●●观察与思考

大数据对科学活动的影响

刘伟榕,王秋君

(厦门大学人文学院,福建 厦门 361005)

[摘 要]科学界正在从数据缺乏时代过渡到数据泛滥时代,大数据处理系统有望成为新一代的科研基础设施。在新的技术环境下,数据成了科研最主要的对象,统计算法成了最有力的科研工具,科研思路也将从假说驱动转向数据探索。为了分享知识生产日益依赖的技术与设备,科学家将结成联盟进行工程化协作,科学进步中的共享化与全球化也将更为显著。有乐观的学者认为大数据有望成为科学研究的"第四范式",不过也可能伴随着科研路径依赖、资源垄断、成果纠纷等负面影响。

[关键词]大数据;科学发现;知识生产;第四范式

[中图分类号]G311 [文献标识码] A [文章编号] 1009 — 2234(2014)05 — 0042 — 02

大数据通常用来指无法在可容忍的时间内用传统 IT 技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合,具有容量大、产生速度快、类型繁多、信息价值大与冗余信息多等四个特征。[1]人类正在进入大数据时代,推动这个时代到来的科学研究活动将不可避免地受到大数据的反作用。本文综合最新的大数据研究、前沿的科研案例及科学哲学理论,尝试对科学活动正在发生与将要发生的变化进行探讨。

一、科学研究工具、方法、对象的变革

一是大数据处理系统将成科研基础设施。在科研信息化的推动下,人类对自然和社会的观察、感知、计算、仿真、模拟、传播等活动产生出大量科学数据。如何存储海量的科学数据成为科学家遇到的首要困难,例如欧洲粒子中心的大型强子对撞机每天都产生好几个千万亿字节(PB),但现在却只能按照可管理的能力限制其数据速率。[2]科学家难以密切关注到任何一项单独的数据,而需要机器进行辅助筛选。跨学科研究的兴起更是加大了数据的规模和复杂性,包含采集、管理与分析工具的大数据处理系统对环境应用科学、海洋科学、生态科学、物理学、天文学、生物学等领域来说已经成为一种基本的科研设施。在大数据科研设施布局方面,美国已经走在世界前列。例如能源部(DOE)将斥资 2500 万美元建立可扩展数据管理与可视化研究所,帮助科学家对数据进行有效管理,促进其生物和环境研究计划、美国核数据计划等的研究成果。[3]

二是科研方法从假说驱动转向数据探索。正如第谷的助手开普勒从第谷对天体运动的系统观察记录中发现了行星运动定律那样,在对所采集并仔细保存的实验数据进行挖掘和分析的基础上建立起新的理论,正是大数据时代科学活动的一个重要特征。大数据技术的巨大魅力在于通过统计算法揭示事物之间的相关性。美国 Wired 杂志主编 Chris Anderson 就认为"理论已终结"、"数据洪流使传统科学方法变得过时"。[4]他相信只要将有相互关系的 PB 级数据丢进巨大的计算机机群中,统计分析算法可以发现过去的科学方法发现不了的新规律、新知识。基于这样的技术,人们有理由相信,未来的科研方法将从传统的假说驱动型转向数据探索型。科学家们不必关心通过什么实验来验证假说,而是追求从现有数据中发现研究对象之间的关联,把多个学科和领域的数据进行融合,或许就能有新的发现。

三是科研对象的双重虚拟与观察渗透。与大数据科研方法相对应,科学研究的对象被以数据的形式二重虚拟化。"海量数据的出现催生了一种新的科研模式,科研人员只需从数据中直接查找或挖掘所需要的信息、知识和智慧,甚至无需直接接触所研究的对象。"[5] 美国的海洋观测站计划(OOI)旨在帮助科学家们通过高清影像设备、传感器控制、遥控潜水器等与海洋实现互动。但要实现该计划,还需要计算机科学家与海洋学家合作,共同提供采用连续数据的模型、自动化的数据质量控制和校准、支持数据分析和可视化方面的新方法。[2](P32–35)这预示着在大数据时代,科学的观察渗透进一步加强;获取数据的方法与设备、处理庞大数据的能力决定科学家能研究什么以及得到怎样的研究结

[收稿日期]2014 — 04 — 26

[基金项目]教育部人文社科课题《构建促进协同创新的人文社科科研评价体系研究》(13JDXF007)。

[作者简介]刘伟榕(1990—),男,福建泉州人。硕士研究生,主要研究方向:科技哲学、科技政策与管理。

果；渗透到观察结果中的不仅有本领域的科学理论，还有来自数据处理领域的理论与算法。

二、知识生产方式和科学进步模式的变化

首先，知识生产对技术与资本的依赖性增强。大数据时代，科学研究与信息技术手段之间的联系越来越紧密。以大数据技术进行的研究需要极多的资源，收集、储存、保留、管理、分析和共享海量数据各个环节都需要设备、技术与人才，获得相当的科研资金才可能进行。先进的数据处理技术既对科学研究提供了有力和有效的手段，又造成了科研路径上的依赖甚至是障碍。研究者若没有相应的技术与设备，就无法获得足够的数据和深入的分析处理。因此，资源以及获取资源的能力决定着科学家事业的前途，资本对知识生产的控制力将得到空前的强化。例如美国和加拿大海洋气象台的海王星项目拨出大约 30%的预算用于信息化基础设施（将近 1 亿美元），而小实验室的科学家只能用免费的 EXCEL 来处理数据。

其次，知识生产更倾向于工程化协作。由于使用大数据系统需要昂贵的技术成本，这使得科学家之间形成合作联盟，共享仪器设备与技术服务。如 LHC 每年将产生 50–100PB 的数据，其中大约 20PB 数据通过国家级网格的全球联盟进行存储和加工，这一联盟连接了 100 万台 CPU。[6] 除了节约成本的考虑外，产生大数据的项目大多本身就是一个大科学工程，需要科研人员进行跨越多个领域的协同工作、各个领域的专家共同解决一些复杂问题。例如海洋观测站计划（OOI）的电缆部件研究由华盛顿大学负责，维多利亚大学领导了在加拿大的工作，美国海洋规划协会管理和整合整个 OOI 系统，伍兹霍尔海洋研究所和加利福尼亚大学圣地亚哥分校分别负责管理项目的沿海–全球部分和网络基础设施部分。[2](P32)

再者，科学进步日益共享化与全球化。在工程化协作中所实现的科学进步，实质上也是一种共享式进步。得益于大规模计算能力、存储能力和科学仪器的共享支持，科学家们能够方便地获得和使用大量的来自其他科研团队的科学数据。例如，2009 年丹麦第一例 H1N1 感染者得到确认的几天之后，H1N1 病毒中的 H1 亚单位序列的全部 1699 个碱基就被提交到了 EMBL–Bank（欧洲分子生物学实验室核酸序列数据库），此后美国、意大利、墨西哥、加拿大、以色列等多个国家都提交了更多的病毒亚单位序列数据。[2](P120)在这样的共享中，研究周期和研究费用将大幅度缩减，从而提高了科学进步的速度与质量。大数据还使得科学进步日益呈现出全球化的效应。例如微软全球望远镜（WWT）作为国际“虚拟天文台”的一部分，现在可以无缝链接到天文学家们已经习惯的定量研究工具上。[2](P41)在这样的研究模式中，科学家足不出户就能获得其他国家的技术设备与科研成果，来自全球的数据和信息能够被用来为某一研究课题服务，得出的成果原则上是一种全球性的成果。

三、反思：大数据的利与弊

从积极的一面来看，大数据或将开创科学研究的“第四范式”。大数据相关的科研方法将在越来越多的领域中发挥重大的甚至是决定性的作用。有了数据处理系统的辅助，科学家可以把精力集中在创造性的劳动上，大数据不会自动产生科学知识，但至少增加了科学家做出科学发现的时间和可能。一批乐观的科学家更是看到了大数据对科学的变革力量。2007 年，已故图灵奖得主吉姆·格雷（Jim Gray）把数据密集型科学从计算科学中区分出来，提出了数据密集型科学研究的“第四范式”。科学研究最早的两种范式是实验型科研与理论型科研，第三种范式即计算型科研通过利用计算能力发挥理论的作用，第四种范式则是在未知规律的情况下，运用计算能力从大数据中发现规律。

依赖大数据也可能带来众多负面影响。一是科研资源垄断可能加剧。科学家能否进入大数据的研究平台，受制于海量的科研数据是否开放，也取决于是否有相应的设备来获取和处理这些数据。大数据与资本紧密结合的特性强化了科研资源掌握者对科研的走向与产出的控制。二是科学家可能形成技术路径依赖。数据技术只能对丰富而且复杂的真实世界提供相对简略的描述。更进一步而言，寻找不同寻常和意料之外的东西需要创造性和洞察力。计算机和数据库不可能自动导致创造性的科学发现，科学家如果过分依赖数据资源和搜索工具，就会造成亲身实践获取“第一手”资料的能力退化。三是科学合作的成果归属易引起纷争。首先，对于数据提供方能否算作合作者并给予一定的署名权存在争议；其次，对于工程化和全球化协作产生的成果是属于集体智慧的，对于成果的所有权该如何分配？2013 年的诺贝尔物理学奖仅颁给两位理论创始人弗朗索瓦·恩格勒特和彼得·希格斯，而发现希格斯玻色子的几千名粒子物理学家却无缘此荣誉，这引起了包括诺奖评委安德斯?巴拉尼在内的抗议。最令人担心的是，随着科学和技术和商业性的开发越来越联系紧密，一些具有商业价值的科学信息和数据为拥有者所不愿意公开，甚至通过申请专利来实施保护，这将带来更大范围的不公平与纠纷。

大数据是对人类信息处理能力的挑战，对科学家们来说则是面临着科研数据爆炸式增长的威胁，如果没有应对好，科学可能就无从进步。科学家们面对数据的泛滥，还应该从根源上去反思，比如实验思路是否出了问题。同时，大数据是应对数据挑战而提出的技术系统，这也使得科学研究与技术手段之间的界限越来越模糊，科学能力甚至在某种意义上转化为了技术能力，这对科学与技术之间的关系提出了新的问题，值得学者们深入研究。

〔参 考 文 献〕

〔1〕Manyika，J，Chui M，Brown J，et al. Big Data: The Next Frontier for Innovation，Competition and Productivity〔R〕.McKinsey Global Institute，2011:1.

〔2〕Tony Hey，等.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.

〔3〕冯海超.透视美国大数据爆发全景〔J〕.互联网周刊，2013，(01)：39.

〔4〕Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete〔J〕. Wired，2008，(07)：16.

〔5〕牛禄青.构建大数据产业环境——专访中国工程院院士、中科院计算所首席科学家李国杰〔J〕. 新经济导刊，2012，(12)：39.

〔6〕A.M.Parker.Towards 2020 Science〔M〕.MicrosoftCorporation，2006.

〔责任编辑：陈玉荣〕