唐文方：大数据与小数据：社会科学研究方法的探讨

2016-09-02 政治学人

点击「政治学人」可快速关注

微信号：zzxrbjtd

学人简介

唐文方，先后毕业于北京大学和芝加哥大学，分别获法学学士学位和政治学博士学位，曾就职于匹兹堡大学、北京大学、清华大学，和斯坦佛大学，现任美国爱荷华(艾奥瓦)大学政治学与国际问题研究斯坦利华夏讲座教授。

【摘要】大数据时代的到来使得学界兴奋不已，传统的小数据研究似乎一下子变得微不足道。虽然大数据有着诸多的优势，但在短时间内无法取代抽样调查和实验研究等小数据研究在社会科学中的地位，原因在于技术发展的限制、社会科学研究的特点和人类社会的复杂性。最好的解决办法是多种方法的结合，但要做到这一点，研究人员必须首先了解什么样的方法能解决什么样的问题。

一、来势汹汹的大数据

随着互联网的普及，信息总量正以空前的速度爆炸性增长，人类社会进入了一个可以用“BB”(brontobyte，千亿亿亿字节)为单位的数据信息新时代，即大数据时代。从社会科学研究的角度来看，大数据是指巨大而多样化的数据集(Armah，2013)，是对全世界每一个人所做的每一件事的即时记录。大数据在网络时代正在成为可能，人们生活中的每一个空间正在越来越多地成为网络空间(Graham andDutton，2014)，例如分居不同地点的家人和朋友可以通过网络视频进行24小时的免费连接，购物、理财、银行业务、工作、娱乐、信息的获取都是通过网络来实现，而这些“cyber”行为都会留下“digital footprint”，也就是说大量的信息数据(Mayer-Schonbergerand Cukier，2013)。而在网络时代之前，这些以往物理意义上的行为很难被观察、测量和记录下来。试想，如果全世界每一个人的生命全过程和

每一天24小时的所作所为都被以图像、文字、视频或音频的方式记录下来并且汇总到计算机上，那么现有的数量词就已经无法形容其量之大，恐怕要创造新的量词了，这就是所谓的“大数据”。

在政治社会科学研究领域，大数据可以提供大量的宝贵信息(Galderisi，2015)，例如网路媒体中民众政治意见的表达、政治信息的传播和获取、社会动员与社会网络联络(Lohr，2012;Cantijoch、Gibson andWard，2014)，选举动员、竞选宣传、选民投票、社会运动与群体行为的产生和发展，以及政府与民众的互动、公共政策的制定等等。

大数据的标志性优势是研究者不必担心数据的代表性问题，即，大数据试图展示的是“全数据”，同时可以进行“大背景”式的可视化展现。以往，社会科学的研究对象即便是整个社会、国家或人口，由于技术条件和人力物力的限制，通常只能在总人口中抽取一个有代表性的样本(随机抽样)，或者在实验室中对更小的人群进行各种测试，从而采集研究数据。但抽样调查无论其样本多具代表性，仍然免不了有误差，实验室的研究对象更是无法代表整个人口。大数据的出现，似乎颠覆性地解决了传统数据不具代表性的问题，因为大数据是总人口的数据，不存在抽样，因此不存在误差的问题。例如，研究者可以对CCTV《新闻联播》全年的节目进行全文本的分析，由此生成年度出现词频最高的词语，从而对特定时期的中国政治语境做出全面的描述。

大数据的显著优势是可以从大海里捞针——也就是海量数据的检索功能。从理论上说，大数据试图收集的是全体人口的信息。由于计算机技术的高度发达，研究者可以在很短的时间内处理海量信息，包括整个人口的银行转账、ATM机的使用、电话通讯、商店购物等等，从而鉴别个别人的非正常行为，这对防止恐怖主义等对人类社会危害极大但又比较罕见的行为会起到很好的效果(Akhgar etal. ，2015)。此外，大数据大海捞针的功能还可以使研究者搜寻具有某种特征但是数量相对稀少并且地域分布广泛的特定人群，例如有吸烟历史的艾滋病女性患者。

在处理海量信息的过程中，大数据还可以揭示出一些被忽视的相关联系。例如，通过对《新闻联播》海量信息的处理，研究者不仅可以发现不同关键词出现的频率，还可以发现这些关键词之间的相互联系，可以在成百上千的关键词之间建立一个语义网络图(邵、张，2015;孟、郭，2015)，从而让我们十分清晰地发现各个关键词之间的关系。

最后，大数据的另一个优点是可以避免社会调查中的敏感问题。例如，社会调查中经常会遇到诸如访问色情网站、堕胎、使用翻墙软件、购买盗版光盘、逃税漏税等等，受访人常

常不会如实回答;但在大数据中，人们“难以启齿”的这些信息可以通过特有的技术手段，诸如用浏览器的cookie统计出来。

二、大数据的局限性

姑且不谈大数据在挖掘、收集和分析上的困难，即使我们通过日益完善的科学技术，可以收集到每个人的每个行为，并且有足够的计算能力来分析这些数据，大数据还是会面临诸多问题:

第一，大数据并不是在所有时候都是“全数据”。大数据有些时候收集的是总人口的信息，而不是像抽样调查那样依靠随机样本，例如人口普查。然而，在公共管理和决策研究中，人们越来越多地利用人的网络行为而产生的“大数据”来分析民意，但网络用户并不是全部人口。虽然网络用户在人口中的百分比正在迅速增长，但远非百分之百。例如，在2012年的世界价值观中国部分的调查中，只有40%的受访者通过网络获取政治社会新闻信息，而通过电视获取同样信息的比例则高达88%;这40%的人更不是能代表中国人口的群体，他们具有高学历、低年龄和白领阶层的显著特征(Tang、Zhang and Martin，2015)。显然，如果用这一群体的网络行为来推算中国人口的信息摄取偏好和民众意见的分布，就很有可能出现偏差。在上述例子中，那60%才是“沉默的大多数”。

第二，大数据并不是大家都可以用。绝大多数的大数据推崇者认为它是比较容易得到的，因为大数据是公开、透明的。的确，在有些情况下大数据是可以得到的，例如一些政府网站为了提高信息透明度而公布的数据，然而在很多其他方面，大数据涉及个人的隐私、商业机密或国家安全(Rahman and Ramos，2013;Webb，2015)，因此不光有侵犯个人隐私的道德问题要处理，更重要的是根本无法得到许多涉及商业或政府行为的数据，即便是某个研究者通过个人渠道获取了此类信息，也只能做自己的项目，无法共享。无法共享就说明你的结果无法被别人验证，因此不具权威性、可信性。此外，研究人员如果只能依靠政府公布的大数据来寻找研究题目，而无法根据本领域理论和现实的需要来设计研究计划，将会大大限制社会科学研究的发展。大数据的可用性将是长期存在的问题，不会因为收集数据的科技手段的提高而消失，并对大数据的广泛使用产生很大的局限性。

第三，大数据并不意味着数据的多样化。大数据的支持者认为，大数据时代的特点不仅仅是数据量上的庞大，还具有数据种类来源多的特点(Armah，2013)。这是因为如果人们生活中对网络的依靠越来越多，那么人们生活的各个方面，包括工作、家庭、社会生活、经济生活、政治行为等等都会被记录下来，从而使数据变得多样化。然而，大数据无论有多么全

面，它只能记录人的行为，却不能确切描述人的思维。在社会科学研究中，很多时候研究的关注点是人的主观态度和价值观，例如人际信任、社会公平观、政治效能感、民族主义情绪等等。这些概念都是社会科学关注的重要理论问题(因变量)，也是社会现象和个人行为的重要解释变量(自变量)，但是大数据在测量这些主观态度方面却显得有些力不从心(Rahman and Ramos，2013)。可能有人会说:网络评论不是也可以反映人们的态度吗?但是，如上所述，网络意见的问题在于网络用户是一个特定的人群(高学历、低年龄的白领)，他们的意见不具普遍性。况且，网络表达意见的范围是有限的，常常是就事论事，无法涵盖社会科学中关心的其他理论问题。

第四，大数据重相关而轻因果(梁，2015;孟、郭，2015)。大数据有些时候可以通过数据在时间上的先后来描述各种现象之间的因果关系，例如日本官员参拜靖国神社导致中国网民的愤怒和民族主义情绪的上涨，食品安全隐患导致民众对厂商诚信度评估的下降，节日使得民众的消费水平上升等等。但在很多其他问题中，大数据却无法确切地建立变量之间的因果关系，有时候甚至会导致虚假的因果关系。例如，如果研究者发现在一个大数据库中，冰淇淋的销量与群体事件之间有很强的相关系数，因此得出结论说吃冰淇淋会导致群体事件，这显然是有问题的，起码是很勉强的结论。在大数据库中，由于数据量庞大，通常很容易得出统计意义上显著的回归系数，但这并不意味着两个变量之间存在因果关系(Marcus and Davis，2014)。

第五，大数据特别是以网络为基础的大数据不能准确反映人的社会政治行为。原因有三:首先，有些人认为，基于网络社交媒体而收集的大数据可以用来预测社会运动等社会政治行为。例如在“阿拉伯之春”期间，人们成功地运用社交媒体建立了社会网络，并利用此网络表达和宣传革命理念，最终成功地组织和发动了一场社会革命。然而，大数据无法回答的问题是:同样有社交网络和网络意见表达的国家中，为什么有些发动了成功的社会变革(例如突尼斯)，而有些却没有(例如埃及)?显然，网络行为不是社会运动成功与否的决定因素。人与人面对面的互动以及在社会组织和环境(例如教会、工作场所、社区等等)中产生的“强联系”(strong tie)，才是社会运动产生的更重要原因。大数据所能检测到的网络联系，只是一种“弱联系”(weak tie)，不足以来预测诸如社会抗争这一类高风险的社会行为(Gladwell，2010)。其次，大数据无法反映言行不一这一问题。例如，在对上述《新闻联播》2013年内容的分析中，“美国”并未成为显著的关键词。这与人们的政治常识不一致，可能是节目编排、时段安排、宣传的目的造成的。最后，大数据分析中研究人员的主观因素会导致分析结果的误差。目前对大数据的分析，很多是对大数据的内容进行归纳分类，从而得出有关数据中的各种趋势的结论。例如，研究者可以对某一事件的所有微博留言进行分类，从而得出公众对此事件的看法(King、Pan and Roberts，2013)。然而，不同研究人员

对不同的留言可能有不同的理解，因此对留言的编码也会不同，从而使研究的结论发生变化。换句话说，大数据中的趋势并不是大数据自身固有的，而是会受到研究人员主观因素的影响。

三、“小数据”的必要性

从上面的论述中，我们既看到了大数据的新奇与优势，也看到了大数据并非全能。事实上，大数据无法取代以抽样调查和实验研究为代表的传统的“小数据”研究，两者的关系是相辅相成的。

首先，大数据只能被动地挖掘、收集已经客观发生了的行为信息，而抽样调查和实验研究则可以“制造”数据。例如，在小数据研究中，研究人员可以根据自己的理论需求设计问卷，并测量受访人对不同问题的看法和态度;而大数据只能局限于每个人对一个固定事件已经表达的意见。此外，小数据研究不仅能收集已经发生的事情的数据，还可以收集并未发生、或发生几率渺茫的事件信息，比如通过情景设置的方式或实验的方法来检验受访者在假设情景中可能的态度和行为，这显然是大数据研究很难做到的。再者，小数据在收集受访人观念、态度和行为方面数据的同时，还可以收集他们各方面的个人基本信息，例如家庭、工作、收入、政治面目、宗教信仰等等，这些信息为解释受访人的其他行为和观念提供了更多的可能性;而大数据研究无法根据研究者的需要来收集个人信息。从这个意义上说，小数据比大数据更适合进行具有理论意义和理论突破的研究。

其次，抽样调查的样本在特定情况下比某些“大数据”更具有代表性。所谓抽样调查，就是以总人口为基础，用科学的方法，随机抽取样本。好的随机样本应该符合总人口的基本特征，例如性别、年龄、教育程度和地区的分布等等。而通过网络收集的“大数据”，无论数量上再庞大，也不过是总人口中的一个特定群体，即网络用户。如前所述，这一群体通常是低年龄、高学历的白领阶层，哪怕他们有成千上万甚至上亿的数量，他们的意见仍然不能代表总人口。往往只有几千人的随机抽样的样本，虽然具有一定的误差，但研究者可以通过数学、统计方法来估算和减少误差，至少使得抽样数据接近理论上所讲是代表总人口的。因此这些人所表达的意见，比大数据更具有普遍性。

第三，小数据研究在因果关系的分析上别有特点。大数据虽然可以通过网民对某一事件的反应来确定此事件和公众态度的因果关系，但却无法确定这两者之间是否有中介变量，比如说不同年龄、学历或职业的区别，而这在小数据研究中却很容易做到。在很多数情况下，大数据中只存在相关性，而不是因果关系，比如前面所举的冰淇淋和群体事件的例子，数据

中无法确定是吃了冰淇淋才去参加群体事件还是群体事件导致冰淇淋销量上升。在这方面，近年来社会科学中越来越流行的实验研究有着独一无二的作用。研究人员可以在实验室中对受访人加入一定的实验条件，然后观测受访人是否受到实验条件的影响，从而确定实验条件与受访人态度或行为之间的因果关系(例如通过观看环境公益广告来确定受访人环保意识的变化)。实验研究的一个问题是受访人一般数量很少，不具代表性。为了解决这个问题，近年来，人们开始将实验研究植入到抽样调查中，从而解决了大数据研究无法解决的因果关系和普遍性的双重问题。

第四，小数据能更好地规避学术伦理的问题。大数据表面上很容易获得，网络是公共场所，谁都可以去，但现实并非如此。如果想真正获得有价值、可以根据自己的理论兴趣做分析的多变量大数据，就会涉及个人的隐私、商业或政府的机密以及个人权利、经济利益和政治敏感性等问题。因此，大数据可能永远也达不到其支持者最初设定的条件，也就是数据完全的公开和透明。没有了这两个前提，大数据的幻想就很难实现。相反，抽样调查和实验研究遵循受访人自愿、匿名的原则，所产生的数据的所有权属于研究者，使用起来不受其他人的限制。

四、结论

大数据时代的来临，表面上打断了原有的以抽样调查和实验研究为基础的社会科学研究方法的发展路径。网络的发展和对人们生活的不断渗透，使得大数据的挖掘和收集成为可能。大数据的特点是它的整体性、即时性、全面性和数量上的庞大。在社会科学研究中，特别是在公共政策和公共管理的领域中，人们越来越多地利用网络媒体产生的大数据来研究选举、民意分布、社会运动、社会网络、政治动员以及恐怖组织的形成和发展等等重要问题。

但是，大数据不是万能的，从一开始就包括至少五个方面的局限性:(1)基于网络用户的“大数据”无法代表总体人口的特征;(2)大数据由于侵犯个人隐私、涉及经济利益和国家安全，而无法实现彻底的公开和透明;(3)大数据无法对人们的思想状况的各个方面做出准确的描述;(4)大数据只能对人们不同行为的相关性做出描述，而在多数情况下无法确立事件之间的因果关系;(5)大数据所基于的行为有时候无法代表人们在社会中的真实行为。

传统的问卷调查和实验研究至少可以从四个方面弥补大数据中的上述不足:(1)抽样调查和实验研究不仅可以被动地收集已经发生的数据，还可以主动为研究者“制造”数据;(2)抽样调查的随机样本具备普遍性和人口的代表性，并从多方面收集受访人的个人信息，使调查数据更具多样化;(3)实验研究可以更准确地确立各变量之间的因果关系;(4)抽样调查和实验

研究数据的所有权归研究者，可以随便使用，而大数据的使用则受到多方面的限制。比较各种方法的优劣，得出表1中的结论。

从长远来看，大数据不仅不会取代小数据，而且必须依靠小数据才能得到发展。例如，大数据可以提供新闻媒体内容的语境描述及其历史变迁，但却无法呈现新闻媒体对受众的影响，受众研究中必须借助问卷调查和实验研究等手段，而问卷调查和实验研究则可以借助大数据所发现的关键词、相关联系等更有针对性地设计问卷和实验条件。

文章来源 《中山大学学报：社会科学版》2015年6期

本期编辑 吉先生

政治学人·专业的学术分享平台

编辑团队微信号：zzxrbjtd

原创投稿、文章荐稿邮箱：zhengzhixueren@sina.com