特邀主持人：潘绥铭（中国人民大学性社会学研究所教授、博士生导师）

主持人的话：近年来，对于大数据的讨论日渐升温。2015年9月，国务院印发《促进大数据发展行动纲要》，系统部署大数据发展工作，表明大数据发展已经上升为国家战略。大数据对各个学科的发展都造成了显著影响。其中，在社会学研究方法领域表现得尤为明显。因为大数据坐拥所有数据，信息的精确性让位于丰富性、强调相关关系而不是因果关系，“理论已死”等论调对社会学传统研究方法造成一定冲击，学界对此展开了丰富的讨论，但并未达成共识。本期刊出的四篇文章，针对“大数据与社会学研究方法”这一争议性问题，提出了各自的看法。潘绥铭认为，大数据已经出现了盲目崇拜，“一切皆可量化”的核心口号和基本理论需要被质疑，所谓大数据，其实只是“量化研究”的最新表现形式，仍然有不可克服的“原罪”；孙秀林旗帜鲜明地表示“社会学应该拥抱大数据”，认为大数据为研究人类行为提供了新的工具，为研究社会互动与社会交往提供了新的可能，为宏观层面的社会测量提供了新的视角，为社会学带来了新的研究方法；鲍雨致力于分析大数据的方法论逻辑，并对其方法论困境进行了深入剖析，认为谨慎使用大数据应是基本态度；张旭和唐魁玉在分析了大数据对社会学的正负后果之后，认为大数据为社会学研究打开了一扇新的大门，但是这些研究方法只能作为传统社会学研究的补充，而不能完全替代传统的小数据研究方法。这些研究或者大胆推断，或者小心论证，都在社会学领域的大数据研究中有所推进，对未来的大数据研究与应用大有裨益。

生活是如何被篡改为数据的？

——大数据套用到研究人类的“原罪”

文／潘绥铭

摘要：目前对于大数据已经出现了盲目崇拜，“一切皆可量化”是其核心口号和基本理论。但是在量化过程中，不可避免地会出现四种情况：剪裁现实生活、忽视社会情境、抹煞主体建构、取消生活意义。这种“原罪”并不能由于数据规模的无限增大而被消除。因此，大数据不能质疑，更不能取代各种非量化的人文社会研究。大数据只有对其“原罪”进行深刻反思，并且予以充分展示，才有资格在人文社会研究中保留一席之地。

关键词：大数据；数据崇拜；量化研究

中图分类号：C91—03 **文献标识码**：A **文章编号**：1006—0138(2016)03—0032—04

近年来，对于大数据已经出现了盲目崇拜，就是无质疑、不反思地跟风颂扬和无限拔高。[1] 本文不涉及任何自然科学领域中的大数据及其应用，仅讨论一个根本问题：大数据能够套用到对于人类的研究中吗？

对这个问题，我国学术界虽然也出现了一些质疑，但是不仅寥若晨星，而且在学理上也主要是在可操作性的层次上争论，并没有击中要害。其实，大数据最值得质疑的，既不是其定义，[2] 也不是其功能或意义，[3] 还不是方法论层次上的“以相关分析取代因果分析”，[4] 而是“一切皆可量化”[5] 这

作者简介：潘绥铭，中国人民大学性社会学研究所教授、博士生导师，北京市，100872。

个核心口号和基本理论。它表述了大数据的三层意思：其一，没有量化，就没有数据，更不可能有什么大数据；其二，物质世界当然是可以被量化的，但是如果仅限于此，那么所谓的大数据就仅仅是数量的增加，性质毫无改变，纯属炒作，例如天气预报一直就在分析海量的数据，却并没有以大数据自居，更没有形成崇拜；其三，现在的大数据之所以被崇拜，要害其实只有一点：把人类的行为及其结果，也给量化了，而且号称无所不包。

这样一来，大数据的性质就变了，从自然科学侵入到人文社会研究，从科学蜕变为“唯科学主义”。这就不仅仅是一个研究工具的问题，而是一个认识论的根本问题。对此进行批评的人文社科著作汗牛充栋，本文不再一一列举，仅在操作的层次上分析一下，人类无限丰富的生活实践，在被“唯科学主义”改造成“数据”的过程中，究竟发生了什么。

一　现实生活被裁剪

大数据崇拜者极力鼓吹“4V”（规模大、种类多、高速度、高价值），[6] 却故意回避了一个根本的问题：在最开始，您收集到的，就是可以用来分析的数据吗？[7]

在社会学的问卷调查中，这是有可能做到的；但是在所谓的大数据中，却绝对不可能。因为大数据并不是研究者主动去收集的人类行为及其结果，而是五花八门的所谓“客观记录”，是人类生活中微乎其微的那一部分“可获得信息”，例如上网活动所留下的痕迹、监控记录等。

可是尽人皆知，在人类活动的全部信息中，可获得的要远远少于不可获得的。后者最典型的就是人类的一切精神活动的信息，在可预见的未来，仍然不但是无法获得的，而且根本就是无法监测的。这样一来，所谓大数据所获得的信息，首先是极端片面；其次是漫无边际；第三是支离破碎；第四是毫无意义；根本不可能直接用于任何量化的分析。

那么，这样的信息怎么才能转化为可分析的数据呢？首先是必须加以“界定”，就是保留什么和舍弃什么；其次是进行“分类”，就是把什么归属于什么；第三步是加以“定义”，就是给某类信息赋予特定的人类意义；最后一步则是“赋值”，就是把不同的定义转换为可计算的数值。

以上网活动的痕迹为例，大数据的生产者，怎么来界定那些痕迹是有意的还是无意的、闲置的还是凝视的、主动寻找的还是被引导而来的？界定之后，到底是根据停留时间长短还是活动的频率，来制造出“活跃”或者“不活跃”这样的类别呢？为什么把“活跃”就给定义为“需求”呢？最后，把“需求”赋值成什么？从“不需求”到“强需求”的不同赋值之间，究竟是什么样的数量关系呢？

显而易见，在这个四部曲的过程中，完完全全是研究者自己在主观地、人为地、强制地“整理”那些“可获得信息”，把人类生活的痕迹，完完全全地篡改为自己的世界观和价值观所能接受的“数据”。往好里说，这叫做无可避免地加工；往坏里说，这就是赤裸裸地伪造。

这就是说，所谓的大数据，其实一点都没有超出“小数据”原有的局限性：裁剪生活，撕碎人生；非要把整体生存的“人”，视为一堆杂乱的零碎。在实际生活中，人类绝对不是，也不可能是这样来“量化地”认知和行动的。因此，大数据其实并不是帮助人类思考，而是企图取代和控制人类的生活经验，是人工智能的噩兆。

二　社会情境被忽视

有人已经发现，大数据记录的都是单独个人的行为，无法发现不同行为者之间的关系；[8] 于是问题就来了：在这个现实世界里，难道真的存在一种与他人毫无关系的个人行为吗？难道个人的一切行为，不都是在一定的人际关系中，才会产生，才会带来某种结果吗？

社会不是个人的简单集合，而是人们通过各种关系有机地组织起来的。同时，人们又是在特定的社会环境中做出各种行为的，不可能天马行空，独往独来。因此，人类活动留下的一切痕迹，必定蕴含着无限丰富的社会内容。如果舍弃之，那么不管什么样的数据，不仅是浮光掠影，而且必定是盲人摸象。

尤其是，每一个人都在特定的社会中，一

点一点地成长为“此时此景中的此人”，然后才会做出“此因此果的此行为”。这就是每个人的社会历史建构过程，其中最重要的就是我们的一切社会背景、生活状况和成长经历。

可是这一切，往往仅仅存在于我们自己的经验与记忆之中；往往难于言表，更往往无法记录。从“客观监测”的角度来说，根本就是“风过无痕”。那么，就算毫无隐私，就算监测可以天罗地网，所谓大数据的信息源又是从何而来的呢？[9] 因此，对于了解人类生活而言，大数据其实根本就是空中楼阁。

如上所述，这样的批评还是很中肯的：“数据不懂社交、不懂背景，会制造出更多噪音，遗漏真正有价值的东西。大数据无法解决大问题。”[10]

三　主体建构被抹煞

“大数据崇拜者”很可能不知道，或者不敢承认：在人类生活中还有一种现象，叫做“主体建构”。即人们对于自己的行为所做出的解释，很可能与监测者的解释大相径庭，甚至背道而驰。最常见的就是，一切人际的误会，盖源于此。

那么，无论大数据监测到多少人类的行为，它究竟是如何分辨出其中主体建构的成分呢？首先，以网购的大数据为例，即使您收集到全部的上网痕迹，而且全都数字化地一览无余，那您怎么知道人家就真的就是这样想的呢？这种“客观测定”，离矿物学很近，可是人却是有主观意志的啊，您是怎么监测到的？连物理学还有个“测不准原理”呢，何况对于人的主观意愿？其次，您知道人类还会“自我呈现”吗？说不好听一些，就是表演。如果连测谎仪的结果，法律都还不予采信，那么您怎么筛除被监测对象的表演呢？第三，难道您就不找被监测对象去核实一下？在司法审判中，就连证据确凿的罪犯，法官也必须听取他的说法，才能做出正确的判断。可是大数据崇拜者却根本漠视主体意愿的存在。这岂不是自欺欺人？第四，您听说过弗洛伊德吗？您知道除了“动机”，还有“无意识”吗？

即使是某些询问对方意愿而获得的数据，也仍然存在着这样一个问题：对方是否具有足够的能力来表述自己的意愿呢？我们不应该忘记弗洛伊德，不应该忽视无意识行为的广泛存在，更不应该否认：人类的一切行为痕迹，无论多么海量，其实并不能容纳和表述人类的生活意义。因此，如果行为者自己都搞不清楚自己是怎么回事，那么您还怎么去核实呢？根据什么来判断真伪与程度呢？

总而言之，一切试图用自然科学或者数字化来了解人类及其社会的尝试，不是都必然失败，而是都无法否定人类“主体建构”的重要性，结果都必然是把真实的生活给削足适履了。

因此，大数据所获得的一切“发现”，其实只不过是某些人在描述其他人的生活。其他人既不知道自己已经被描述了，也没有渠道去修正这种描绘。结果，大数据其实只不过是一帮技术分子所构建起来的新的认知霸权，其崇拜者也只不过是急于使用这个霸权而已。

四　生活意义被取消

人文社会研究的至少两千年历史告诉我们：人类的一切行为，不仅蕴含着他们的人生意义，而且是为了追求其人生意义而行动的。这是人与物的根本区别。

可是，大数据所谓的一切“可记录的痕迹”，如果没有获得对方的主诉，那么就不可能包含该行为意义的信息。例如一切上网活动，行为主体都不会表述自己为了寻求什么才这样做的，也不可能表达出这样做带来了什么样的价值与意义。

以购物网站记录下来的数据为例，它确实可以容纳数千万人在购物时不知不觉地留下的近乎无穷无尽的痕迹；但是，这就能反映出这些人的购物偏爱吗？难道这些人就再也不在实体商店中买东西了吗？难道他们在一时一事上表现出来的偏爱就永恒不变吗？难道他们的每一次上网购物都能得到自我满足吗？

那么，您怎么能够确定：他们在不同的渠道中，在不同的情境之中，都会做出一模一样的选择呢？如果您无法证明这一点，那么您的“大数据”就只能是“大垃圾”，一点儿也不冤。

交通监控录像、医疗记录、通讯记录等等，

都足以号称自己是“大数据”。可是，所有这些数据，都仅仅是记录下了人们生活中的一个个零散的侧面。因此，这样的“大数据”再怎么大，也无法解决以下一系列常识性的问题：首先，人在生活的某个侧面里的表现，与他／她的整个人格与人生，难道不存在紧密的关联吗？农民工吃20元的盒饭都嫌贵；富豪买上千万的汽车也不眨眼，这难道仅仅是所谓的“消费选择”吗？其次，人类生活的各个侧面之间，难道不是相互影响着的吗？农民工吃20元的盒饭，却可以搭上200元的礼钱；富豪买上千万的汽车，却不肯做一点儿慈善，这也仅仅是所谓“购买习惯”吗？第三，任何一个人的生活，难道不是被社会、文化、历史等因素制约着吗？农民工之所以要吃20元盒饭，绝不仅仅是因为工资低，还因为他的抚养系数、失业可能性、职业风险等等都比富豪要高出很多。这，难道也是“可支配资金”吗？

如此这般，数据越大，岂不是错误越大？

五　结语：原罪就是原罪

本文所论述的一切，其实都是来自人文社会研究中，久已存在的对于“量化研究”的批评。[11]大数据崇拜只不过是这种思潮的最新表现，只不过是披上了更为光鲜亮丽的外衣而已。

在基督教教义中，原罪不但是与生俱来的，而且是背负终身的，不能通过人自己的救赎而被消除。很可惜，量化研究也是如此。无论其技术手段如何发达，无论其数据多么大，一旦应用于人文社会研究，其缺陷与弊病就根本无法避免，充其量也不过是程度的减轻而已。说到底，“大数据崇拜”，其实就是“唯科学主义”在人类历史面前一败涂地后的末日哀鸣。如果科学没能阻止希特勒的统治，也没能预测出此后人类的一切发展，那么就绝不是“艺不精”的问题，而是用错了地方，是越界跑到了自己无能为力的领域。

当然，这并不是说，量化研究和大数据就一定不能用，而是表达三层意思：首先，它们都不能质疑更不能取代各种非量化的人文社会研究；其次，只有对这些先天缺陷进行深刻反思，并且予以充分展示的量化研究，才有资格在人文社会研究中保留一席之地；第三，两种研究就像是两条铁轨，缺一不可，但又平行延伸，永不交叉。

注释：

[1]王程韡：《“大数据”是“大趋势”吗：基于关键词共现方法的反事实分析》，《科学学与科学技术管理》2015年第1期。

[2]李天柱、王圣慧、马佳：《基于概念置换的大数据定义研究》，《科技管理研究》2015年第12期。

[3]钟瑛、张恒山：《大数据的缘起、冲击及其应对》，《现代传播》2013年第7期。

[4]张晓强、杨君游、曾国屏：《大数据方法：科学方法的变革和哲学思考》，《哲学动态》2014年第8期。

[5]道格拉斯·W. 哈伯德：《数据化决策——大数据时代，〈财富〉500强都在使用的量化决策法》，邓洪涛译，广州：世界图书出版广东有限公司，2013年。

[6]Bill Franks：《大数据：不是技术难题》，《成功营销》2013年第4期。

[7]阎光才的《教育及社会科学研究中的数据——兼议当前的大数据热潮》（《北京大学教育评论》2013年第4期），已经哲学化地论证了数据与真实世界之关系，但是仍然缺乏具体的分析。

[8]谢然：《大数据社会的具体场景》，《互联网周刊》2014年第22期。

[9]有论者提到了这一点，但是仍然囿于“数据源”，不足为训，参见黎争：《从数据源看大数据》，《IT经理世界》2013年第14期。

[10]转引自刘宏伟、徐翠英：《拷问大数据》，《企业管理》2013年第9期。

[11]潘绥铭：《社会学问卷调查的边界与限度——一个对“起点”的追问及反思》，《学术研究》2010年第7期。

责任编辑　刘秀秀