社会学应该拥抱大数据

文／孙秀林　施润华

摘要：大数据的快速发展，大大扩展了社会学定量研究的领域：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。当然，在正面看待大数据带来的积极意义的同时，也要意识到大数据分析失灵的可能性，要理性认识大数据的优势与劣势，处理好大数据与小数据之间的关系。

关键词：大数据；定量研究；研究方法

中图分类号：C91—03　**文献标识码**：A　**文章编号**：1006—0138(2016)03—0036—06

今天，我们生活在一个数据急剧膨胀的时代。它不仅改变了我们生活的世界，同时也在改变我们看待这个世界的方式。一夜之间，“大数据”成为商界、学界、政界的时髦词语，无人不谈大数据，无事不涉大数据。在商界，从尿布与啤酒的关联，到亚马逊（Amazon）和奈飞（Netflix）的推荐系统，无数例子已经证明了大数据的应用前景。在政界，各国政府相继制定大数据发展战略，2015 年 9 月，中国国务院印发《促进大数据发展行动纲要》，将大数据发展提升为国家发展战略工作。但是，在学界，关于大数据的应用与发展，争论却持续不断。有人为大数据的发展加油呐喊，认为这是社会学未来发展的方向，一种新的计算社会学的研究范式正在急速崛起之中；同时也有人认为这只是一种新的数据“玩具”而已，不可能取代数百年来社会学已经发展出的理论框架与研究范式。本文无意也无力对上述争论做出一个详细的评判，因为这两方面都各有其道理。本文仅仅从社会学定量研究的角度出发，探讨大数据发展对社会学带来的机遇，以及我们应该如何应对这种变化和影响。

一　社会学为何需要拥抱大数据？

大数据是指巨大而多样化的数据集，是对全世界每个人所做的每一件事的即时记录。大数据的出现与快速发展，为社会科学的发展带来了前所未有的研究机遇，大大扩展了原有的研究领域。

（一）大数据为社会学研究人类行为提供了新的工具

在大数据时代，我们日常生活中的一切，都已经进入一个数据化的过程中。人们每天在微博、微信上发表的评论，忠实记录了个人偏好，包括个人

基金项目：国家社会科学基金项目“我国新社会群体研究”（14BSH026）

作者简介：孙秀林，上海大学社会学院教授，上海市，200444；施润华，上海大学社会学院研究生，上海市，200444。

对于美食的评论、对他人意见的评论、对公共事件的评论等。人们每天的通话记录，可以清晰刻画人们的联系人记录和社交网络。人们每天的消费记录，保留在各大银行和电商的数据库中，通过对于这些信息的分析，可以充分展现城市不同阶层的消费模式。上班族每天上下班的公交卡信息，构成了研究城市生活的重要数据库。更重要的是，这些数据格式都不是一次性的，而是实时变化的。相对于以前我们通过调查问卷来间接测量人类行为，这些新的数据形式，对于我们理解人类的行为提供了前所未有的机遇。哈佛大学的金加里（Gary King）教授甚至认为，这种新的数据方式对于社会科学研究而言，其意义不亚于显微镜对于生物学、天文望远镜对于天文学发展的意义。[1]

在社会学的研究中，理解与解释同样重要。我们不仅需要解释人类的行为，同样也需要理解人类的行为。在这一点上，社会学不仅需要观测到具体的人类行为与交往情况，同样也需要理解不同行为背后的原因。因此，我们不仅需要获取人类的行为模式，同时也需要获取主观意识方面的认知、想法、观念等。对于定量研究而言，这些观念性的、文化性的、理解性的数据，是非常难以量化和测量的。庆幸的是，在大数据时代，利用新的测量手段，已经有学者开始进行了一些尝试。彭特兰从孔德的实证社会学出发，关注人的想法（idea）。他把问题聚焦在“想法流”(idea flow)上，将其作为看待人类关系建构、社会结构演进的新视角。在这样的语境下，他认为社会学习是想法流的关键，多样性是想法萌生的土壤。他利用可穿戴设备，把数据获取的方式从测量（如传统的问卷调查、访谈、观察等）上升为感知（可穿戴设备记录的心理学、生理学、生物学特征），使得利用大数据对于人类互动行为意义的理解和分析成为可能。[2]

（二）大数据为研究社会互动与社会交往提供了新的可能

在社会学的研究领域中，如何测量人与人之间的社会交往与社会网络，一直是个非常重要的研究议题。在传统的研究中，虽然我们都承认社会网络是个非常复杂的社会结构，但是，由于传统测量手段主要是通过对于个体的问卷调查来进行，所以多数对于社会网络的研究都采用了简化的测量。一种社会网络的测量集中于个体网络，如个人的拜年网、餐饮网、交谈网等等。[3]另外一种对于整体网的社会测量则多集中于界限比较清楚的、规模较小的测量，如一个班级、一个企业部门、一个村庄等。[4]

这种数据获取方式的局限性，极大地限制了社会学对于人类交往与社会互动的深入研究。大数据时代，各种社交网络平台（如国外的脸书、推特，国内的微博、微信、豆瓣、人人网等）的发展，使得研究者们轻易突破了上述限制，可以在一个更大规模上研究人们之间的社会网络与社会互动，甚至可以研究全球网民之间的社会交往情况（如 Facebook、Twitter 等）。[5]相对于国外学者利用社交媒体进行的社会网络分析，国内学者也开始利用本土的社交媒体，将社会网络分析的研究领域进行了拓展。如通过比较广州 118 个业主论坛和上海 199 个业主论坛的社会网络图，黄荣贵等人研究了上海的业主网络与广州的业主网络之间的差异，以及这种差异对于基层治理的影响。[6]新近一篇文章以一个业主论坛为切入点，利用网络技术抓取 6 万多条业主发言，从全体网的分析角度探讨了不同类型的虚拟社区用户参与虚拟社区讨论对社区在线参与的影响。[7]甚至有研究利用大数据获取的手段，研究了中国千人学者的合作者网络与社会资本转化情况。[8]

虽然社会计算领域的一项研究表达了对于社交媒体收集和使用大量数据所产生的潜在危害的担忧，[9]对这些大数据进行分析的时候，社会选择和测量问题使得一些理论本身变得“可疑”，[10]但是这种数据获取方式，可以使我们对于人们之间的社会互动有一个更深入的了解。而且最近的研究显示，社交网络和机器学习的快速发展为我们打开了新的图景，我们通过技术革新，使用云技术进行机器学习获取人脸信息这些非结构性数据，从而更有效地解决大数据中数据获取的难题。[11]还有学者基于 MapReduce 等方法，运用并行随机迭代方式搜索社会网络编码状态空间中的最佳编码方法，从而挖掘出大数据社会网络中的最佳社团划分。[12]

（三）大数据为宏观层面的社会研究提供了新的测量手段

随着中国城市化进程的快速发展，对于城市议题的研究，将是未来社会学研究的一个重要方向。但在以往的城市研究中，由于社会学家往往难以获得微观的城市数据（如观测单位具体到街道、居委会的数据），严重阻碍了城市议题的量化研究，如城市中的居住隔离问题、贫困问题、职住分离等。[13] 在大数据时代，随着“基于位置服务”技术的发展（如手机定位信息、出租车轨迹、交通卡信息、消费卡信息等），为研究城市社会学研究提供了新的视角与可能。如利用公交卡的刷卡数据，不仅仅可以分析大都市的通勤状况与职住分离情况，而且可以更好地理解人们在城市中的不同行为模式与空间特征，对中国的城市社会学研究具有重要的启示作用。[14]

在这一背景下，汤森以一种前瞻式的视角为我们解读了城市的未来。他用活生生的现实案例向我们展现了，随着数据的开放、移动智能设备的普及、互联网时代的来临，智慧城市不再是一个空洞的名词，它有了全新的意义。[15]

（四）大数据产生了新的数据分析方法与分析技术

大数据时代，数据的产生方式发生了变化。在传统情况下，对于国民经济指标的统计，基本是依赖于国家行政力量的统计系统来进行的。这种传统的统计方式，需要通过科层体系的层层上报，并逐级汇总，比较费时费力。大数据的发展，可以使得很多传统的统计数据在很短时间内获取，一个最著名的例子是谷歌的“谷歌流感趋势”，通过汇集人们在谷歌上搜索的关键词，谷歌可以迅速标示流感疫情的发展、扩散情况，通过与美国疾病预防控制中心的监测报告进行比较，谷歌认为自己利用网络搜索做出的结果非常可靠。重要的是，谷歌的“谷歌流感趋势”只需要 1 天就可以生成一份最及时的报告，而不是美国疾病预防控制中心的 2 周。[16] 虽然针对谷歌的这一研究争论持续不断，但不可否认的是，谷歌的这一研究思路，极大地促进了“大数据”中“用户生成数据”（User-generated content）的研究在学术界快速发展。

大数据时代，数据的获取方式发生了变化。如在一篇对于上海市社会组织的空间分析中，作者通过“网络爬虫”，获得了上海所有（一万多条）在册社会组织的详细信息，包括组织名称、组织注册代码、注册时间、证书有效时间、组织类型、注册地、主管单位、法人代表、地址、邮编、电话、网址、主要业务内容以及奖惩情况等。这种数据获取方式，相对于传统的方式，无疑极大降低了学术研究的交易成本。[17]

大数据时代，数据的分析技术发生了变化。研究者可以通过社交网络、社交媒体等方式，大规模随机设定、发送不同的信息，以此形成随机实验中的“参照群体”与“实验群体”，通过这两个不同群体的反应情况，来进行科学研究的因果推论。[18] 人类学也在大数据时代发展出来“虚拟民族志”的研究方法，对虚拟社区中的社会互动进行追踪观察，以更好地关注和探究信息时代的社会生活。[19]

二　社会学如何拥抱大数据？

在大数据时代下，很多学者认为将社会学与计算机科学结合起来，将为社会学研究带来革命性的改变。社会计算作为一种新的计算范式，会产生一个新的跨学科研究与应用领域，具有广阔的研究与应用前景。[20] 甚至有学者认为，大数据时代产生的新计算社会学会引发一场社会学范式革命，社会学的“计算范式”会成为一种在社会学研究中占主导地位的范式。[21] 在这种情况下，社会学应该如何面对大数据所带来的挑战？社会学如何利用大数据的优势，促进自己学科的实质发展？下面仅仅根据笔者的研究经验，提供一些管窥之见。

（一）理性认识大数据的优势与劣势

从大数据实际应用的发展前景来看，一方面要看到数据本身带来的积极意义，另一方面也要意识到大数据分析失灵的可能。在对于大数据的“崇拜”或曰“幻觉”中，最需要一提的是“大数据傲慢”（Big Data Hubris）的问题。在谷歌发表其“谷歌流感趋势”的研究后不久，另外一篇发表在《自然》杂志上的文章发现，如果使用 2013 年最新的数据进行检验，谷歌的预测结果存在非常严重的偏误。研究者

认为，造成这种结果有两个重要原因。其中一个最重要的原因就是“大数据傲慢”，即大数据科学家们认为大数据是传统数据收集方法的终结而非补充，因此可以完全忽略传统的数据收集方式。在这个案例中，谷歌的工程师无法证明在网上进行搜索的群体等同于流感涉及的群体。如果我们无法判断这两个群体的具体情况，那么大数据所收集到的数据是一个有偏的样本，而一个有偏的样本其规模越大，做出错误判断的概率也就越高。同时，用户搜索行为的改变也会影响关键词的搜索结果。另外一个是算法变化，谷歌的工程师对算法会进行不断地调整和改进，而搜索引擎算法的改变会影响预测结果，比如媒体对于流感流行的报道会增加与流感相关的词汇的搜索次数，进而影响“谷歌流感趋势”的预测。[22]

在对于大数据的争论中，最令社会学家诟病的是，大数据对于社会学理论的态度。大数据的教父级人物舍恩伯格宣称，在大数据时代，理论不再是我们分析和理解世界的必备武器，数据分析本身就可以揭示一切问题。对此，我们需要警惕。大数据的优势在于不用担心数据的代表性问题，可以弥补传统数据中不具代表性的问题。通过计算机巨量的运算方式发现相关关系，包括已知的和未知的，这种方式可以帮助研究者发现更有效的事实（比如超市中尿布与啤酒的关系）。但是，数据本身只是一种材料，大数据本身并不构成、也不能回答特定问题。大数据是寻找问题的一种方式，但其本身不构成对象，它只是一种工具，适用于一些特定用途，切忌将其盲目地神圣化。社会科学领域另外一位重量级的人物金加里教授，在谈到这一问题时一再强调，在任何社会科学领域，甚至在任何科学领域，都必须尊重理论，从事理论的学者与从事经验研究的学者，都是必不可少的。大数据革命在经验研究方面不管取得如何大的成绩，都无法降低理论研究对于我们社会科学研究的意义和价值。[23]

（二）处理好大数据与小数据之间的关系

在大数据时代，传统的小数据仍然具有不可或缺的价值。[24] 相对于大数据，小数据的优点仍然非常明显，比如变量定义清晰、数据生成机制可控、检验评估成本较低等。最重要的是，小样本数据对于可能推论的研究总体具有比较明确的认知，从而可以对社会现象之间的因果关系具有更好的判断。大数据虽然具有收集快速、数据颗粒更细、数据总体量巨大等优点，但由于大数据通常并不是通过专门的理论设计和测量工具产生，而多数是政府部门和企业的业务流程数据沉淀而来，所以虽然其规模巨大，但其样本的代表性往往是有偏的。

虽然大数据的规模往往很大，但是，在很多时候大数据并不是“全数据”，比如网络用户并不能包括全部人口。在上述对于“谷歌流感趋势”的研究中，谷歌做出错误判断的一个重要原因就是忽略了样本可能存在的偏误，从而得出了错误的推论。一个有偏的样本，不管其规模多大，对于我们做出预测都没有真正的帮助。一个最著名的例子来自 1936 年美国总统大选的预测。当时，为了提前进行总统大选结果的预测，《文学文摘》杂志给自己的读者群寄出了 1000 万份的调查问卷，但因为没有考虑到杂志订阅群体在美国总体选民中并不是一个代表性样本，所以在这次预测中，《文学文摘》惨败给盖勒普公司，而后者使用了一个具有代表性的样本，规模仅为 5000。

最近也有学者提出，虽然大数据的有偏性备受质疑，但学者们可以充分利用大数据的有偏性，重点关注特定人群（如经常使用公交卡系统出行的低收入人群）、局部人群（如数据更易获得的大学生群体），期待与其他有偏的数据互补，慢慢将特定研究领域的拼图补齐。[25]

大数据和小数据的关系如果处理得当，可以彼此取长补短。在一项关于时间利用的调查中，研究者发现，大数据的引入可以有效弥补小数据收集信息不全的弊端，为传统调查提供了新的数据收集方式；通过可移动穿戴设备，可以在第一时间获取受访者的时间利用情况；此外，互联网提供的关于时间利用的相关记录可以作为调查数据来源的一个重要组成部分。[26]

（三）大数据需要新的研究技能与团队合作

在大数据时代，由于新的数据来源和分析方法快速发展，对于任何一个作为个体的研究者来说，完全掌握快速发展的新技能都成为一

项不可能的任务。仅就数据采集来说，就涉及编程、数据库、网络传输、文本解析甚至分布式计算等等各种技术环节，这些技术对于社会学研究者提出了新的技术要求。至于在分析阶段，一些新近发展出来的模型，如主体建模、文本分析、深度学习、复杂网络建模等等，都将进入社会学研究者的视野。对于层出不穷的分析软件，也没有人能够完全精通。比如现在应用越来越广泛的R软件，已经有超过5000多个包（package）在其镜像网站（CRAN)上面发布，而且每天都会有基于新模型、新算法的包（package）加入进来。

在这种情况下，社会学需要积极调整，才能紧跟发展趋势，而不会成为被淘汰的学科。首先，要鼓励年轻学者持有一种开放的心态，对于一些传统上属于自然学科的知识技能，如网络爬虫技术、网页分析技术等，也能有一定的了解和掌握，只有这样，才能将其他学科中一些有益的学术热点纳入社会学的分析中。其次，要鼓励团队合作，在大数据时代，单个研究者掌握所有的技能是不现实的，只有通过社会学研究社群的合作，才有可能跟上这种发展趋势。整个学科的评价标准应该鼓励多人合作，在高校与科研院所的科研管理和职称评定中应该承认多人合作的贡献。再次，要调整我们传统的人才培养体系。面对大数据的发展趋势，突破传统的学科培养人才体系，培养具有交叉学科分析能力的研究者。在传统的课程体系外，如何将大数据时代需要的一些技能纳入人才培养体系，是一个长期、艰巨，而又刻不容缓的任务。

三 结 语

社会学界对于大数据的应用和发展前景产生了较大的分歧。有学者为大数据的发展欢欣鼓舞，认为这是产生一种新研究范式的萌芽；也有学者对大数据不以为然，认为这种对新技术的过度崇拜只是某些学者的猎奇而已，终为昙花一现。这种争论，对于一个学科的发展，是必不可少的，只有在真正的学术争论中，一个学科才可能获得实质性的发展。

本文认为，面对大数据对社会学研究带来的挑战与机遇，社会学的研究者应该敞开双臂，用一种开放的心态来对待这一新生事物，并利用大数据的优势，促进自己学科的实质发展，而不仅仅将之视为一种数据玩具。当然，要实现这一点，需要无数学人进行大量的实证研究，从理论、议题、方法、技术等每个方面来推进这一领域的研究，而非仅仅停留在哲学思辨与逻辑辩论层面。

注释：

[1]Gary King, “Ensuring the Data Rich Future of the Social Sciences”, *Science*, vol.331 (2011), pp.719–721.

[2]阿莱克斯·彭特兰:《智慧社会》,汪小帆、汪容译，杭州：浙江人民出版社，2015年。

[3]边燕杰、张文宏：《经济体制、社会网络与职业流动》,《中国社会科学》2001年第2期；边燕杰、Ronald Breiger、Deborah Davis、Joseph Galaskiewicz、伊洪：《中国城市的职业、阶层和关系网》,《开放时代》2005年第4期；边燕杰、张文宏、程诚：《求职过程的社会网络模型：检验关系效应假设》,《社会》2012年第3期。

[4]彭建平：《员工社会网络结构特征对关系绩效影响的比较研究——基于中外两个研发事业部员工整体社会网分析》,《社会》2011年第4期；孙秀林、陈华珊：《1940年代苏南地区借贷市场的网络分析》,《学术研究》2015年第1期。

[5]David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert–Laszlo Barabasi, Devon Brewer, Nicholas Christakis, et al., “Computational Social Science”, *Science*, vol.323 (2009), pp.721–723.

[6]黄荣贵、张涛甫、桂勇：《抗争信息在互联网上的传播结构及其影响因素——基于业主论坛的经验研究》,《新闻与传播研究》2011年第2期；黄荣贵、桂勇：《为什么跨小区的业主组织联盟存在差异——一项基于治理结构与政治机会（威胁）的城市比较分析》,《社会》2013年第5期。

[7]陈华珊：《虚拟社区是否增进社区在线参与？——一个基于日常观测数据的社会网络分析案例》,《社会》2015年第5期。

[8]杨张博、高山行、刘小花：《近朱者赤：基于社会网络分析方法的归国者跨国社会资本转移研究》，《社会》2015年第4期。

[9]A.Oboler, L.Cruz, K.Welsh, "The Danger of Big Data: Social Media as Computational Social Science", *First Monday*, vol.17, no.7 (2012).

[10]J.W.Patty, E.M.Penn, "Analyzing Big Data: Social Choice and Measurement", *Political Science & Politics*, vol.48, no.1 (2015), pp.95—101.

[11]A.Vinay, V.S.Shekhar, J.Rituparna, et al., "Cloud Based Big Data Analytics Framework for Face Recognition in Social Networks Using Machine Learning",*Procedia Computer Science*, vol.50 (2015), pp.623—630.

[12]邓波、张玉超、金松昌、林旺群：《基于MapReduce并行架构的大数据社会网络社团挖掘方法》，《计算机研究与发展》2013年第2期。

[13]孙秀林:《城市研究中的空间分析》，《新视野》2015年第1期。

[14]龙瀛、张宇、崔承印：《利用公交刷卡数据分析北京职住关系和通勤出行》，《地理学报》2012年第10期。

[15]安东尼·汤森：《智慧城市——大数互联网时代的城市未来》，赛迪研究院专家组译，北京：中信出版社，2014年。

[16]J.Ginsberg, M.H.Mohebbi, R.S.Patel, et al., "Detecting Influenza Epidemics Using Search Engine Query Data", *Nature*, Vol.457 (2009), pp.1012—1014.

[17]孙秀林：《社会科学中的空间分析：概念、技术和应用实例》，《山东社会科学》2015年第8期。

[18]具体的例子，详见：Robert M.Bond, Christopher J.Fariss, Jason J.Jones, Adam D.I.Kramer, Cameron Marlow, Jaime E.Settle & James H.Fowler, "A 61—million—person Experiment in Social Influence and Political Mobilization", *Nature*, Vol.489 (2012), pp.295—298；Gary King, Jennifer Pan and Margaret E.Roberts, "How Censorship in China Allows Government Criticism but Silences Collective Expression", *American Political Science Review*, Vol.107, no.2 (2013), pp.1—18; Gary King, Jennifer Pan and Margaret E.Roberts, "Reverse—engineering Censorship in China: Randomized Experimentation and Participant Observation", *Science* Vol.345 (2014), pp.1—10.

[19]卜玉梅：《虚拟民族志：田野、方法与伦理》，《社会学研究》2012年第6期；卜玉梅：《从在线到离线：基于互联网的集体行动的形成及其影响因素——以反建X餐厨垃圾站运动为例》，《社会》2015年第5期。

[20]孟小峰、李勇、祝建华：《社会计算：大数据时代的机遇与挑战》，《计算机研究与发展》2013年第12期；C.Cioffi—Revilla, "Computational Social Science", *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol.2, no.3 (2010), pp.259—271.

[21]罗玮、罗教讲：《新计算社会学：大数据时代的社会学研究》，《社会学研究》2015年第3期。

[22]David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, Vol.343 (2014), pp.1203—1205.

[23]Gary King, "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science", *Political Science and Politics*, vol.47, no.1 (2014), pp.165—172.

[24]沈艳：《大数据分析的光荣与陷阱——从谷歌流感趋势谈起》，2015年10月27日，http://www.nsd.edu.cn/teachers/professorNews/2015/1027/24272.html?from=timeline&isappinstalled=0，2016年3月25日。

[25]龙瀛：《新数据境下的城市研究、规划与设计》，《城市规划学刊》2015年第3期。

[26]蒋萍、马雪娇：《大数据背景下中国时间利用调查方案的改革与完善》，《统计研究》2014年第8期。

责任编辑　刘秀秀