大数据给社会学研究带来了什么挑战？

邱泽奇*

编者按：

这篇文章是根据2015年5月29日邱泽奇教授在北京大学社会学系的一个讲座整理而成。为了缩短篇幅，在整理中删除了重复的、缺乏信息的内容。

今天跟大家分享我的研究成果，我对大数据的观察，不是扫盲。为了让大家听起来尽量没有障碍，也加入了一些知识性的东西，因此，也是和各位交流。我想和大家讨论三个问题：第一，什么是大数据？人们说的很多，错误的概念也非常多，我想澄清大数据是什么。第二，大数据和社会学研究到底有没有关系？对这个问题，人们也有比较多的想法，同样也有很多误解，我要说说我的观点。第三，重点谈一谈，大数据对社会学研究的重点带来什么挑战？大数据带来的挑战特别多，对社会学研究而言，到底有什么样的挑战呢？

一、什么是大数据？

首先讨论大数据到底是什么？

大家听的很多，了解的却不是特别系统和具体。对社会学家而言，最熟悉的是社会活动。我称之为人类活动的造痕。人类的任何活动都会留下痕迹。考古学研究在各地挖墓，挖各种各样的东西，那些东西都是人类社会生活留下的痕迹，我们拿它作为证据，探讨当时的社会生活。历史中，人类社会生活留下的痕迹绝大多数都消失了，挖出来的墓，在整个人类墓地的亿分位数都不

* 邱泽奇，北京大学社会学系教授。

到。因此,如果你说你掌握了过去人类社会的多少痕迹,我觉得千万不能大胆讲,是因为你真的不知道你到底掌握了多少。

我举一个例子,譬如周原。我有一个博士生,我让他回答一个简单却不能简单回答的问题:中国的村庄为什么三千年不散,如今却突然就散了?在过去三千年里,村庄始终是人类社会生活、人类聚集生活的一个状态。我希望他借助考古数据来做。北京大学考古学文博学院一直在探索陕西省的周原遗址。周原,过去三千年来一直有很多村庄,如今依然还是村庄状态,但很快就会消失。三千年来,村庄生活留下了痕迹。能够保留下来的痕迹,通常被称为证据。考古学、历史学都用证据,社会学也用证据。社会科学其实都用证据。这些证据,通常也被称为数据。不仅考古发现是人类活动的数据,历史档案也是人类活动的数据,譬如人口普查。不少人以为是美国人发明创造了人口普查,其实不是。中国在两千多年前"废井田、开阡陌"就开始登记人口了。在两千多年的行政历史里,户口登记是一项重要的、涉及众多公共事务的制度。

数据既然很早以前就有了,怎么就冒出来大数据了呢?

一个简单的回答是,实时地网络化汇集、网络化存储和网络化运用人类行为的痕迹,这才构成了大数据。

什么叫大?麦肯锡从行业和业务以及价值链的角度给了一个定义,说大数据是生产力的来源。如今,各行各业都在讲"互联网+","互联网+"背后有一个非常重要的概念大家可能容易忽略,叫"数据驱动"。在社会学研究中,过去,我们很熟悉"理论驱动";现在,数据驱动已经变成了非常重要的概念了。

麦肯锡定义的关键点叫消费者盈余浪潮。过去,我们从石油里找财富,后来从机器里找财富,再后来从其他东西里找财富,现在可以从数据里来找财富了。

其实,业界流传的故事说,"大数据"概念是从 IBM 来的。从学术研究的立场出发,可以对大数据概念的出处存疑。不过,IBM 的确用 4 个维度给大数据概念下了一个明确的定义:数量(volume)、形态(variety)、价值(value)、速度(velocity)。我认为,这是从数据出发的定义。

学术研究通常要按照学科规训理解,我也按自己的方式来理解,我给大数据概念一个定义:痕迹数据汇集、存储和运用的并行化、在线化、生活化和社会化。前面我之所以交代痕迹数据,希望说明的是,数据从来不缺。大数据是把过去数据的汇集、保存、利用方式做了一个很大的改变。不能说颠覆,现在颠覆为时太早,但它的改变确实非常重大。

汇集、存储和运用的并行化是一个计算机和网络科学的概念。什么叫并行?其实很简单,北京四环上的四条车道同时跑车就叫并行,如果只有一条车道跑,就不叫并行,叫串行。并行,指同时运行 2 个或多个线程。在计算机学科里叫线程,在交通学科里叫车道。

在线化也是一个计算机和网络科学的概念,指始终在网络上,数据的汇集、存储和运用都是在线状态。社会学的人都知道组织结构的科层制特征。可是网络里的组织结构则不同,总体上看起来是科层制的,实际运行却是网络状的,且不同的网络结构混杂在一起。在线化意味着数据的汇集、存储和运用,都在混乱结构的网络上。

生活化则是一个社会学的科学概念,是说数据的汇集、存储和运用已经渗透到了社会生活的方方面面,无处不在、无时不在。不仅生产活动在汇集、存储和运用数据,如企业产品生产、商店产品销售;生活活动也在汇集、存储和运用数据,如大家日常生活对计算机、手机、网络、家用电器的使用等。

社会化也是一个社会学的科学概念,指社会的大多数成员都参与了数据的汇集、存储和运用。系统和科学地搜集数据,是社会学的专长之一。过去,都是由机构、科学家去搜集。如今,每个人都是数据提供者、存储者,同时也是数据的运用者。譬如导航,你在运用道路数据的同时,也在提供和存储道路数据。

不过,理解痕迹数据汇集、存储和运用并行化、在线化、生活化和社会化的前提是理解 IBM 概念的 4V。下面,我先沿着 IBM 的 4V 概念做一个简单的说明,让各位对大数据在外观上有一个感知。

首先是量。大数据指其超出了任何个人在可接受的时间和范围内汇集、存储和运用数据的能力。我给大家一个基本概念,2012 年,单一数据集已经从兆级(MB),跃升到 TB 级,从 MB 到 TB,中间还有 GB。如果谈大数据,至少是 PB 级数据。任何个人计算机、小型服务器、大型服务器,没有单机可以处理 PB 级数据。为汇集、存储和运用数据,并行化和在线化是其目前的解决方案。

在进一步讨论前,普及一下信息计量单位。字节(bytes)是基本计量单位,相当于货币里的一分钱,每满 1 024 个单位,向上提升一级,上一级为 KB,之后有 MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB 等,简单地说,以 2 的 10 次方晋级。

从直立行走到 2013 年,整个人类积累的可利用数据量大约为 5EB,可 2013 年生产的数据量却达到了 800 个 EB。据统计,全球 90%的数据是在过

去两年生产的,其中社交网络、传感器、科研、金融都在产生越来越多的数据,几乎是每两年数据量翻一番。

其次是形态。传统的调查数据通常是结构化数据。结构化数据也是一个计算科学的术语。如果熟悉 SPSS,就比较容易理解,通常可以形式化为一个二维表,第一行是变量(又叫字段),从第二行开始到结束,就是每一个变量的案例值,形成了一个规整的变量值矩阵。熟悉调查数据的都知道,如果一个值没有对应的变量,就麻烦了,没办法处理了。结构化的特点就是这样。

大数据不是结构化数据,是混合形态的数据。什么叫做混合形态数据?指既有结构化数据,也有其他形态的数据。结构化的数据指各类结构化的数据库表,工业计算和科学计算常见的都是结构化数据,像甲骨文和 ERP 都有自己的结构库表,随时可以通过输入字段查询,比如说在北京大学要找人,找郭志刚,依据结构库表的约定,输入郭志刚三个字的首字母马上可以定位到郭志刚。逻辑是,在姓名字段里给了两个值,一个值是郭志刚的汉字,一个值就是郭志刚的汉语拼音首字母缩写,也许 GZG 三个字母对应很多人名字,其中一定有郭志刚,这是结构化的。

大数据不完全是结构化的,有一部分是结构化的,如姓名、账号、存款余额、消费记录等等,但大多数是非结构化的数据,比如说日志,查了几回,刷了几次卡,每次在哪里刷的,不是结构化的,刷了多少钱却是结构化的,刷了几次不是。每一位用户都有使用日志,有的还有音频,比如说微信中的语音,音频数据不是结构化的,图片不是结构化的。用户应用活动的很多数据都是非结构化,这就让数据变成了混合形态,这是不同于传统数据的非常重要的区别。

接下来,从商业视角来看数据的价值。传统的数据通常是分析目标导向的数据,有非常明确的价值取向。譬如我做中国家庭跟踪调查(CFPS),非常明确,搜集与人类社会生活、未来成就、幸福相关联的各种变量数据,有非常明确的价值指向。

大数据是记录导向的,是一个颠倒。大数据是为了技术活动、获得人类社会活动的痕迹而记录数据,获得是造痕者留下的并行数据(parallel data);不是为了解释某个现象、分析某个结果来记数据。在数据获取上,这又是一个非常重要的变化。

影响这个变化的因素,第一是记录的便捷化,无须研究者花钱花资源去搜集数据,每一个用户自己就主动提供了数据。第二是存储的便宜,存储的价格在过去的一段时间里呈指数曲线下降。

正因为大数据不是有目的的测量，而是造痕者留下的痕迹，因此，它的价值密度与社会学的调查数据比较便低得多。如果希望用大数据来证明什么，就需要从数据中去挖掘、去发现，而不是用假设检验的方式来检验。跟传统的调查数据比较，其基本的出发点是有区别的。通常认为，大数据价值密度比较低，从商业角度来看，的确如此；从学术角度，却不一定。

最后，非常重要的特征是速度。传统的数据，从设计、调查、清理到可用需要相当长的时间。举一个例子，1887—1890 年，赫尔曼·霍尔瑞斯为统计 1890 年人口普查的数据，发明了读卡机，把原本需要 8 年人口普查活动用一年的时间完成了。再譬如 CFPS，发动了几百位访员，用计算机采集数据，从调查结束到可用也用大概 2 年的时间，其中数据清理的时间非常长。

大数据，那么大的量，怎么处理？这是非常大的挑战。此外，大数据不同于传统数据的另一个特点是没有数据概念，只有“数据流”概念。这是社会学研究需要换脑子的关键点。什么意思呢？数据每时每刻都在产生、记录，没有一个时间节点的数据是完整的数据，因为，它根本就不是以完整数据为目的的数据，每时每刻都有数据可用，也都有它的约束性。其中的一个约束性是，它不是针对具体研究问题的可用数据。如果要研究一个问题，可以截一段数据出来，却不是马上就可用的数据，而是可以挖掘的数据。

不管大数据有什么样的特征，本质上，它还是数据，是人类社会生活包括私密生活留下痕迹的数据化。痕迹数据变成大数据有一些条件。第一个条件是行为的监测化，一旦造痕者的行为与数字化设备关联在一起，就具有了可检测性，比如说银行数据、社交数据、健康数据、家居数据等等。很多人喜欢戴手环，手环就是一个监测设备。如果你有什么自己不愿意让人知道的行为，建议你最好把手环摘掉。手环，不仅可以监测你的身体参数，也可以记录你活动的地理位置参数。

第二个条件是监测和检测的网络化。如果只是局部监测，问题不大，天知、地知、你知、我知而已。一旦监测设备具有网络功能，监测活动便让任何造痕活动变成了网络活动，甚至是在你不知情的前提下。比如说手机，现在每个人都在用智能手机，你们把设备上的位置选项打开看一看，默认状态是开启的。你说不愿意让自己的活动变成网络活动，问题是设备的功能你不一定完全了解，它可能随时随地都在把你的活动变成网络活动，监测的网络化就是社会活动的网络化过程，也是这个世界的连通过程，一个典型的例子是微信的朋友圈。

第三个条件是网络的数据化。如果仅仅是造痕活动的网络化倒也罢了,最多是知晓范围的扩大。问题是,网络化的过程也是数据化的过程。造痕活动的网络化首先是活动的数据化,其次是活动数据的网络化。单个节点的数据,常常不具有社会意义,节点数据的汇流便让造痕活动具有了社会意义。比如说,某个老师每周到办公室来两次,根据 GPS 信息,可以知道他什么时间到,什么时候离开,中间离开几次。如果这个老师有一个特别去处,每周固定的时间都要去。作为同事,我不知道,可手机运营商完全了解。依据也是这个老师手机提供的位置数据。当把所有人的位置数据汇集起来,可以知道的事情就多了。不仅可以知道有多少人有特别的去处,也可以知道每个的生活习惯、工作习惯、身体状态等等。

大数据其实与人类的社会行为相伴随,与网络同在,与社会一体。我想,从社会的视角来看,这就是大数据。

简单归纳一下,大数据,形态是数字化的、非结构化的、在线的、流动的数据;容量都在 PB 级以上,是单个计算设备无法处理的数据;来源,不是专门搜集的数据,而是与行为相伴生的、通过传感器、设备获取的数据、通过网络汇集的数据;不过,并非系统、也非完整的数据。

对社会学而言,大数据是一种新的研究数据来源,一种永不停歇流动的数据,目前还不是对过去其他来源数据的全面替代。

我给大家几个例子,大家了解、体验一下什么是大数据。

2014 年双十一。阿里自己造了一个云,叫 ODPS 云,这个云和世界上其他云不一样,用几十万台个人电脑阵列,运行着自己的系统,在 6 个小时内处理 100PB 数据,相当于处理一亿部高清电影。在零点以后,支撑了每一秒有 7 万瞬时订单,让 5 万个人同时抢 1 千件商品不超卖;3 分钟成交额 10 亿人民币,不出任何差错;在 570 多亿的交易中,支持了 243 亿的交易额在手机上完成,产生了 2.78 亿个物流订单;全球有 217 个国家和地区加入交易。这些事情如果不了解的,甚至都不敢想象,而且都是智能化的。

阿里还造了一个数据系统,叫聚石塔。这个聚石塔干什么呢?直接管订单,2013 年的双十一只有 75%的订单在聚石塔上处理,没有丢单;2014 年处理的比例上升到 95%;2015 年的双十一,估计全部都在这上面。

所有这些活动,都在实时发生,也在实时处理。发生的便成了数据,处理的也是数据。流动着的数据量,是传统社会学想象不到的量级。能够完成这些工作的就是计算能力,这个能力是人类在两年前都无法想象的。

二、大数据和社会学研究有关系吗?

接下来讨论大数据和社会学研究有没有关系? 我的观点是:有关系,目前还没那么紧迫。

咱们都是社会学的老师和学生,却常常"只缘身在此山中",忘记了社会学基本范式的差别。为理解大数据与社会学研究的关系,需要简要回顾社会学的基本范式,然后再说明,如果大数据与社会学研究有关系,那么,与什么范式、有怎样的关系。

在社会学的想象力下,我把社会学的基本范式分成三大类,与传统区分的实证、诠释、批判,不大相同,纯粹是为了叙述的方便。第一类,我叫做思辨的社会学,比如说帕森斯(T. Parsons)的宏大社会系统,甚至福柯(M. Foucault)的多种理论,甚至吉登斯(A. Giddens)的社会结构理论等。这些社会学大家,都是从概念到概念的思辨,基本上可以完全隔绝数据。再譬如布迪厄(P. Bourdieu),早年做教育社会学研究时用数据,后来也不怎么用数据了,抽象了,思辨了。

第二类,我称之为诠释的社会学,从胡塞尔(E. G. A. Husserl)以降,舒茨(A. Schutz),甚至到格拉霍夫(R. Grathoff),这些人都围绕意义在做研究。对他们来讲,一个现象本身的代表性是没有意义的,他们观察的是一个现象本身,要阐释这个现象的意义,他们认为的意义。他们也可以不用数据。不过,我认为对意义的挖掘也会面对意义社会性的挑战。

第三类,我叫做实证的社会学,源于法国年鉴学派和美国社会学对帕森斯的反动。在第二次世界大战以后,获得了空前的发展。如果要在实证社会学与前两类之间进行区分,很简单,有没有假设检验是一个关键特征。实证社会学强调假设检验,强调用经验事实检验理论假设。由于在检验中要使用数据和统计方法,也因此被贴上了"定性"或"定量"的标签。

大数据与社会学关系最密切的是最后一类。实证社会学离不开数据,不管是什么类型的数据,什么形态的数据。刚才说,实证社会学在二战以后有一

个大发展,大家可以看一个趋势。我用了两份文献,一份是普莱特的一部著作[①],她对美国社会学三份主流期刊(ASR, AJS, Social Force)的研究显示,1915—1924 年期间,35%的研究用个案,53%的用统计;1955—1964 年期间,用个案的下降至 18%,用统计的上升到 76%,其中 ASR 和 AJS 基本上排除了纯粹的社会理论文章,只要涉及社会事实的,都要有数据,不管是什么形态的数据。一份是中国的文献,北大社会学系的林彬教授和他硕士研究生王文韬的研究显示,2000 年,实证化的趋势在迅速加强[②]。现在的《社会学研究》没有证据的文章基本上发不出来。

对经验事实的刻画需要测量,对理论假设的检验需要测量数据,实证和数据密切地关联在一起,实证研究需要数据。可是,当我们对数据本身进行系统考察时却发现,数据并非因研究需要而产生。我的观察和探讨显示,数据最早源于管理活动的需要,后来慢慢地渗透到了社会科学的研究,直接影响了实证社会学的研究。

实证社会学过去的研究数据主要来自调查活动。二战以后,密歇根大学建立了社会研究院(ISR),调查数据开始逐步成为社会学研究的基础设施。在运用调查数据进行社会学研究的发展中,还有过一场辩论。基什(L. Kish)认为,与其花很多的钱进行人口普查,不如花少量的经费进行抽样调查。基什把自己对抽样调查的思考和想象,在世界各地传播,他也到过北京大学,我记得是 20 世纪 90 年代中期,专门在北京大学讲了一周的抽样调查,我学了整整一周。

调查数据还是社会学家手里的一类资源、一种权力。在大数据之前的数据,主要有三个来源,分别代表了三种资源和三个群体中手中的权力。第一是行政数据,各个政府掌握了身份数据、流动数据、登记数据、家庭数据,等等。第二是商业数据,譬如过去几百年的金融数据,都在商业公司手里。社会科学家到 20 世纪 30 年代才认识到数据的重要性,开始找数据、调查数据、运用数据,在搜集和运用数据的经历中,也认识到数据是研究者手中的资源,是让社会学声音具有独立性的支持力量。进而,与行政数据和商业数据一起,形成了三足鼎立之势。

---

① PLATT, JENNIFER. A History of Sociological Research Methods in America: 1920—1960 [M]. New York: Cambridge University Press, 1996.

② 林彬,王文韬. 对当代中国社会学经验研究及研究方法的分析与反思——90 年代社会学经验研究论文的内容分析[J]. 社会学研究,2000(6):38-48.

大数据是痕迹数据的一种，与实证社会学研究有非常密切的关系。哥德尔和梅西 2014 年的文章认为，数据脚印是社会学研究的挑战，同时也是机会[①]。有兴趣的可以找来读一读。我则认为，总体来讲，大数据的确给社会学研究带来了挑战，不过，现在还没有直接构成威胁。到底带来了什么样的挑战呢？接下来，我们做一些讨论。

三、大数据给社会学研究带来了什么挑战？

（一）还需要做调查吗？

我想，人们有兴趣的第一个问题是，还需要做调查吗？数据来源于问题。的确，大数据无须调查，只需选择。数据调查是有目的、有研究假设的数据搜集和研究活动。对于大数据而言，没有任何人做研究假设，也没有任何人有能力做大数据的研究假设。在这个意义上，与调查数据关注如何搜集数据不同，对大数据，对研究而言，关注的是如何应用数据。

前面讨论过大数据的特征，使得个体研究者不具备接触大数据的机会。对大数据的应用，现在主要是机构性的应用，尤其是商业机构，商业机构站到了第一线，阿里巴巴的大数据应用在世界范围内也是一流的。阿里有人曾经在一个内部会议上说，如果不顾及中国，不待在中国这块土地上，完全可以把中国的银行淹死掉。为什么呢？阿里手里有超过四亿消费者的金融信息，他知道谁要买什么，怎么买，花多少钱，大概什么时间段买。

与商业应用不同，学术研究还没有走到 PB 级数据的台阶。如果有谁告诉你说他在用大数据做研究，你先问问多大的数据量。一个问题，你就知道他是不是用大数据在做研究。对大数据，社会学家们既然没有可及性，或许也没有相应的技能，还能干什么呢？就我所知，自称在用大数据的，通常是大数据中的数据。社会学家不是像网络科学家和计算科学家那样，把建好的模型直接放到网络上运行，譬如百度导航的交通状况，而是从大数据中截取了一段数据在做研究，是大数据的一个小样本。即使如此，也与传统的调查有了很大的区别。我们依然可以把这样的数据看作是调查数据，不过，有诸多的不一样。

---

① GOLDER, SCOTT A, MICHAEL W M. Digital Footprints: Opportunities and Challenges for Online Social Research[J]. Annual Review of Sociology, 2014(40):129.

“访员”,传统的调查数据是访员询问受访对象,搜集数据;现在没有访员了,而是让机器自己汇集数据,研究者直接挑数据。

我举几个例子。第一个是哈佛大学金教授(Gary King)及其同事做的一项研究[①]。他们用社交媒体的数据来观察中国的沉默表达。数据从哪儿来呢?用网络爬虫直接从不同网站爬数据,获得了1 382个社交媒体网的数据。这项研究的影响还是很大的。

接下来,是我做的一项研究。2012—2013年,我跟淘宝做了一个好玩的研究,没有写文章,有一份很有趣的报告。淘宝希望了解店家的成长可能性,譬如年销售额10元的是不是可能做到100万,我希望了解谁在开网店[②]。我们从600万个淘宝店家数据中抽取6万个店家。从大数据中提取了6万个店家的交易数据,还对6万个店家进行了网络问卷调查。我得到的结论是:居住在沿海、城镇、年轻、中高学历、家境殷实、价值观居中的人在开网店。一年换三分之一的店家,能够坚持干的人,是用淘宝来谋生的人。在所有店家中,三分之一玩票,三分之一投机,三分之一谋生。

第三个例子,是我指导并参与的一项研究,通过分析并行数据,我们发现一个调查行为:臆答[③]。什么叫臆答?臆答是指,调查员找到了受访对象,并且跟受访对象聊了半天,不过,不是按照访问规程询问和填答,而是根据闲聊获得的信息,凭借自己的猜想代替受访者填问卷。这种填答作弊的方式,在传统的调查质量控制中是发现不了的。并行数据对访问行为的记录,让研究者有机会在访问行为数据挖掘中获得一种快答模式,通过对访员的询问,才获取了臆答作弊模式。

这三个例子都说明,即便是大数据中的数据,对社会学研究而言也是重要的。

(二)大数据来自哪里?

如果想做研究,从哪里可以获得大数据的数据呢?要回答这个问题,我们首先需要知道大数据到底来自哪里?

---

① KING G, JENNIFER P, MARGARET E, et. al. How Censorship in China Allows Government Criticism but Silences Collective Expression[J]. American Political Science Review, 2013,107(2):326-43.

② 在网上搜索“谁在开网店”即可以获得研究报告的各种版本。

③ 严洁,邱泽奇,任莉颖,丁华,孙妍.社会调查质量研究:访员臆答与干预效果[J].社会学研究,2012(02):168-81.

第一个是传感器。2005 年大约是 1.31 亿个，2010 年增加到了 30 亿个。总数是多少，不知道。因为，传感器的用途越来越广泛。什么叫传感器呢？马路边上的各类探头，刚才讲到的手环、手表、手机、电脑、汽车、空调、电饭煲、插座、灯等，只要是器具，都可以变成传感器，任何可以做数据监测、整合、传输的工具都是传感器。不过，它的基本原理来自射频原理，所以叫射频传感器。

第二个是互联网。根据不同来源的数据，我们了解到谷歌每天要处理大概 24PB 的数据，百度每天新增 10TB 数据，处理 100PB。

第三个是社交网络，像 Facebook 每天要 23TB，推特每天 7TB，腾讯每天大概增加 200～300TB，数据总量大概 100PB。

第四个是电信。中国移动产生 10TB 以上的话单数据、30TB 以上的日志和 100TB 以上的信令数据。其中，话单是结构化数据，有姓名、接入基站、通话时间、计费等，是结构化的数据。但日志不是，日志是非结构化数据。信令也是非结构化数据。

第五个是金融。每一个交易周期，纽交所捕获的数据量只有 1TB，没有很大的数据量。

第六个是网络销售。淘宝每日订单量大概 1 000 万，阿里巴巴已经积累的数据超过 100PB。

第七个是科研。比如说，欧洲核子研究中心强子对撞机每秒产生大约 40TB 数据。

第八个是政府。美国政府大概拥有 800PB 以上数据。在美国，除了商业公司，美国政府大概是第二位拥有海量数据的数据源。

分行业的列举，只是希望给各位一个印象，从比较中认识到，与传统的三足鼎立之势比较，在大数据时代，科学研究，尤其是社会科学的数据量是非常可怜的，你想找一个 PB 级数据的机构都找不到，几乎没有。要找一个 PB 级的社会学研究数据，我估计你在全世界都找不着。

为进一步让各位了解数据的来源，给大家两个图示。第一幅（图 1）是一分钟在因特网上有多少事（what happens in an internet minute），第二幅（图 2）是一天的每一分钟互联网上人们做什么（every minute of the day）。两幅图，大家在网上都可以找到。给大家举一个例子，比如说苹果，一分钟会有 4.8 万个 APP 被下载，你就知道数据量有多大了。

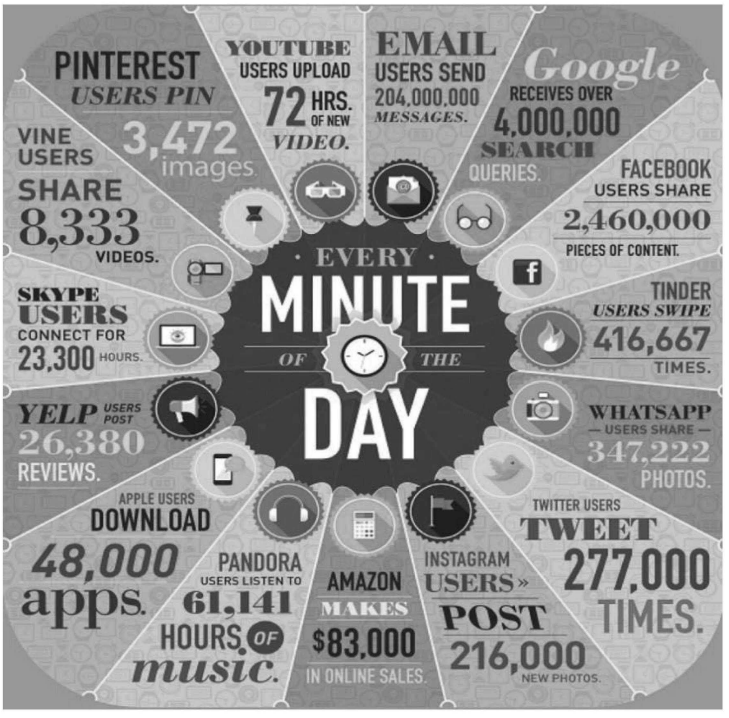

图 1 “一分钟在互联网上有多少事”

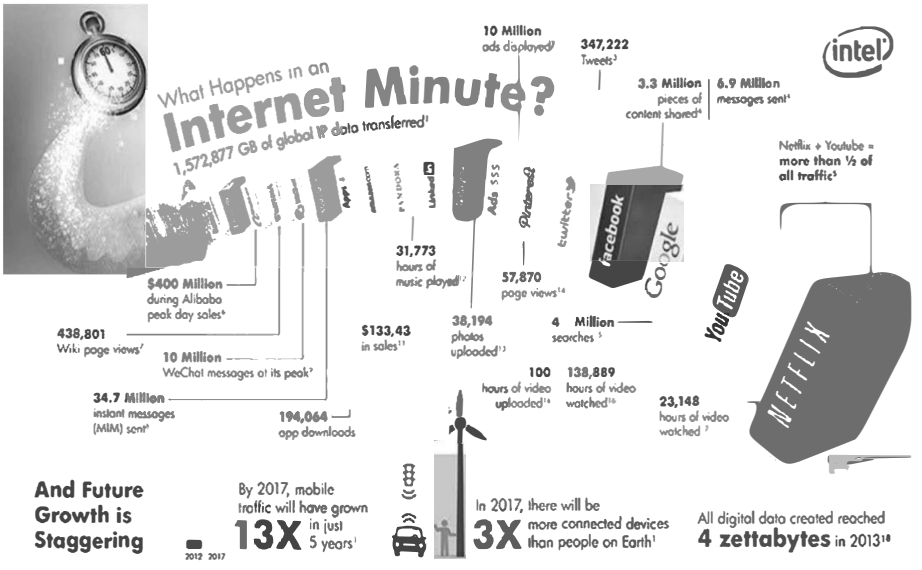

图 2 一天的每一分钟

(三)大数据的挑战到底在哪里?

我的观点是,大数据对于调查数据的挑战取决于它对调查数据的替代程度和扩展程度。

常用的调查数据,是小数据。大数据跟它有交集,现在问题在哪里呢?这两个数据都是可用的研究数据。对于社会学研究而言,至少是现在,我们用大数据的机会非常小,我们没有大数据。好在,我们还有小数据。接下来的问题是,两个数据的交集重叠的部分会怎样发展变化(参见图 3)。如果调查数据完全被取代,则社会学研究的技能包括思想便需要重来,这将是最大的挑战。否则,社会学研究还可以依据小数据继续发展。

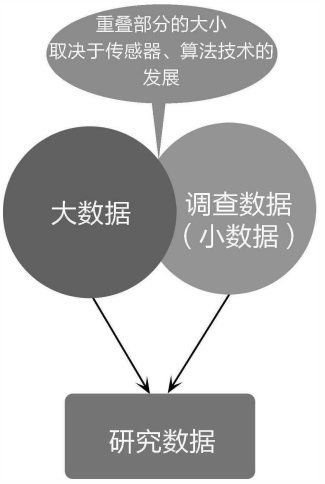

图 3　大数据、小数据与研究数据的关系

大数据对小数据的替代取决于两个因素,一个是传感器的应用,一个是算法技术的发展,两者的发展都会直接影响社会学未来的发展走向。

对于调查数据而言,譬如人口普查、民意调查、社会调查、健康调查,等等。这些调查通常用于做什么呢?对个体,用于研究人的行为、健康、教育、成就、幸福,大概就是这些事;对群体,用于研究群体的行为、结构和动态;对社会,研究社会的状态、结构和动态。大数据对社会学研究的影响在于,大数据有没有可能替代调查数据用于我们要研究的内容。

那么,大数据可以用于什么研究呢?譬如社交网是人的基本人情网络或人际网络;然后生活网,买东西、卖东西、刷卡;工作网络,每天上地铁、上班,在

办公室面用电脑;还有健康网,大家手里戴的手环,等等。貌似个体和群体的数据都在了,只是这些数据现在都在商业公司手里,不在研究者手里;而且,只要不与商业公司发展利益发生冲突,商业公司也不在意学者们在说什么。还有,这些数据还没有互联互通,如果互联互通,商业公司的力量将更大。万物皆比特[①],数据就在那里了,只是看怎么用、谁在用,最重要的是,社会学家们有没有机会和能力利用这些数据。

未来社会学研究对数据的利用,除了取决于机会和能力,还取决于数据化覆盖的范围。如教育,在线教育,大家现在可能还没有感受到危机,坦率地说,危机已经在我们身边了,各种内容产业的兴起就是对教育潜移默化的挑战。

如健康。未来的健康将是完全数据化的健康。中医,过去是望闻问切、号脉,现在中医也依靠检测设备,也依赖数据了,西医的检查数据,中医一样要看,看完了中医给另外一个解释,不是西医的解释。从穿戴式设备到专业社会,都在把人类的健康状态数据化,不少商业公司都在这个领域努力,包括互联网巨头们。

如物联网。什么叫物联网?大家不要误会了,以为是物流。物流不是物联网,物流传递叫物流网。物联网指器物之间的连接与互动。谢老师手里拿着手机,说我马上要回家了,把家里的空调打开,手机上有空调的应用,按一下,通过网络,指令就传到了家里的空调上。万物之间的连接和互动,就是物联网,是未来社会社会生活非常重要的一个领域,也是非常大的一个领域。

其中一个,就是无人机。无人机应用已经非常广泛,从军事到民用已经非常普及了,最近有几个小伙子在深圳搞了一个四翼无人机,是做拍摄的。其实,不止做拍摄,什么都可以做,有一部电影大家可以看一下,是一个真实故事,讲在美国佛罗里达有一个军事基地,在这个基地里面驻扎着一批无人机空军。他们干什么呢?像运用游戏操纵杆一样,操纵无人机。在阿富汗、也门,实施精准轰炸。里面有一个镜头很有意思,里面有一个阿富汗塔利班成员,那人是一个独狼,经常干坏事。他隔段时间会跑到一个人家里强奸女人。操纵无人机的上校第一次在屏幕上看到时很郁闷;第二次,也放了他;第三次,他预估这个人又要实施强奸了,便对操作员们说,没事了,你们休息吧。他一个人操作三台机器,把这个人干掉了。

操作无人机,运用的就是物联网。社会学研究通常只关注社会的逻辑,不

① 詹姆斯·格雷克.信息简史[M].高博,译.北京:人民邮电出版社,2013.

关注器物之间的逻辑，可如果不了解器物之间的逻辑，未来我们就无法理解社会，这也是对社会学研究非常严峻的挑战。

如硬件，智能硬件。任何硬件都有它的智能特征，只要可计算，背后都带着智慧。智能空调是硬件，无人机也是硬件。社会学家们无须像计算、网络、工程学家们那样精通硬件，可也不能不了解背后的互动逻辑。

如工程。现在很多工程都不用人干了，直接用数据来干了。举一个例子，比如说手术，过去的手术都是靠医生操作的精准度；现在，手术的一半要借助仪器，未来可能完全用机器人了。比如说切半页肺，病人躺在手术床上，马上开始 3D 扫描，把所有数据输入计算机网络，大数据开始调集所有相关案例数据，医生输入一个参数，如病灶在什么地方，从哪儿切到哪儿，机器人便开始操作，把你胸打开一个口子，直接把东西切完，止血之后，把打开的口子缝上，手术就完成了，就这么简单。

建筑工程也一样。大家看到，越来越多的智能工程机械在从事建筑活动。

制造也一样。中国制造 2025 的目标就是智慧制造。大家千万不要想象着，那只是制造业升级，不是，那是用智能替代人的一个过程。

还有农业。我刚大学毕业的时候在农场工作，每年有三个月时间需要飞机洒农药，我在飞机上干了两年，每年干一两个月。干什么呢？帮助飞行员洒农药。飞行员说，你打开阀门，我就打开阀门。农用飞机，机舱温度高达 40 多度，热得要死。现在，都是无人机。无人机一航拍，说产量怎么样，什么地方有虫灾，什么地方有病灾，该怎么洒农药和施肥，清清楚楚，农业也完全改变了。这不是简单的互联网，而是多种科学数据的联合应用，背后都是数据之间的逻辑。

再说金融。大家最有兴趣，昨天股票跳水将近 7%。其实，谁知道股票会跳水，上交所和深交所的人都知道，证监会也知道。事先他们都知道，只是鉴于交易规则不便透露而已。

简单地说，大数据在社会生活中的渗透是完全彻底的。回到第一个问题，还要不要调查的问题，除了前面分析的交集的面积大小、大数据对小数据的替代以外，还有两个概念，分享一下。

第一个是“转换”，转换数据、转换思维。数据来源完全变了，不是要你去调查，或者说，需要调查的越来越少。社会学家们掌握在手里的资源越来越少了，由资源带来的权力也越来越小了，我们不得不转化思维，学习与有资源的行动者合作；不得不转换技能，学习如何利用大数据。

第二个是“替代”。数据的来源完全变了,有可能未来完全不需要做大规模调查了,可能个别小事情需要做调查,调查的重要性会越来越下降将是一个大趋势。社会学家作为一个科学群体可能会像社会学家曾经掌握的数据一样,被其他学科所替代。社会学原本就是回应工业时代的需要产生的学科,一个时代结束,或许一个学科也会跟着被替代。在过去的一百年里,学科消失并不是一件稀奇的事情,或许,社会学也免不了自己的宿命。

不管怎么样,现在认识这些问题,还不算晚。眼下,还是要倾注全力,适应大数据时代的需要,发展社会学的能力,拓展社会学的边界。

(四)社会学研究范式还有用吗?

刚才讲,大数据对社会学的第一个挑战是:还需要社会调查吗?通过上面的分析,答案应该是清楚的。第二个挑战是:社会学研究范畴还有用吗?

我尝试着回答。要回答这个问题,不得不提到一本书,叫《大数据时代:生活、工作与思维的大变革》[①]。大家要读英文标题,英文标题橙色部分叫“革命”。什么是革命?他自己提了一个,抽样、精确、因果这三个过去我们为之努力奋斗的范式,正是革命的对象。

是不是真如此,我觉得,可以争论。不过我认为,这至少是一个信号,值得社会学家们认真讨论,因为从事实证社会学研究的人最熟悉这一套。我的问题是,大数据对社会学研究的影响,难道真的与调查数据的基本假设不一样吗?在调查数据的时代,我们用假设检验。如果真的要从假设检验转换为大数据的总体归纳,这两者之间难道一个是白天、一个黑夜吗?我觉得不是,两者之间必然有着千丝万缕的关系。

大家知道,自然科学用的是重复检验。我学生物学出身,生物学研究的要求是重复检验。你说某一个规律存在,至少要做三遍,得到的参数不超过误差值,就说明规律暂时是存在的、逻辑暂时是有效的、模式暂时是有效的,否则你就得重新思考,重新做。

社会科学没有重复检验的基础和条件,故,我们做假设检验。不过,我认为,即便是归纳的模式也要满足重复检验或(和)假设检验的基本要求。运用大数据的社会学研究,我认为,其范式也许重在发现,而不是重在推论。但是,

① 维克托·迈尔—舍恩伯格,肯尼思·库克耶.大数据时代:生活、工作与思维的大变革[M].盛杨燕,周涛,译.杭州:浙江人民出版社,2013.

基本的目标没有变,我们还是要把握事物之间的关系模式。

大数据分析的一些技术性技能是社会学研究缺乏的,我快一点过了。尽管对理解大数据和小数据之间的范式差异非常重要,由于社会学家们通常对技术问题兴趣不大,为了不至于让各位打瞌睡,我还是过快一点。

刚在说大数据分析重在发现,而不是重在推论,在方法上也有证据。大数据的非分析目的性,让对大数据的利用在方法上重视数据挖掘。什么叫数据挖掘?简单来讲说,就是从杂乱的混合数据中发现有意义的事物之间的模式和规则。挖掘是针对大数据分析的一个基本策略,但不是具体方法。我简单介绍一下什么叫做大数据挖掘。

大数据,首先是乱的。面对混乱,怎么办?大数据挖掘有一些基本的步骤,就是混乱的东西先归类,再降低它的维度,降为若干类别以后,便让大数据和调查数据的形态差不多了(见图 4)。图上有四个步骤,第一步拿到数据,非结构化的和结构化的混合数据;第二步梳理数据,用 HPCC,高性能计算系统,通过映射—降维(Map-Reduce),把混合数据就变成分类数据,可分析数据;第三步分析数据,作模式分析,获得初步的结果;第四步应用结果。

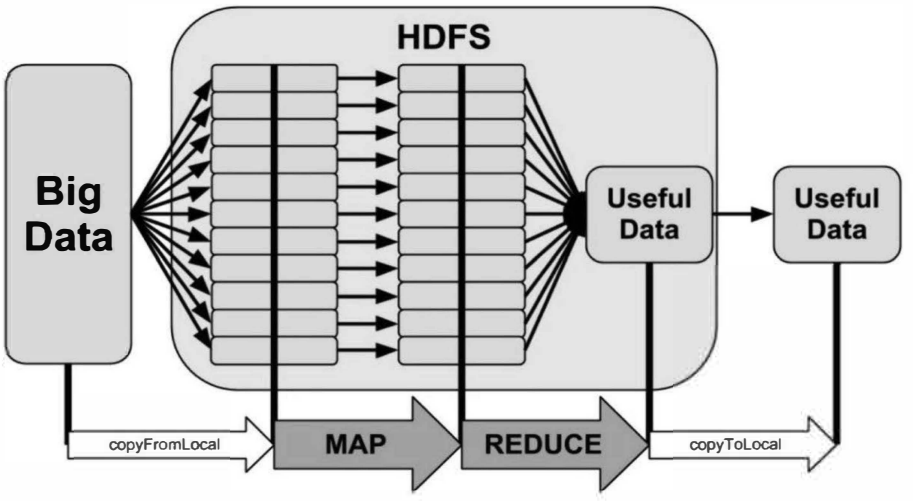

图 4　数据挖掘的基本步骤

我们把这幅图的步骤归纳一下。做大数据分析,第一步获得数据,通过映射—降维,形成可分析的数据;第二步选择要分析的降维数据,选择变量,降维以后的数据变量是可选的;第三步进行数据变换,比如说数据类型的变换,数

据模式的变换等等;第四步模式发现,数据挖掘就是要发现模式;第五步模式评估,对已经发现的模式,评估其信度和效度;第六步知识表达,社会学的最终产出在这(见图 5)。

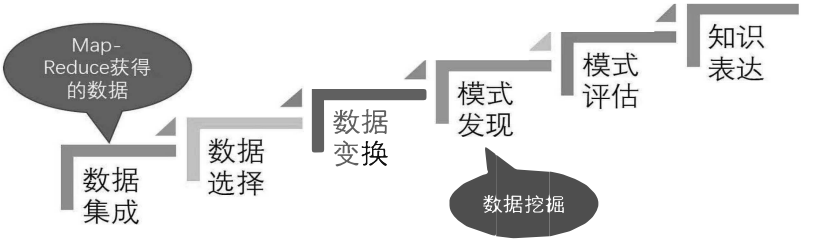

图 5　数据挖掘流程

当然,数据挖掘跟社会学研究一样,也有描述性研究,也有预测性研究,描述性研究同样是探讨特征、探讨属性。预测性研究同样探讨变量之间的关系。

大数据分析的描述性研究,大概是四大类工作,第一是做特征分析,特征分析就是点分析。第二是做关联分析,可以理解为双变量和多变量之间关系的分析。第三是做聚类,聚类主要是做多特征的综合聚类。最后是做离群点分析,调查数据叫极值,在大数据里叫离群点,两个不一样。描述性分析的目的是什么呢?也是用数据刻画,获得研究对象的数字画像。比如说要描绘一类人,性别、身高、生活规律,比如每天几点睡觉、几点起床、深睡时长、醒着的时长,做噩梦的时长,都可以用数据刻画。

简单介绍一下特征分析。特征分析类似于针对调查数据做的单变量分析,刻画研究对象的基本特征,譬如手机用户的年龄、性别、身份、行为、消费偏好、习惯、表情等;淘宝店家的年龄、性别、身份、家庭社会经济地位等;微博传播的网络结构如星形网络、结构洞网络等。

关联分析,类似于调查数据的双变量、多变量分析,是基于事物不同特征之间的相关性分析。不过,其分析的基本思路却大不相同,以频繁项集为例,其基本思路是:将某个频繁项集 Y 划分成两个非空的子集 X 和(Y　X),使得 $X\rightarrow$(Y　X)满足置信度阈值。如果规则 $X\rightarrow$(Y　X)不满足置信度阈值,则类似于 $X_1\rightarrow$(Y　X_1)的规则,一定也不满足置信度阈值,这里,X_1 是 X 的子集。根据这一特征,假设由频繁项集{a,b,c,d}产生规则,且规则{b,c,

$d\}\rightarrow\{a\}$具有低置信度，则可以丢弃包含 a 的所有规则。有点晕了，对吗？不晕才怪，这是计算思路，不是社会学的假设检验思路。

用一个例子试试。比如说，只要发现某教师哪天早晨五点钟起床了，可预测其要出门，这就是频繁项集的应用。注意，数据挖掘会运用其既往早起后的行为预测其会不会出门，并给出预测正确的概率。

聚类分析，原本就是调查数据统计分析方法的一种，用分类原则，筛选因子，减少变量的数量，又称“降维”。在数据挖掘中，点集数据是适合聚类分析的数据类型，通过聚类，让原本混杂的数据归入各自的类。再强调一遍，对大数据的聚类分析，采用的依然是计算思维：可算，计算有效率。

接下来看看预测分析。预测分析的技术对我们来说复杂了一些，这里不讲。只讲与调查数据分析根本不同的部分。调查数据是先建模，再搜集数据，最后检验模型。大数据分析是先有数据，建模的基础是数据，因此被称为数据建模。数据建模是基于数据归纳的，在数据里发现、挖掘，通过描述性分析建立简单模型，用简单模型让机器学习。

还是举刚才的例子。某个老师每周有哪几天早晨五点起床，机器可以预测他到学校来，还是到另一个地点。也许会有离群点，不过没关系，机器会自动调整预测概率。经过一段时间的数据积累和模型修订和迭代，便可以准确地预测。如果某老师在周五的早晨五点起床，他到北京大学社会学系办公室的概率有多大，通常，这个预测是精准的。这就是机器学习，不是人干的事，完全交给机器了。

举一个经典例子，谷歌流感模型。前面的故事大家应该都知道，即使不知道，网上搜索一下也可以知道。我要讲的是，2007 年谷歌处理了 4.5 亿个模型，最后筛选出一个综合模型，在这个模型基础上，跟随数据的积累，2012、2013 年又修订了新预测模型。现在，谷歌流感模型的预测比美国 CDC 的预测还要准。

大数据的数据建模，通常有两类。一类是分类模型，一类是回归模型。分类模型分析事物的类别，关注特征值；回归模型分析变量之间的关系模式，做预测。

在这个基础上，数据挖掘是多种技术应用。首先是统计学，郭志刚老师不会失业，统计学你得继续教，没有统计学知识，大家玩不转。其次是算法，如何让机器可计算和计算的有效率。我觉得社会学的学生未来至少要懂一些算法，我们可以不写代码，但不能连基本原理都不懂。在算法中还涉及一系列的

理论与技术如数据库、可视化、机器学习、模式识别等等。

此外,数据挖掘还会用到一些其他的技术,这里就不多说了。

先说统计技术,运用调查数据的统计技术,描述统计、推断统计、假设检验、统计模型等,在大数据分析中,技术不一定会用到,思想却不可或缺。大数据分析最常见的是回归分析。当然,大数据对统计技术和思想的应用与拓展也在发展,懂基础是发展的前提。

再说算法,相对复杂一些,也是数据挖掘中的核心,他不仅用于建库,也用于做所有与数据挖掘相关的工作,比如说机器学习。从初始数据建模到模型迭代、稳健,都依靠算法的效率。

前天,有个老师告诉我说,早上一来,发现计算机死机了。我问:为什么?他说,做了一个回归模型。我问:你做多少?他说,做 50 步。50 步?在大数据里面是完全小儿科,而且一个数据量级,还记得谷歌的流感模型,初始模型 4.5 亿个!初始建模、模型迭代、稳健化,都需要用算法。

机器学习是一个新兴的知识领域,知识性问题我不讲了,直接给大家例子。

淘宝 2014 年双十一,让 TCL 狠赚了一把,原来预订量,TCL 预计只有 8%,机器学习的结果预测 4K 电视机会热销,结果是一天上升了 60%。还有一个更搞笑的,服饰公司 A21,双十一前通过阿里数据锁定了 1 000 个老客户,公司只想试一下一种新的、依据数据的营销方式。比如说你是 A21 的客户,今年我根本不通知我要给你做衣服,也不要你在网上预定,而是把你的衣服做好了,直接送到你家门口,如果你认为不错,你就收单;如果你不喜欢或不需要,就拒收。结果是:90%的客户买单。

菜鸟网络,这是马云 2012 年说自己退休以后干的事。这是一个物流网络,对不同线路订单的预测准确率也达到 90%。说的是什么呢?各位知道,双十一的订单量惊人,如果不事先布置地方性的仓储,是无法在一周之内让所有客户拿到货品的。问题是,谁知道哪儿的客户需要什么?需要多少呢?大数据知道。依据大数据建模,菜鸟网络事先把货品部署到各地的仓储,一旦有订单触发,快递网络便直接从离订单地最近的仓储取货和送货。2014 年,截止到 11 月 14 日下午 14 点,双十一期间的物流已经被签收 4 000 万个,双十一商家当天发单率达到 20%,揽收率 60%多。对商业应用来讲,预测的重要性可想而知。

对于社会学研究,其实没那么着急。不过,依然非常重要。

我们再举一个例子，百度做的，春节期间的人口迁徙图。作这样的图，对于有大数据的商业公司很简单，可对社会学家们来说，貌似一个难题。说简单，是说原理的确简单。手机在中国的普及率非常高，有能力使用手机的人几乎人手一部。手机之间的通话、短信、微信等，有一个中介，那就是基站。每一部手机只有接入一个具体的基站，才算是上网了。每部手机都有唯一识别码，每个基站也有一个唯一识别码，运用手机在基站上的移动，就可以定位人口的迁徙了。春节期间，百度的人口迁徙图，就是应用这个原理让机器自己做的。其中，既有统计学原理，也有计算机的算法。

除了人口流动，社会学家感兴趣的还有人类行为。我再讲一个例子，男人一看球，女人就网购。2012 年欧洲杯期间，女性网购成交量上升 10%。真的是因为男人看球导致的吗？你必须有进一步分析，更加有力的证据。大家看图(见图 6)，在世界杯期间网购的人当中，男性所占的比例是 38%，女性所占的比例是 62%，女性网购的几乎是男性的一倍。但怎么就能确认男人一看球，女人就网购呢？大家再看下面的柱状图，左边的柱子是 6 月 1 日的网购成交量，右边的柱子是 6 月 15 日的网购成交量，下面是一天 24 小时。6 月 1 日欧赛没开始，成交量的高峰在晚上 9 点，就是左边那个气泡标注的地方；6 月 15 日球赛开始了，成交量峰值的时间往后推了，推到晚上 11 点，也就是了右边的气泡标注的地方。这就能证明女人真的是在男人看球时网购的吗？是的。大家注意一下图片最右边的三柱数据，也就是每天晚上 10 点到 12 点。在这个时间段，右边代表 6 月 15 日成交量的柱子都比左边代表 6 月 1 日成交量的柱子要高，而这时候正好是球赛最热闹的时候。我们之前有提到过，和男人相比，女人才是网购的主力。特别是当男人的注意力都在球赛上的时候，自然就是女人在创造网购峰值了。要获得这种证据，采用调查方法是比较困难的，可对大数据的挖掘来说就是分分钟的事。

最后再举一个例子。2014—2015 年跨年夜的上海踩踏事件。大家看下面的图，图里面的两条曲线一条是地图搜索量，也就是搜索外滩跨年夜的地点；另一条是对应的人流积聚量。通常来说，这两条曲线都是交叠在一起的。而且人群积聚曲线一般要略高于地图搜索曲线，就像 12 月 25 日至 30 日这段时间显示的一样，很平稳，也很有规律(见图 7)。但是大家注意图的最右边，也就是 12 月 31 日这部分。我们可以看到，上面这条代表地图搜索量的曲线迅速抬升，很快就超过了人群积聚曲线。然后人群积聚曲线也紧跟着搜索量曲线爬升。还有两幅热成像图我没有放上来，人流热度的移动也非常明显，跟搜索

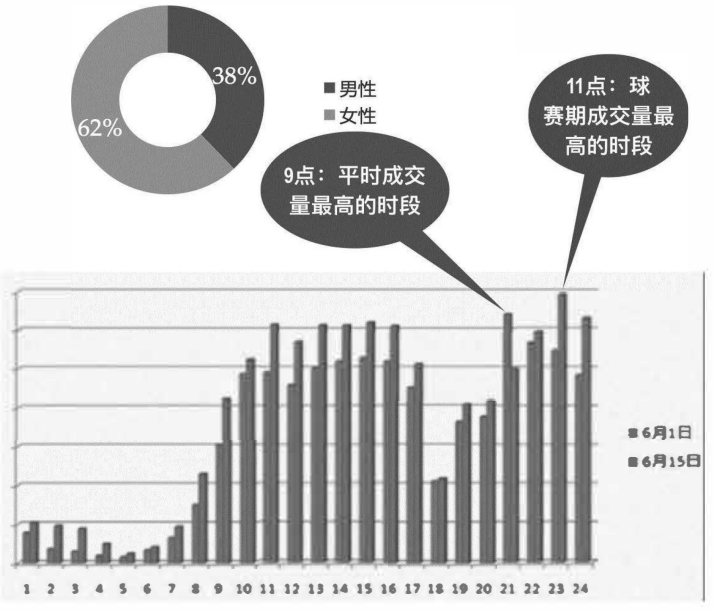

图 6　2012 年欧洲杯期间的网络成交量

图完全重叠。31 日下午的搜索量陡增就已经预示了晚上人流会激增。可是，上海市警方并没有注意到大数据的力量，手里有数据，却不布置警力。

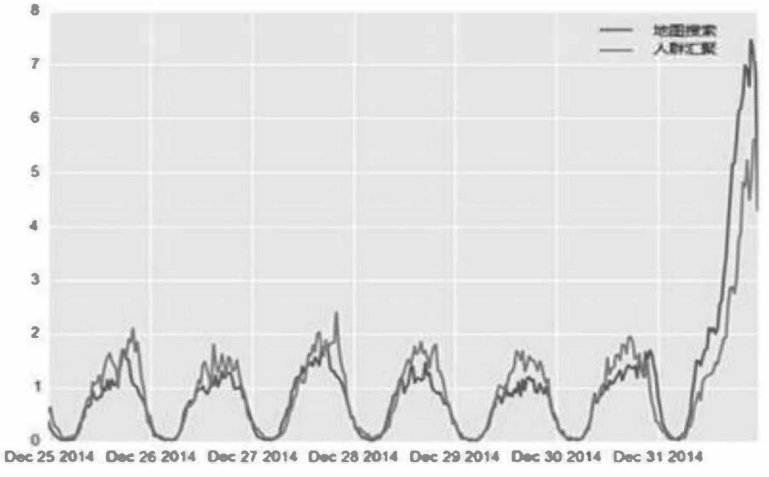

图 7　上海踩踏事件前夕“跨年夜”关键字的地图搜索量

运用这些例子，我想说明的是，大数据在渗透进我们社会生活的方方面面，其中的一些方面是社会学传统的调查方法处理不了的，无论是方法还是时效，都难以应付的。但是，在大数据的挖掘和利用中，我们又常常看到社会学研究范式的影子。它意味着社会学范式不仅有用，而且有大用！

（五）社会学的优势在哪里？

社会学曾经的优势有调查数据、有分析工具、有知识积累。这三块是社会学最核心的优势。调查数据、政府数据、商业数据各自有自己的专业领域，也因此形成了各自的话语权，也保障了社会学家们的独立性！除了数据以外，保障社会学家们话语权独立的还有社会学的分析工具和知识积累。社会学家们用自己的数据、科学的分析工具，形成了针对社会的知识积累，形成对社会有益的一股力量。

大数据的发展，使社会学曾经拥有的优势变了，社会学家们依然掌握着调查数据，可大数据对调查数据的冲击越来越大，调查数据的局限性越来越明显，大数据对调查数据的替代趋势也越来越强，将来会不会完全替代，现在下判断还为时尚早。尽管如此，调查数据的话语权变弱是不争的事实。

社会学家们剩下的优势只有知识积累了。问题是，知识积累也依靠数据，在数据受到冲击的前提下，社会学的知识积累也可能会坐吃山空，我想，这才是社会学研究面对的真正挑战。未来，社会学如果不能寻找替代，在新的分工图谱中找到自己的位置，没有独门秘籍，没有超人的创新能力，面对的结局可能是大家非常不愿意接受的，譬如做知识劳工。如果我们回顾自己的职业生涯，有一条线索非常清晰：从 20 世纪 90 年代开始，教授们的工作开始逐步劳工化，先是做政府的劳工，帮政府做课题，哪一个政府找到你，请你做一个课题，你高兴得要命。接下来做商业公司的劳工，商业公司请你开个会，给你一两千块钱，你也非常高兴；让你发表一个观点，你也很高兴。我们可能从来没有想过，如何开发自己的脑力、知识力、社会学的知识力，形成一股独立的力量，让社会学家们再次成为一股独立的社会力量。我认为现在是时候了。

归纳起来讲，如果说大数据对社会学研究有什么挑战，其实不是大数据的挑战，而是社会变迁的挑战，我们生活的这个社会变了。社会学的先祖们曾经面对的是从农业社会到工业社会的变迁带来的挑战，我们如今面对的是从工业社会到信息社会的变迁带来的挑战，这个挑战的基础部分是社会的数据化。我们的先祖们把握了工业社会的特征，让社会学成为一个学科；如今，如果我

们没有能力把握信息社会、数据化社会的特征,社会学就有可能毁在我们手里。因此,让社会学再学习,向社会学习,把握数据化社会的特征,可能是社会学进一步发展的希望。

对从事社会学教育的我们来说,与数据化社会相伴随的更大挑战还在于教育模式的革命性变革。这两年教课,我有一个体会,跟大家分享。请大家看图(见图 8),上面有各种应用,非常形象地隐喻了一个学生的脑袋。他在干什么呢?数字化学习(e-learning),这个词 15 年前在美国开始流行,今天依然流行。

图 8　数字化学习

大家知道大英百科全书是最大的百科全书,我记得,过去社会学系要买一套大英百科全书要请示校长,为什么呢?太贵了!可如今,还需要买大英百科全书吗?完全不需要!维基百科是众多人修订的成果,知识正确率 96%,只要有错,就会有人修订,只要有错,就会有人改,完全免费,比 300 部大英百科全书总量还要大,你为什么还要买大英百科全书?为了适应信息社会的情境,大英百科全书不卖纸版了,也是世界上第一个不卖纸版的百科全书。

我用这个例子希望说明,学生的学习方式变了,纯粹的知识性内容,完全可以通过搜索获取和积累,完全不需要老师,有的时候,老师在课堂上讲的还是错的,为什么要老师。学生在课堂上之所以需要老师,是希望老师带给他们在网上搜索不到的东西,那么,这个东西是什么呢?我相信老师们对这个问题都有思考,可能也没有直接的答案!

因此,更大的挑战在于整个教育模式的革命转变。挑战不在于你当不当老师,而在于整个大学的教育模式,整个教育模式的未来,比如说班级模式还会不会继续存在。我举一个例子,初等教育的例子,有一个学生,应该是2011级的,休学了,自己去创业。做教育,做了一个小应用,很简单。把各地的优质教师汇集到平台上,学生付费进来。你说要什么?学什么?系统自动匹配,一对一。这就是一种新的教育模式。教育平台,像马云做淘宝一样。这样模式能坚持多久,不知道。不过,在当下,社会是认可的,他差不多拿到了一亿多的投资。用这个例子同样希望说明,学习在变,初等教育在变。

高等教育难道不变吗?美国人弄了两个课程平台,其中一个是斯坦福大学弄的,叫Coursera,7 000多门课,我比较大胆,我放了一门在上面。如果你真的有信心,就需要在世界范围内竞争,你讲的不对,立马有人吐槽,这就是教育模式的革命。我想,现在只是一个开始,更大的挑战还在后面。

而这一些,都源于大数据作为一种新的社会资源带来的挑战。

四、归　纳

最后,我大致做一个归纳。

简单来说,大数据是一个并行化、在线汇集整个人类社会生活的、包括个人隐私生活的大规模、混合结构的数据,传感器是大数据搜集的主要工具,人类行为,无论是社会性的还是私密性的,都是大数据的来源。

大数据与社会学研究密切相关,与传统的调查数据不一定是竞争关系,可在事实上,我们观察到了大数据对小数据的替代,也观察到了大数据对数据覆盖范围的扩展。

大数据给社会学研究带来的挑战不仅在于数据源的替代,更在于社会学想象力和技能的转换,甚至是扩展。适用于传统调查数据的社会学能力在面对大数据时已明显不足,社会学需要拓展想象力和技术能力,才可能把大数据作为一种新的研究资源纳入社会学学科。

不仅如此,我认为,大数据带来的更大挑战在于对大学教育模式的冲击。课堂上,知识性的传授已经为大数据资源所取代,创造性的启发和智力挖掘可能是教育的未来。