

实证社会科学

(第六卷)

主编 钟 杨

大数据给社会学研究带来了什么挑战？

邱泽奇

公众慈善捐赠行为的影响因素研究——以上海市为例

郭俊华 刘琼

村民自治会影响党的领导吗？——乡村居民与地方干部意见的分析

肖唐镖 涂曼冰

因果推理中的科学模型——反事实、选择性偏差与赫克曼结构计量经济学模型

李文钊

制度设计、实验与中国国家治理的『地方性方案』——杭州市余杭区基层协商机制创新实验区研究

郎友兴

亚洲的边界与想象

翟一达

Social Science
Research



上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

实证社会科学

Social Science Research

(第六卷)

钟 杨 主编



上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

图书在版编目(CIP)数据

实证社会科学. 第六卷 / 钟杨主编. —上海: 上海交通大学出版社, 2018

ISBN 978-7-313-10316-1

I. ①实… II. ①钟… III. ①社会科学-文集 IV. ①C53

中国版本图书馆 CIP 数据核字(2018) 第 301455 号

实证社会科学(第六卷)

主 编: 钟 杨

出版发行: 上海交通大学出版社

邮政编码: 200030

出 版 人: 谈 毅

印 刷: 上海景条印刷有限公司

开 本: 787mm×1092mm 1/16

字 数: 150 千字

版 次: 2018 年 12 月第 1 版

书 号: ISBN 978-7-313-10316-1/C

定 价: 59.00 元

地 址: 上海市番禺路 951 号

电 话: 021-64071208

经 销: 全国新华书店

印 张: 9.25

印 次: 2018 年 12 月第 1 次印刷

版权所有 侵权必究

告 读 者: 如发现本书有印装质量问题请与印刷厂质量科联系

联系电话: 021-59815625

实证社会科学

Social Science Research

主办单位：上海交通大学国际与公共事务学院

主 编：钟 杨

副 主 编：吴建南 樊 博

编委会成员：（按姓氏笔画排列）

边燕杰（西安交通大学）

杨开峰（中国人民大学）

吴建南（上海交通大学）

陆 铭（上海交通大学）

陈 捷（上海交通大学）

庞 珣（清华大学）

钟 杨（上海交通大学）

唐世平（复旦大学）

敬义嘉（复旦大学）

蓝志勇（University of Arizona）

李连江（香港中文大学）

肖唐镖（南京大学）

何艳玲（中山大学）

陈映芳（上海交通大学）

邱泽奇（北京大学）

赵鼎新（University of Chicago）

唐文方（University of Iowa）

阎学通（清华大学）

谢 宇（Princeton University）

樊 博（上海交通大学）

编辑部成员：

钟 杨 吴建南 樊 博 陈映芳 刘帮成 陈永国 黄琪轩

陈慧荣 陈 拯 魏英杰 杜江勤 韩广华 杨 姗 奚俞颀

目 录

学者评论

- 大数据给社会学研究带来了什么挑战？ 邱泽奇 / 3

研究文章

- 公众慈善捐赠行为的影响因素研究——以上海市为例
郭俊华 刘 琼 / 31
- 村民自治会影响党的领导吗？——乡村居民与地方干部意见的分析
肖唐镖 涂曼冰 / 49
- 因果推理中的科学模型——反事实、选择性偏差与赫克曼结构
计量经济学模型 李文钊 / 72
- 制度设计、实验与中国国家治理的“地方性方案”——杭州市余杭区
基层协商机制创新实验区研究 郎友兴 / 89
- 亚洲的边界与想象 翟一达 / 103

数据库介绍

- 全数据研究美国政企关系 冯一郎 / 129

投稿须知 / 136

CONTENTS

Scholar Review

What Challenges Do Big Data Bring to Sociological Research?

Qiu Zeqi / 3

Research Article

Research on the Influencing Factors of Public Charity Donation: A Case

Study of Shanghai

Guo Junhua Liu Qiong / 31

Will Villagers' Self Rule Affect Communist Party's Leadership?

Xiao Tangbiao Tu Manbing / 49

A Scientific Model of Causality

Li Wenzhao / 72

System Design, Experiment and Local Plan of Chinese State Governance

Lang Youxing / 89

The Boundary and Imagery of Asia

Zhai Yida / 103

Introduction of Social Science Data

Data Sources on Government-Business Relations in the United States

Feng Yilang / 129

学者评论

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。为了让大家都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，在整个人类基地的亿分位数都不

* 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。在计算机科学里叫线程，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.