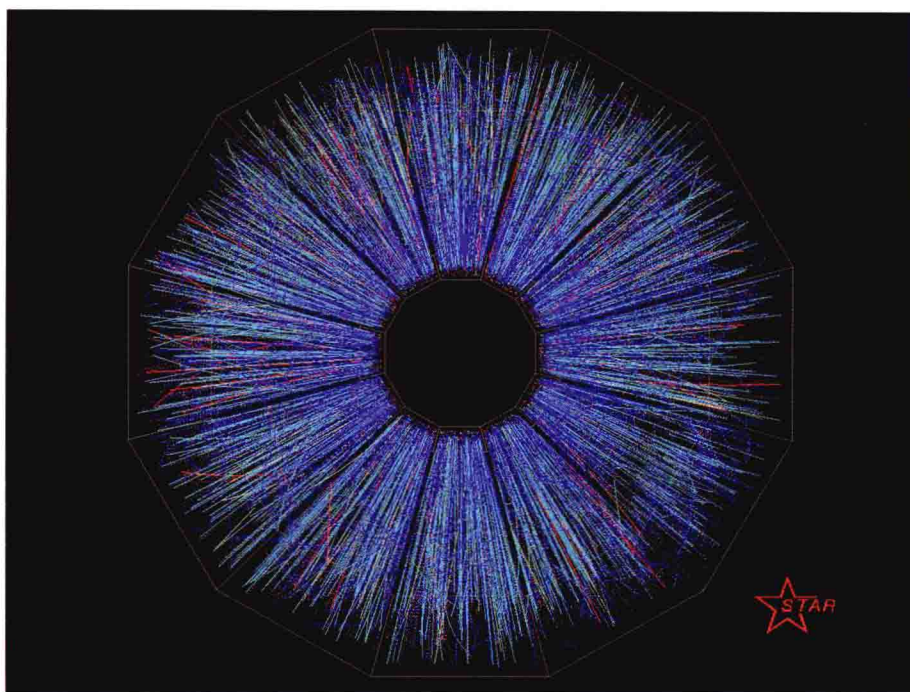
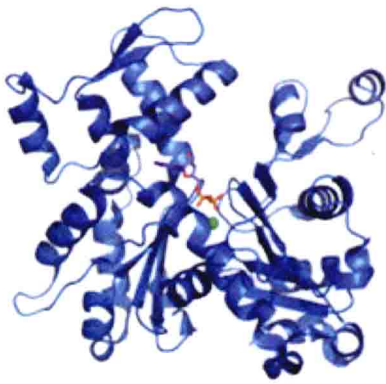
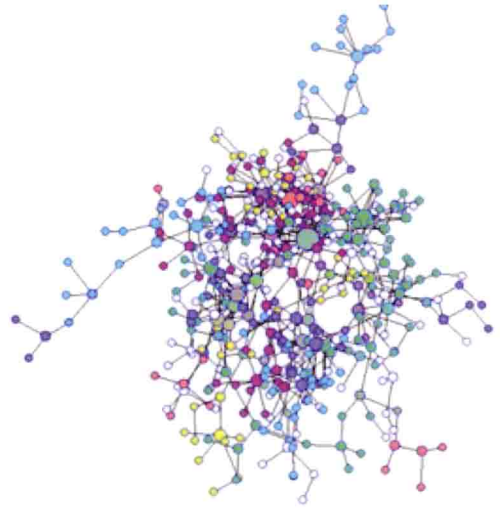


图2 布鲁克海文实验室螺线管式径迹探测器( Solenoidal Tracker at RHIC, STAR )实验金核与金核对撞产生的粒子在探测器中的径迹。相对论重离子对撞机( Relativistic Heavy Ion Collider, RHIC )是布鲁克海文实验室的大型对撞机



(a)蛋白质三维结构图



(b)蛋白质相互作用图

图3 蛋白质数据的图结构

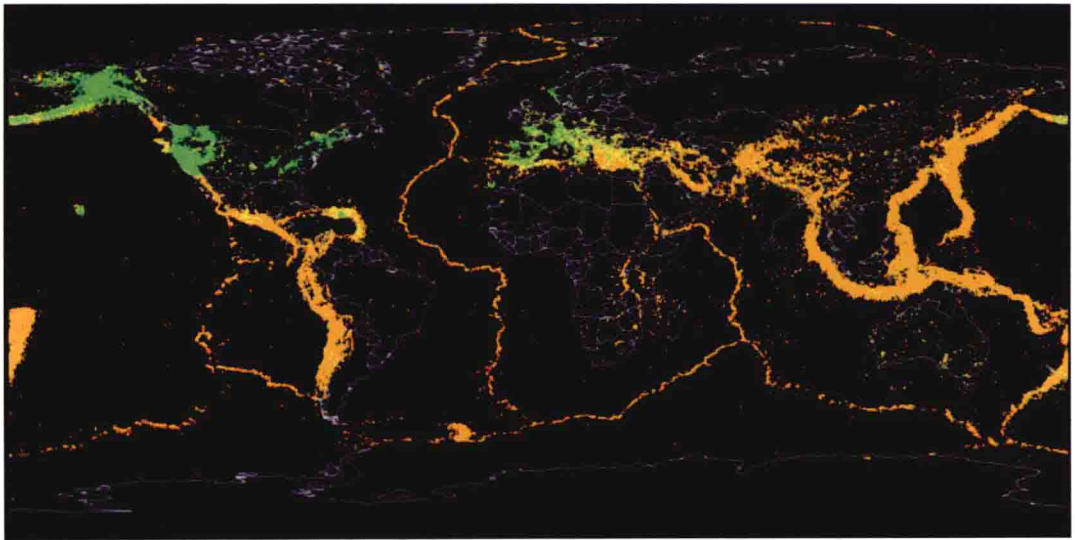


图4 全世界范围内2004—2011年7年时间内地震数据可视化

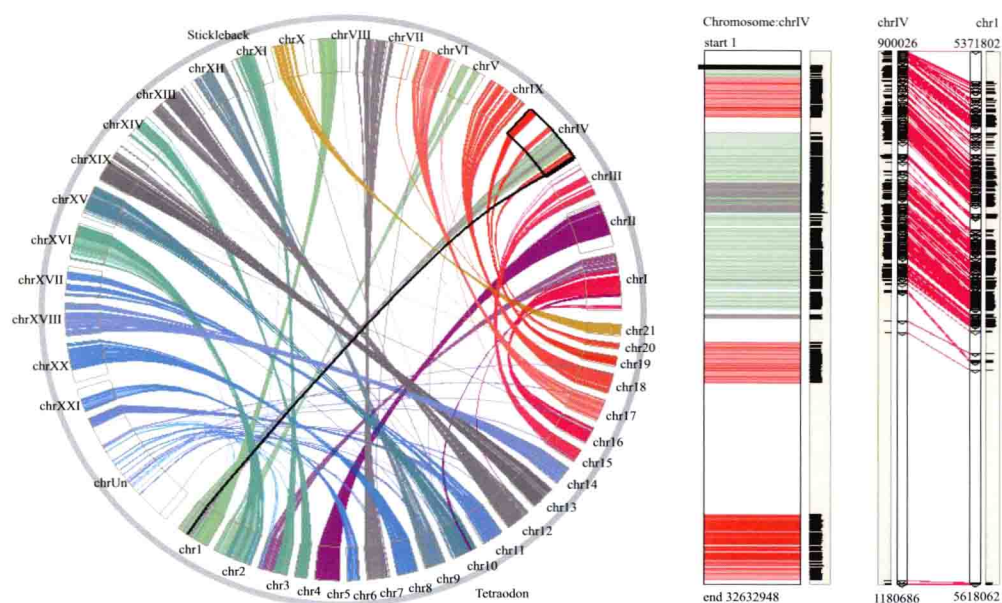


图5 两个物种全基因组比对结果的多层次可视化

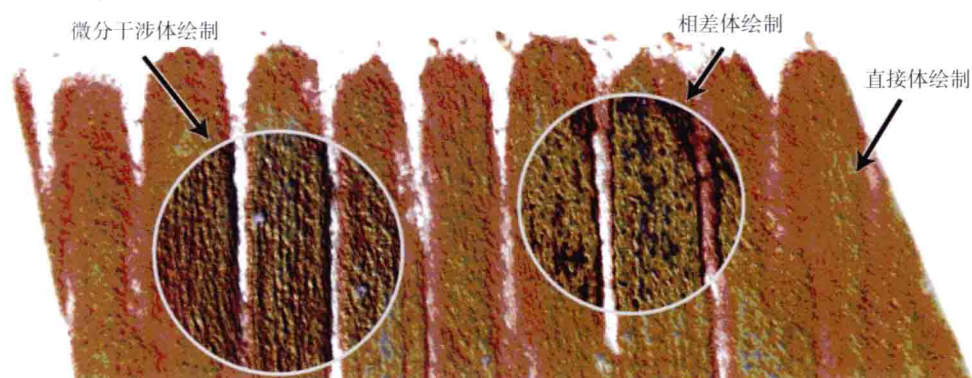


图6 鼠小肠微绒毛数据的直接体绘制结果

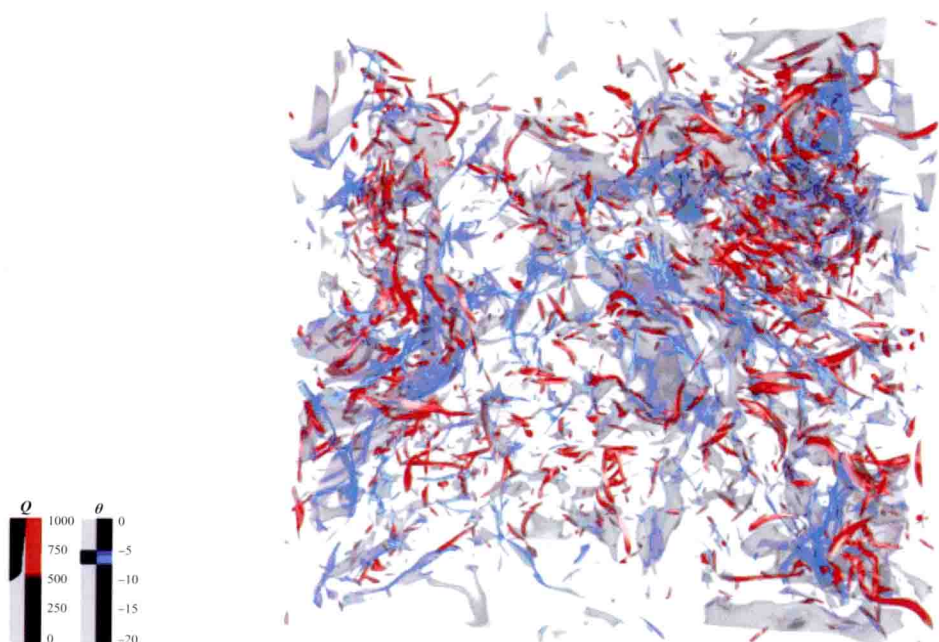


图7 湍流直接数值模拟数据可视化

# 目 录

|          |           |
|----------|-----------|
| 序 .....  | 丁仲礼 (i)   |
| 前言 ..... | 黄向阳 (iii) |
| 概述 ..... | (v)       |
| 引言 ..... | (1)       |

## 第一部分 以数据为核心的科研活动 .....

(17)

|                                     |                      |
|-------------------------------------|----------------------|
| 导言 .....                            | 李国庆 马俊才 (19)         |
| 一、高能物理中的数据和挑战 .....                 | 陈 刚 (22)             |
| 二、数据密集型时代的天文学 .....                 | 崔辰州 李 建 赵永恒 (33)     |
| 三、大数据时代的空间科学数据应用环境 ...              | 邹自明 马文臻 姜文凭 余继周 (42) |
| 四、大数据时代的对地观测科学研究 .....              | 李国庆 郭延辉 (54)         |
| 五、大数据时代的生物多样性研究 and 应用 .....        | 许哲平 (64)             |
| 六、大数据时代微生物学研究新趋势——数据的整合和应用<br>..... | 马俊才 吴林寰 杜晓萌 (76)     |
| 七、大数据时代的生物医学及其面临的隐私保护问题 .....       | 刘 雷 (82)             |
| 八、大数据与社会计算 .....                    | 王飞跃 曾大军 郑晓龙 李林静 (90) |
| 九、大数据背景下的经济管理与金融研究 尚 维 鲍 勤 张 珣 王 帅  | (100)                |

## 第二部分 大数据时代支撑科研的共性技术 .....

(109)

|                    |                   |
|--------------------|-------------------|
| 导言 .....           | 黎建辉 袁晓如 刘 莹 (111) |
| 一、科学数据获取的新方式 ..... | 周园春 郭 亮 宋 欣 (113) |

二、发展科研网络，提升科研大数据传输能力 ..... 葛敬国 (119)

三、大数据存储与处理技术

.....周园春 韦远科 冯 凯 沈 庚 张 洋 邵 靖 (128)

四、科学数据管理技术的新进展 ..... 沈志宏 胡良霖 王卫华 (139)

五、科学数据与文献的互操作 ..... 张晓林 沈志宏 刘 峰 (149)

六、大数据挖掘 ..... 刘 莹 (159)

七、大数据时代面向科学研究的可视化 ..... 袁晓如 (171)

**第三部分 大数据时代的数据科学 ..... (185)**

  导言.....朱扬勇 黎建辉 (187)

  一、数据科学发展与展望 ..... 朱扬勇 熊 赞 (188)

  二、数据政策的发展 ..... 侯艳飞 胡良霖 (199)

  展望和建议..... (210)

  附录1 编写人员 ..... (213)

  附录2 编写历程 ..... (215)

# 引言

黎建辉 侯艳飞 张丽丽  
中国科学院计算机网络信息中心

我们正处在一个聚合点上,在这里,数据、科学以及技术联合起来共同对抗那个最大的谜题——我们的未来,既是个人的又是社会的<sup>1</sup>。

——艾伯特-拉斯洛·巴拉巴西

源于人类测量、记录、分析和认识世界的渴望,信息技术持续地飞速发展并被广泛应用,使得人类的数据化能力和范围得到不断扩张。进入21世纪,这种扩张所导致的数据爆炸越来越显著,现在世界上充斥着比以往更多的数据,而且数据的增长速度在不断加快并存在将世间万物转换成数据形式的可能,包括科研、商业等在内的各个领域都被推到了一个前所未有的“大数据”时代<sup>[1]</sup>。量变引起质变,不断增加的数据引发了人们思维和行为模式的变革。在科学研究领域,科研模式正在发生极大的改变,继实验科学、理论科学、计算科学之后出现了被称为“数据密集型科学”的第四种研究范式,采集、传输、存储、管理、分析和可视化数据的能力成为科学研究必须适应的新事实<sup>[2]</sup>。

我们有理由相信大数据为科学研究带来了重大的机遇:当所能利用的数据量增加时,人们将可以做很多在小数据基础上无法完成的事情,基于对大数据的分析能更好地理解世界,解决从前难以解决的或甚至不可能解决的很多科学问题,产生意料之外的科学发现。然而大数据也意味着多方面的挑战,既包括技术、基础设施、政策和法律方面的,也包括机构、人才及文化等方面的。唯有克服这些挑战,人们采集、传输、存储、管理、分析和可视化数据的能力才能切实得以提高,从而最大限度地把握住机遇。

一个国家的科学研究在国际上的地位将越来越多地取决于其在数据方面的优势以及将数据转换为信息和知识的能力。美国、欧洲等发达国家和地区的政府认识到了这一点,将大数据和对大数据的开发利用能力提高到了国家科技战略的高度,并已经采取措施来落实这一战略。

下面主要介绍和分析科研领域的大数据现象和日渐兴起的数据密集型科学研究范式,阐述为把握机遇而需要应对的一系列挑战,并概述欧美等发达国家和地区相关的战略和计划以及我国的有关情况。

1 艾伯特-拉斯洛·巴拉巴西. 爆发:大数据时代预见未来的新思维. 马惠,译. 北京:中国人民大学出版社, 2012.

## （一）大数据时代的到来

在科学发展的历程中,数据作为对自然、社会现象和科学试验等的定量或定性记录,一直是科学研究的重要基础。然而在计算机出现以前,受限于数据的存在形式(以模拟数据的形式存储于纸张、胶片、磁带等介质上)和测量、记录、存储、分析和传播数据的工具的能力,科研人员往往只能收集少量数据进行分析。计算机的发明不仅提高了人类的计算能力,而且带来了新的数据存在形式(以0和1二进制码表示的数字数据是机器可读的,并能够被计算机处理和分析)和存储介质,这为信息技术的快速发展奠定了基础<sup>[1]</sup>。随后的几十年中,计算机技术、以互联网为代表的通信技术和传感技术的持续创新和广泛应用使人类的数据化能力和范围快速扩张:一方面我们能够测量和记录的东西越来越多,另一方面我们对事物、现象等的测量、记录等更加频繁和细致。进入21世纪,这种扩张让我们看到大量从宏观到微观、从自然到社会的观察、计算、传播等仪器设备和活动(例如,高通量科学仪器、望远镜、卫星、加速器、传感器网络、医疗成像设备、DNA测序仪等各种科研仪器和装置、模拟方法、智能终端以及各种Web应用等)正在快速产生着海量、多样的数据,几乎每个领域都面临着空前的数据爆炸<sup>[12]</sup>。

以天文学领域为例,2000年斯隆数字巡天(Sloan Digital Sky Survey, SDSS)项目启动的时候,仅仅几个星期从新墨西哥州的望远镜收集到的数据量就超过了整个天文学历史的积累数据量,它每晚产生的数据约200GB天文学领域的数据开始从GB级跃升到TB级;2008年投入运行的“泛星计划”(Pan-STARRS,全景式巡天望远镜和快速反应系统)每年在运行中捕获的数据超过1PB;预计2018年在智利投入使用的大型视场全景巡天望远镜(Large Synoptic Survey Telescope, LSST)将能在五天内就获得与SDSS运行十年同样多的数据<sup>[2,4]</sup>。又如基因学领域,第二代测序技术带来了基因组数据的跨越式增长,相比于2000年,2010年的基因组数据产量增大了8个数量级。以国内基因组研究机构华大基因为例,其目前拥有150台以上测序仪,每天产生约15TB数据。而在21世纪初,华大基因使用的数台基于Sanger双脱氧法原理的ABI系列测序仪每天的数据通量仅约为20GB<sup>[5]</sup>。20世纪世界上最大的大型正负电子对撞机(Large Electron Positron Collider, LEPC)在1989—2000年的实验期间积累的数据不到20TB,而最新的欧洲粒子物理研究中心(European Organisation for Nuclear Research, CERN)的大型强子对撞机(Large Hadron Collider, LHC)每年产生的实验原始数据就达15PB左右。利用原始数据进行事例重建以及物理分析所产生的数据规模更大,以其中的ATLAS(A Toroidal LHC Apparatus)实验为例,仅2011年产生的总数据就达40PB<sup>[6]</sup>。刚刚退役的Landsat 5卫星的数据获取能力为67GB/天,而2012年发射的ZY3卫星的数据获取能力可以达到10TB/天以上。类似能力的传感器现在已经大量部署在卫星、飞机等飞行平台上,未来十年全球部署的遥感平台和传感器的数据获取能力将超过10PB/天<sup>[7]</sup>。包括化学、材料科学等在

1 “我们周围到底有多少数据? 增长的速度有多快?”据估计,在2000年,数字数据只占全球数据量的1/4;到2007年,所有数据中只有7%的数据是存储在传统介质上的模拟数据,其余全部为数字数据;到2013年,非数字数据只占不到2%<sup>[1]</sup>。在2011年,全球创建和被复制的数据量达到1.8ZB,在过去五年增加了5倍,而在未来十年将增长50倍,数据的增长速度符合摩尔定律<sup>[1]</sup>。

内的“长尾”科学<sup>1</sup>——在这样的学科中,有成千上万个普通实验室在进行着研究,它们之间没有或者很少有基础设施方面的合作——也在经历着数据爆炸,虽然每个数据集合在规模上并不大(一般不会超过1TB),但这种小数据集合数量众多且异构性强,它们汇聚后的规模和复杂性都不容小觑<sup>[10]</sup>。除自然科学领域外,得益于互联网应用和智能移动终端的发展,个人的活动得到了前所未有的记录。这种记录的粒度很小,频度在不断增加,以“众包”(crowd sourcing)<sup>2</sup>的模式为社会科学的定量分析提供了极为丰富的数据。社会科学也在经历着空前的数据爆炸,而且大数据之所以成为一个全社会话题,在很大程度上要归功于这类数据的增长<sup>[11]</sup>。

“大数据”的概念与数据爆炸紧密相关。最先经历数据爆炸的学科提出了“大数据”的说法,如今它几乎已被应用到了各个领域<sup>3</sup>。然而,“大数据”并非是一个确切的定义。常见的定义是那些规模和复杂度非常大,以至于很难通过目前主流的数据存储、管理和分析工具在合理的时间内对它们完成处理的数据集合<sup>[14]</sup>。在谈到“大数据”的特征时,常见的说法是3个V:大数据量(volume)、快速的产生速度(velocity)和多样的数据类型(variety)。以上定义和解释强调数据规模和复杂度的改变对相关数据技术的挑战和对技术发展的推动。相对于以上常见定义和解释,Viktor对“大数据”提出了一种独到的见解,他认为“大数据是人们在大规模数据的基础上可以做到的事情,而这些事情在小规模数据的基础上是无法完成的”<sup>[1]</sup>。对大数据的这种定义强调数据规模的改变对于知识发现、决策等的价值<sup>4[1,12,13]</sup>。

最早明确提出大数据时代来临的是在2012年2月11日《纽约时报》上发表的一篇文章为“大数据时代”的专栏文章。文章中称大数据正在对每个领域造成影响,大数据时代正在到来。科学、商业、政治和其他领域都在朝着数据驱动型的知识发现和决策的方向发生转变,这是一种变革。我们确实正在进行这场变革,庞大的新数据来源所带来的从量变到质变的变化,将在学术界、企业界和政界中迅速蔓延开来,没有哪个领域不会受到影响<sup>[16]</sup>。

大数据时代始于何时?虽然目前并没有一个明确的界定,但大数据时代的到来已是不可逆转的事实。而大数据之所以成为一个“时代”,不仅是因为数据爆炸在众多领域广泛发生,而且是因为数据爆炸积累到了一个开始引发变革的程度。这变革不仅是相关数据技术方面的,更是人们进行知识发现和决策等的思维和行为模式方面的。人

1 “长尾”原是统计学中幂律和帕雷托分布特征的一个口语化表达。2004年,Chris Anderson提出了长尾理论,用“长尾”来描述众多小型商业市场的集体力量<sup>[8]</sup>。后来,Dempsey巧妙地将这一概念用于图书馆,并指出这种模式的成功需要消费者(或读者)能够访问和了解长尾产品(或资料),也需要与按需供应相匹配的服务到位。这种模式也适合于科学数据。长尾科学产生小型但数量众多的数据集合,关键是如何释放这些数据集合的潜力,降低访问和重用它们的门槛<sup>[9]</sup>。

2 互联网的特点之一是人人参与。特别是在Web 2.0时代,互联网上的每个用户都是信息的提供者,互联网应用正是利用这一特点,依靠用户的群体力量快速产生着大规模的数据。

3 “大数据”并非一个新名词<sup>[12]</sup>,20世纪90年代中期,Jim Gray就指出大数据的挑战将会首先来自科学而非商业<sup>[2]</sup>。在1997年10月的第8届IEEE可视化会议上,Michael Cox和David Ellsworth发表的论文*Application-controlled demand paging for out-of-core visualization*中提到:“可视化给计算机系统提出了一个有趣的挑战:数据集通常很大,使主存储器、本地磁盘和远程磁盘负担过重。我们称之为大数据问题。”<sup>[13]</sup>

4 大数据定义是一个以相对性为特点的定义,这使得对于“具体多大的数据才是大数据”很难有一个普遍适用的界定。很多人认为大数据的数量级应该是TB级的。麦肯锡认为并不需要给“什么是大”定出一个具体尺寸,因为随着技术进步,这个尺寸本身会不断增大。此外对于不同的领域,“大”的定义也是不同的,不需要统一<sup>[15]</sup>。Viktor认为,“大”并不是以字节数作为衡量标准的,而是以是否利用有关被分析事物的所有数据或至少是尽可能多的数据为衡量标准的<sup>[1]</sup>。