对大数据的理解与误解*

邱泽奇**

摘　要：大数据已经成为新的学术资源，它是经由数字设置和设备伴随行动者的社会行动而产生的、在互联网络中汇集的、需有并行计算设施处理和挖掘的资源，它不会取代调查数据，也不是对行动模式的篡改，更不会主动忽悠。大数据是数据的一个类型。

关键词：大数据　互联网技术　痕迹数据　皮书研究

一　对大数据的理解

（一）暖场：大数据热？

2012 年 7 月，《国务院关于印发〈“十二五”国家战略性新兴

---

* 本文根据作者 2019 年 5 月在贵州召开的第五期全国皮书研创高级研修班上的发言录音整理而成，已经本人审阅。

** 邱泽奇，教育部长江学者特聘教授、北京大学博雅特聘教授、北京大学中国社会与发展研究中心主任。研究领域：技术应用与社会变迁、组织社会学、社会研究方法。

产业发展规划〉的通知》明确提出支持海量数据存储、处理技术的研发和大数据产业化。实际上，早在2013年8月，国务院办公厅公布的《国务院关于促进信息消费扩大内需的若干意见》就推动了企业加快信息基础设施演进升级，增强信息产品供给能力，形成行业联盟，制定行业标准，构建了大数据产业链。

2014年是中国大数据元年，大数据首次写入政府工作报告。2015年是中国大数据顶层战略架构设计元年，3～9月陆续推出方案和计划。同年9月，国务院印发《促进大数据发展行动纲要》。

2016年，大数据战略布局逐渐完善。《中华人民共和国国民经济和社会发展第十三个五年规划纲要》中首次公开提出“国家大数据战略”，其中对“国家大数据战略”的定位，成为各级政府在拟定大数据发展规划和配套措施时的重要指导。2016年被称为中国的“大数据之年”，国家推行“大数据战略”，各行各业都在谈论大数据。同年5月25日，李克强总理出席在贵阳召开的中国大数据产业峰会暨中国电子商务创新发展峰会。

从2017年开始，政策趋严。6月，《中华人民共和国网络安全法》发布；10月，人工智能写入党的十九大报告，推动互联网、大数据、人工智能和实体经济深度融合；12月，《信息安全技术个人信息安全规范》落地，明确提出在收集、保存、使用数据等过程中个人敏感信息的范围。

2018年，国务院先后针对教育、医疗、金融等大数据使用较密集行业发布计划和指引，大数据产业化加速落地。4月，《国务院办公厅关于印发〈科学数据管理办法〉的通知》明确中国科学数据管理的总体原则、主要职责、数据采集、汇集与保

存、共享利用、保密与安全等方面的内容，从不同侧面提出了具体管理措施，进一步加强和规范科学数据的管理，保障科学数据的安全，提高开放共享水平，为国家科技创新、经济社会发展和国家安全提供支撑。5 月，中国银行保险监督管理委员会发布《银行业金融机构数据治理指引》，突出强调数据汇集能力建设、新产品评估要求，有效评估和处理重大收购和资产剥离等业务对数据治理能力的影响，明确监管机构的监管责任、监管方式和监管要求。

由此可见，中央政府大数据政策出台及发展趋势为：2017 年之前以促发展为主，2018 年以后以规范市场、提升数据安全意识、加速产业落地为核心。此外，数字经济的重要性具体体现为：在国际流动中，数据流动是全球 GDP 增长的第二大贡献来源，43. 4% 的世界人口（约 32 亿人）在线，12% 的全球商品贸易通过电子商务进行，全球 50% 的服务贸易是数字化的。

（二）笔者理解的大数据

1. 人类活动

数据是在人类活动中产生的。人类活动一定会留下痕迹，这个痕迹曾经被当作证据。在数字化进程中，这些证据被数据化。当大多数行动被数据化之后，人们用一种特别的方式来汇集这些数据，在运用这些数据时，才产生了“大数据”。人类活动会产生痕迹，产生痕迹的活动称为“造痕”。“造痕”是刑侦学概念，也是考古学概念。痕迹如果留存下来，就变成了证据。证据如果被数据化，就变成了数据。如果数据通过网络汇集、存储、挖掘，就变成了大数据。从数据发生学角度看，这是我们理解大数

据最简单、最直接的一个方式。

人类活动分为两大类：一是生产与生活，二是人际交流。这两类活动都会留下痕迹。考古学和历史学有个边界，一般而言，地下的归考古学，地上的归历史学，不过，很难区分。考古学和历史学在面对研究对象时，各自形成了自己的研究体系。比如，考古学重断代，重器物；历史学重文献。二者又不是截然分割的，有着千丝万缕的联系，有共同点，面对的都是历史痕迹。把痕迹作为历史证据，是早期社会科学运用数据的一种手段，人文科学对数据的研究和社会科学在某种意义上有着非常相似的地方。尤其在历史学中，一定要有佐证，这是科学研究对痕迹利用的方式和方法。

2. 痕迹

现代社会科学对过去痕迹证据的应用，形成了历史学与考古学交叉的研究问题以及研究领域。人类的社会活动都会留下痕迹，留存下来就是“社会人为事实”，是社会科学研究的对象之一，遍地都在。在刑侦领域，痕迹证据也是侦查研究的对象之一。如果能做交叉比对，就变成了清晰的证据。人类的社会活动，时时刻刻都留下了痕迹，现在留下的痕迹，形成了痕迹社会。从人类诞生至今，人类活动的绝大部分痕迹随时间而消逝。无论是物化的痕迹还是数据化的痕迹，留下来的是极少一部分。

3. 证据

痕迹证据变成如今证据的数量是非常有限的。祖先们活动的痕迹并没有很好地留下来，留下来的是极少的一部分。数字化存储技术的发展，让存储变得快捷，交易成本更低。痕迹证据已转化为痕迹数据，转化为数据化了的痕迹，也是社会学研究常见的

文献调查。

4. 数据

互联网技术的发展，让痕迹数据变成了大数据。在互联网进入普遍应用之前，痕迹数据是分散的。互联网技术的应用促进了数据通过网络汇聚和存储。个案性的痕迹数据，汇聚成了社会性的在线数据流、大数据。现实的发展是，痕迹数据在“造痕”的同时就被数据化了，这是大数据最突出的和传统数据的不同之处。

（三）什么是大数据？

数据，已经渗透到当今每一个行业，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波消费浪潮的到来。国际商业机器公司（International Business Machines Corporation，简称 IBM）给大数据的定义纳入了四个维度（速度、价值、形态、数量）的测量。结合这四个维度，笔者的理解如下。

1. 形态

传统的数据是结构化的数据。大数据是混合形态的数据，既有结构化的数据，例如 SQL 数据；还有非结构化的数据，如日志、音频、视频、图片等。

2. 容量

大数据的基本定义围绕着容量。数据集的大小，超出了依靠人力在可接受时间内搜集、利用、管理和处理的能力。2012 年，单一数据集的大小，从 MB 级跃升至 TB 级，以及 PB 级。这个量级超出了普通桌面计算机乃至大型计算机的处理能力。

3. 来源

如果传统的数据是目标导向的，大数据则是记录导向的，它没有明确应用取向。大数据不是为某一个人做调查，而是为整个生产和运营做记录。正因为如此，大数据并不是有目的的测量数据，而是造痕留下的并行数据。大数据其实是痕迹数据的汇集，它的价值并非用来证明什么，而是需要挖掘的。IBM 认为，大数据的价值密度较低。这也是数据行业通行的观点。

4. 特征

从开始测量到数据可用，传统的调查需要相当长的时间。1890 年的人口普查活动促使美国统计学家赫尔曼·霍尔瑞斯发明了读卡机，他用 1 年时间完成了原本耗时 8 年的人口普查活动。中国家庭追踪调查（China Family Panel Studies，简称 CFPS）从结束调查到数据可用，也需要 1～2 年的时间。大数据几乎随时可用，人们每时每刻都在记录数据，每时每刻都有数据可用。不过，可用，不是针对具体研究问题的可用，而是在给定研究问题时，用于数据清理的可用和机器学习的可用，其体现在以下三个方面。第一，行为数据化。造痕者的行为与设备关联在一起，比如，学习、社交和饮食。第二，行为网络化。造痕者与行为关联在一起的设备，只是网络的一个节点。第三，网络数据化。数据与人类的社会行为并行，与网络同在、与社会一体。

简单地说，可以用形态、容量、来源、特征四个方面归纳大数据。第一，形态上，大数据是数字化、非结构化的，也是在线的流动数据。第二，容量上，大数据基本上是在 PB 级别的。第三，来源上，大数据的来源与人类社会的行为相伴，通过网络汇集。第四，特征上，大数据非常完整，但不系统。所以，研究人

员要理解这些数据的特征。对社会科学而言，大数据是现代研究资源，需明确，大数据不是对原来数据的全面替代，有些地方的确有替代性，但不是全面替代，它有自己的特征和弱点。

二　对大数据的应用

（一）大数据可以帮我们变“不可能”为“可能”：敌对分子聚集的在线生态

研究议题：ISIS 成员聚集在线生态模式。数据来源：类似于 Facebook、VK 等在线社交数据。搜集方法：人工设定种子（支持 ISIS）ID，跟踪 ID，滚雪球，直到形成闭包（Closure）。数据周期：2015 年 1 月 1 日 ~ 8 月 31 日。数据规模：196 个聚集，108086 个跟随者；峰值：134857 人。

现象观察：最终可能激发没有极端主义经历、没有正式成员资格或没有与领导层直接联系的行动。驱动这一网络支持的是一种超快社会生态。

研究结论：这一社会生态的特征是自组织聚集，这种聚集在现实世界活动开始之前会急剧增加，并通过适应机制保证其存活。一种可控的预测是，通过打击小型潜在准 ISIS 组织，可以阻止大型潜在准 ISIS 组织的发展。

（二）大数据可以帮我们变“推测”为“确证”

1. 小世界的真相

设计：选择一个随机起点，观察需要经过多少个中间人能够

到达目标点。规则：参与者只能将信件转发给能直呼其名的熟人，并请他继续转发；如果一个参与者不认识目标收信人，则他不能直接将信寄给他；参与者需力争让信件尽早送达目的地。第一次：从堪萨斯州的威奇塔市到哈佛大学神学院某位学生的妻子；第二次：从内布拉斯加州的奥马哈市到波士顿的股票经纪人。

结论：中间人数为5～7人。

此外，2012年1～6月，笔者主持议题为“网店店主的社会特征和执业特征”的项目，运用“执业数据+抽样问卷数据”，有效地预估了网店店主的行为特征。

2. 情绪传染实验

主持：Facebook数据科学家亚当·克莱默领衔。议题：社交网络影响（积极与消极）的传播方式。方法：挑选689003名Facebook用户，分成两组，将其中一组用户中带消极情绪的帖子剔除，将另一组用户中带积极情绪的帖子剔除。周期：2012年1月11～18日。

观察：当看到朋友带积极情绪的帖子变少后，自己发布的积极帖子也会变少，消极帖子则会增多；反之亦然。

结论：面对面互动和非语言线索并非情绪传染的绝对必要条件。

3. 回音壁实验

主持：Facebook数据科学家艾唐·巴克什。议题：影响社交网络回音壁（Echo Chamber）效应的因素，回顾1010万个美国匿名账户，搜集这些账户的好友信息和反应意识形态的偏好。周期：过去6个月。

观察：用户不会总处于相同观点的环境中，在大部分时间

里，用户的观点是不同的。Facebook 的算法只会让用户有 5～8 个小时的时间处在单一观点的环境中。

结论：用户通常活在自己的世界里，形成回音壁的真正原因是用户自身。

（三）大数据可以帮我们变“不可预知”为“可预知”

1. 各种疾病

（1）埃博拉：从移动测绘着手，美国疾病控制与预防中心（CDC）和瑞典非营利组织（NPO）刻画出典型的人口流动模式，并且估算出病例似乎会在哪里出现。通过跟踪拨打热线电话，了解和预测疫情的爆发地点。

（2）败血症：通过查阅败血症的大数据存储库以及一系列指标，可基于床边监护仪搜集的数据创建出预测模型。把仪器连接到一套云系统上，便把含糊的传统诊断方法替换成了准确的、注重细节的循证分析方法。

（3）脊髓灰质炎：匹兹堡大学第谷计划（Project Tycho）的负责人用 1888～2011 年间的 87950807 宗病例创建数据集，试图绘制这些数据与疫苗运动之间的关系，探讨每次运动的效果。第谷计划获得了比尔及梅琳达·盖茨基金会和美国国立卫生研究院的资助。

2. 上海外滩踩踏事故

如果我们重视大数据提供的信号，2014 年 12 月 31 日跨年夜上海外滩的踩踏事故其实是可以避免的。客流数据和移动基站接入数据都清楚地预示了外滩人流与空间之间的张力。可惜，当时的警方还没有意识到大数据的预测意义。

三　对大数据的三种误解

人们对大数据存在三种误解。第一，大数据是大泡沫，认为大数据炒作多，真正用得少。第二，大数据有大价值，也有局限。第三，大数据是大篡改。

克里斯·安德森说过，“有了数据，就不要模型了，或者很难获得具有可解释性的模型，那么，模型代表的理论也没有意义了”。微软研究院首席研究员、麻省理工学院公民媒体中心客座教授凯特·克劳福德说过，“有了足够的数据，数字就可以自己说话，大数据将使我们的城市变得更加智能和高效，它对不同的社会群体不会厚此薄彼，大数据是匿名的，因此，它不会侵犯我们的隐私，它是科学的未来”。

由此可见，大数据的本质是数据，是测量的产出。数据没有对错，运用却有对与错、好与坏。皮书研究正是运用了大数据的资源特征，研创人员在开发皮书内容的同时，也在积累大数据，正在为挖掘中国社会发展与变迁的新内容提供难以替代的资源。笔者以为，这是皮书生命力的又一个重要支撑。