大数据研究中的两个流派及两类大数据——基于案例的研究

薛永红[1],董春雨[2]

1. 华北科技学院理学院,三河 065201
2. 北京师范大学哲学学院,北京 100875

摘要　关于大数据的研究,学界已经形成了泾渭分明且针锋相对的两个大数据流派——激进派与保守派。通过对2个经典大数据案例的研究,发现“大数据”实际上指称两类既有区别又有联系的对象,一类是“用数据的方法研究科学”,另一类是“用科学的方法研究数据”。两类大数据及二者存在的显著差异,是形成激进派与保守派两种阵营的原因。在归纳了两类大数据各自特点的基础上,提出了从根本上消除目前这种对立且混乱的认识现状,并将大数据研究推向深水区的路径。

关键词　大数据;激进派;保守派;谷歌流感预测;人类数感

大数据研究方兴未艾。分析大数据认识论研究中的各种观点,虽然错综复杂,但是在整体上又呈现出泾渭分明的两个流派——激进派与保守派。通过研究发现,这种现象的产生,并不是源于对同一对象的不同认识,而是因为“大数据”这一概念所指称的对象本身就存在差别。美国科学哲学家汉弗莱斯(P. Humphreys)[1]在《Big data: Thick mediation and representational opacity》(《大数据:深调制与不透明表征》)中就指出,客观上存在两种大数据,一种是大写的大数据(BIG DATA),另一种是小写的大数据(big data)。小写的“大数据”指的是与数据科学相关的活动和方法,是拥有海量数据的组织机构所面临的技术问题。而当这些活动、方法尤其关于处理海量数据的技术向社会各领域渗透并迅速发展时,便产生了大写的大数据。这种区分虽然非常有价值,但是由于更多的是侧重于文化演进的宏观层面,因而缺乏更为直观和具体的视角,尤其是认识方法层面。本文通过对2个典型的大数据案例的分析,发现所谓的“大数据”实际上指称2种与数据有关的研究方式,或者可以认为存在2种

收稿日期:2020-06-09;修回日期:2020-09-16

基金项目:教育部人文社会科学研究青年基金项目(20YJC720025);国家社会科学基金重点项目(18AZX008)

作者简介:薛永红,副教授,研究方向为科学哲学、科学文化传播,电子信箱:aristotle@ncist.edu.cn

引用格式:薛永红, 董春雨. 大数据研究中的两个流派及两类大数据——基于案例的研究[J]. 科技导报, 2021, 39(13): 125-133; doi: 10.3981/j.issn.1000-7857.2021.13.014

大数据的研究路径:一种是用数据的方法研究科学(to study science in a data way,SSD),另一种是用科学的方法研究数据(to study data in a scientific way,SDS)[2]。前者是随着科学研究中海量数据的出现而产生的,目前已经形成了生物信息学、天体信息学、数字地球等研究领域;后者是数据科学和技术的发展与应用,包括统计学、数据挖掘、数据库以及机器学习等领域。这种区分也为汉弗莱斯的两类"大数据"思想提供了科学依据。正是两类"大数据"的存在及其二者之间的差异,才是形成激进派与保守派两个大数据阵营的原因。

1 大数据研究中的两个流派

1.1 大数据激进派

激进派的代表人物是舍恩伯格(Viktor Mayer-Schnberger)。他概括了大数据在认识论上引发的3种变革:"更多"(全体优于部分)、"更杂"(杂多优于单一)、"更好"(相关优于因果)。他认为,有了大数据,科学家不再需要进行有根据的猜测来构造假设和模型,并用基于数据的实验和实例来测试它们。相反,他们可以挖掘完整的数据集,以揭示效果,从而产生科学结论,并且也不需要进一步的实验验证[3]。克拉克(L. Clark)[4]认为,人们可以从实验数据中"蒸馏"自由形式的自然规律,即借助于一些自动化工具,"无论问题的复杂性如何,在没有问题的情况下,程序可以自动地发现洞察,就像一种'数字意外'"。对于模型在研究中的作用,斯普林格(J. Sprengery)[5]认为:"科学和统计推理主要建立在明确的参数模型之上,往往有很好的理由。然而模型的有限性和现代科学研究系统的日益复杂性增加了错误定位的风险。因此,基于大数据的推理技术是一种可行的替代模式。"此外,在认识论道路上走的最远的是安德森(C. Anderson)[6],他宣称,由于技术可以捕捉到关于对象的任何数据,而对这些数据的分析又能产生非常准确的结果。因此,传统的抽样调查(小数据)的方法将被彻底淘汰。在这种研究方式下,只需要关心相关关系,不再需要去探究现象背后的机制和原因。可以说,有了大数据之后可以不需要科学或模型,理论将被大数据研究终结。克拉克[4]说得更加直白:"它(大数据)的意图是完全删除进入数据挖掘的人类因素以及所有人类的偏见,而不是等待被问到一个问题或被引导到特定的现有数据链接,系统将提供给人类自己可能没有想到的要寻找的模式。"舍恩伯格[3]对此也持类似的观点,他认为,大数据的相关分析方法更准确、更快,而且不受偏见影响……因为它不受限于传统思维模式和特定领域里隐含的固有偏见,大数据才能为我们提供更多的新视野。

总之,大数据激进派的基本观点可以概括为:数据可以客观地表征世界;只要数据量足够大,就不需要模型、问题及相关的理论,只要在数据的驱动下,数据可以自己发声;相关性是世界的本质;由于大数据可以完全避免人类的主观因素进入科学研究,大数据知识发现的模式更客观、更自由。

1.2 大数据保守派

大数据保守派一方面承认大数据的独特性,另一方面对大数据是否能客观反映实在等保持理性的怀疑态度,并且通过案例和证据,对激进派的各种论调一一进行反驳。这一派的代表人物冯启思(Kaiser Fung)[7]认为:"大数据所谓的'N=all'全样本承诺只是一种理想,而不会是现实。"信息哲学家弗洛里迪(L. Floridi)[8]认为大数据时代真正的认识论问题是如何寻找数据中的"小模式"。"小模式"代表的是新竞争领域:从科学到商业、从治理到社会政策。他警告说:"大数据有风险,因为它们改动了可预测的范畴和界限"。虽然大数据在寻求相关性方面具备其他方法难以企及的优势,但是大数据改变了科学研究的方式,使我们在很多情况下能像谷歌(Google)一般,很容易获得关于事物之间的内在联系。但是若不借助模型和因果机制,人类对事物的理解根本达不到谷歌这样的级别。此外,更为重要的是,也绝不会有人愿意将自己对事物的理解水平停留在这个层次,并乐此不疲[9]。蒂莫(J. Timmer)[10]则认为,对相关关系的研究只是为了引起科学家的注意,相关关系通常会很吸引人,因为它可能产生有效的预测,但是模型和相互作用机制所能做的不仅是实现准确预测,最为关键的价值在于,

它可以推动科学发展和应用。布鲁克斯(D. Brooks)[11]认为,基于大数据的很多相关性研究只是“白噪声”,因为对海量数据的分析必然会制造出更多、更大的“干草垛”(相关关系),而其中必然存在很多伪相关甚至虚假相关,它们的数量也会随着数据量的增加呈现指数式的增长。其结果是,人们要想从这个巨大的“干草垛”里找到关于事物本质的联系将变得非常困难。

保守派极力反对所谓的数据驱动和理论自由。克勒曼(W. T. Coleman)曾指出,除非人们对自己正在思考的事情创造出一个模型,否则在面对大数据时也不能提出任何问题,而且当在问问题的时候,已经必然存在了某种偏见。尽管大数据可以力求详尽无遗,捕捉整个领域并能透视每一个角落,但它既是一种表征,也是一种样本。数据由技术和平台所产生,由使用和监管的环境所决定,并且受到抽样偏差的影响。事实上,如同克劳德(K. Crawford)[12]所言:“所有数据都提供了关于世界的独断性的观点,它在使用特定工具的某些优势,因而不是一个看不见的、绝对可靠的上帝之眼。”因此,在这一派看来,数据并不是简单地以中立和客观的方式从世界中抽象出来的自然的和必要的元素,数据是在一个复杂的组合中产生的,它在主动地塑造着自身。

当然,这一派也坚决反对大数据将会使理论终结的观点,因为识别数据中的模式的策略不会发生在科学的真空中,它受限于先前的发现、理论和训练,或者说这种猜测是以先前的经验和知识为基础的[13]。卡纳利(S. Canali)[14]和皮耶奇(W. Pietsch)[15]等通过案例研究,不但发现了大数据可以寻求因果:“认为因果知识对数据驱动是多余的看法是有缺陷的,因果知识应该被认为是大数据科学的必要元素”,而且还发现大数据所使用的算法作为“关于研究问题的框架,在外部意义上是负载着理论的。”因果知识不仅对大数据研究的投入很重要,对实现项目目标同样重要,因为在没有理论的情况下,数据也不会被捕捉和汇总,如果没有理论,数据也就没有意义[16]。在保守派看来,这种不需要模型与理论的论调实质上是混淆了基础理论和现象建模的关系。科学不仅仅是用来产生一个简单机械的模型来寻找各种相关性预测,相反,它的目标是使用那些从数据中抽取的规律,建构一个统一的方法来合理地理解它们[17]。

对于大数据方法与经典方法的关系,贝瑞(D. Berry)[18]认为,大数据中存在着一种傲慢的倾向——其他分析方法太容易靠边站了。传统方法在大数据面前的缺席,说明这是一个不欢迎旧有智能“工艺”的体系;而且由大数据所提供的知识和信息也缺乏哲学的调节能力,即缺乏康德哲学所追求的那种知识的理性基础。类似地,剑桥大学的斯皮格汉特(D. Spingelhalter)[19]也认为,在大数据研究中存在着许多小数据的问题,它们不但不会消失,而且还会随着数据量的增加变得越来越突出……,统计学的难题并没有因为大数据及其相关技术的出现而得到解决,如对因果关系的理解、对未来的预测以及如何对一个系统进行干预和优化。因此,尽管大数据和相关的新分析方法快速地增长,但小数据依旧是研究事物的重要组成部分。而且在将来,不会出现范式的转变,即大数据研究取代小数据的研究,小数据和大数据将互相补充[20]。

通过上述对2个大数据流派的梳理,可以将大数据认识论中的关键问题大致归纳为:大数据与现象之间的关系——主要是现象如何被大数据表征的问题;大数据与理论(模型)的关系,可归结为大数据到底是数据驱动还是理论驱动的;相关性与因果性的关系;大数据是否是理论自由的;大数据方法与经典方法(小数据方法)的关系。以下通过2个经典的大数据案例来分析二者的分歧及其本质。

2　案例研究

2.1　人类数感(number sense)研究

心理学研究表明,人们对物体或事件数量存在一种非言语的表征方式,这种表征区别于通过言语或数字符号对数量的精确表征,具有近似性和不精确性[21]。心理学家将这种表征系统称为近似数量系统(approximate number system,ANS)。ANS被认为是一种与生俱来的结构,是人类数感和形成数学

能力的基础,并且在理论上服从韦伯定律。此外,无论人还是动物都有ANS,它不仅体现在视觉任务中,也能体现在听觉任务中。脑科学研究表明,脑区双侧的顶内沟处大致为ANS系统所处的位置。目前,心理学对该领域的研究成果被广泛应用到教育教学实践中。但这一领域的研究一直以来缺乏对ANS在整个生命周期内的研究,因为实践中很难对每一个样本进行终生的追踪研究。大数据的出现为心理学家研究这一问题提供了可能。约翰霍布斯大学的心理学家哈尔伯达(J. Halberda)[22]的研究方案是,通过已有的ANS理论,构造测试模型,然后向全球征求志愿者,在线完成测试任务。结果在短短的几个月时间里,便收集到了分布在全球不同地区的13000名年龄在11~85岁的测试者。测试过程中,志愿者除了被要求完成基本的测试外(图1)[22],还要回答年龄、数学能力等问题。通过科学的数据筛选之后,对数据做相关性分析,得到研究结论如表1[22]所示。

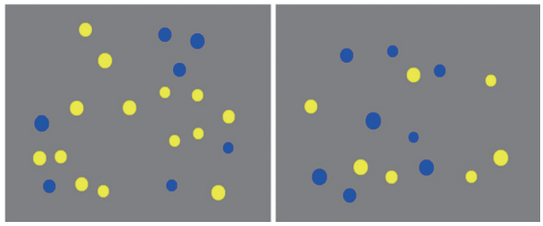


图1 ANS点测试的试验(判断两种颜色的点的数量关系)

2个反应ANS精确度的量分别为韦伯系数(*W*)和响应时间(*RT*)。韦伯系数(*W*)越小,ANS的敏锐度越高;响应时间越短,ANS的精确度越高。

在整体上,这个研究测量了超过10000多名被试的ANS的精确度,并用可视化的方法表现了这个核心认知系统的变化。通过对这些数据的分析,哈尔伯达不但完成了对人类数量感知力发展的整体描述、验证了前期对于不同年龄阶段ANS与数学水平之间的理论假设(如ANS与数学能力之间呈正相关等),填补了这一领域的研究空白,而且还发现了之前没有发现的"意外"规律。

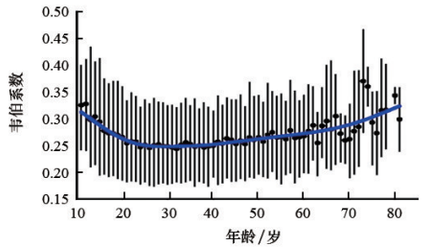
表1 自我报告的学校数学能力与ANS精度的相关性

年龄段	人数	*W*	*RT*	*W*和*RT*
所有年龄段	10584	$-0.19/0.72/7.3\times10^{-83}$	$-0.09/0.98/3.1\times10^{-21}$	$-0.22/0.77/1.5\times10^{-111}$
11–17 y (1st decile)	994	$-0.13/0.73/3.5\times10^{-5}$	$-0.11/0.98/5.4\times10^{-4}$	$-0.19/0.74/1.9\times10^{-8}$
18–20 y (2nd decile)	1267	$-0.21/0.74/5.4\times10^{-14}$	$-0.04/0.98/1.1\times10^{-1}$	$-0.22/0.75/7.0\times10^{-15}$
21–22 y (3rd decile)	919	$-0.19/0.70/5.5\times10^{-9}$	$-0.09/0.98/7.0\times10^{-3}$	$-0.23/0.73/3.3\times10^{-11}$
23–24 y (4th decile)	1017	$-0.19/0.72/6.4\times10^{-10}$	$-0.06/0.98/5.8\times10^{-2}$	$-0.21/0.75/8.7\times10^{-11}$
25–26 y (5th decile)	1013	$-0.23/0.70/5.6\times10^{-14}$	$-0.08/0.98/7.1\times10^{-3}$	$-0.26/0.74/1.7\times10^{-16}$
27–28 y (6th decile)	868	$-0.17/0.71/9.2\times10^{-7}$	$-0.05/0.98/1.4\times10^{-1}$	$-0.18/0.76/5.7\times10^{-7}$
29–32 y (7th decile)	1310	$-0.19/0.69/1.4\times10^{-12}$	$-0.04/0.98/1.8\times10^{-1}$	$-0.20/0.76/1.6\times10^{-12}$
33–37 y (8th decile)	1066	$-0.14/0.69/2.8\times10^{-6}$	$-0.05/0.98/7.4\times10^{-2}$	$-0.15/0.77/2.9\times10^{-6}$
38–44 y (9th decile)	991	$-0.20/0.70/2.2\times10^{-10}$	$-0.07/0.98/3.6\times10^{-2}$	$-0.21/0.78/7.8\times10^{-11}$
45–85 y (10th decile)	1103	$-0.20/0.69/5.8\times10^{-11}$	$-0.11/0.99/2.4\times10^{-4}$	$-0.23/0.79/2.0\times10^{-13}$

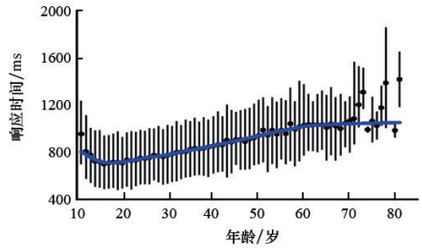
例如,图2(a)显示,拟合曲线存在极小值,这个最小值处在30岁左右。年龄小的时候*W*值高说明在学龄期通过教育可以逐渐提高ANS的精确度,而到约30岁时达到最佳精确度。图2(b)显示,响应时间也存在极小值,说明通过教育可以缩短反应时间,但是最佳时间在15岁左右;在整个生命周期内的大量重叠的黑线表明,即使在发育改善之后,*W*和*RT*的个体差异仍然很大。

2.2 谷歌流感预测(GFT)

流感尤其是季节性流感是人类社会长期面临的一个世界性的威胁和问题。据统计,全球每年约有250000~500000人死于季节性流感。早期发现流感疫情,然后迅速做出反应,可以有效控制疫情的发展、降低病亡人数。美国疾病控制和预防中心(CDC)、欧洲流感监测计划(EISS)所使用的流感预测系统,都是依据病毒学理论(如病毒的致病原理、



(a)　韦伯系数随年龄的变化关系



(b)　响应时间随年龄的变化关系

图2　ANS随年龄的变化关系

病毒的传播与进化等),使用临床监测数据(确诊的流感病人、疑似流感病人的就诊数据等),对流感进行预测,并向公众发布预测报告,但预测报告通常会滞后1~2周。

研究人员发现,在某一地区,某些词的互联网搜索频率与流感样疾病(influenza-like illness,ILI)病例的就诊比率高度相关。2008年,谷歌便建立了一种通过分析大量谷歌搜索查询来跟踪、预测流感疫情的系统。这一系统是以一些通过自动化方法来研究搜索查询与流感关系的已有研究为基础,通过5年的谷歌搜索日志中的数10亿次个人搜索数据为训练样本而完成的[23]。此前,已经有用类似方法对流感预测的研究文献和数据积累,例如:电话分流咨询热线的呼叫量和非处方药销售与流感的研究、网络搜索查询与疾病控制的研究、基于对美国健康网站搜索数据的跟踪与流感关系的综合监测研究、基于加拿大相关网页访问的日志分析与流感监测研究、雅虎搜索查询中"流感"字样搜索与多年来的病毒学和死亡率监测数据相关的报告以及每年约9000万美国成年人在网上搜索有关特定疾病或医疗问题的数据集[23]。

在谷歌的预测模模型中,自变量为同一地区与流感样疾病相关的检索词的检索频率。对ILI就诊比例与ILI检索频率取对数,然后做线性拟合从而生成预测模型见式(1),线性拟合如图3[23]所示。

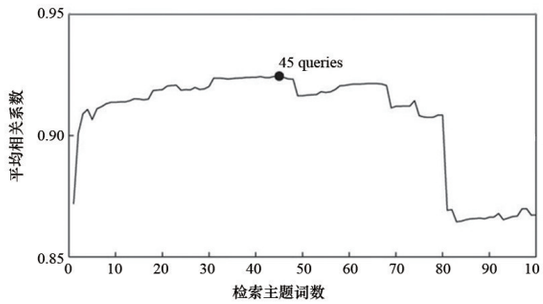


图3　线性拟合图

$$\mathrm{lg\,it}\left(I(t)\right)=\alpha\,\mathrm{lg\,it}\left(Q(t)\right)+\varepsilon \tag{1}$$

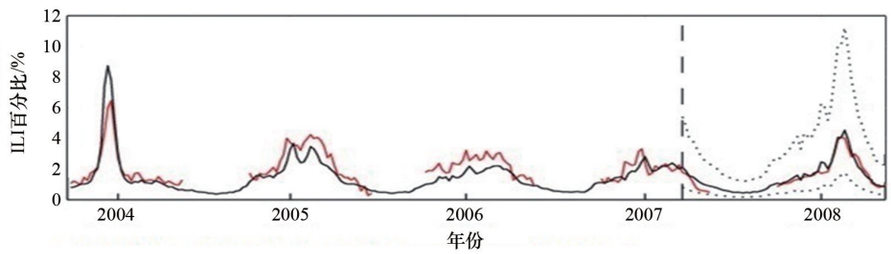
以美国CDC流感监测网络数据作为模型的因变量,对谷歌5000万个常用检索词分别进行拟合,并根据拟合效果评分,然后对评分由高到低排序,结果如表2[23]所示。

表2　检索主题词的评分统计

检索主题词	前45个词		后55个词	
	n	加权	*n*	加权
Influenza complication	11	18.15	5	3.40
Cold/flu remedy	8	5.05	6	5.03
General influenza symptoms	5	2.60	1	0.07
Term for influenza	4	3.74	6	0.30
Specific influenza symptom	4	2.54	6	3.74
Symptoms of an influenza complication	4	2.21	2	0.92
Antibiotic medication	3	6.23	3	3.17
General influenza remedies	2	0.18	1	0.32
Symptoms of a related disease	2	1.66	2	0.77
Antiviral medication	1	0.39	1	0.74
Related disease	1	6.66	3	3.77
Unrelated to influenza	0	0.00	19	28.37
Total	45	49.40	55	50.60

通过与CDC的监测数据对比,确定当n=45(即前45个检索词)时,预测结果与检测结果高度相似,如图4[23]所示。

将这45个检索词作为监测对象来预测ILI的趋势,将预测结果与CDC的结果相比较(图5[23]),发现对2008年各季度预测的结果与美国CDC的监测结果的相关系数达到0.97。最为关键的是,由于可快速处理搜索查询,谷歌的模型产生的ILI估计值始终比疾病预防控制中心ILI监测报告提前1~2周。



红线为CDC报告结果;黑线为谷歌预测结果;虚线为95%的预测区间。

图4 大西洋区预测结果与CDC监测结果比较

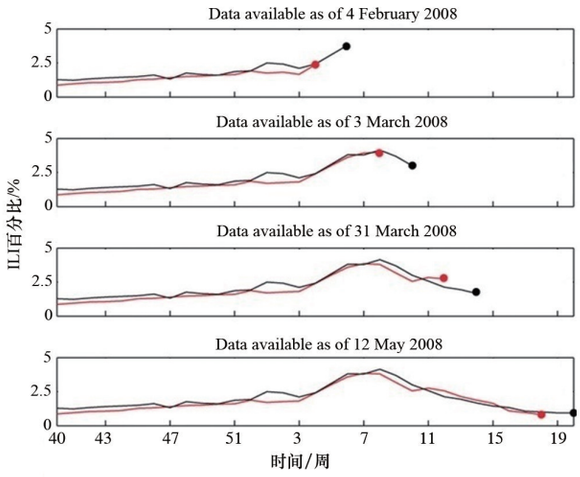


图5 大西洋区2007—2008年流感季的预测对比

3 两类“大数据”

对以上两个案例的共同点和不同点做区分,结果如表3所示。

两个大数据案例恰好反映了两个大数据流派对大数据的不同认识。比如当研究者基于“案例一”进行分析时,必然会得出诸如大数据研究离不开模型、问题驱动、相关性不能代替因果性等结论,而影响人类“数感”的机制是什么仍旧悬而未决;如果以“案例二”为依据时则可以得出,大数据不需要具体问题,只需要相关性就足够了,因为其研究目的就是为了预测和控制。

表3 两种大数据案例特征对比

	相同点	不同点
案例一 人类“数感”研究	数据量大 传统方法无法完成处理 使用新算法 具有一定的自动化程度 使用人类已有知识	使用本学科的基本理论和模型 有具体问题 问题驱动 追求相关性,但引发了对因果性追求 使用人类已有知识理解和解释结论 目的在于拓展研究宽度,揭示更多未知规律,识别因果结构 应用领域为科学
案例二 谷歌流感预测	数据量大 传统方法无法完成处理 使用新算法 具有一定的自动化程度 使用人类已有知识	使用数学模型 无具体问题 数据和算法驱动 追求相关性 不使用人类已有知识理解和解释 目的在于预测与控制 应用领域为社会

对大数据的认识之所以存在两种针锋相对的观点,是因为“大数据”所指称的对象并不同一。也就是说,客观上存在着两类“大数据”,或者可以认为有两种大数据研究路径:一种是用数据的方法研究科学,如案例一;另一种是用科学的方法研究数据,如案例二。两类大数据从“从属的关系上讲,都属于数据科学”[2]。

1)用数据的方法研究科学。它的研究对象首先是一个科学问题,采用的是将科学对象数据化,然后运用列表、排序、回归、聚类、计算机模拟等方法,研究数据背后的规律。由于技术的不断进步,使“万物皆可数据化”成为可能,从而使这一方法迅速在科学研究中兴起。历史地看,开普勒对行星运行轨迹的研究就是一个SSD的经典案例,只不过当时的数据完全可以用手工处理。而到大数据背景下,数据量大到无法用传统方法处理,需要大数据相关技术的介入。深圳华大基因研究院、生物信息系统国家工程研究中心及中国科学院北京基因研究所的科学家于2007年完成的中国人基因图谱绘制工程,就是典型的案例。

2)用科学的方法研究数据。它的研究对象首先是海量数据,采用的方法是关于数据收集、处理和分析的方法,包括在计算机和数据库出现以来所形成的数据库、数据挖掘以及机器学习等方法。这种方法在互联网时代迅速崛起,尤其是在机器学习技术获得突破性进展以来,得到了快速的发展。AlphaGo的成功是这一方法的经典案例。

3)虽然两类大数据有所区别,但随着二者的不断融合,他们之间的界限越来越模糊。SDS的发展,一方面是来源于互联网技术的发展,但其所使用的方法、模型很多都是来自于SSD的研究成果,因为科学家在处理基于模型而产生的海量数据问题时,所创造的个性化的行之有效的方法和算法,被逐渐地扩展和一般化,从而成为一般的数据研究方法[24]。正如汉弗莱斯所指出的,当小写的大数据向社会各领域渗透并迅速发展时,便产生了大写的大数据[1]。

结合以上的具体分析,以下3点结论是可以接受的。

第一,SDS虽然也使用模型,但是模型在其中的重要性完全低于数据本身。因为大量的数据可以消除模型不精确带来的偏差(最简单的方法是进行参数调整)。所以激进派宣称大数据不需要模型,虽然比较极端,但也不无道理。

第二,SSD是以数据之间的相关性为出发点的,但终极目标在于识别因果结构[25]。因为对于复杂现象,虽然通过参数变化映射因果结构与科学本身一样古老,但在计算机中执行它可以解决以前所不能接近的因果分析问题。这种数据密集型科学提供的新手段,用于探索高度复杂现象的因果结构,将对科学实践产生不同的影响[15]。也就是说,通过研究海量数据之间的相关关系,是探索复杂现象背后的因果机制的有效手段。

第三,虽然两种大数据都使用了已有的理论或知识,即大数据并非理论自由,而是负载理论的,但是从两个案例可以看出,他们负载理论的方式有所不同。在SSD中,从问题的提出、模型的建立,再到对结论的解释,都以现有的与研究问题相关的科学理论为基础。而在SDS中,在模型建立之初才会使用相关的理论(有些案例中甚至不需要理论,如“垃圾邮件分类”“啤酒与尿布”等)。这类大数据更多使用的是算法。算法虽然负载理论,但是其负载的理论与具体问题无关,这也正是沃尔夫冈将其称为“外部负载”的原因[15]。

4　结论

关于大数据的两种相互对立、不可调和的认识其根本原因在于指称的对象并不同一,也就是客观上存在着两类大数据。因此,在后续的研究中,研究者首先要明确大数据概念的所指:是“用数据的方法研究科学(SSD)”,还是“用科学的方法研究数据(SDS)”。如此,才能从根本上消除目前这种对立、混乱的认识现状。其次,由于大数据的主要价值在于它有可能表征和解释复杂系统,而非线性是造成系统复杂性的主要原因,因此借用经典统计理论对复杂数据进行相关性分析时就存在适用性的问题(例如在大数据处理中经常出现的过度拟合问

题),因此,研究经典统计理论在大数据中的适用性问题,包括明确它的适用条件和边界、对研究结论在方法层面进行反思等是必不可少的研究内容和程序。当然,要从根本上解决适用性问题,就需要发展出基于全样本的统计理论以及基于大数据的全新的数据处理方法[26]。再次,需要深入探讨两种大数据都面临的一个难点问题——不透明性,尤其是算法的不透明性。由于大数据所使用的复杂算法以及计算处理的速度造成了大数据分析过程的无法跟踪性与可回溯性,使不透明性成为人们对大数据保持警惕的关键[26]。因此,如何保证适度的透明性,提高对数据处理的可解释程度,将是大数据研究的重点课题。

参考文献(References)

[1] Alvarado R, Humphreys P. Big data, thick mediation, and representational opacity[J]. New Literary History, 2017, 48 (4): 729-749.

[2] 欧高炎, 朱占星, 鄂维南, 等. 数据科学导引[M]. 北京: 高等教育出版社, 2017.

[3] Schönberger M V, Cukier K. Big data, a revolution: that will transform how we live, work, and think[M]. Boston: Houghton Mifflin Harcourt, 2013.

[4] Clark L. No questions asked: Big data firm maps solutions without human input[EB/OL]. [2020-04-10]. http://www.wired.co.uk/news/archive/2013-01/16/ayasdi-big-data-launch.

[5] Sprenger J. Science without (parametric) models: The case of bootstrap resembling[J]. Synthese, 2011, 180(1): 65-76.

[6] Anderson C. The end of theory: The data deluge makes the scientific method obsolete[J]. Wired, 2008, 16(7): 1-3.

[7] 冯启思. 大数据统治世界[M]. 曲玉彬, 译. 北京: 中国人民大学出出版社, 2013.

[8] Floridi L. Big data and their epistemological challenge[J]. Philos and Technol, 2012, 25(4): 435-437.

[9] 董春雨, 薛永红. 从经验归纳到数据归纳:特征、机制与意义[J]. 自然辩证法研究, 2016, 32(5): 9-16.

[10] Timmer J. Why the cloud cannot obscure the scientific method[EB/OL]. [2020-04-10]. http://arstechnica. com/uncategorized/2008/06/why-the-cloud-cannot-obscure-the-scientific-method.

[11] Brooks D. What you'll do next: using big data to predict human behavior[N]. The New York Times, 2013-04-16.

[12] Boyd D, Crawford K. Six provocations for big data[J]. Social Science Electronic Publishing, 2011, 123(1): 1-17.

[13] Sabina L. Integrating data to acquire new knowledge: Three modes of integration in plant science[J]. Studies in History & Philosophy of Biological & Biomedical Sciences, 2013, 44(4): 503-514.

[14] Canali S. Big data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS[J]. Big Data & Society, 2016, 3(2): 1-11.

[15] Pietsch W. Aspects of theory-ladenness in data-intensive science[J]. Philosophy of Science, 2015, 82(5): 905-916.

[16] Frické M. Big data and its epistemology[J]. Journal of the Association for Information Science & Technology, 2015, 66(4): 651-661.

[17] Hey T, Tansley S, Tolle K. The fourth paradigm: data-intensive scientific discovery[C]. Microsoft Research, 2009.

[18] Berry D. The computational turn: Thinking about the digital humanities[EB/OL]. [2020-04-10]. https://culturemachine. net/wp-content/uploads/2019/01/10-Computational-Turn-440-893-1-PB.pdf.

[19] Harford T. Big data: Are we making a big mistake?[J]. Significance, 2015, 11(5): 14-19.

[20] Kitchin R, Lauriault T P. Small data in the era of big data[J]. Geojournal, 2015, 80(4): 463-475.

[21] 曹贤才, 时冉冉, 牛玉柏. 近似数量系统敏锐度与数学能力的关系[J]. 心理科学, 2016, 39(3): 580-586.

[22] Halberda J, Ly R, Wilmer B, et al. Number sense across the lifespan as revealed by a massive internet-based sample[J]. PNAS, 2012, 109(28): 11116-11120.

[23] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009, 457(7232): 1012-1015.

[24] Hey T, Tansley S, Tolle K. 第四范式:数据密集型科学发现[M]. 潘教峰, 张晓林, 译. 北京: 科学出版社, 2013.

[25] Pietsch W. The causal nature of modeling with big data [J]. Philosophy & Technology, 2016, 29(2): 1-35.

[26] 朱迪亚·珀尔, 达纳·麦肯齐. 为什么:关于因果关系的新科学[M]. 江生, 于华, 译. 北京: 中信出版集团, 2019.

Two schools of big data research and two types of big data: A case study

XUE Yonghong[1], DONG Chunyu[2]

1.College of Science, North China University of Science and Technology, Sanhe 065201, China
2. College of Philosophy, Beijing Normal University, Beijing 100875, China

Abstract In the research of big data, two distinct and diametrically opposed academic schools have emerged, namely radicalism and conservatism. Through an analysis of two typical cases, this article finds that the so-called "big data" actually refers to two types of "big data", one is to study data in a scientific way, and the other is to study science in a data way. It is the existence of the two types of big data that forms the two camps of activism and conservatism. The types of big data and their significant difference are the reasons of the formation of radicalism and conservatism camps. On the basis of summarizing the characteristics of the two kinds of big data, this paper puts forward the only approach that may eliminate the antagonism and confusion and push the big data research further.

Keywords big data; activism; conservatism; GFT; number sense

（责任编辑　王丽娜）