

大数据时代的公众理解科学

李大光

(中国科学院大学人文学院, 北京 100049)

[摘要] 大数据时代对公众理解科学研究提出了挑战。应用技术已经渗透进公众生活的最微小之处, 对公众的产生的怀疑已经不能仅仅归结为科学素养水平低或者所谓的“素质低”。美国近些年对科学素养测试题的争论以及科学知识 with 态度之间的非线性趋势, 意味着不能仅依靠主观的问卷设计和抽样调查就得出可信结果。大数据为观察公众对科学技术的理解形成的模式的多种因素, 以及对态度的影响的因素可能会有新的发现。文章提出大数据的社会模式以及其他建议, 其目的是引发更多讨论。

Public Understanding of Science in Big Data Age

Li Daguang

(University of Chinese Academy of Sciences, Beijing 100049)

Abstract: Big Data Age poses challenges for studies on public understanding of science research. Application technology has penetrated into all aspect of public life. The public skeptical of science and technology could not only be due to low level of scientific literacy, or the so-called “low quality”. The NSF of United States has been in the debate recent years, on science literacy test and nonlinear trend between scientific knowledge and attitude. The research in the past 30 years warns us that we could not just rely on questionnaire and sampling survey to get credible results. Big Data could suggest a model for observing the change of public understanding of science and technology and the attitude, and the influential factors. The purpose of this paper gives a idea of the social mode of Big Data in China and other recommendations, and bringing more discussions.

1 大数据时代的特征

2014 年《自然》的一篇论文中, 科学家绘制出地球附近成千上万个星系, 他们称其为“拉尼凯”(Laniakea)^①。这个星系团远比天文学家以前认识的大得多。银河系是这个令人惊叹的巨大“超星系团”的部分星系。

“拉尼凯”内有超过 100 000 个星系, 纵横 5 亿光年, 银河系只是一粒位于右边边缘的一颗小颗粒。

加利福尼亚大学的拉里·斯马正在追踪研究所有人的身体功能。你想过粪便含有多少信息吗? 每克粪便上大约有 1 000 亿个细菌, 每

收稿日期: 2015-02-15

作者简介: 李大光, 《科普研究》编委, 中国科学院大学人文学院教授, Email: ldaguang@vip.sina.com。

① Laniäkea 源自火鲁努努语, 意思是“开放的天空”(open skies) 或者“广阔视野”(wide horizon)。

个细菌都有自己的 DNA。

2014 年 48 个最重要的科学发现中, 天文学研究居第一位的是罗塞塔 (Rosetta) 在跟踪 67P 彗星 10 年后, 分离“菲莱” (Philae) 彗星登陆器, 成功着陆彗星, 实现了人类天文学研究的伟大壮举。最令人吃惊的是, 即使在信息时代, 地球人也只能在“菲莱”探测器上的摄像仪器回收以后, 才能解读发现新数据。但是, 在大数据时代, 即使“菲莱”在碰撞翻滚掉落至狭小的沟坎, 无法在接受太阳光充电的情况下, 仍然能够将数据发射回地球。地球人几乎在同一时刻“观看”到着陆实况和彗星表面状况。

大数据对生命科学研究也带来令人惊异的成果。美国宇航局“好奇号”一直在探索火星上生命迹象。目前已经有成堆的数据表明这颗红色星球曾是一个适合微生物生存的环境。探测器收集的样本表明水绝对存在。人类对宇宙生命起源的好奇心在“好奇号”的成功探索下, 正得以满足。美国生命科学家、“科学坏小子”克里格·文特尔 (Craig Venter) 不仅成功地实现了个人基因图谱解读, 而且开始向生命之谜进军。在他的《光速中的生命: 从双螺旋到数据生命的黎明》(Life at the Speed of Light: From the Double Helix to the Dawn of Digital Life) 一书中, 已经设想通过国际太空站对月球生命样本基因解读, 将数据发至地球, 他将其还原为基因组, 使月球生命复活。

大数据在社会文化和信息管理中的作用已经使所有人都感觉到了。每天人们在 Twitter 上所发的信息相当于 1 000 万页鸿篇巨制, 大约是 8 163 本《战争与和平》。高度大约为 1 470 英尺, 大约相当于台北 101 大厦从基座到屋顶的高度^②。

除了科学技术研究中大数据发展, 使得宇宙间几乎所有的领域都可以通过大数据计算、储存和传播以外, 大数据已经在国家和城市管理中逐步实现。

大数据时代数据管理主要是“云”。“云”应该成为水、电、气社会服务体系的第四部

分。云储存和云计算是管理的主要手段, 个人的信息储存和所有的能量消费通过“云”及时计算和通告。城市交通管理中, 汽车无人驾驶和寻找停车位都实现云计算和通告; 空气质量管理中, 城市的能源利用和空气质量控制都可以通过大数据管理达到最佳^②。

20 世纪人类已经迈入信息化时代。20 世纪 80 年代, 美国副总统阿尔·戈尔提出信息高速路 (Information High-Way) 的计划, 美国在短短的 20 年时间内实现遍布全国的光缆覆盖。信息化使得美国的科学研究和工业化速度得以极大提升。

1998 年, 美国前副总统阿尔·戈尔再次提出“数字地球” (Digital Earth) 概念。2008 年, IBM 董事长彭明盛 (Samuel J. Palmisano) 提出“智慧地球” (Smart Earth)、“智慧城市” (Smart City) 的理念。大数据时代借助数据分析来改变人的行为和决定。世界各国开始了大数据时代的竞争。

2 大数据时代观察公众理解科学社会模式

从半个世纪前 (约 20 世纪 60 年代) 公众与科学之间关系的研究构成了大约三个范式。

第一个范式是开始于 20 世纪 60 年代的“公众科学素养” (Scientific Literacy)。这个范式主要研究解决的是公众科学知识缺失 (Public Deficit Model) 的问题。研究方向主要是科学素养调查和加强科学教育。

第二个范式是从鲍默爵士 (Sir Walter Bodmer) 于 1985 年发表其“公众理解科学”报告^②作为起端, 引发了公众理解科学 (Public Understanding of Science) 运动。从那时到现在, 其主要研究目标是公众对科学的态度问题。研究内容主要是知识与态度之间的关系, 观察态度变化的原因。研究结果发现, 公众知识与态度之间呈现出非线性趋势。即科学与态度转变并非知识水平越高, 对科学技术越支持。主要解决方法是建立科学家和科学家团体的形象和教育。

第三个范式为英国上院发布的《科学与社

② Bodmer, W. (1987) “The Public Understanding of Science”. Science and Public Affairs. 2: 69-90。

会》(Science and Society, 1985)^③为起端而开始的现代科学传播模式研究。主要研究问题为公众对科学的信任缺失(trust deficit)、专家信任缺失(expert deficit)和公众对科学的信任危机(crisis of confidence)。解决问题方案:公众参加科学决策;公众审议(deliberation);科学的“天使”形象传播以及传播效果评估。

20世纪90年代,科学与公众之间关系研究已经进入“科学与社会”(Science and Society)之间关系研究。^④公众的科学素养已经不再仅仅是脱离整个社会发展的孤立的研究领域。脱离社会环境和国情,脱离文化环境和地域文化特征的科学素养研究,除了能够得到一些数据以外,其价值与意义已经变得微不足道。这就是为何这些年来,中国的科学素养研究遭受学术界和公众一些质疑的主要原因。

以马丁·鲍威尔(Martin Bauer)^⑤为主要代表的欧洲学者和部分其他国家学者提出了建设科学文化环境中的公众理解科学研究设想。但是,科学文化中的公众理解科学研究的提出不是突来奇想的创意,而是基于早在20世纪60年代开始的以接近科学文化基本研究框架的研究成果提出的。他认为,仅仅靠调查公民的科学知识除了满足政府机构的一些要求以外,对于了解数据后面的原因没有任何帮助。^⑥而在科学文化环境中观察公众对科学技术的理解以及寻找影响科学文化形成的因素,从而有针对性地制定政策具有重要的意义^[3]。

构成一种文化的主要要素是信息粒。这种信息粒由数据构成。大数据时代对公众对科学技术的理解和态度应该放置于科学文化角度进行观测,仅仅将公众的科学素养或者公众对科学技术的理解单维角度进行探究,不仅不会得出真正的结果,而且可能会误入歧途。建议在中国,根据中国的国情,建立中国的科学文化

社会结构形成观察模型,并在经过长期观察后了解在中国科学文化的发展规律(见图1)。

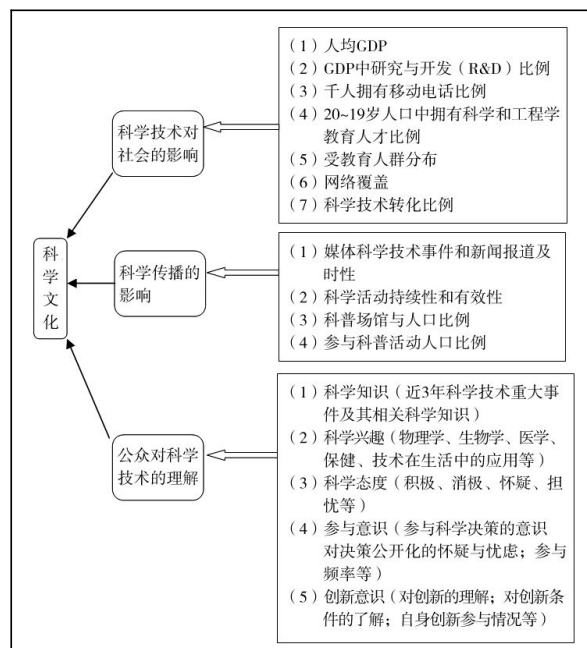


图1 科学文化模式

任何人,任何一个民族,对任何事情的看法和观点的形成都不能脱离其文化。在分析公众对科学技术的理解中,必须考察构成科学文化的所有要素,以及各种要素对公众的认知、理解和行为的影响。这种观察和分析只有在大数据时代才能完成。

3 大数据观测公众对科学技术态度形成模式

1839年,美国统计协会(ASA)在波士顿成立。19世纪30—70年代,美国的数据文化进入到一个新阶段。这个统计阶段的出现与南北战争期间,议员们需要通过社会学详尽的统计数据争取利益与决策权力。严格讲,统计,从其诞生之日起就具有民主的意味。不知国情,无法决策。

在1824年到1936年的100多年时间内,民意的调查一直追求对调查群体的大面积覆盖,那时,无人怀疑“覆盖越大数据越准”的

③ <http://rsnr.royalsocietypublishing.org/content/early/2010/06/16/rsnr.2010.0035.full.pdf+html>.

④ Bauer et al: 25 years of PUS research, p85, What we learn from 25 years of PUS survey research? Liberating and expanding the agenda. Public Understanding of Science, 16 (2007) 79-95.

⑤ 伦敦经济学院教授。1989年与John Durant共同进行欧洲共同体(欧盟前身)公众对科学理解的调查。其中科学素养部分采用的是米勒设计的题目,只是在科学概率问题上进行了修改。

⑥ Bauer, Martin et al, Construction and Validation of “Science Culture Index”, <http://www.ncaer.org/Downloads/WorkingPapers/WP100.pdf>.

看法。

统计学的革命应该追踪溯到 1895 年挪威国家统计局的安德斯·尼克拉·凯尔 (Anders Nicolai Kiaer, 1838—1919) 在国际统计年会上提出的抽样的观点。但是, 抽样的观点在经过长达 30 多年的时间才逐步达成学术界的共识。只有乔治·盖洛普 (George Gallup, 1901—1984) 在 1936 年成立不久的美国舆论研究所的具体实施下, 仅仅用了 5 000 个样本的问卷调查就成功预测罗斯福竞选总统的成功。而坚信大量覆盖统计结果的《文学文摘》(Literary Digest) 在综合了 240 万人的意见之后, 仍然错判。从此, 盖洛普的“科学抽样”风靡全球以及各个领域的调查。在 1936 年到 2012 年的 19 次总统大选预测中, 仅仅失败两次。盖洛普开创了抽样调查的伟大时代。

抽样调查方法通过对具有代表性的人群进行抽样, 通过精心设计的问卷采集被访者的看法和意见, 在精心设计的数学模型的归纳后得出结论, 这种方法直至今日都是普遍采用的方法。在公众理解科学和对科学技术的态度的研究中, 通过长达半个世纪的讨论, 学术界认为公众理解科学最重要的是公众的科学素养, 只有科学素养提升了, 公众才能理解科学家的工作。科学素养的研究成为这个领域的焦点。从 20 世纪 50 年代进行的由科学作家协会做的第一次流产的公众对科学技术态度调查后, 1979 年, 美国 NSF 委托乔恩·D. 米勒总结出科学素养的三个维度以及将 2061 计划 (Project 2061) 中 120 个科学概念通过项目反应理论 (IRT) 设计出符合美国和西方国家知识水平的科学素养测试问卷。至今仍然在很多国家使用^⑦。

在大数据时代, 对于抽样调查, 也就是通过对具有主观意识而抽取的“具有代表性”的样本进行的调查, 在学术界产生了怀疑与讨论。这些讨论虽然相当大程度上集中在工业界和商业界的数据采集, 与此同时, 学者们也开

始对社会学一贯采用的抽样数据在大数据时代的价值产生了怀疑。

争论可以追溯到 2010 年负责发布每两年一次的《科学与工程学指标》(Science & Engineering Indicators) 中的第七章“公众对科学技术的理解和态度”中调查结果的争论。

“如果一个人不接受‘进化论’和‘大爆炸’学说是否具备基本的科学素养? 大多数科学家和教育家都会毫不犹豫地回答‘当然不具备科学素养’”这个问题在美国和某些欧洲国家的社会学研究领域一直存在争论。

2000 年, 美国科学基金会召开了专家会议, 会议建议将“就我们目前所知, 人类是从早期动物进化而来”和“宇宙产生于大爆炸”这两个问题前面加上前提陈述: “根据进化理论……”, “据天文学家说……”。理事会决定要求科学基金会在下次调查的时候使用新修改的问题对半数被访者进行调查, 分析结果。米勒听说后十分生气, 他说: “我们从不类似什么‘某些经济学家说, 我们遇到经济衰退’, 或者‘气象预报员说, 发生了海啸’这些前提说明。”^⑧总统任命的科学理事会决定将“进化论”和“大爆炸”知识测试题从 2010 年《科学与工程学指标》调查问题中删除。^⑨问卷的修改引发了激烈的讨论。奥巴马政府“白宫科学技术政策办公室”(White House of Science and Technology Policy Office) 要求科学理事会提供完整和准确的调查数据, 并追问为何在调查问卷中删除了这两个问题。理事会事后的解释是因为这两个测试题设计不成熟, 所以放弃了修改调查问卷文本。目前, 理事会已经承认撤销那两个测试题是错误的, 并决定在 2012 年的调查报告中保留这些问题的分析。

2011 年, 一直支持公众对科学技术理解和态度调查的美国国家基金会 (NSF) 领导机构国家科学理事会 (National Science Board) 不能确定, 因为宗教信仰的原因而否认进化论

⑦ Yudhijit Bhattacharjee. New NSF Survey Tries to Separate Knowledge and Belief. p.394, 22 July 2011, Science.

⑧ Yudhijit Bhattacharjee. NSF Board Draws Flak for Dropping Evolution Indicators. p.150, 9 April 2010, Science.

是否会影响公众的科学素养。科学理事会试图在调查题中将科学与宗教信仰区分开来。这种尝试立即遭致学术界的批评。批评者认为这种试图修改测试题的想法是放弃科学立场，向宗教投降。美国《科学》援引康奈蒂克大学政治学教授乔治·毕晓普（George Bishop）的话：“美国文化中圣经传统使得那个问题确实是在测试信仰，而非知识。”对此，乔恩·米勒坚持知识问题本身确实能够测试一个人的科学知识理解水平，他说：“如果一个人说地球确实是宇宙的中心，你难道说这个人是具备科学素养的吗？”

尽管理事会认为删除那两个问题是一个错误，但是，在2010年10月的研讨会上，由人类学家克里斯·托米（Chris Toumey）领导的社会学家提出了一个超越基于传统的公民权利科学素养（civic scientific literacy）观点的测度公众科学素养的知识的新框架。新框架将采用多种问项，其中包括基因型变异和微生物进化等知识。

2010年11月，由社会学家托马斯·古特伯克（Thomas Guterbock）主持的另一个科学基金会支持的研讨会上，与会学者们提出涉及到进化论的测试题应该包括植物进化、适者生存和其他进化过程的知识，避免直接涉及人体器官等热门议题。

米勒认为，两个会议报告目的是将多数美国人持续不断的否认进化论的基本事实变得并不那么引人注目。“其主要思想是，美国人在这个问题上得分不高，因此，这个问题应该取消。这是对科学调查最基本原则的公然挑战。”他认为删除这两个问题是“愚蠢的尝试”，“没人愿意谈论我们的婴儿死亡率数据，但是，即使你不说，事实仍然存在。”^⑨美国国家科学教育中心的约苏尔·罗森纳（Joshua Rosenau）长期以来坚持将“创世论”（creationism）排除教育课程之外。他认为，这两个报告令人沮丧。他说：“无论什么文化背景和原因，否认进化论都会给一个人将新科学

进展和已有的科学成就与他们自己的生活，参与医疗护理以及适应21世纪的经济生活产生深刻的影响。”“如果国家科学基金会在调查中删除这个基本科学观点，那么，他们的科学素养调查将空图虚名，无实际价值。”

在科学素养第三维关于伪科学和迷信的认识也有巨大的争论，这也是为何直至今日，第三维一直不具备国际比较的主要原因。以西方基督教和多维思想占据主要地位的研究界认为，迷信与科学本身并不矛盾。反对者认为，五大科学天才相信魔法。伽利略（Galileo Galilei, 1564—1642）相信占星术改变了一切；牛顿（Isaac Newton, 1642—1726）认为炼金术将改变未来，同时，在30岁创造了万有引力定律后将余生都用来探索上帝的存在，以解释第一推动力的来源；第谷·布拉赫（Tycho Brahe, 1546—1601）使每个人都相信他是一个巫师；卡尔·林奈（Carl Linnaeus, 1707—1778）相信生物分类中包括神奇的动物，像九头蛇、美人鱼和凤凰；帕拉塞尔苏斯（Paracelsus, 1493—1541）相信可以医治疾病的是自然魔法和他自己的魔力。^⑩这也是为何西方学者不认为迷信与科学探究相矛盾的主要原因。

在抽样设计中，样本的“代表性”是主观的，问卷中问题的设计也是主观的。尽管经过长时间的研究和设计以及预调查，甚至增加质化调查做补充数据参考。但是，有一点无法改变，那就是，所有的问题都是调查者主观设定的。

大数据时代主要特征是：大数据通过海量数据的存储、计算和即时图形，显示数据之间的关联关系，而非因果关系。最著名的案例就是沃尔玛在遍布全球的8800多家超市中统计的大数据中发现尿不湿与啤酒的销量同步增长的数据，通过观察，发现购买尿不湿的都是年轻的丈夫。他们在买了尿不湿后，顺便给自己买了啤酒。大数据的发现，使得沃尔玛改变商品的摆放，迅速地增加了啤酒销量。

⑨ Udhijit Bhattacharjee. p.150, 9 April 2010, Science.

⑩ “These 5 men were scientific geniuses. They also thought magic was real.” <http://www.vox.com/2015/2/3/7969843/scientific-genius-magic>。

大数据专家涂子沛认为：“当前人类的数据约 75%都是非结构化数据，大数据的表现形式主要就是非结构化数据，而大记录、非结构化数据要体现出价值，当前主要的处理方法，还是把它们转化为有严整结构的数据，即传统的小数据，因此，我认为，大数据的价值维度主要体现在传统的小数据和结构化大数据之上，而大数据的容量维度主要体现在现代的大记录和非结构化数据两个方面^[5]。”大数据是通过信息技术手段，通过云储存和云计算，在即时的图形形式下最及时地反映出某一个数据与其他数据的关联度，从而发现其用抽样方法和问卷无法得知的自然数据。

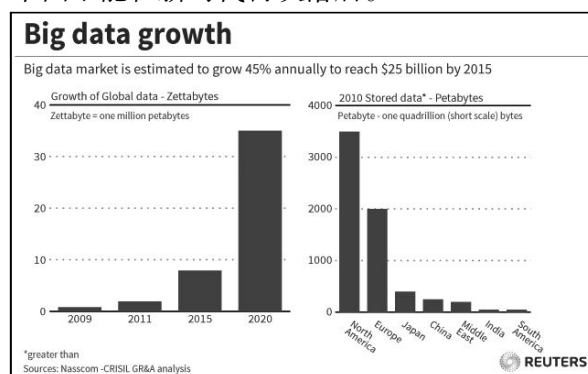
公众对科学和技术的看法在每一个不同的文化的国度中都存在差异。这种差异除了前面谈论到的宗教因素以外，还有随着社会的发展和变化，我们不可能随时发现的现象和趋势。

近些年来，科学技术事件层出不穷，不断引发公众的争论。大概记录一下：SARS、禽流感、圆明园防渗膜事件、“汉芯”事件、太湖巢湖蓝藻事件、西部水利大开发争论、怒江大开发的争论、厦门PX事件、松花江污染事件、中西医争论、伪科学的争论、珍奥核酸事件、华南虎事件、三鹿奶粉事件、转基因农作物、纳米技术、食品添加剂、核能利用、蓬莱石油泄漏、大连PX事件等。随着技术的不断应用以及网络的发展，公众对技术事件的舆论会在瞬间形成，在网络上形成乌合之众效应。对于公众的舆论的形成需要大数据的手段探究公众舆论形成的基本数据以及数据之间的关联性。我们大概不能仅仅用“科学素养低”这样简单的结论做最终判断。

4 我们需要做什么

对于大数据的发展以及中国在这个领域中的地位，我们应该有清醒认识。大数据市场预计每年以 45%的速度在增长，到 2015 年将达到 250 亿美元。数据量到 2020 年将比 2015 年增长将近 3 倍。截至 2010 年，大数据存储量最高为北美两个国家，欧洲其次。亚洲国家日本居先，而中国仅排第四位(见图 2)。大数据专家们在呼吁：叫醒中国，超越中国模式，大

数据将是下一个社会发展阶段的石油和金矿。中国不能在新时代再次落后。



(数据来源: Reuters graphic/Catherine Travethan, 2012-5-10)

图 2 路透社关于大数据增长的趋势统计

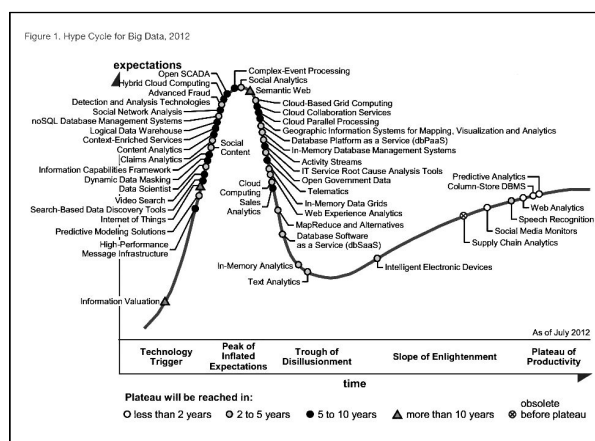
全国各个部门，地方政府和企业，都要数据开放。这是一个艰难，但是必需的工程。2008 年 1 月 21 日，奥巴马就职第一天发表了一份总统备忘录，命令美国联邦机构的负责人尽可能多地开放数据。他说：“面对怀疑，公开优先。”他的指令导致了 data.gov 网站的建立。2012 年 7 月，公开数据已经达到 45 万个，涵盖 172 个机构。中国很多学者和科学家也都在呼吁信息开放。目前我们的政府和地方政府以及国家所属机构已经开放了部分信息，但是，距离大数据的要求仍然还有距离。

厘清公众理解科学影响因素，在没有实现真正意义的大数据之前，与有关机构合作，开放数据，共同研究，数据共享。

大数据创始人基登·嘉特纳 (Gideon Gartner, 1935—) 提出了大数据的 3V (Volume, 大容量; Velocity, 高速率; Variety, 多样化) 的特征，而且他成立的研究所研究出大数据时代数据的处理技术和过程在时间段的周期变化规律(见图 3)。

确立数据使用伦理道德观念，建立数据使用法律和监督系统。数据使用、保密限制和权限。

关于大数据只有数据之间的关联性，没有因果关系的认识，可能需要时日才能下最终结论。美国一个数据发现冰激凌的销售与鲨鱼伤人事件呈正相关趋势。初步猜测是冰激凌的味道导致鲨鱼伤人。但是，再进一步分析，冰激凌夏季销售量大，而夏季去海边游泳的人比例



(来源: <https://www.gartner.com/doc/2100215/hype-cycle-big-data->)

图3 大数据技术成熟曲线图

会增加很多,自然鲨鱼伤人的比例会增加。相关性的数据应该是追究因果关系的基本依据⁹。从这个角度讲,从海量的大数据中分析关联性,在关联性中探究因果关系,大概是今后大数据在社会学研究方面能够取得突破的地方。

参考文献

- [1] 涂子沛. 大数据:正在到来的数据革命,以及它如何改变政府、商业与我们的生活[M]. 桂林:广西师范大学出版社, 2012.
- [2] 维克托·迈尔-舍恩伯格, 肯尼斯·库克耶. 大数据时代:生活、工作与思维的大变革[M]. 盛杨燕, 周涛, 译. 杭州:浙江人民出版社, 2013.
- [3] 涂子沛. 数据之巅:大数据革命,历史、现实与未来[M]. 北京:中信出版社. 2014.
- [4] 埃里克·西格尔. 大数据预测[M]. 周昕, 译. 北京:中信出版社. 2014.
- [5] Jon D. Miller, Rafael Pardo, Fujio Niwa. Public perception of science and technology: A Comparative Study of the European Union, the United States, Japan and Canada[R]. Foundation of BBV.
- [6] Martin W. Bauer, Nick Allum and Steve Miller. What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda? [J]. Public Understanding of Science, 2007, 16: 79-95.

(编辑 丁雪)

2015 年度《科普研究》欢迎投稿

《科普研究》(双月刊)是由中国科普研究所主办、中国科学技术协会主管的专业学术期刊。它的刊文范围包括科技传播、科学教育、科普展教、科学与文化等领域。本刊以“促进科学技术普及理论的研究,推动科普事业的发展”为办刊宗旨,以“百家争鸣、求是创新”为办刊方针,坚持聚焦中外科技传播前沿动态,促进科技的传播与发展,为广大科技传播和科普工作者搭建一个活跃学术思想、发表学术论文的平台,努力为提高公民科学素质、促进科技创新、经济发展及人文进步作贡献。

2015 年,本刊栏目设置有专稿、专题、研究论文、综述等,欢迎广大作者投稿。

投稿网址为: <http://kpyj.crsp.org.cn>