

大数据专题研究 (九)

找回失去的传统: “大数据”研究范式的反思与重构

文 军 吴晓凯

(华东师范大学 中国现代城市研究中心暨社会发展学院, 上海 200241)

摘 要 : 作为一项正在兴起的新技术, 大数据从根本上改变了传统科学论证研究的架构模式, 为科学研究、商业活动、政府决策和社会生活都带来了便捷的技术支撑, 越来越成为现代社会不可或缺的一部分。然而, 在大数据日益繁盛的背后, 却隐藏着“数据崇拜”的深刻危机。其所推崇的工具理性在一定程度上抑制了人文情感的表达。文本从社会语境的角度, 试图从数据产生的机制、创新的方式和价值的建构等角度重构大数据作为符号的文化属性和社会意义, 以适应现代人所具有的自觉意识和反思能力。与此同时, 指出“大数据”研究方法同抽样调查、观察与访谈、实验等方法相互融合、相互补充、相互渗透后的复合方法对于人类认知系统的升华具有重要的现实意义。

关键词 : 大数据; 范式反思; 人文属性; 社会意义; 复合方法

中图分类号 : C3

文献标识码 : A

文章编号 : 1005-9245 (2018) 01-0063-09

自 2008 年 9 月《自然》(Nature) 杂志推出了“大数据” (Big Data) 的专栏以后, 大数据便逐渐成为互联网技术行业中的热门词汇^[1]。同时, 与大数据相关的研究和著作大量涌现。与以往的抽样数据 (小数据) 相比, 大数据不再使用随机抽样法, 而是使用所有数据进行分析处理^[2], 在数据采集、处理方式、类型构成等诸多方面都有本质的区别^[3]。当前, 由于大数据在理论研究与社会实践上的巨大影响, 有关大数据的讨论已经渗透到每一个行业和领域之中, 带来了各行各业对大数据的热捧。大数据正在从一种信息符号称谓、信息技术体系或者信息技术的应用, 提升到足以影响一定区域内各种社会活动主体思维、价值判断和行为选择的文化高度^[4]。正如有学者指出的, 大数据开启了一次重大的时代转型, 它正在改变我们的生活以及理解世界的方式, 成为新发明和新服务的源泉。与此同时, 大数据可

能带来的风险与挑战也是不容忽视的, 只有正视各种风险与挑战, 才能更好地推动大数据向前发展。

一、大数据: 一项范式转换的技术革新

在过去的近 20 年时间里, 随着信息技术与设备的狂飙式发展, 人类对于数据的驾驭能力也达到了前所未有的高度, 数据几乎无处不在、无处不有。“大数据”作为一项崭新的技术已经被应用于多个领域, 成为重要的生产因素。当前, 全球大数据解决方案不断成熟, 各领域大数据应用全面展开, 为大数据发展带来强劲动力。据 Wikibon 数据显示, 2015 年全球大数据市场规模达到 384 亿美元, 同比增长 34.7%, 其中硬件、软件和服务所产生的收入占比分别为 38%、23% 和 39%, 大数据孕育了诸多新兴产业 (见表 1)^[5]。尽管从目前相关研究来

收稿日期 : 2017-04-02

基金项目 : 本文系国家社科基金重大项目“有序推进农业转移人口市民化研究” (13 & ZD043)、“现阶段我国社会大众精神文化生活调查” (12 & ZD012) 的阶段性成果。

作者简介 : 文军, 华东师范大学中国现代城市研究中心暨社会发展学院院长、教授、博士生导师, 华东师范大学—纽约大学社会发展联合研究中心 (上海纽约大学) 主任; 吴晓凯, 华东师范大学社会发展学院社会学系博士研究生。

看,学术界对大数据的历史梳理与概念界定仍然处于争议之中,但是关于大数据的表现形式,则形成了较为流行的“4V”学说,即规模性(Volume)、多样性(Variety)、高速性(Velocity)以及价值性(Value)^①。大数据依托其迅速、庞大、多样、细微的计算优势将整个世界编制为空间数据符号,从自然界到人类社会再到行为个体都能够凭借其精湛的技术量化为丰富的数理模型。

大数据的兴起并不仅仅是一项技术领域的革新,更是与人类认知和科学研究密切相关的范式转换^②。大数据改变了以传统抽样数据为主流的科学研究方法,推动科学研究范式从知识驱动向数据驱动转型。正如图灵奖得主吉姆格雷(Jim Gray)所言,科学将进入继实验、理论、计算模拟之后的第四种范式:数据密集型科研(Data-Intensive Scientific Discovery)^[6],全球正在步入数据化社会阶段。数据密集型科研主要表现在以下几个方面:

第一,在科研人员的角色上,数据密集型科研推动研究者的工作重点由“如何获取经验材料”转向“如何分析海量数据”并获取知识。在大数据时代,伴随着信息技术的普及,数据的规模呈几何增长态势。以中国Web数据为例,中国互联网络信

息中心发布《中国互联网络发展状况统计报告》显示,截至2016年12月,中国网民规模达7.31亿,相当于欧洲人口总量,互联网普及率达到53.2%。其中,手机网民占比达95.1%,规模达6.95亿,增长率连续三年超过10%^[7]。尽管人类利用大数据获取资源的能力不断增强,但是大数据受设备仪器精确度、人工操作、自然环境、统计口径等方面的影响,管理难度异常大,数据不准确、不一致、不吻合现象时有发生。那么,如何从巨量数据中发现知识,寻找隐藏在数据中的模式,揭示世界规律成为科学研究者不得不面临的一项挑战。从科学研究的角度来看,科学研究者的角色将发生改变,他们不再思考如何获取研究数据,而是将研究工作重点转向从海量数据中挖掘有效信息,获取知识。

第二,在科学知识的生产方式上,数据密集型科研推动演绎式逻辑转向归纳式逻辑。作为认识论的核心,知识的来源问题在哲学史上主要可以分为经验论与唯理论两大阵营。二者的核心分歧在于人类的知识来源于已经被前人证明的正确理论、规律与命题还是人类观察到的现象^[3],并由此产生出两种知识生产的路径——演绎与归纳。与演绎主导的科学论证逻辑不同,归纳并不需要任何先验前提,

表1 2011-2017年全球大数据细分市场规模分析(亿美元)

	2011	2012	2013	2014	2015	2016e	2017e
云	3.6	6.2	11.9	18.2	25.2	30.5	36.5
行业解决方案	28	44.2	61.5	101	135	160	172
应用	5.2	9.9	16.9	34.5	52.9	66.5	77.5
非关系型数据库	0.7	1.3	2.9	5	8	10	12
关系型数据库	6.2	8.8	13.1	17.5	22.5	24.5	27
基础软件	1.4	4.4	8.3	10.8	12.5	16	19
网络	1.5	2.3	4.2	6.5	8.5	10.1	11.5
存储	11	17.5	30.9	42	55	64	69.5
计算	15.3	22.9	36.5	49.2	64	71	76

①美国IBM公司最早将大数据的特征概括为四个方面,具体表现为:第一,数据体量巨大。大数据的起始计量单位至少是P(1000个T)、E(100万个T)或Z(10亿个T)。第二,数据类型繁多。比如,文本、视频、图片、传感器、地理信息等。第三,价值密度低、商业价值高。第四,处理速度快,这也是和传统数据挖掘有着本质不同的地方。参见:《大数据究竟是什么?一篇文章让你认识并读懂大数据》,2013年11月4日,中国大数据网:<http://www.thebigdata.cn/YeJieDongTai/7180.html>。

②科学史学家库恩(T.S.Kuhn)自1962年提出范式(Paradigm)概念以后,范式就成为科学研究十分重要的概念。按照库恩对“范式”的理解和论述,“范式”是由从事某种特定学科的科学家们在这一学科领域内所达成的共识及基本观点,是一个学科的学者在研究准则、基本方法、概念体系等方面的某些共同约定。“范式”一般包括三个方面的内容,一是共同的基本理论、观点和方法;二是共有的信念;三是某种自然观。参见文军:《范式整合:全球化时代社会学研究的变革》,《学术论坛》,2001年第3期。

也无需研究结论证实或证伪相关研究假设，而是在无相关假设的情景下通过对有限经验材料的技术加工和抽象思维概括出普遍性命题或结论。不难看出，经验材料越丰富，那么归纳所得结论准确的概率也就越大。在抽样数据时代，受样本容量和抽样逻辑的限制，归纳仅仅可以定义为不完全归纳法，那么据此所得结论常常受到各方面的质疑。因此，归纳更加适合大规模数据作为支撑的宏观性科学论证，而大数据资源的广泛性正弥补了归纳式推理在抽样数据时代应用的不足。

第三，在科学理论的检验手段上，数据密集型科研推动经验检验转向预测检验。无论是通过有限的经验材料还是通过知识思辨论证亦或是两者相互结合所得出的理论，理论检验对于理论的发展都起到十分重要的作用。受逻辑实证主义、证伪主义、科学研究方法论等影响，检验理论的手段千差万别，形态各异。但归结起来，基本上都是依赖经验完成理论检验工作。然而，在大数据时代，由于规律是从过去和现在的海量数据中挖掘而来的，解释范围通常比较大，所以基本无需再回到经验世界进行检验。但是，大数据的“去因果性”特点对于未来预测的精准性则受到了方方面面的挑战，即大数据在具有较高预测力的同时，却只拥有较弱的解释力^[1]。为此，基于大数据所获得的数据规律更应该从其对未来的预测性进行检验。数据密集型科研的兴起助推了科学经验图景的改变和研究范式的跨越，颠覆了科学哲学一贯的认知方式。

第四，在科研方法的影响上，数据密集型科研导致学科数据化趋势日渐增强，由此带来了科研教学的深度交流与交叉学科的迅速衍生。长期以来，数据分析在传统人文社会科学类学科研究中普遍处于缺席状态，甚至部分人文学者认为人文社会学科应该坚持尊重人文关怀和美学意义的价值定位，抵制和批判大数据所引发的技术理性泛滥^[8]。事实上，大数据的来临不仅仅冲击了自然科学研究范式，人文社会学科也开始注重大数据在教学和科研中的量化思维与拓展式功能：一是在文献的整理与采集上，人文学科开始逐步利用大数据思维，广泛而全面地搜集世界各地的文献资料与经验素材，致使人文科学文本库的容量越来越大，学科跨领域、跨地区、跨文化交流向纵深发展。例如，截至目前，至少已经有120多个中文词汇被收入到《牛津词典》中，大数据助推语言学研究国际化^[9]。二是在教学模式上，自媒体时代的学习者不再仅仅局限于固定的教学空间，

而是能够通过信息传导迅速接受全球各地的教育资源，打破时空的物理束缚。三是在研究类型上，传统以抽象理论为主要研究内容的人文学科开始逐步关注经验事实，并尝试借助于自然科学方法论与统计技术向着精确性、标准化方向发展。有利于破解长期以来人文社会科学研究者因数据的残缺而无法有效地描述客观世界的现实性难题。与此同时，大数据的广泛应用导致各个学科资源的开放与共享，自然科学与人文社会学科研究资料的数据化统合了“文献梳理”“田野观察”与“实验室实验”之间外在形式的同时，也在一定程度上消解了研究方法之间的张力，并在更大范围内增强了数据使用的确定性与包容性，从而使经验资料的量化为越来越多交叉学科的出现提供了方法论基础。

无论如何，大数据时代的到来重构了人类对于世界认知与理解的方式，并开启了数据时代的大门。伴随着研究范式的转向，大数据被广泛应用于远程感应、天气预报、金融市场以及计算机技术、因特网运用等众多领域之中。在当今全球竞争日趋激烈的大环境中，大数据已经上升到国家战略高度^[10]，成为国家之间竞争的重要手段。以至于有学者提出，“除了上帝，任何人都必须用数据说话。”^[11]

二、范式反思：“大数据”研究范式引发的冷思考

大数据作为一项正在兴起的新技术革命，是计算机技术、智能技术、网络技术和云技术等多种科学技术的迅猛发展所促成的一项重要成果^[12]。大数据的出现从根本上改变了传统科学论证研究的架构模式，为科学研究、商业活动、政府决策和社会生活都带来了便捷的技术支撑，越来越成为现代社会不可或缺的一部分。然而，在大数据日益繁盛的背后，却隐藏着“数据崇拜”的深刻危机，其所推崇的工具理性不免会抑制人文情感的表达，掩盖个性特征、生产数据垃圾以及消解个体作为“人”的本真符号意义等。因此，持续的“大数据”热现象也为我们引发许多现实性的思考^[13]。

（一）防止大数据的“去人性化”忽视人的主体价值和数据的社會文化意义

大数据作为从传统社会的经验结构中脱域（Disembedding）出来的现代理性存在方式，当理性无限扩展到可能达到的极限，就会导致某种张力和“二律背反”^[14]。对于大数据而言，一个根本缺陷

就在于只关注“客观数据”的呈现,而缺少对“主观数据”的反映,忽视数据所蕴含的社会文化意义。从社会学角度看,所有的数据说到底都是关于“人”的符号,是一种具有社会文化意义的建构行动^[15]。因此,数据符号所蕴含的社会文化意义是真正需要了解和掌握的内容。如果脱离了人及其所在社会文化的价值本性,而无限放大大数据的效应及其影响力,就可能本末倒置,最终产生许多难以预期的后果。数据来源于社会又作用于社会。无论是可以数字化的数据还是不可数字化的数据,其比较、分析和归纳其实反映的都是人在符号层面上的一种互动。大数据汇集了行为主体、时间、地点三个维度的“数据宇宙”。在大数据时代,数据符号几乎可以囊括一切丰富信息,个人的信息更多的是以数据为载体来呈现并进行互动的。如果人及其各类社会行为都仅仅被看作一个个孤零零的数据符号,我们就很难从数据中解读出充满朝气、富有激情的人,也就很难从数据塑造出来的没有个性特征的人去推导群体样态和社会构成。

（二）防止大数据的“巨量化”掩盖数据个性特征、强化数据霸权

犹如互联网一样,大数据也是一把双刃剑。尽管大数据能够提供丰富、巨量的信息,但其基本来源仍然是个体单位。巨量化的大数据在给我们研究和分析社会现象带来便利的同时,实际上也掩盖了许多具有个性特征的数据的意义。大数据重点在于向广度拓展,从巨量化数据中揭示出具有普遍化的规律(Universal Regulation)以实现其方法论意义上的价值;而并非向深度延伸,寻求量化自我(Quantified Self)的纵向解析^①。在连续的巨量数据海洋中,许多本身具有个性特征的数据已经变得无足轻重,它们在大数据霸权面前会成为微不足道的牺牲品。由此,大数据简化了人们对数据差异性的认知。而且大数据的复杂性不仅在于其数据样本巨量,更在于其多源异构、多实体和多空间之间的交互动态性。多种多样的数据类型使信息载体更加复杂化,同时也丰富了“数据暴力”的手段。从这个

意义上说,防止大数据滋生出“数据暴力”是大数据发展中需要注意的问题。与此同时,现代性制度的特性与抽象体系中的信任机制紧密相关,对非个人化原则的信任,以及对匿名他人的信任,成了社会存在的基本要素^[16]。大数据的广泛延展性客观上要求数据的共享与透明,实现信息的相互匹配。事实上,当前只有少数大型企业掌握控制和处理这种复杂的巨量化大数据的技术,这样一种单向化透明的数据极容易对普通民众产生信息盲区。为此,特别需要警惕其对个体信息的安全性带来的威胁。

（三）防止大数据的“碎片化”导致人们难以全面认知复杂事物

随着互联网、物联网、云计算等信息技术的发展,大数据通常以“非结构化”形态存在,表现为离散型数据状态。从规范意义上说,任何数据本身并没有意义,只有同具体的需求相互关联才能产生相应的效应。大数据的价值在其能够被规范成为可供分析的形式之后才能最大程度地被挖掘出来,才能更好地发挥其应有的价值。然而,在巨量的大数据面前,实际上可供规范分析的数据只是极少数,许多数据都停留在“碎片化”阶段而难以被真正挖掘和分析,较容易产生片面经验图景和认知效果。从系统论的视域看,任何系统都是一个有机的整体,它不是各个部分的机械组合或简单相加,系统的整体功能是各要素在孤立状态下所不具有的新质^[17]。那么,对这些碎片化的数据进行清理,形成规整的结构,就成为大数据能够得到充分利用的根本所在。在大数据中,其研究范式重在发现而不是推论,运用大数据进行各种分析不是像传统的社会调查方法那样通过假设检验来进行推论,而是通过数据的总体归纳来达成对社会现象的总体分析。这不仅要求我们在数据清理技术、存储结构上进行优化,而且要在数据采集环节尽可能地做到规范,否则就很难获得对复杂事物的系统认知。

（四）防止大数据的“模糊化”产生大量信息垃圾、为数据造假与非法传播打开方便之门

在大数据时代,信息数量快速激增并形成了分

^① “量化自我”(Quantified Self)是指“运用多种多样的高新技术手段,对个人生活中有关生理吸收(Inputs)、当前状态(Status)和身心表现(Performance)等方面的数据进行获取。量化自我可分为广义量化自我和狭义量化自我。狭义量化自我又包括日常记录(Everyday Tracking)、运动追踪(Activity Tracking)、睡眠管理(Sleep Management)、情绪监测(Mood Tracking)和机能监测(Body Tracking)。参见刘振声:《“量化自我”——从数据化个体的角度重新审视“大数据”》,2014年3月21日,人民网, <http://media.people.com.cn/n/2014/0321/c150620-24705386.html>。

门别类的专业数据库,有效地推动了现代社会发展。但是,对于人类而言,真正需要的并不仅仅是大数据本身,而是大数据背后所隐藏的各种信息资源,而这些信息资源可能只占数据总体的少数^①。从这个角度说,大数据所蕴含的大量信息其实都缺乏实用价值,对人类科学研究、商业活动和决策咨询等并未发挥应有的功能。

不仅如此,许多大数据本身就存在模糊性,其中含有大量虚假和有害的内容,特别是在缺乏绝对隐私的大数据面前,各类信息的传播呈现出随意性、短暂性和盲目性等特征,各种色情、赌博、诈骗等不良信息接踵而至。如果我们纯粹凭借数据来判断和分析人类行为,甚至不假思索地去利用和传播某种数据,就极易导致许多严重后果。因此,未经前期调研论证和规范分析所获取的信息,其数据量规模越大、效果可能越模糊,得到的垃圾信息也就可能越多,这也为各种数据造假和不当传播打开了方便之门。对此,必须学会辨识和判断数据的真实性,避免因盲目的数据崇拜造成误判,甚至迷失在大数据的洪流之中。

三、大数据的价值复归：作为科技符号的一种社会文化转向

大数据作为认知与实践范式的系统转换的后果,已然成为实证主义方法论的新取向,并被建构为科学主义现代化的表征方式。尽管大数据为科学研究的范式转换提供了诸多可能性,但对于科学研究范式所带来的潜在革命性影响却包含了许多负面因素^②。大数据的精准分析和标准化归纳在尽可能地排除价值关联的同时,也夸大了纯粹理性的中立原则,从而使数据所关联的行为主体缺乏深度在场呈现。大数据理性的普遍价值追求和扁平化拓展功能为其所承载的信息难以从多维向度完整地诠释个体行动的主体意义和本真价值。然而,现代性在本质上规范了人作为主体具有的自觉意识和反思能力。可以说,大数据倡导的规则理性和人的自由自

觉构成现代社会内部一对基本矛盾。因此,大数据的挖掘和开发需要内化文化属性与社会意义,实现社会语境中大数据实践的现实价值。

从数据产生的机制来看,大数据作为文化外化的一种有形符号,是人类文明的重要表现,同时也是现代性对于“时-空”分离强化作用的一种后果。现代性作为“脱域”之后的理性化社会的主导性文化模式必然作为自觉的制度安排而构成社会运行的主导机理和图式^③。理性化的制度安排与抽象化场域设置为标准化、精确化、收敛化数据符号的诞生提供了外部条件。凭借数据的符号性(Symbolism)使意义系统与实在系统相互关联,并充当心智结构中信息交换媒介的角色,形成复杂社会结构网络。在实证主义侧重符号化系统形式语境的直接关怀下^④,数据的产生无外乎是对“物”符号代表的理性化建构,并通过与之关联的实在世界在多种社会语境中互相映衬,能够与其所隐喻的文本形成异质文化中的意义共振。换言之,任何数据符号都是通过文化编织而成的网络节点,数据的运算与解读都需要在具体的社会文化语境中完成。

从数据发展的方式来看,社会语境的复杂性赋予了大数据不断被挖掘、创新和循环使用的潜质。一方面,大数据既能够表征实在世界,但同时也限制着实在属性意义的符号转换能力。主要表现为,在认识复杂社会现象时,大数据往往基于机械决定论和还原论的价值预设,虽然能够将具有物理特征的社会事实普遍以量化符号的方式直观呈现出来,但不可否认的是大数据的表征能力始终与人类心智结构之间存在一定的时差,这也决定了其对于社会的某些隐晦事实往往难以完整表达。因此,只有不断推动大数据再生产与创新才能使数据持续充满活力。另一方面,社会系统内部的关联性助推数据在运动中循环使用。在现代社会系统中,子系统之间相对独立的关系已经被现代性危机所打破^⑤,经济、政治、文化等社会子系统相互嵌入在彼此内核之中。大数据所表征的符号意义通常处于各个子系

①有观点认为,大部分企业“大数据”并未产生实际作用,其中90%的数据都是无用数据和垃圾数据。有关内容请参见评述网:《大部分企业“大数据”并没有卵用,90%都是垃圾数据》, <http://www.idcps.com/news/20150706/85388.html>。

②这里的现代性危机主要指1929-1933年资本主义世界经济危机。经济危机爆发后,在全球主要资本主义国家中盛行的以马歇尔为代表的新古典学派自由放任经济学理论开始逐渐式微,而以约翰·梅纳德·凯恩斯为代表的“国家干预经济”学说成为全球资本主义各国化解经济危机的主要理论支撑。参见王晓升:《社会性的终结与现代性理论面临的主要挑战》,《南国学术》,2014年第2期。

统相互叠加之上，同一数据在不同领域均能发挥相应功能，数据的重复利用和循环使用已成为不可避免的常态。另外，不同于物质资源，大数据虚拟化的外显特征为其在形态上拓展、重组、开放和创新带来了优越条件，从而使数据在结构上灵活变换，并伴随高密度利用释放出更大的价值。

从数据价值的建构来看，大数据的根本价值在于其与世界关联所生成的意义构成系统。在社会语境的情景化状态下，对于大数据的理解必须与文化构成相契合，否则其极有可能转换成一种缺乏能动性的标识性记号（Sign）。文化生态毕竟是以文化为核心的系统，不能只让数据技术来定义未来文化经济的游戏规则，对数据的迅猛发展应保持警觉^[20]。社会语境结构性地引入政治、经济、民俗、自然环境等要素，在动态中具体化的现实要求必然不断地调整大数据的修辞方式和诠释路径以适应系统的平衡，实现抽象系统中规范性量化数据与描述性意义哲学有机统合。与此同时，在差异化的复杂社会系统中，符号意义的开放性赋予了大数据在时空连续体（Space-time Continuum）中发展变化的文化特质和道德意涵^①，在具体语境下意指方式的差别化、层次化，无疑对海量数据的处理提出了更高的要求。正如前文所述，无论是涉及人的大数据还是关乎物的大数据，都应该走出技术理性急剧膨胀的误区，避免落入以抽象运算取代解释性理解的陷阱，重视大数据的人本解读与文化复归。

四、比较与重构：复合方法的兴起与“大数据”的发展趋势

作为一项研究方法的技术创新，“大数据”研究方法填补了有限的经验材料支持无限的宏大理论论证之间的鸿沟，加快信息时代到来的步伐。但是，这并不意味着传统的抽样数据、访谈、实验、观察等研究方法的落后与淘汰。从本质上来看，任何一种研究方法都有侧重点，都是从有限的视角观察事物，这在一定程度上阻碍人类系统性的对世界复杂现象的发觉。因此，打破单一化研究方法几乎成为科学研究的一种共识和发展方向。同样，大数据本

身所无法逾越的缺陷迫使我们不得不反思其超越限度所引发的后果。从这个角度而言，大数据的出现对人文社会科学的实证基础带来了重大变化，甚至加速了定性与定量研究的大融合^[21]。而多学科研究方法在空间内相互融合、相互补充、相互渗透后的复合方法对于人文社会科学理念与思维方式的变革和人类认知系统的升华具有重要的现实意义^[22]。

首先，与大数据相比，抽样数据在某些状态下更加精确，更能准确描述事物的客观状态。作为现代数理统计学的重要支撑，抽样逻辑已经广泛地应用于自然科学、社会科学以及生产、销售等多个领域。抽样调查并非漫无边际地从总体中提取样本，而是要遵循“明确研究目的、概念操作化、设计问卷、确定总体、编制抽样框、决定抽样方案、实际抽取样本、评估样本质量”等比较严格和规范的程序，尽可能降低样本的误差^[23]。多样化的抽样类型从不同角度完善了人类认识世界的内容与手段，其灵活性、准确性、可测性、经济性等特点受到了各行各业普遍青睐。在抽样法则的作用下，所获取的抽样数据通常具有针对性，并能够根据需要限定在可控制的范围内，可以较为有效地支撑相关研究。特别是经过严格的现代数理统计程序处理后的抽样数据，不仅可以比较科学地分析变量之间的关系，而且从代表样本推导出的基本结论几乎与总体状况相吻合。统计学家已经证明：样本分析的精确性会随着采样随机性的增加而提高，而与样本数据的增加关系较小。也就是说，样本的随机性要比样本的数量更加重要^[2]。相反，尽管大数据所含信息量大、资源丰富，但是未经关联与整合的大数据所含信息往往较为杂乱、无法实现精确性，甚至有些繁冗的数据还会对人类的认知起到妨碍作用。对于具体研究而言，在某些特定场域，抽样数据会更加容易把握和控制，更加便于开展相关研究。因此，打破抽样数据与大数据之间的鸿沟，实现抽样数据与大数据之间的彼此交融、相互补充、相互印证，既有利于控制单个研究变量、拓展研究范围，也有利于防止大数据带来的资源泛化和信息虚假现象。

其次，与大数据相比，观察与访谈更能深入理解和探究主体行为及其生活环境。长期以来，观察与访谈型构了定性研究方法多样式、多元化、多层

①社会语境对于大数据的解读和理解至关重要。一方面，相同的数据符号在不同的文化模式中所表现的意义指涉可能不相同，甚至还有可能出现完全相反的诠释路径；另一方面，不同的数据符号在同一文化模式中又有可能表现出相同的意义。

次的架构形态^①，同时也合宜了社会环境的复杂现实。不同于数据，观察与访谈法的一个很明显的优势在于从封闭性量化资料转向开放性编码资料，以人与人互为主体性的经验感知和设身处地的观察与理解克服了数据资料的表面化、简单化、横向化缺陷，具有较高的测量效度。特别是对于人类情感、情绪等难以量化为可控指标的变量，具有标准化的数符号不可避免地暴露出浅层描述的固有缺陷，难以凭借“解释性的理解”生动地呈现研究对象的深层思维与行动逻辑。当然，观察与访谈法一般适用于小规模群体的纵向挖掘，其深入、历时、理解等特征对于探究社会发展变化大有裨益。可以说，观察与访谈法与大数据形成了两种截然不同的分析路径。因此，在大数据日渐流行的今天，我们更应该反思肤浅化的数据与回归“人”的本真及其符号意义之间的内在张力。从社会学的角度来看，在海量的平面化数据面前，借助观察与访谈等方法深度探解数据符号潜在的内容具有重要意义。

最后，与大数据相比，实验法更容易把握变量之间的因果关系。实验研究最早起源于自然科学，后逐渐拓展到社会科学领域，并在心理学、教育学等学科研究中受到广泛推崇。与抽样数据不同，实验法基于实证主义研究范式，借理想函数（Ideal Function）自变量与因变量之间的平衡关系，在高度控制某些条件下，以研究对象在场（Presence）的动态监测揭示出二元或多元变量之间的因果关系。自社会学家涂尔干指出所有社会现象都合乎因果关系的法则之后^[24]，实证主义社会学形成了比较统一的分析路径。在以后的研究中，探寻社会现象的因果性基本成为社会科学的终极目标。然而，在大数据时代，我们无需探究事物现象之间的必然联系。一方面，由于海量数据不便于控制，难以寻求个体数据之间的必然关系，这就让概括因果关系变得十分困难；另一方面，在大数据时代，很多学者认为明确事物之间的“相关性”在某种程度上比“因果性”更加重要。大数据致力于实现从因果性到相关性的

历史性跨越^②，其“去因果化”直接挑战了传统唯理论和经验论共同的哲学基础。然而，尽管历史上不少哲学家也曾对“因果关系”发出过种种质疑和猜测，并试图解构科学认识论对于因果关系追求的偏执，但是最终仍然没有否定“因果性”对于认知世界的普遍意义。因此，如果能够以大数据普查资料为基础，借鉴实验法的“控-变”原理，降低实验控制存在的干扰，深入探测变量关系之间的因果关系，将有利于进一步提升人类认识世界的广度和深度。

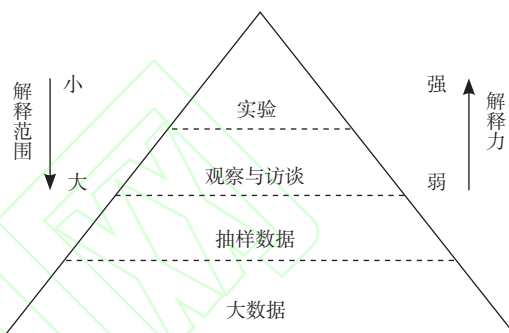


图1 研究方法比较示意图

综合图1可以看出，无论是实验、观察与访谈、抽样数据还是大数据，都始终难以突破“解释范围-解释力”这一张力的困境，在精确化与整体化之间存在着天然的对峙^[25]。即解释力（精确化）较强的研究方法常常是以牺牲解释范围（整体性）为代价的，而解释范围较大的研究方法又难以兼顾解释的精确性。这种相互的矛盾性镶嵌于每一种研究方法之中。同样，大数据蕴含的海量的经验信息在本质上难以形成较为深入的解释类型。而抽样数据、观察与访谈和实验却能有效补充大数据在数据信度、主体呈现和因果解释等方面存在的缺陷。因此，研究方法的相互融合和复合方法兴起有助于人类形成系统认知方式。

五、总结与讨论

当前，大数据浪潮正在以惊人的速度席卷全球

①目前，国内外学术界对于观察法的类型划分仍然没有取得一致的看法，比较流行的看法是依据观察者“是否参与背景的活动”，可以分为四种角色类型，即作为观察者的参与者、作为参与者的观察者、完全的参与者和完全的观察者。有关内容请参见：[美] 艾尔·巴比：《社会研究方法》（第11版），邱泽奇译，北京：华夏出版社，2009年版。

②需要指出的是，“因果性”是“相关性”的一种特殊形式。相关关系旨在探究两变量的关联程度，但不足以说明引起变量与被引起变量的辩证逻辑关系。而因果关系则是在确定两变量相互关联的基础上，进一步确定自变量与因变量之间的动态平衡，存在因果关系的两个变量通常能够建立数学模型。

各地。作为一种科学图景,大数据以全景式数据规律弥补了以往因果规律的不足,从而拓展了人类对于科学规律的认识^[1],为人类视野开辟了新的世界。与此同时,在信息化时代,大数据凭借其自身的优势为现代社会留下了“数据万能”的心理烙印,数据似乎成为经验表征唯一可信赖的符号和科学研究合法性的必然象征。显然,这种对于数据的盲目崇拜和数据效应的无限放大极为容易导致诸如“数据霸权”“数据至上”“数据垃圾”“数据安全”“数据区隔”等现实问题。对于这些潜在风险,应当有针对性地建立起大数据风险防范、监控预警等机制,提高人类数据筛选、处理能力。

对于大数据而言,其不仅面临着功能发挥的外部风险,而且面临着系统内部构成之间的冲突风险。伴随数据内涵与外延的扁平化扩张,使得数据不再仅仅指代“有根据的数字”,而且还可以指存储在网络中多种多样的信息,包括文本、图片、视频等,数据几乎成为“信息”的代名词^[26]。现实中,多元化的内容构成让大数据呈现出错综复杂的事实意向,甚至大数据不同形式的信息载体呈现出信息相左的现实而使数据结构紧张。从根本上看,大数据是在后现代主义的庇佑中逐渐发展起来的。大数据对现代性规则意义进行“反叛”的同时,也极力适应后现代“去中心”“去权威”“去逻辑”“去结构”等特征。当然,后现代社会的信息技术革命既能在网络真空内体现出“自由、平等”的幻觉,也容易导致数据空洞、缺乏主体和意义构成。基于以上反思,笔者认为,大数据不能简单地作为科技符号承载信息内容,而是更应该将社会文化内涵融入到数据之中,实现科技符号的社会文化转向。因为任何数据的产生、发展和评价都是在社会文化系统内完成的,只有在社会语境中重拾数据的文化属性和社会意义,才能避免大数据沦为脱离社会主体的孤立符号。

总之,大数据既是一种生活方式,也是一种范式转换的技术革新,更是一种理论和研究方法变革,其基础和核心是服务人类对于世界本真的探索与认识。大数据将人类纳入到了统一的网络虚拟世界中,并重构人类社会关系网络^[27]。其引发的信息革命在日益压缩全球“时-空”距离的同时,也在拓展人类的思维架构和认知领域,催生了大量的新兴学科、产业形态和认知模式,向传统认识论提出了多种多样的颠覆性挑战。然而,单向化的认识论往往形成片面化视角的经验图式和认知惯习,大数据在本质上所难以突破的缺陷不断地拷问人类对于突破

局限认知世界的可能性。对于复杂系统而言,语境、数据、虚拟、计算等都将无一例外纳入到复杂性科学方法体系之中^[28]。从这一意义上说,大数据时代的到来,并没有对传统研究方法形成巨大冲击、甚至到了可以替代传统研究方法的地步。相反,如果能将大数据同传统的抽样数据、观察与访谈、实验等方法有机结合,探索形成“大数据+抽样数据+观察与访谈+实验”复合方法和立体式的观察视角,这无疑会为人类认知世界开启另一扇大门。

参考文献:

- [1] 张晓强, 杨君游, 曾国屏. 大数据方法: 科学方法的变革与哲学思考[J]. 哲学动态, 2014, (8).
- [2] [英] 维克托·迈尔-舍恩伯格, 肯尼思·库克耶. 大数据时代——生活、工作与思维大变革[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.
- [3] 黄欣荣. 大数据对科学认识论的发展[J]. 自然辩证法研究, 2014, (9).
- [4] 陈德权, 林海波. 大数据的文化解构与演进路径研究[J]. 电子政务, 2016, (12).
- [5] 2016 年全球大数据产业规模及细分市场现状[EB/OL]. 中国产业信息网, <http://www.chyxx.com/industry/201610/455960.html>.
- [6] T. Hey, S. Tansley, K. Tolle. The Fourth Paradise: Data-Intensive Scientific Discovery[J]. Microsoft Research, 2009.
- [7] 2016 年中国网民数量达 7.31 亿 相当于欧洲人口总量[EB/OL]. 凤凰网, http://news.ifeng.com/a/20170122/50614799_0.shtml.
- [8] 郝岚. 大数据与世界文学教学[J]. 中国比较文学, 2016, (1).
- [9] 被收入牛津词典的中国词汇, 你知道? [EB/OL]. 搜狐网, <http://mt.sohu.com/20160616/n454705577.shtml>.
- [10] 习近平. 建全国一体化的大数据中心[N]. 大众日报, 2016-10-10.
- [11] 涂子沛. 大数据: 正在到来的数据革命, 以及他如何改变政府、商业与我们的生活[M]. 桂林: 广西师范大学出版社, 2015.
- [12] 黄欣荣. 大数据技术的伦理反思[J]. 新疆师范大学学报, 2015, (3).
- [13] 文军. 大数据热引发的冷思考[N]. 人民日报, 2016-10-20.
- [14] 衣俊卿. 现代性的维度及其当代命运[J]. 中国社会科学, 2004, (4).
- [15] 马翀炜. 文化符号的建构与解读——关于哈尼族民俗旅游开发的人类学考察[J]. 民族研究, 2006, (5).
- [16] [英] 安东尼·吉登斯. 现代性的后果[M]. 田禾, 译. 南京: 译林出版社, 2011.
- [17] 黄少华. 评贝塔朗菲的系统认识论[J]. 兰州大学学报, 1995, (3).
- [18] 方环非. 大数据: 历史、范式与认识伦理[J]. 浙江社会科学, 2015, (9).

- [19]郭贵春.“语境”研究纲领与科学哲学的发展[J].中国社会科学, 2006,(5).
- [20][加]林青.大数据应用与文化发展趋势[J].江西社会科学, 2014,(3).
- [21]King, G. The Changing Evidence Base of Social Science Research[C]//In G. King, K. Lehman Schlozman & N. Nie (eds.), The future of political science : 100 Perspectives. New York : Routledge, 2009.
- [22]罗伟, 罗教讲.新计算社会学: 大数据时代的社会学研究[J].社会学研究, 2015,(3).
- [23][美]艾尔·巴比.社会研究方法[M].邱泽奇, 译.北京:华夏出版社, 2009.
- [24][法]埃米尔·涂尔干.社会学方法的规则[M].胡伟, 译.北京:华夏出版社, 1999.
- [25]夏国美.大数据时代的社会学审视[J].新疆师范大学学报, 2016,(1).
- [26]涂子沛.中国如何应对大数据时代的挑战[N].南方日报, 2013-01-07.
- [27]沈浩, 黄晓兰.大数据助力社会科学研究: 挑战与创新[J].现代传播, 2013,(8).
- [28]王英.从后现代主义、后工业社会到大数据时代[J].新疆师范大学学报, 2017,(3).

Retrieve the Lost Tradition: Reflections and Reconstruction of “Big Data” Research Paradigm

WEN Jun WU Xiao-kai

(The Center for Modern Chinese City Studies and School of Social Development,
East China Normal University, Shanghai 200241)

Abstract : As an emerging technology, big data has become an indispensable part of modern society by fundamentally changing the traditional mode of scientific research and creating convenient technical support for business activities, government decision-making and various aspects of social life. However, there is a deep crisis of "data worship" hidden in the increasing prosperity of big data. To a extent, the instrumental rationality rooted in the application of big data is inhibiting the expression of human emotions. From the perspective of social context, this paper attempts to reconstruct the cultural and social significance of the big data as symbol from such aspects as the mechanism of data generation, the modes of innovation and the construction of value so as to help us adapt to the self consciousness and self reflection of modern people. At the same time, the paper points out that the "big data" research is of great practical significance to the elevation of human cognition when it is wisely combined with sampling investigation, observation and interviews, laboratory experiment and so on.

Key words: Big Data; Reflection on Paradigm; Humanistic Attributes; Social Significance; Integration of Methods

[责任编辑: 马瑞雪]

[责任校对: 李 蕾]