规律与因果：大数据对社会科学研究冲击之反思

——以社会学为例

刘林平　蒋和超　李潇晓

摘　要：在社会科学中，大数据研究还刚刚起步，但也取得了一定成果。大数据为社会学和社会科学重新发现社会历史发展规律提供了可能性：它提供了认知宏观社会、检验社会现象的“异质性假设”和“结果稳定假设”的数据基础；它以实时记录的特点较大程度上排除了获取数据时的人为干扰；它将抽样数据中被排斥的极端值重新纳入统计分析。在因果关系上，大数据有助于从根本上克服由于抽样偏颇所引起的样本选择性偏误；匹配数据可以克服或缓解变量遗漏问题；作为面板数据和分层数据，大数据对确定因果效应、检验因果关系比抽样数据更为有利、稳健和可靠。大数据也许可以重构社会学和社会科学的研究目标。

关键词：大数据；规律；因果关系；冲击；反思

中图分类号：C91-03　文献标识码：A　文章编号：0257-5833(2016)09-0067-14

DOI:10.13644/j.cnki.cn31-1112.2016.09.007

作者简介：刘林平，南京大学社会学院教授、博士生导师；蒋和超，南京大学社会学院博士研究生；李潇晓，南京大学社会学院博士研究生　（江苏　南京　210023）

一、大数据特征与社会学相关研究

“数据”是系统收集到的关于世界的信息要素①。“大数据（Big data 或 Megadata），或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的形式的信息。”②由于互联网的普及和相关设备的广泛使用，人类活动的痕迹几乎都可以转化为可以储存的数据，如日常起居、运动、购物、旅行、休闲、人际交往、写作（发表意见、评论和文章等）等等莫不如此。在社会和国家的层面，经济、政治、军事、科学、教育、社会和文化活动，及人类对自然界的影响，都会留下可储存的海量数据。这些数据可以用来分析人类活动的特点和规律。因而，大数据必然会对传统社会科学的研究方式产生巨大冲击、挑战并提供新的机遇。本文从社会学和社会科学研究的基本目标入手来进行反思。

大数据不同于传统数据之处在于：它不是通过抽样调查所获取的样本数据，而是人类活动的

收稿日期：2016-05-24

① ［美］加里·金、罗伯特·基欧汉、悉尼·维巴《社会科学中的研究设计》，陈硕译，格致出版社、上海人民出版社 2014 年版，第 21 页。

② https://zh.wikipedia.org/wiki/大數據，2015-10-01。

实时记录，并大都可以通过互联网存储、获取、交换和分析。大数据是“由科学仪器、传感设备、互联网交易、电子邮件、音视频软件、网络点击流等多种数据源生成的大规模、多元化、复杂、长期的分布式数据集”①。大数据有多方面的来源，一般而言可以分作五类：企业公司数据，指来自公司企业的销售、交易等数据，比如阿里巴巴的销售数据、证券公司的交易数据等；网络数据，主要是指来自互联网、社交媒介的数据，比如 Facebook、Twitter、新浪微博等；期刊图书数据库，是指取自某一个具体的数据库的数据，比如 CNKI 期刊数据库、Web of Science、Google 图书等；政府数据，是指源自政府的总体数据，比如人口普查数据、全国用水用电数据等；其他，是指除上述四类数据之外的其他数据，但不包含抽样调查数据。有关大数据的基本特征，我们可以在与传统数据的比较中进行描述和分析。

1．样本与总体

和以往抽样调查获得的数据不同，大数据不是抽样数据而是一个总体数据。但是，这个总体是一定范围里的总体，而不是绝对总体。比如，人们通过京东商城购物，所有的购物过程都可以转化为数据，所得到的总体就是在京东商城发生购物行为的总体。这个总体不是所有网上购物者的总体，更不是包括线下购物者的全部购物者的总体。不过，有一些数据的总体，就是一个完整的总体。比如，美国国防气象卫星计划（Defense Meteorological Satellite Program）的夜间灯光图像数据，就是每天对地球进行扫描的数据，其平均灯光强度可以作为代表区域社会经济发展的指标，现有研究表明这一指标与 GDP 的相关度非常高②。这个数据的总体，就是整个地球。因而，我们不能笼统地说总体，而要具体看该数据所代表的总体是什么样的总体。这样所得结论的界限就比较明确。

作为总体的大数据，在统计上至少有两个意义：其一，它可以给抽样数据提供参照，纠正其偏差。抽样调查采用抽样数据推断总体，实际上，很大程度上对总体认识不清，并不知道推论的实际效果，只是根据统计的显著性来进行检验。大数据的出现为抽样数据提供了总体的基本特征，抽样数据可以与大数据进行比较，看到底有没有偏差，偏差有多大。所以，大数据给小数据（抽样数据）提供了一个标杆和判断的标准。其二，运用大数据进行统计时，显著性检验可能就是不必要的了，实际数据差异是多少就是多少，因为它就是总体。

2．结构化与非结构化

与人们的一般想象不同，大数据其实主要不是结构化的数据③，而是非结构化（含半结构化）的数据。“据统计，只有 5% 的数据是结构化的且能适用于传统数据库。”④ 非结构化的数据对社会科学研究提出了如下问题：其一，它对数据的分类、整理提出超越以往任何时候的技术要求和理念更新。其二，精确性与模糊性并存。“大数据要求我们有所改变，我们必须能够接受混乱和不确定性。”⑤ 大数据是精确性与模糊性并存的数据，可能精确的更精确，而模糊也是能够接受的。

3．单一与匹配

一般说来，大数据的数据比较单一，它仅包含有限的变量。比如家庭和企业的用电、用水记

① http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767，2015-10-29.

② Elvidge C. D.，Baugh K. E.，Kihn E. A，“Relation between Satellite Observed Visible-near Infrared Emissions，Population，Economic Activity and Electric Power Consumption”，*International Journal of Remote Sensing*，Vol. 18，No. 6，1997，pp. 1373-1379；Ghosh T.，Anderson S.，Powell R. L，“Estimation of Mexico's Informal Economy and Remittances Using Nighttime Imagery”，*Remote Sensing*，Vol. 1，2009，pp. 418-444；Chen X.，Nordhaus W. D，“Using Luminosity Data as a Proxy for Economic Statistics”，*Proceedings of the National Academy of Sciences*，Vol. 108，No. 21，2011，pp. 8589-8594.

③ 所谓结构化数据是指有机组织起来的、可识别、可搜索的格式化数据，参见 http://www.entity.co.uk/solutions/structured-data/。

④ ［英］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》，周涛译，浙江人民出版社 2013 年版，第 64 页。

⑤ ［英］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》，周涛译，浙江人民出版社 2013 年版，第 66 页。

录，通讯公司的手机消费记录，等等。但是，这些数据是可以匹配起来的。比如，通过通讯公司的手机（或座机）记录、网上购物记录和快递公司的送货记录，我们可以分析手机用户的网络消费情况，进一步也可以将其人际交往情况匹配起来，等等。如果匹配是可能的，那么将有可能改变大数据目前变量较少的状况。这种匹配，在技术上是可行的。问题在于不同数据的产权可能归属于不同的公司或部门，怎么解决数据交换的问题，就是一个市场交易的问题，是一个经济学和法学的问题。

在个体的层次上，将数据匹配起来，牵涉到个体的权利、隐私等问题。在组织（如企业、公司、学校、科研机构、政府组织和非政府组织等等）层次上、地区（如社区、城市、行政区划等）层次上，也同样存在上述问题。不过，非个体层次对隐私的要求没有那么严格，在现有条件下，数据的获得、使用主要受限于信息的不公开。

4. 容量、记录与面板

数据容量巨大也是大数据的基本特征。当前，大数据是指容量超过 1TB 或 1PB 的数据集，容量的界定是相对的，它会随着时间和数据类型有所不同，随着存储能力的提高，大数据容量的阈值也会提高，对大数据容量给出一个确切的阈值是不切实际的①。

大数据是人类活动的实时记录，与抽样数据如问卷调查数据相比较，它往往不是回顾性的，基本不受到人的记忆的干扰，所以，在这个意义上，它比问卷调查数据更准确。由于大数据是实时记录，所以它又具有时效性。由于大数据源源不断地产生，它又是面板数据，而且是间隔时间非常短暂的面板数据，这是抽样调查数据，哪怕是其中的面板数据所难以比拟的。

总而言之，作为人类活动实时记录的大数据是一个总体数据，它包含结构化数据和非结构化数据，一般容量较大，现实中单一的数据变量较少但可以进行匹配，许多大数据是源源不断涌现的面板数据。这些特征使得它区别于传统数据，并对人类活动和科学研究产生了难以估量的影响。2008 年，在《自然》杂志出版的专刊“大数据（Big Data）”中，费利斯·弗兰克尔（Felice Frankel）和罗莎琳德·里德（Rosalind Reid）指出，巨大的数据流埋藏着对新科学的启示，但是，我们需要发现的工具，比如透镜②。当然，大数据对科学的启示或冲击是从自然科学领域里开始的，但这种影响必然延伸到社会科学。

我们在收录了 SSCI 期刊的 WOS（Web of Science）数据库中，对社科类文献中涉及大数据的社会科学（包含社会学）文章进行检索，结果发现：在社会科学研究中，涉及大数据的研究还很少，2010 年至 2015 年 12 月以标题检索的总计仅为 249 篇，采用大数据进行实证研究的则更少，仅为 43 篇。这说明大数据研究刚刚起步，涉及大数据的文章 80% 以上还在讨论概念、特征和研究框架等初步问题。

在社会学学科中，共有 30 篇有关大数据的论文发表，其中 2011 年到 2014 年有 9 篇，2015 年则有 21 篇。其中实证研究仅有 2 篇，所用数据来自网络中的 Twitter，研究方法采用时间序列分析，所用软件为 R。这说明，在英文文献中，社会学的大数据研究也是刚刚起步。尽管大数据的研究并不多，但还是取得了一定的成果。根据斯科特·戈尔德（Scott A. Golder）和迈克尔·梅西（Michael W. Macy）的归纳，西方学界对大数据中的网络数据的研究（主要涉及传播学、心理学、社会学和政治学等学科）在三个方面有所进展③。

一是社会网络与传播研究。借助 Facebook、Twitter、邮件、电话通讯等数据提供的丰富的人口学特征和社会网，学者们验证了格兰诺维特（Mark Granovetter）“弱关系假设”和博特

① Amir Gandomi, Murtaza Haider, “Beyond the Hype: Big Data Concepts, Methods, and Analytics”, *International Journal of Information Management*, Vol. 35, No. 2, 2015, pp. 137-144.

② Felice Frankel, Rosalind Reid, “Distilling Meaning from Data”, *Nature*, Vol. 455, No. 4, 2008, p. 30.

③ Scott A. Golder, Michael W. Macy, “Digital Footprints: Opportunities and Challenges for Online Social Research”, *Annual Review of Sociology*, Vol. 40, 2014, pp. 129-152.

（Ronald Burt）“结构洞”理论。伊格尔（Eagle N.）等对 6500 万电话用户的通讯记录的研究表明，社区成员社交网络的多样性与其经济发展呈正相关，证实了社会网络理论①。乌干达（Ugander J.）等使用 Facebook 的社交网络数据发现，随着用户社交网络规模的不断扩大，用户之间的分割由 2008 年的 5.3 步下降到了 2011 年的 4.7 步，验证了“六度分隔理论”②。巴克什（Bakshy E.）等对 2.5 亿 Facebook 回帖数据的研究表明，新信息的传播主要通过弱关系③。相反，奥涅拉（Onnela J-P）等对 460 万手机用户的通讯记录的研究发现，尽管弱关系使社交网络联系了起来，但是大多数信息的传播都是通过中等强度的联结实现④。

二是社会交换、合作与信任的研究。巴克斯卓（Backstrom L.）和克莱因伯格（Kleinberg J.）随机抽取了 130 万成年 Facebook 用户来测试恋爱关系建立和维持过程中社交网络的“嵌入效应”。令人惊奇的是，他们发现“离散”（较少的重叠）而不是“嵌入”更容易产生恋爱关系，这与“嵌入关系理论”相悖，但却与博特的“结构洞”理论相一致，即那些能够填补结构洞的人更能吸引同伴⑤。

三是集体行动与社会运动研究。用户的网络互动数据为研究者检验集体行动理论、公共物品和博弈论的相关假设提供了良好的机会，并受到政府机构的极大重视。比如，冈萨雷斯（Gonzalez-Bailon S.）等使用 Twitter 和 Facebook 提供的数码痕迹来追踪“阿拉伯之春”中的抗议信息和公众舆论，因为通过追踪用户发布内容的转变可以用来衡量抗议动员的速度和程度⑥。迪格瑞齐亚（Digrazia J.）等关于地方选举的研究表明，当地共和党的选票与 Twitter 用户消息中出现“共和党”名称的次数呈正相关关系，社交网络数据为传统舆论调查提供了一个重要的补充⑦。

在国内的研究中，大数据研究相关文献并不太多。我们采用同样标准对 CNKI 数据库中收录在 CSSCI 中的文献进行搜索，结果发现：中文文献中大数据（共 1359 篇）相关的文章远远超出英文文献，但实证研究仅为 30 篇，少于英文文献。这说明，中文文献更是处于介绍、讨论基本概念、特征等初步阶段上。在中文文献中，社会学有 54 篇，但实证研究仅有 4 篇。王程韡使用 CNKI 数据库搜索“大数据”关键词，进行反事实分析，认为暂不能判断“大数据”是否能引领新科学范式的“大趋势”⑧。陈云松等人使用谷歌图书和社交媒体 Twitter 的数据，研究了中国城市的知名度和社会学百年来的发展情况⑨。这些研究具有一定的价值，但研究方法还比较简单。

总体来说，在中英文文献中，大数据的研究刚刚起步，多数文章还在描述大数据的特征，确定研究大数据的基本框架。实证研究非常少，也比较简单，不过也取得一定的成果。

当然，大数据对社会科学及其社会学的影响并不是表现为刚刚开始发表的少量论文，更为根本之处在于：它冲击或挑战了社会学和社会科学的基本理念、研究逻辑、研究方法与技术，或者

---

① Eagle N., Macy M. W., Claxton R, “Network Diversity and Economic Development”, *Science*, Vol. 328, 2010, pp. 1029-1031.

② Ugander J., Karrer B., Backstrom L., Marlow C, “The Anatomy of the Facebook Social Graph”, 2011-11-18, http://arxiv.org/abs/1111.4503, 2015-10-28.

③ Bakshy E., Rosenn I., Marlow C., Adamic L, “The Role of Social Networks in Information Diffusion”, *Proc. 21st Int. Conf. World Wide Web*, New York: ACM, 2012, pp. 519-28.

④ Onnela J-P, Saramaki J., Hyvonen J., Szabo´ G., Lazer D., et al, “Structure and Tie Strengths in Mobile Communication Networks”, *Proc. Natl. Acad. Sci.*, Vol. 104, No. 18, 2007, pp. 7332-7336.

⑤ Backstrom L., Kleinberg J, “*Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook*”, *Proc. 17th ACM Conf. Comput. Support. Coop. Work (CSCW)*, New York: ACM, 2014, pp. 831-841.

⑥ Gonza lez-Bailon S., Borge-Holthoefer J., Rivero A., Moreno Y, “The Dynamics of Protest Recruitment through an Online Network”, *Scientific Reports*, Vol. 1, 2011, p. 197.

⑦ Digrazia J., McKelvey K., Bollen J., Rojas F, “More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior”, *PLoS ONE*, Vol. 8, No. 11, 2013, p. e70449.

⑧ 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。

⑨ 陈云松 《大数据中的百年社会学——基于百万书籍的文化影响力研究》，《社会学研究》2015 年第 1 期；吴青熹、陈云松 《我国城市国际关注度的总体结构与特征——基于互联网搜索引擎和社交媒体的大数据分析》，《南京大学学报（哲学·人文科学·社会科学版）》2015 年第 5 期；陈云松、吴青熹、张翼 《近三百年中国城市的国际知名度——基于大数据的描述与回归》，《社会》2015 年第 5 期。

说，大数据对以往社会学及社会科学的研究范式形成很大挑战。

二、找回规律：古典社会学是否可以重生？

自 17 世纪牛顿力学和 19 世纪达尔文进化论以来，近代和现代自然科学逐渐成型并给社会科学带来深刻影响。如同自然科学家发现自然界的规律一样，社会科学家也力图发现人类社会历史发展规律。恩格斯在《在马克思墓前的讲话》一文中宣称“正象达尔文发现了有机界的发展规律一样，马克思发现了人类历史的发展规律。……马克思还发现了现代资本主义生产方式和它所产生的资产阶级社会的特殊的运动规律。”①

经济学家亚当·斯密、大卫·李嘉图和哲学家康德、黑格尔等人都在某种程度上认为经济学和哲学的基本目标或使命就是发现人类社会历史发展规律。亚当·斯密将功利主义视为“永劫不移的……原理”②。李嘉图“相信经济学的某些结论与‘万有引力原理同样确定’”③。康德相信，“大自然即使在混沌中也只能有规则有秩序地进行活动”④。而心中的道德律使他认识到“处于普遍必然的联结中”⑤。黑格尔认为，“‘理性’是世界的主宰，世界历史因此是一种合理的过程’。“‘景象万千，事态纷纭的世界历史’，是‘精神’的发展和实现的过程。”⑥

社会学家也是规律的探寻者。孔德认为“作为我们智慧成熟标志的根本革命，主要在于处处以单纯的规律探求（即研究被观察现象之间存在的恒定关系）来代替无法认识的本义的起因。”⑦ 迪尔凯姆认为“社会学研究方法的最基本规则是，要将社会现象当做客观事物来看待。”社会现象又可以分为“规则现象”和“不规则或病态现象”，其中内涵着规律⑧。马克斯·韦伯的看法有所不同，他认为，“社会学……应该被称之为一门想解释性地理解社会行为、并且通过这种办法在社会行为的过程和影响上说明其原因的科学”。但韦伯并不否认规律，他认为，如果统计结论能证明，那么“将来的科学研究也能……发现……规律性”⑨。

这种寻求社会历史发展规律的努力后来遭到许多学者的质疑，其主要理由在于质疑者提出，社会现象具有与自然现象不同的特征。

其一，整体性。波普尔认为，社会科学寻求宏观社会历史发展规律的企图，受到社会整体的困扰，“‘整体’绝不能成为科学研究的对象”。“如果我们要研究一个事物。我们就不得不选择它的某些方面。我们不可能观察或描述整个世界。”整体的理论或假设没有办法进行检验，“如果没有检验的可能性，那么，声称采取了任何一种科学方法，都是白说的。整体主义方法与真正的科学态度是不相容的”⑩。社会科学只能通过局部去研究整体，通过对个人的了解去研究社会。与整体主义方法论相反，波普尔认为，“社会理论的任务是要……依据每个人以及他们的态度、期望、关系等情况来建立和分析我们的社会学模式——这个设定可以称为‘方法论个人主义’”⑪。

其二，异质性。欧内斯特·内格尔（Ernest Nagel）指出，社会文化是相对的。社会现象具

① 〔德〕恩格斯《在马克思墓前的讲话》，载中共中央马克思恩格斯列宁斯大林著作编译局《马克思恩格斯选集》第 3 卷，人民出版社 1972 年版，第 574 页。
② 郭大力、王亚南《译序》，载亚当·斯密《国富论》，郭大力、王亚南译，上海三联书店 2009 年版，第 2 页。
③ 〔德〕克劳斯·迈因策尔《复杂性思维：物质、精神和人类的计算动力学》，曾国屏、苏俊斌译，上海辞书出版社 2013 年版，第 390 页。
④ 〔德〕康德《宇宙发展史概论》，上海外国自然科学哲学著作编译组译，上海人民出版社 1972 年版，第 14 页。
⑤ 〔德〕康德《实践理性批判》，邓晓芒译，人民出版社 2003 年版，第 186 页。
⑥ 〔德〕黑格尔《历史哲学》，韦卓民译，上海世纪出版集团、上海书店 2001 年版，第 8、451 页。
⑦ 〔法〕奥古斯特·孔德《论实证精神》，黄建华译，凤凰出版传媒集团、译林出版社 2011 年版，第 10 页。
⑧ 〔法〕埃米尔·迪尔凯姆《社会学方法的规则》，胡伟译，华夏出版社 1999 年版，第 13、45 页。
⑨ 〔德〕马克斯·韦伯《经济与社会》，林荣远译，商务印书馆 1997 年版，第 40、43 页。
⑩ 〔英〕卡尔·波普尔《历史决定论的贫困》，杜汝楫等译，世纪出版集团、上海人民出版社 2009 年版，第 59、62、56 页。
⑪ 〔英〕卡尔·波普尔《历史决定论的贫困》，杜汝楫等译，世纪出版集团、上海人民出版社 2009 年版，第 107 页。

有“‘受历史约束的’或‘文化上决定的’特征”①。受此影响，社会科学理论具有严格有限的应用范围，对一个社会的样本资料研究所得出的结论可能不适合另一个社会。“人类社会的差异性与特质性造成了社会科学具有情境性（contextual）和相对性的特征。”“情境差异和社会变迁在所有社会科学学科中都是两个重要的参量。这两个参量都反对雄心勃勃的范式性的概化。”②

其三，能动性或意向性③。社会生活中的人不是物体，他具有主体性或能动性，并具有特定的阶级立场和利益。波普尔认为，社会科学的客观性受到人的因素的干扰。“社会科学涉及社会偏见、阶级偏见和个人利益，所以在社会科学里，缺乏科学的客观性就至关重大了。”“在绝大多数的，或者在全部的建构社会理论中，人的因素将仍然是一个非理性的成分。”④ 内格尔认为，“人类由于获得了对他们所参与的事件，或对他们作为其成员的社会的新知识，因而经常更改他们习惯的社会行为方式”⑤。由于人的能动性，社会科学的预言可能改变人的行为：或者与之作对，或者有所加强；由于人的能动性，在进行实验、回答问题时，可能依情境而发生变化，这会影响到社会科学“应用数量方法的特殊困难，尤其是测量方法”⑥。克劳斯·迈因策尔（Klaus Mainzer）指出“在社会科学中，人们通常在生物学进化和人类社会历史之间作出严格的区分。原因在于，国家的、市场的和文化的发展被假定是由人的意向性行为所指引的，即人的决策是以意向性和价值为基础的。”⑦

因而，以波普尔为代表的一些学者，将马克思等人寻找社会历史规律的追求称之为“历史决定论”：这“是探讨社会科学的一种方法，它假定历史预测是社会科学的主要目的，并且假定可以通过发现隐藏在历史演变下面的‘节律’或‘模式’、‘规律’或‘倾向’来达到这个目的”⑧。波普尔认为，历史决定论根本行不通。他的观点产生了广泛而深远的影响。

在社会科学中重要的一门分支学科——社会学中，一些学者认为应该抛弃发现普遍规律的企图，尤其是与包罗万象的“帕森斯主义”决裂，回到墨顿的“中层理论”，关注具体的“因果机制”。就如埃尔斯特（Jon Elster）所说，“社会科学的重点将会有一个从理论推定到机制的重要转变……对所发生事情的描述会进入中观或微观层次。”⑨ 机制性解释的核心理念是，“不通过提出放之四海而皆准的社会规律或者寻求统计相关的因素来解释社会现象，而是通过探求那些可以展示出社会现象如何产生的机制来进行解释”⑩。

那么，大数据对社会学及社会科学寻求规律的研究宗旨或目标会带来什么样的影响呢？我们认为，波普尔对社会科学中的整体主义方法论的批评是有问题的：姑且不论是否存在整体主义方法论（他在这里可能对马克思主义存在误解），以人们只能认知局部而不能认知整体来推论不能得到关于社会历史规律的逻辑是错误的，因为自然科学也存在同样的问题，人类社会是一个整体，但作为自然界的地球、太阳系乃至宇宙也是一个整体，自然科学能，为什么社会科学就不能？

千百年来，尤其是近现代社会科学发展成型以来，人类对于社会的认知，从经验事实的角度

① ［美］欧内斯特·内格尔《科学的结构》，徐向东译，上海译文出版社 2005 年版，第 517 页。
② ［法］马太·杜甘《比较社会学：马太·杜甘文选》，李洁等译，社会科学文献出版社 2006 年版，第 19、278 页。
③ “意向性……常常被当作人类思维的根本性特征”“意向性是精神状态对于外部世界的事物对象或状态的参照性：我明白了某事物，我相信某事物，我期望某事物，我害怕某事物，我想要某事物，等等。有意向性的精神状态可以与没有任何参照物的无意向状态区别开来：我紧张，我害怕，我开心，我沮丧，等等。”参见克劳斯·迈因策尔《复杂性思维：物质、精神和人类的计算动力学》，曾国屏、苏俊斌译，上海辞书出版社 2013 年版，第 213、206 页。
④ ［英］卡尔·波普尔《历史决定论的贫困》，世纪出版集团、上海人民出版社 2009 年版，第 123、124 页。
⑤ ［美］欧内斯特·内格尔《科学的结构》，徐向东译，上海译文出版社 2005 年版，第 524 页。
⑥ ［英］卡尔·波普尔《历史决定论的贫困》，世纪出版集团、上海人民出版社 2009 年版，第 112 页。
⑦ ［德］克劳斯·迈因策尔《复杂性思维：物质、精神和人类的计算动力学》，曾国屏、苏俊斌译，上海辞书出版社 2013 年版，第 454 页。
⑧ ［英］卡尔·波普尔《历史决定论的贫困》，世纪出版集团、上海人民出版社 2009 年版，第 2 页。
⑨ 参见理查德·斯威德伯格《经济学与社会学》，安佳译，商务印书馆 2003 年版，第 331 页。
⑩ ［美］彼得·赫斯特洛姆《解析社会：分析社会学原理》，陈云松等译，南京大学出版社 2010 年版，第 26 页。

来说，首先来自于个体生活经验，其次来源于有限个案（质性研究），第三来源于抽样调查，第四来源于普查（例如人口普查）。但在互联网及相关设备发展普及之前，关于社会总体的数据少之又少。巧妇难为无米之炊，要从有限的经验事实或数据中得到关于宏观社会的总体认识，的确是盲人摸象，难之又难。以抽样数据推断总体的做法，总是难以避免偏差。因此而放弃对社会历史发展规律的探求而关注较为微观和具体的因果机制，也是可以理解的。互联网及相关设备（如传感器和微处理器等）的发展，对人类活动进行实时记录并储存起来，形成大数据，提供了认知总体社会的数据基础。在波普尔时代，从总体上来认知社会的确不可能，但现在具有了可能性，这是因为我们有了总体的大数据——“样本”＝“总体”。

以社会现象异质性较强而自然现象同质性较强来否定对社会历史发展规律的追求，也是建立在有限经验事实基础上的。自然界丰富多彩，人类社会也具有同质性①。“复杂性和非线性是物质、生命和人类社会进化中的显著特征。”② 异质性和同质性都是较为抽象的概念，笼统地说异质性（如文化异质性）并没有扎实的基础。应该将社会现象的异质性作为一个假设而不是一个前提，并通过对人类社会生活、历史发展和文化异同的大量经验事实的分析来检验这一假设。社会现象的异质性或同质性，绝对不是有限个案可以确证的。而大数据提供了检验这一假设的可能性。

人的确具有能动性、意向性、情境性和逆反心理，这是产生质性研究、抽样调查以及人口普查所得数据之误差的重要来源之一。但大数据恰好在这方面具有一定的优势或长处。“随着大数据分析取代了样本分析……当记录下来的是人们的平常状态，也就不用担心在做研究和调查问卷时存在的偏见了。”③ 大数据是人类活动的实时记录，和通过访谈等方式得到的数据不同，它更能排除获取数据时人的不诚实、记忆误差及环境干扰等因素导致的误差。

关于人的能动性和意向性对社会历史规律的影响，恩格斯早有论述。他认为，在社会历史领域内进行活动的人具有意识、激情，经过思虑、追求目的。但是，历史进程受内在规律支配。“无数的个别愿望和个别行动的冲突，在历史领域内造成了一种同没有意识的自然界中占统治地位的状况完全相似的状况。行动的目的是预期的，但是行动实际产生的结果并不是预期的，或者这种结果起初似乎还和预期的目的相符合，而到了最后却完全不是预期的结果。这样，历史事件似乎总的说来同样是由偶然性支配着的。但是……这种偶然性始终是受内部隐藏着的规律支配的，而问题只是在于发现这些规律。”④

恩格斯的论述可以概括为“结果稳定假设”。如前所述，迪尔凯姆也十分强调社会现象外在于个人的客观性。现代社会学和社会科学的众多研究证明，个人的意向性只是增强了社会现象的随机性、偶然性，而并不是没有规律可循。比如，人的迁移行为（国际移民，如中国人移民美国建立唐人街⑤；国内移民，如农民外出打工⑥）是有意向性的，单个人的迁移也可能是偶然的，但大规模的迁移行为则是有规律的。“国家的和国际的迁移效应不可能用单个人的自由意志来解

① 人类社会的语言丰富多彩，异质性很强。富特雷尔等人建立了包含 37 种语言的数据库，通过分析发现了“语言共性”（language universals），即相关联的概念在句子中会尽可能保持最近的距离，如此可以降低句子含义的理解难度。几乎所有的语言都以这种方式组织语句。Futrell，R.，Mahowald，K.，Gibson，E，“Large-scale Evidence of Dependency Length Minimization in 37 Languages”，*Proceedings of the National Academy of Sciences*，Vol. 112，No. 33，2015，pp. 10336-10341.

② ［德］克劳斯·迈因策尔《复杂性思维：物质、精神和人类的计算动力学》，曾国屏、苏俊斌译，上海辞书出版社 2013 年版，第 20 页。

③ ［英］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》，盛杨燕、周涛译，浙江人民出版社 2013 年版，第 42 页。

④ ［德］恩格斯《路德维希·费尔巴哈和德国古典哲学的终结》，载《马克思恩格斯选集》第四卷，人民出版社 1972 年版，第 243 页。

⑤ ［美］周敏《唐人街——深具社会经济潜质的华人社区》，鲍霭斌译，商务印书馆 1995 年版。

⑥ ［美］范芝芬《流动中国：迁移、国家和家庭》，邱幼云、黄河译，社会科学文献出版社 2013 年版。

释。”① 当然，恩格斯的“结果稳定假设”还需要证明或证伪。如果有足够的经验材料，比如大数据，这个假设就会得到进一步的检验。

大数据对于探讨人类行为和社会历史规律并且更为准确地进行预测还有一个非常有利的方面，那就是：在抽样数据中往往被删节的少量极端值在大数据中成为可以分析的个案或变量。帕特里克·塔克尔（Patrick Tucker）指出，“大数据可以帮你实现的，是找到拥有特定的行为模式和性格的人，而在小样本中你很难遇到——或许永远不会遇到，因为周围噪音太多了”。“当你的数据中有了足够的点，即便异常事件也可能显示出某种特征。”② 在小范围里的小概率事件，在一个大范围里可能就不是小概率事件，或者至少有较多的个案可以进行统计分析。这样，大数据就超越了小数据，将在小数据里被排斥的个案重新纳入分析框架之中。

总而言之，作为总体、实时记录和面板的大数据为重新发现宏观社会历史发展规律提供了以往所不具备的数据基础和可能性，在这个意义上，大数据可以重构社会学和社会科学的研究目标：它使得社会学、经济学和其他社会科学研究者至少可以发现或寻找人类活动的行为规律，并在此基础上发现社会历史的发展规律。

需要进一步澄清的是，承认社会历史发展具有规律并以此作为社会学或社会科学的研究目标，并不必然导致决定论。波普尔等人对古典社会科学的决定论性质的批评并非毫无道理。受限于当时的科学理念与发展水平，大多数古典哲学家、经济学家、社会学家及其他社会科学家将人类活动与社会历史规律看作是必然的，这当然具有决定论的特征③。区别决定论或非决定论，关键在于将规律理解为是必然的还是概率性的，而不是是否具有规律。人类行为、社会现象、历史进程的变化是有规律的，但不是决定论意义上的必然性，而是概率论意义上的可能性。

在大数据的研究中，一些学者指出了人类行为的可预测性、规律性。帕特里克·塔克尔指出，“人类行为的可预测性比任何人想象中的都要强。”④ 艾伯特-拉斯洛·巴拉巴西（Albert-László Barabási）同样认为，“人类行为遵循着一套简单并可重复的模型，而这些模型则受制于更加广泛的规律。”⑤

三、大数据要放弃对因果关系的追求吗？

任何科学都要追求因果关系解释，缺乏因果关系解释就没有规律。反过来，追求发现规律就必然要追求因果关系。休谟认为，因果关系“是我们从经验中得来的关系”⑥。发现因果关系的必要条件是：第一，“凡被认为原因或结果的那些对象总是接近的”；第二，“在时间上因先于果”；第三，原因和结果之间的“恒常结合”之“必然联系”⑦。休谟奠定了科学对于因果关系的基本理解。休谟所谓的“恒常结合”就是事物之间统计上的强相关关系。此后，经过密尔等人的发展，关于确立事物之间因果关系的标准就基本稳定了：“两个变量间存在因果关系，即一个变量导致另一个变量，如果（1）在时序上，因先于果。（2）两者间有实证的相关性，而且

① ［德］克劳斯·迈因策尔《复杂性思维：物质、精神和人类的计算动力学》，曾国屏、苏俊斌译，上海辞书出版社 2013 年版，第 17 页。

② ［美］帕特里克·塔克尔《赤裸裸的未来大数据时代：如何预见未来的生活和自己》，钱峰译，江苏凤凰文艺出版社 2014 年版，第 83、32 页。

③ “传统意义上的概率是定义在确定事件上的。在古典主义者看来，我目前患前列腺癌的概率要么是 0，要么是 100%。但是我们现在都是概率论者。专家们以前常说“是”或“不是”。现在我们为自己观点辩护时必须使用估计和概率。”参见伊恩·艾瑞斯《大数据：思维与决策》，宫相真译，人民邮电出版社 2014 年版，第 167 页。

④ ［美］帕特里克·塔克尔《赤裸裸的未来大数据时代：如何预见未来的生活和自己》，钱峰译，江苏凤凰文艺出版社 2014 年版，第 30 页。

⑤ ［美］艾伯特-拉斯洛·巴拉巴西《爆发：大数据时代预见未来的新思维》，马慧译，中国人民大学出版社 2012 年版，第 13 页。

⑥ ［英］休谟《人性论》，关文运译，商务印书馆 1980 年版，第 85 页。

⑦ ［英］休谟《人性论》，关文运译，商务印书馆 1980 年版，第 91、92、105 页。

(3) 因果关系不是第三个变量的结果。完全符合上述三个条件的关系，就是因果关系。”①

上述标准中的第 3 条，实际上就是要排除其他因素的干扰，确认就是原因对结果的影响而不是其他因素的影响。“研究特定原因的理想状态是什么？那就是所有其他的‘干扰’因素都消失的状态。……当所有的其他的干扰都不复存在的时候，原因就在它的行为中清楚地展示它的力量。”② 对于第 3 条的理解，也可以从反事实的角度进行：当有原因 A 时，会导致结果 B；当没有原因 A 时，则不会导致结果 B。因而，“因果关系问题实际上是一个反事实问题”③。

当然，社会科学通常是在概率的意义上来理解因果关系的。“统计学对因果关系表述为：在相等条件下，如果 A 发生，则 B 发生的概率提高，或者 X 变化导致 Y 平均值的变化。因果关系的必然性不表述为个体事件，而表述为群体概率或平均值和随机组试验的可重复性。”④

基于第 3 条标准，实验法成为确立因果关系最成熟的方法和手段。因为实验可以将实验对象随机分配到控制组和实验组，并排除外界其他因素的干扰。可是，社会科学的研究对象是人，造成了“进行实验的特殊困难”⑤。“在社会研究题材上进行受控实验的可能性极为狭小。”⑥ 因而，社会科学主要采用统计方法并结合其他手段来探寻因果关系。

从统计的意义上来探讨因果关系，就不是两个变量（一个因变量，一个自变量）之间的关系那么简单的事情，因为社会生活中几乎不存在单因单果的现象。统计控制就是要将可能对因变量（被解释变量）和自变量（关键解释变量）有影响的变量纳入模型。从统计的角度来说，因果关系的问题就转变成了因果效应。“当解释变量被赋予两个不同的值时，因果效应就是这些值对应的观察值中系统部分间的差异。”⑦ 在统计模型中准确估计因果效应主要受制于三个因素：

其一，样本选择性偏误。样本选择性偏误是由于缺乏科学的研究设计、非随机抽样、客观条件限制等因素引起的。这既可能是由于研究者的主观选择所导致（比如力图证明某一假设而只选取有利证据），也可能是由于客观条件限制（如没有好的抽样框⑧导致缺乏随机抽样的基本条件），还可能是尽管有一个好的研究设计，但由于操作过程中的失误所致。样本选择性偏误有两种基本的形式：其一，缺乏参照组或对照组，不能进行反事实分析。其二，只看到有限样本，而且是一个非随机样本⑨。就缺乏参照组来说，当下的许多大数据也是如此。比如，由京东商城购物者行为所形成的大数据，就是一个线上购物者的数据，且不说还有其他的线上购物（如亚马逊），如果要完整研究消费者的购物行为，那就缺少线下购物者这一参照组，即使只研究线上购物，也可能由于缺乏线下购物的对比而导致认识偏差。显然，大数据只对由于抽样引起的有限样本的选择性偏误具有一定的纠正作用。因为大数据就是一定范围里的总体，在理论上可以“收集所有的数据，即‘样本’=‘总体’”⑩。因此，它也纠正了对于这一总体抽样所导致的偏差。

统计学家们想出种种方法来解决样本选择性偏误，但最根本的解决方案之一是不需要抽样，换句话说，就是具有一个总体样本。而这恰好是大数据的优势所在。大数据如果是总体或全部样本的数据，那就从根本上解决了由于抽样偏颇所引起的样本选择性偏误。

---

① ［美］艾尔·巴比《社会研究方法（上）》，邱泽奇译，华夏出版社 2000 年版，第 100 页。
② ［美］D. 韦德·汉兹《开放的经济学方法论》，段文辉译，武汉大学出版社 2009 年版，第 346 页。
③ 谢宇《社会学方法与定量研究》，社会科学文献出版社 2012 年版，第 69 页。
④ 彭玉生《社会科学中的因果分析》，《社会学研究》2011 年第 3 期。
⑤ ［英］卡尔·波普尔《历史决定论的贫困》，世纪出版集团、上海人民出版社 2009 年版，第 111 页。
⑥ ［美］欧内斯特·内格尔《科学的结构》，徐向东译，上海译文出版社 2005 年版，第 503、507 页。
⑦ ［美］加里·金、罗伯特·基欧汉、悉尼·维巴《社会科学中的研究设计》，格致出版社、上海人民出版社 2014 年版，第 78、79 页。
⑧ 关于无抽样框的条件下怎么进行概率抽样，这方面的研究已经取得进展，可以参见刘林平、范长煜、王娅《被访者驱动抽样在农民工调查中的应用：实践与评估》，《社会学研究》2015 年第 2 期。
⑨ ［美］加里·金、罗伯特·基欧汉、悉尼·维巴《社会科学中的研究设计》，格致出版社、上海人民出版社 2014 年版，第 124—135 页。
⑩ ［英］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》，盛杨燕、周涛译，浙江人民出版社 2013 年版，第 37 页。

其二，变量遗漏。现实生活中的大数据往往只有几个简单的变量，其中一些数据只有客观变量（缺乏态度或评价性的主观变量），如果采用单一数据，变量遗漏问题会非常严重，甚至远远不如精心设计的抽样数据。不过，如果将不同的数据匹配起来，那么这一问题将在一定程度上得到缓解。匹配大数据在技术上是可以解决的，现实的问题主要在于数据的产权交换和数据使用的伦理，经过充分的讨论，这些问题是可以解决的。某一大数据变量简单或较少的问题，是测量标准、技术和设计的问题，这些问题是可以逐步改进的。现有测量，社会科学很少介入，在一定程度上导致社会关系指标或变量较少。假以时日，由于社会科学的进步和公众对此的认可，一些社会性的变量被列入、重视和普及，也是很有可能的。

其三，内生性问题。内生性问题涉及对于因果关系的基本理解。所谓内生性问题，是指“在一些情况下出现反向因果问题：解释变量受到被解释变量影响，而不是我们假设的影响被解释变量”①。我们认为，在简单、封闭、稳定和局部的系统，因果关系较易确定；在复杂、开放、动态和庞大的系统中，因果关系难以确定。因果关系之所以难以确定，主要是互为因果或因果关系相互纠缠的问题，也就是“内生性”问题。“许多社会变量具有相互作用的效果，因而因果关系通常是不能简单累加的。”②“预期的作用会导致因果关系难以在许多人类互动中定位。……由于人们会根据对他者如何行动的预期以及对自己行动结果的信念来调整自己的行为，因此经验性调查乃至因果概念的界定都变得非常困难。”③

所以，在复杂、开放、动态和庞大的系统中，因果关系的内生性问题较难解决，而在简单、封闭、稳定和局部的系统中，在统计模型中可以尽量避免内生性问题。过去和现在的社会科学研究模型，就是将纳入模型的有限变量视为与其环境相对隔离或独立的因素。由这样的模型所得出的因果关系，如果将其放入或回归社会环境中，很有可能发生变化。这也是社会科学研究预测难以准确的基本道理。进一步说：由有限数据得出的因果关系要接受大数据的检验。以往的研究缺乏大数据，所以检验就要多次重复地进行。

尽管在复杂、开放、动态和庞大的系统中，因果关系的内生性问题较难解决，但并不是说就一定不能解决，复杂与简单、开放与封闭、动态与静态、庞大与狭小、全局与局部，都是相对而言的。避免在抽样数据中所设置模型的内生性问题的原则与技术，也可能在大数据中能够得到应用，或者有所改进。而在大数据中能确立的因果关系，其稳定性应该远超于抽样数据的结果。

大数据对确定因果效应的有利之处还在于：大数据中的多数数据是面板数据，并且具有层次性，可以进行分层处理。

基于上述分析，我们不能同意所谓大数据不需探求因果关系而只是追求相关关系的说法。如维克托·迈尔-舍恩伯格（Viktor Mayer-Schönberger）和肯尼斯·库克耶（Kenneth Cukier）认为的，不是因果关系，而是相关关系④。

我们认为，相关关系和因果关系不是对立的，相关关系是因果关系的必要条件，因果关系是表明事物间作用之方向性的一种特殊的相关关系。事物间具有较强的相关关系，其中必然蕴含着因果关系，只是谁是因、谁是果，需要甄别，并要弄清楚因果关系的作用机制。当然，我们同意“相关关系分析本身意义重大，同时它也为研究因果关系奠定了基础”⑤。

① ［美］加里·金、罗伯特·基欧汉、悉尼·维巴《社会科学中的研究设计》，陈硕译，格致出版社、上海人民出版社 2014 年版，第 180 页。

② ［美］保罗·汉弗莱斯《社会科学中的数学模型》，载斯蒂芬·P. 特纳，保罗·A. 罗思主编《社会科学哲学》，杨富斌译，中国人民大学出版社 2009 年版，第 189 页。

③ ［美］罗伯特·杰维斯《系统效应：政治与社会生活中的复杂性》，李少军等译，上海世纪出版集团 2008 年版，第 X 页。

④ ［英］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》，盛杨燕、周涛译，浙江人民出版社 2013 年版，第 67 页。

⑤ ［英］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》，盛杨燕、周涛译，浙江人民出版社 2013 年版，第 88 页。

实际上，使用大数据是可以探讨因果关系的。约翰·格林（John Gerring）等人曾收集了一个覆盖国家、地域和区县的多层次的选举档案（the Multi-Level Election Archive，MLEA）来研究政体大小与民主的因果关系[①]。该文档记录了从 18 世纪到 2013 年间，88 个国家、2344 次选举、79658 个选区、超过 400000 场竞选活动的数据，是典型的大数据。通过普通最小二乘法（OLS）对样本总体进行的一系列检验表明，在多党竞选的地区，较大的选区能在更大程度上鼓励民主选举，选民规模对各政党的竞选力有正向的显著影响。这一发现与传统观点——政体大小与民主呈负相关关系相矛盾，为了证明作者的研究结论，文章对可能影响选举竞争力的其他因素，比如选民对反对党的偏好、竞选者的供给、选区的文化多样性以及候选人与选民的关系等因素进行了干预，在排除干预效应的模型里，选民规模仍然对政党的竞选力有正向的显著影响。随后，作者又用部分国家的议会选举数据和投票权改革数据验证了这一结论。网络数据同样可以用来探求因果关系，拉塞尔·纽曼（W. Russell Neuman）等人在 2014 年曾根据美国国家选举研究网站中涉及的 29 个议题，从传统媒体与社交媒介中获取了美国 2012 年全年各个议题的数据资料，这 29 个议题包含经济、外交事务、政治、公共秩序、社会问题和环境六大方面，其中，平均每天有 13362 条社交媒介的评论数据和 4573 条传统媒体的新闻报道[②]。作者运用格兰杰因果关系检验了传统媒体、社交媒介与公共议题之间的因果关系，研究发现，社交媒介是社会问题和公共秩序议题的动力，而在经济、外交事务、政治和环境议题方面没有一种媒体主导这些公共议题，社交媒介和传统媒体呈现出复杂的、动态的领先与滞后模式。此外，约翰尼斯·本德勒（Johannes Bendler）等人对 Twitter 用户数据的研究发现，某一兴趣点（Point of Interest，POI）（比如餐厅、酒吧、银行、博物馆等）与用户在该兴趣点发布的 Twitter 消息之间存在着因果关系[③]。

四、结论与讨论

基于上述描述和分析，我们可以得出如下结论：

（1）和以往抽样调查所得到的数据不同，作为人类活动实时记录的大数据基本不受人类记忆、偏好和情感干扰；大数据是一个总体数据，但大部分数据不是全球或全国范围里的完整总体，而是一定范围里的总体；大数据包含结构化数据和非结构化数据；现实中单一的大数据变量较少但可以与其他数据进行匹配，匹配的困难主要不在于技术，而是产权和伦理问题；大数据具有时效性，大数据大多是面板数据。

（2）中英文文献检索结果表明，大数据的研究并不多，但还是取得了一定的成果。比如，西方学界对大数据中的网络数据的研究就有所进展。有的研究验证并支持了社会网络理论中格兰诺维特的“弱关系假设”和博特的“结构洞假设”，有的研究验证了“六度分隔理论”，有的研究则提出了新的理论假设。

（3）古典社会学和社会科学理论力图发现人类社会历史规律。后来的学者基于人类社会活动的整体性、异质性、能动性或意向性对此提出质疑。在社会学领域，主流观点认为应该抛弃发现普遍规律的企图，回到中层理论，关注具体的“因果机制”。

大数据为社会学和社会科学重新发现宏观社会历史发展规律提供了可能性：它以“总体”数据提供了认知宏观社会的数据基础；它为社会现象的“异质性假设”检验提供了较为全面的数据；它以实时记录的特点排除了获取数据时的人为干扰；它也为恩格斯的“结果稳定假设”

① John Gerring, Maxwell Palmer, Jan Teorell, Dominc Zarecki, “Demography and Democracy: A Global, District-level Analysis of Electoral Contestation”, *American Political Science Review*, Vol. 109, 2015, pp. 574-591.

② W. Russell Neuman, Lauren Guggenheim, S. Mo Jang, Soo Young Bae, “The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data”, *Journal of Communication*, Vol. 64, 2014, pp. 193-214.

③ Johannes Bendler, Sebastian Wagner, Tobias Brandt, Dirk Neumann, “Taming Uncertainty in Big Data: Evidence from Social Media in Urban Areas”, *Business & Information Systems Engineering*, Vol. 6, No. 5, 2014, pp. 279-288.

提供了检验所用的充分的经验材料；它超越抽样调查的小数据，将小数据中被视为极端值并且往往被删节的个案或变量重新纳入统计分析。

作为总体、实时记录和面板的大数据也许可以重构社会学和社会科学的研究目标：它使得社会学、经济学和其他社会科学研究者至少可以发现或寻找人类活动的行为规律，并在此基础上进而发现社会历史的发展规律。但这种重构不是回到历史决定论，不是对规律作决定论的理解，而是概率论的理解。

（4）追求因果关系解释是科学包括社会科学的必然目标。由于作为社会科学研究对象的人的特殊性，社会学和社会科学很少采用实验法而主要采用统计方法并结合其他手段来探寻社会现象之间的因果关系。在统计模型中准确估计因果效应主要受制于三个因素：样本选择性偏误、变量遗漏和内生性问题。

大数据作为总体或全部样本的数据，有助于从根本上克服由于抽样偏颇所引起的样本选择性偏误。单一大数据变量较少，如采用单一数据，变量遗漏问题会非常严重；如果将不同的数据匹配起来，可以克服或缓解变量遗漏问题；尽管在复杂、开放、动态和庞大的系统中，因果关系的内生性问题较难解决，但大数据对因果关系的检验比有限样本的抽样数据更为稳健和可靠，避免在抽样数据中设置模型的内生性问题的原则与技术，在大数据中也能应用，甚至有所改进；大数据作为面板数据和分层数据，对于确定因果效应极为有利。因而，我们不能同意大数据不需探求因果关系而只是追求相关关系的说法。大数据对于社会学和社会科学追求因果关系的努力比抽样数据更为有利。

总体来说，我们认为，大数据是可以用来重构社会学和社会科学的研究宗旨和目标的。不仅如此，大数据对社会学和社会科学的研究逻辑、方法和技术、研究的组织方式及人员素质等都会产生深远影响。对此，我们略加讨论。

其一，研究逻辑。一般说来，传统社会学和社会科学定量研究的基本套路是假设检验，即提出假设，然后用数据去检验，这种逻辑被视为演绎逻辑。从抽样数据推论总体的角度看，也有人认为是归纳逻辑。在逻辑实证主义看来，真正使用演绎逻辑的是数学和逻辑学本身，得出的是先验知识；其他从经验事实中得出结论的都是归纳逻辑，科学就是这样，得出的是经验知识①。

我们认为，关于归纳还是演绎的争论并不是特别有意义，对于大数据来说，事先不提假设，直接从数据得出结论，是完全可以的；提出假设，比如从抽样调查的数据或理论演绎提出假设，再用大数据去检验，也是可以的，前文所述用大数据验证了“六度分隔理论”就是一例。

不管是归纳还是演绎，只要遵循科学的推理过程，都是可以的，在大数据研究中都可以使用。当单一大数据的变量较少，主要使用描述统计时，就主要是归纳逻辑；当某一大数据可以使用模型进行统计分析时，演绎逻辑可能就更为重要。

从演绎逻辑出发，使用大数据进行检验，可以称之为“理论驱动”；从归纳逻辑出发，使用大数据进行描述和分析，可以称之为“数据驱动”。两种逻辑并存，理论驱动和数据驱动并存，可能是使用大数据进行研究的一个特点。以往的抽样数据研究，主要是理论驱动和演绎逻辑；而质性研究，主要是数据驱动和归纳逻辑。大数据将两种逻辑结合起来，可能是其优势所在。

其二，研究方法和技术。大数据对定量研究方法的挑战目前可能主要是对当下定量研究所使用的工具（比如软件）形成冲击。主要用于抽样数据的传统软件不足以容纳这么大的数据量，难以进行计算，更难直接获取或抓取数据。所以，大数据对统计分析技术会有很大的冲击和促进，对计算机及其软件的发展有要求。

社会科学发展一个重要的推动力就是技术手段的进步。从技术的角度，对大数据的获取、存

① ［美］D. 韦德·汉兹《开放的经济学方法论》，段文辉译，武汉大学出版社 2009 年版，第 80 页。

储、交换、匹配、分析、建模，大数据分析对统计理念、技术和软件的要求，都会形成冲击①。

需要指出的是，大数据对质性研究方法挑战可能更为尖锐：一是，大数据给质性研究提供了源源不竭的数据，对以往质性研究限于有限个案的做法影响极大，至少不比对定量研究的影响小，只不过质性研究者对此往往认识不足。二是，传统的质性研究之所以有一席之地，原因之一是可以研究统计中的极端值。② 在抽样数据中，极端值个案数太少，难以单独进行统计分析，给质性研究留下空间。大数据可以提供大量极端值的个案数，因而可以进行统计分析。在这个意义上，大数据又可能压缩了质性研究的空间。三是，大数据中绝大部分是非结构性数据，也就是质性研究的基本材料，怎么对这些数据进行分类、处理，既是定量研究的难题，也可求助于质性研究深入、细致的分析；四是，大数据提供了总体的基本特征，对于质性研究将个案类型化、进而选择个案（抽样）提供了很大的帮助；五是，大数据的非结构化特征迫使人们从简单的二值逻辑走向多值逻辑，走向人工智能，大数据为人工智能的训练提供了数据基础。

其三，社会条件。当人类进入大数据时代，使用大数据进行社会科学研究时，也对其组织方式、管理方式、文化条件提出了新的要求。

我们不想抽象地谈论这些问题，而是结合中国国情进行讨论。我们认为，在大数据时代，中国具有一些有利条件，有助于使用大数据进行社会科学研究。这些条件是：（1）中国历史悠久，留下了丰富的史籍和其他文献③。（2）中国是一个人口大国，互联网发展较快，网民众多④，网络数据异常丰富。（3）中国社会变迁剧烈，人口流动迅速，变迁轨迹会产生很多新的数据。比如一个农民一辈子待在家里，就缺乏移动的轨迹，但是一旦外出务工，就会流动，并产生数据。（4）中国的市场经济已经激活了一批民营企业，他们对数据比较敏感，对于数据的获取、储存、分析产生了巨大的市场需求。但是，中国也有一些不利于使用大数据进行社会科学研究的条件。这些条件是：（1）中国传统哲学和文化观念不重视数据，坐而论道，以圣人之言为评判言论对错的标准，往往进行注释式的讨论，不重视实证研究。（2）中国从普通人群到专业人员大多以个人经验去做判断，从数据视角观察、分析和处理问题的人还比较少。（3）中国的社会科学更强调与自然科学的区别，较少强调要向自然科学学习，对科学理念、研究方法和技术的学习都不够。在社会科学领域没有形成定量研究传统⑤。（4）中国的大学、科研机构的层级组织机构，不利于建立扁平、横向、跨学科的大数据研究组织形式。因而，大数据时代对中国的科研体制、人的素质和文化观念都提出了新的要求并产生巨大冲击。

当下，大数据在商业、社会管理和科学研究等众多领域里蓬勃发展、方兴未艾，深刻地改变了我们的时代。这种发展还是初步的，却在科学研究领域构造了近乎无限的想象空间：它可能根本上颠覆千百年来人类从个体经验逐渐归纳进而认知宏观社会和自然界的思维逻辑，而以总体特征作为我们认知和思维的出发点；它以源源不断的实时记录给我们留下了人类活动的巨量数据，这些数据具有类似于实验数据的特征；它渗透到社会生活的各个领域，从而使得人们不能视而不见，听而不闻，而对经院哲学的纯思辨模式提出根本性质疑；它对传统的学科分类、学科版图形

① 关于研究方法和技术手段，可以参见 Cioffi-Revilla，C，“Computational Social Science”，*Wiley Interdisciplinary Reviews*：*Computational Statistics*，Vol. 2，No. 3，2010，pp. 259-271.

② 佩蒂格鲁指出，考虑到案例研究的数目有限，有理由选择那些极端情境和极端类型的案例。参见 Pettigrew，A.，“Longgitudinal Field Research on Change：Theory and Practice”，*The National Science Foundation Conference on Longitudinal Research Methods in Organizations*，Austin，1988.

③ 中国现存汉文古籍约有 19 万种。参见杨琳《古典文献及其利用》，北京大学出版社 2014 年版，第 321 页。

④ 截止 2015 年 6 月，中国网民约为 6. 68 亿人。资料来源 《CNNIC 发布第 36 次〈中国互联网络发展状况统计报告〉》，中国互联网信息中心，http：//www. cnnic. cn/gywm/xwzx/rdxw/2015/201507/t20150723_52626. htm，2015-07-23。中国网民占世界网民的比例约为 22. 27%。世界网民（2014 年底）近 30 亿。资料来源：ITU，“Measuring the Information Society Report”，2014，p. ⅲ，http：//www. itu. int/dms_pub/itu-d/opb/ind/D-IND-ICTOI-2014-SUM-PDF-E. pdf，2014-11-24。

⑤ 瓦特指出，借助于社交网络和计算机分析技术，21 世纪的社会科学有可能实现定量化研究，从而成为一门真正的科学。见 Watts，Duncan J，“A Twenty-first Century Science”，*Nature*，Vol. 445，No. 7127，2007，pp. 8-13.

成巨大冲击，并要求科学研究，尤其是社会科学研究必须与数据科学相结合，并创造新的研究和学习的组织形式；它对科学研究的定量化和工具化提出了不断发展的客观需求，并强力推动研究人员重新学习、终身学习；它将过去一切似乎是定论的东西重新变成假设，并且要接受其检验，由此可能颠覆以往的真理或常识，并生产出新的知识。

如果中国的社会科学还亦步亦趋跟随西方社会科学走的话，那就难以发展。直接进入大数据时代，是中国社会科学跳跃式发展的机遇。在大数据时代，中国社会科学和西方社会科学几乎在同一起点上起步，关键在于，更新理念、努力学习、改造和革新社会科学研究的组织方式。

（责任编辑：薛立勇）

Regularity and Causality: Rethinking the Challenge of Big Data for Social Sciences Research — A Case of Sociology

Liu Linping Jiang Hechao Li Xiaoxiao

Abstract: The big data studies in social science fields have just started and some preliminary achievements are obtained. Big data provides a probability of discovering the development rules of social history once again by sociology or social sciences. First, it offers the data basis of "heterogeneity hypotheses" and "stable results hypotheses" to understand the macro society and examine the social phenomena. Second, factitious disturbances in data collection process might be mostly overcame as it is characterized by instantaneity. Third, extreme values which are usually ignored in the sampled data are adopted in the statistical analysis. While from the perspective of casual relationship, big data is helpful to radically overcome the samples´ selective bias which caused by sampling bias. The matched data can solve the problem of missing variables. And, as the panel and stratified data, it is more beneficial, stable and reliable to examine the causality using the big data instead of sampled data. Therefore, this study has different opinions on the statement indicating that big data research only needs to study the correlation relationship rather than the causal relationship. Big data can reconstruct the research objectives of sociology or social sciences. At least, the regularity of human behaviors could be identified by sociology, economics and other social sciences studies according to the big data. Furthermore, the law of development of social history might be discovered. However, this type of reconstruction is neither the revisit for the classical historical determinism, nor the inevitable understanding of the regularity. It advocates a probabilistic pattern. The research logics, methods, techniques and relevant social conditions of sociology and social sciences are also discussed in this paper.

Keywords: Big Data; Regularity; Causality; Challenge; Rethink