大数据时代的社会治理与社会研究:现状、问题与前景

冯仕政
中国人民大学社会与人口学院, 北京 100872

摘要
区分大数据社会研究中科学和应用这两种取向及其相互关系;考察大数据给社会研究带来的机遇、挑战和面临的困难;揭示大数据所具有的数据、社会和技术的三重面相以及相应而来的不同学科在大数据社会研究中的地位和关系;剖析社会科学、统计科学和计算科学3个学科合作中的困境及其原因。

关键词
大数据;社会研究;社会治理;社会计算;社会统计

中图分类号:C91-0 文献标识码:A doi: 10.11959/j.issn.2096-0271.2016014

Big data and social studies in China's governance: status quo, problems, and prospects

FENG Shizheng
School of Sociology and Population Studies, Renmin University of China, Beijing 100872, China

Abstract
The two potential orientations that are scientific research and social application, in big data studies, and their relationship were clarified. In light of this, the chances, challenges that big data brings and its weakness towards social studies were explored. Based on specifying the three properties of big data as statistical numbers, as social processes, and as technical functions, their respective positions of different disciplines in big data-based social studies, and both the difficulties and their sources in cooperation between different disciplines were shed light on.

Key words
big data, social study, social governance, social computation, social statistics

1 引言

随着大数据的兴起，基于大数据的社会治理也成为热门话题。社会各界和多个学科莫不跃跃欲试，与大数据社会治理有关的研究项目、会议和组织一时如雨后春笋般地涌现。然而，即使在大数据时代，社会治理也离不开社会研究。社会治理是操作，社会研究获得的是原理，离开原理的操作不可能走得太远。大数据为揭示社会活动的规律提供了新的可能性，如能取得突破，其科学和应用价值不可估量。但目前进展并不乐观。正如杜克大学教授、TED创始人阿利里（Dan Ariely）所调侃的：“大数据就像青少年性行为，每个人都在说，实际都不知道怎么做；每个人都以为别人正在做，于是声称自己也在做。”这句话形象地道出了大数据在大热之下的大虚。然而，调侃归调侃，真正需要反思的是：以大数据为基础的社会研究是否必要和可能？目前存在什么问题？这些问题是怎样形成的？又该如何突破？本文试图回答这些看似离社会治理很远，实则高度相关的问题。

2 大数据开发的两种取向：应用与科学

关于大数据研究，迈尔-舍恩伯格和库克耶在《大数据时代》[1]一书中的论述在国内外流传甚广，影响极大。该书的核心观点是，大数据的崛起将给人类的信息分析工作带来三大转变：一是不再依赖随机采样，二是不再追求精确性，三是不必寻找因果关系。在他们看来，代表性、精确性和因果性都是“小数据”时代的思维观念。在“小数据”时代，由于生产力和技术水平的限制，获取和分析数据的成本比较高，所以人们倾向于用尽可能小的数据预测尽可能多的现象，代表性、精确性和因果性等追求应运而生。而在大数据时代，数据的获取是如此快捷和低成本，能够获取的数据又是如此全面，追求代表性、精确性和因果性也就没有必要了。

这一观点可以说全面颠覆了以往社会科学的主流观念。相应地，它也引发激烈的争议。那么，究竟应该怎样看待这一观点呢？仔细观察会发现，当前大数据开发中同时存在两种取向：一种是应用取向，一种是科学取向。迈尔-舍恩伯格和库克耶的观点虽然以纵论大数据的面目出现，实际只是应用取向的表现。见**表1**，这两种取向存在多个方面的区别，混淆两种取向之间的关系将给大数据开发造成严重的不良后果。

应用和科学这两种取向的分野，从根本上讲，源于它们对大数据分析的价值期待不同，即应用取向追求实际功用，而科学取向追求一般原理。这是两种既有联系又有区别的追求。说有联系，是因为需求驱动创新，实际需要经常成为促进科学发展

表1 大数据开发中的两种取向

取向	应用取向	科学取向
价值期待	实用	原理
条件约束	时效	永恒
评价原则	完成	完美
工作标准	粗放	精确
工作内容	相关	因果

2016014-2

的强大动力，而科学原理则有利于更好地满足实际需要；说有区别，是因为人类对实际效用的追求并不必然引起甚至可能妨碍对科学的追求，反过来，人类付出不菲代价求得的科学原理常常没有什么即时的应用，以致给世人造成一种不中用的感觉。这两种取向之间的关系，就如同学术界争论已久的应用研究与基础研究之间的关系，其中的道理很明显，不赘述。

由于追求的目标不同，两种取向面对的约束条件也就不同。应用取向的大数据研究，由于重点是满足实际需要，而需求又是时时变动的，所以对时效性要求比较高；而科学取向的大数据研究志在获得一般原理，而一般原理必须经得起时间的检验，所以更重视永恒性，对时效性不那么敏感。

相应地，在评价原则上，应用讲求结果导向，即完美与否是次要的，关键是在规定的时间内完成规定的任务；而科学则尊重探索，既然是探索，就允许试错，所谓试错，就是目标、任务和行动路线都可以根据新的发现不断调整。在这个意义上，科学无所谓完成不完成，或者说永远不会有完成，完美才是决定性的。

基于不同的评价原则，两种取向在工作标准上也存在显著差异。应用讲求时效和绩效，因此，只要边际效益递增即可接受，并不追求最优解，对工作结果的容错率较高。体现在大数据分析上，就是宁可粗放一些，也不能错过时机。而科学基于完美原则，一定会不懈地追求最优解，因此对工作结果的容错率比较低，对边际改进只能暂时接受。体现在大数据研究中，就是倾向于不惜代价地提高分析精度，不愿浅尝辄止，“小富即安”。

最后，从工作内容来看，揭示事物之间的因果关系是科学的本质所在，止步于相关关系对科学来说是不可接受的。但从应用的角度来看，效益才是第一位的，其他的都不重要。而效益的获得并不总是依赖于对因果关系的掌握，如果了解相关关系即可带来足够高的效益，就没有必要去探究背后的因果关系，尤其是当这个过程的代价比较高的时候。就像大数据分析发现，很多顾客在超市买婴儿尿布时会连带买啤酒，那么，将尿布和啤酒摆在一起，一定可以同时提高两种商品的销量。对商家来说，知道这一点就够了，至于为什么顾客在买尿布的同时会买啤酒，大可不必追究。也就是说，应用取向的大数据分析大可知其然而不知其所以然，但科学取向的大数据分析则必须揭示“所以然”，这是两者追求的目标不同决定的。

综上所述，科学和应用对大数据分析有着不同的价值期待，进而决定了它们工作的内容、标准、约束条件以及对工作的评价原则也有所不同。显然，人类既需要应用，也需要科学，因此，两种取向的大数据分析都是人类所必需的，二者只是分工不同，并无高下之别。关键是怎样处理两种取向之间的关系，处理得好可以相得益彰，处理不好则会两败俱伤。

毋庸讳言，在大数据开发中，当前占主导地位的是应用取向。这样一种局面的形成，与应用取向的大数据研究相对来说难度更低、见效更快，同时更容易获得市场和资本的青睐有关系，这无可厚非。但一种值得忧虑的倾向是，许多人因此而轻视甚至否定科学取向的大数据研究。迈尔-舍恩伯格和库克耶的观点是这一倾向的典型代表。该观点的广泛流行表明这一倾向的影响不容小觑。应该说，这是一种短视而危险的倾向。人类不能满足于眼前的实用而放弃对科学的追求。且不说科学探索本身是一种乐趣，即使出于实用目的，放弃科学，最终也会损害人类的福祉。就像中国古代，曾经有着遥遥领先的实用技术，最终

却因为没有发展出物理、化学等纯粹的科学而落到西方国家后面。历史的教训应该记取。

中国当前方兴未艾的大数据社会治理，虽然涉及的是公共议题，主角是政府或公共事业组织，但从其思维和行为方式来看，也非常强调应用，急于事功而对发现事实背后的一般规律缺乏兴趣，应用取向色彩非常浓厚，比商界有过之而无不及。这是一种危险的倾向。没有理论指导的实践是盲目的，短期或许有一定效果，长期来看一定不可持续。特别是，社会治理的对象是人，而人是有反思性的，即可以根据对未来的预测而调整当下的行为。这就要求大数据研究不仅能够实时监测社会当下的状态，更要求其能够在一定程度上预测社会未来的状态，以便未雨绸缪。这就要求从当前的、已知的事实中发现带有一般性、普遍性的规律。而发现规律，正是科学的兴趣和本职所在。因此，基于大数据的社会治理必须尽快扭转应用取向主导一切的局面，大力发展科学取向的大数据研究。

3 大而不精——关于大数据科学价值的疑虑

要发展科学取向的大数据研究，就必须重视社会科学的理论和方法。“社会科学（social sciences）”指用现代科学的思维和方法去探究社会运作规律的所有学科，是复数而非单数，通常包含社会学、经济学、政治学等。也就是说，社会科学不等于社会学。不过，社会学有一个突出的特点，对于考察大数据与社会研究之间的关系是极有意义的，即它除了高度重视在研究中使用数据之外，还通过问卷调查、个案调查、参与观察、社会实验等方法亲自采集数据。在这个意义上，社会学可能是社会科学中对数据的环节涉猎最完整、体验最丰富的学科。因此，下面在讨论大数据与社会研究之间的关系时，会较多地援引社会学的观点、方法和事例。

社会学素来重视数据的采集和使用，但面对如火如荼的大数据热潮，却似乎有点无动于衷。截至目前，无论国内还是国外，应用大数据的社会学研究屈指可数。其中固然有大数据兴起时间不长，进入社会学研究尚有一个过程等客观原因，也与社会学家对大数据的科学价值心存疑虑有关。这些疑虑集中在4个方面，即大数据不够真、不够全、不够整齐、缺乏代表性。

不够真，是指大数据中的许多数值并不是真实社会过程的表示，比如微博数据中存在的大量假账号、假粉丝、“灌水帖”和虚假的个人注册信息等。造成数据失真的情况很多，大体可以分为两种：一种是由于技术失误或不成熟而产生的错误数据，另一种则是出于某种目的，故意操纵而产生的虚假数据。相对而言，前一种数据失真还好处理，后一种数据失真则比较麻烦，因为在技术较量中并不能保证优势在研究者这一边。任何数据的形成都存在失真的风险。但长期以来，社会学对数据采集中的失真风险已经形成一套较为成熟的控制体系，而大数据目前尚无与之相埒的办法。这是社会学家对大数据缺乏信心的原因之一。

不够全，是指大数据虽然大，实际上展现的社会信息十分有限，以致难以以之为基础进行复杂、严密的逻辑演算。社会学本质上是“群学”，在研究方法上特别注重分群比较。表现在统计上，就是倾向于根据个体的社会特征，比如性别、年龄、政治面貌、宗教信仰、教育程度、收入水平、职业、职级、所在行业等，将研究对象分成

若干组，然后比较组内差异和组间差异，并通过分析这些差异的原因和后果来揭示社会规律。这样，研究对象具有的社会特征就成为社会学推理中不可或缺的变量。然而，大数据常常只有总和层次（aggregate level）的变量，并且不是很多，个体层次（individual level）的变量更是严重缺乏，致使社会学的大量理论构想难以通过大数据进行检验和修正。这是社会学家对大数据不感兴趣的原因之二。

不够整齐，指大数据中变量的取值往往非常杂乱、发散而不够收敛，甚至存在大量缺失。因此造成的一个后果是，当进行社会学需要的分组比较时，大量组别内的个案数太少，以致统计结果不稳定，甚至无法进行比较。也就是说，大数据虽然体量巨大，从社会统计的角度来说却有些中看不中用。传统的社会学数据则不存在这个问题，因为这些数据中变量的赋值都是按照事先确定的统一标准进行的，即使是开放式调查，也可以通过后编码的方式实现取值的标准化。尽管从理论上说，大数据中各变量的取值也可以通过后编码的方式实现标准化，但正如后文将要指出的，由于技术、组织等多方面原因，事实上实现起来非常困难。这是社会学家对大数据态度冷淡的原因之三。

最后，是质疑大数据缺乏代表性。不少人认为，大数据就是全样本，样本代表性的思维已经过时。《大数据时代》一书就持这种观点，这是一种错误的看法。从科学的角度来说，研究网络社会最终还是为了探索整个社会生活。特别是社会学，揭示社会整体而非局部的运行规律是其作为一门学科的核心关切。而社会治理，更是要面向全社会，不能只面向网络社会。很显然，无论信息技术如何发达，来自网络社会的大数据永远不可能覆盖整个社会；技术再加上法律、伦理等诸多限制，使得电子数据永远只能展现社会生活的局部。换言之，从社会研究和治理的角度来看，大数据再大，也只是社会总体的一个样本，不可能是“全样本”。更何况，被大数据遗漏的那些部分往往并不是随机偏差，而是系统性偏差。如果大数据的代表性问题得不到解决，探寻社会整体运行规律从而推动全面善治的追求注定将遭到挫折。这无论对社会研究者，还是对社会治理者，都是不能接受的。大数据虽然以大著称，但它与社会总体之间的关系仍有许多依靠大数据本身无法得到澄清的问题。这是社会学家对大数据保持疑虑的原因之四。

比如，互联网上的各种意见，集合起来堪称海量，是当之无愧的大数据。但是，这些声音与全体国民的意见之间是什么关系？从社会学的角度来说，这个问题非常重要。因为一个社会中，有大量民众是不想上网、不能上网或上不起网的，而这批人的意见恰恰是最容易被剥夺、被忽视的。如果简单地以网民意见代替国民意见，造成的偏差及其后果将是十分严重的。要避免这样的偏差，就必须追问网民意见在多大程度上、在什么意义上代表着国民意见。不澄清数据的代表性，理论分析就难免陷入就事论事或过度推论的困境。

上述4个方面其实都是关于数据质量的担忧。一言以蔽之，就是大数据大而不精，难以满足社会学推理对于变量的丰富程度、变量值的精确和标准化程度以及样本代表性的要求。

4 大数据对社会研究的机遇与挑战

不少学者因为大数据在真实、系统、整齐和代表性等方面存在问题而怀疑其科学价值，进而对大数据研究持观望态度。

这些问题固然是事实，但同时应该看到，大数据也有相对于传统数据的优势。其中最突出的一点是，传统数据基本是拟态数据，而大数据基本是实态数据。所谓拟态数据，是指数据并非社会行为之实时的、原始的印迹，而是研究者通过某种研究设计去观测和捕捉的结果。由此造成以下3个问题。

第一，数据的形成高度依赖于研究设计。任何研究设计都是理论构想的产物，很显然，一个研究者无论多么追求客观，理论构想都不可避免地存在偏见（bias），由此造成所搜集的数据存在误差，甚至是严重的、系统性的误差。尽管经验社会学力图通过可重复的“假设—检验”过程不断消除理论构想中的偏见，但仍然难以彻底摆脱自证预言陷阱，即基于某种研究假设而进行的数据采集，可能把一些能够证伪这些假设的数据排除在外，从而使这些假设永远不会被证伪。

第二，数据的形成高度依赖于研究对象对研究设计的反应。社会研究的对象是人，而人是有反思能力的，会主动理解外部环境并相应调整自己的行为。同样地，在社会研究中，研究设计作为一种外部因素，也会影响研究对象的反应，从而导致测量不准。比如，调查问卷中的问题设置可能对受访者形成某种心理暗示，调查者的举止客观上会对受访者形成某种压力，从而诱导或迫使受访者往特定方向作出反应，如此等等。简而言之，在社会研究中，研究设计的介入会在不同程度上干扰研究者本来的状态，从而使通过该设计获得的数据出现误差，此即“霍桑效应”。

第三，传统数据无论多么真实、系统、整齐和有代表性，相对于观测的社会行为，它永远都是事后构拟的结果。即使是参与式观察，数据的发生与行为的发生也不是同步的，同样存在时差，只不过时差相对较小而已。至于抽样调查等数据采集方式造成的时差就更大了。假设研究者和被研究者都有前后两种状态，在前的记为S1，在后的记为S2，时差的存在意味着S2会影响对S1信息的捕捉，从而造成数据误差。比如，一个劳动者在失业后回忆失业前的职业状况时，受失业后精神状态的影响，可能夸大失业前的职业地位。

总之，在传统的社会研究中，数据多是研究者基于一定的研究设计对社会行为进行观测的结果，获得的只是拟态数据，并且由于多种因素影响，拟态数据对社会现实的观测总是存在误差，甚至发生严重的系统性误差。而大数据则不同，它是实态数据。这表现在，它或者是社会行动者主动生成的（比如微博），或者是自动生成的（比如应用所记录的活动轨迹），总之是社会行为的实时印迹，而非事后的构拟。这样，首先是真正实现了数据与行为同步发生，避免了延时观测或记录所造成的误差；其次，数据在形成过程中没有研究设计的介入，避免了研究设计不周全以及霍桑效应所造成的误差。从这个意义上讲，大数据对社会研究不啻是天赐良机。

然而，更重要的是，对社会研究来说，大数据不仅意味着机遇，而且是一个无法回避的挑战，因为互联网的出现已经深刻地改变了社会生态。这表现在，随着互联网应用的日益广泛和深入，一方面是“社会的数字化”，即社会中各色人等有意无意留下的数据足迹越来越丰富，现实社会活动于是越来越多地以数据的形式表现出来；另一方面是“数字的社会化”，即数据足迹及其结构本身就成为社会结构和过程的一个环节，从而不断塑造着新的社会秩序和关系。这两个过程连绵不绝地相互作用，使数据不再是现实社会的虚拟和映射，而是彻底与社会融为一体。这样，只要

研究社会，就必须研究数据，因为数据已经不再是研究者可以自主选择的研究方法和手段，而是研究者无法选择，也无法回避的社会本体的一部分。

典型的例子是网购。消费者在网购之后，部分人会留下网评。众所周知，首先，这些网评没有代表性，因为并不是所有消费者都会通过互联网购物，即使通过互联网购物，也不是所有人都会留下网评；其次，网评所对应的实质含义并不清晰：同样是给五星，有的是对商品质量的评价，有的是对快递速度的评价，有的是对商家态度的评价，如此等等，不一而足，有些商家尽管已经在设计上把上述几个方面分开，但消费者未必按照设计的板块去回答；最后，有些网评甚至是商家或其他行动者恶意操纵、造假的结果。但是，不管怎样，后来的消费者在购物时都会不同程度地参考这些网评。换言之，不管这些网评的真伪、含义和代表性如何，它都会影响实际的购物行为；数据可能是虚假的、含糊的，但造成的结果却是真实的、确定的。这样一种现象意味着，网评作为大数据已经与现实的消费行为高度融合，只要研究消费行为，就绕不开大数据。消费会影响生产，将来关于生产的研究恐怕也得研究这些网评数据。

现在流行一种观点，说互联网世界是对现实世界的映射，是与现实社会相对应的“虚拟社会”。这种观点是不对的。它只看到了“社会的数字化”，而未看到同时存在着另一个方面——“数字的社会化”，更未看到这两个方面已经实现高度融合，即以互联网为中介，社会不断地演变为数据，数据又不断地演变为社会。这样一种社会形态的出现决定了社会研究不面对大数据已经不可能了；要面对大数据已无需讨论，需要讨论的只是怎样面对大数据。

5 大数据的三重面相与不同学科的角色

大数据通常是指复杂程度大到超出常规处理能力的数据。大数据何以复杂？是因为它具有传统数据所不具有的独特特征。关于大数据的特征，分别有“3V”、“4V”和“5V”之说。所谓“3V”，是指大数据具有规模大（volume）、变化快（velocity）、结构复杂（variety）3个特点。“4V”则是再加一个特征——价值密度低（value），即相对于传统数据，同样单位大数据中的价值含量要低得多。4V再加上veracity，即是“5V”。veracity意为“真实性”。关于“真实性”怎么理解，可能有歧义。据笔者理解，这里所谓的“真实性”，不是指大数据中没有造假。由于技术、利益或道德原因，大数据中的错误和操纵比比皆是。这里说的“真实性”，应该指大数据是行动者根据本人意图而独立形成的，不受研究者的干涉和干扰。即使其中有造假，也是行为人基于自己独立的原因而造假，不是出于对某种研究设计的反应而造假。换言之，数据的形成与研究者的意图是相互独立的，不存在相互反馈；相对于特定的研究意图来说，大数据是真实的、无欺的。不难发现，这个意义上的“真实性”，其实就是前面所指出的：大数据是实态数据，而非拟态数据。

无论3V、4V，还是5V，都对大数据的特征做了很好的概括。但在这些概括之外，基于推动学科合作的目的，本文更想指出大数据的三重属性。

首先，如其名称所示，大数据具有数据属性，即它表现为一组有意义、有逻辑、可追寻、可计量的数值，可以用来揭示特定事物发生和演变的规律。这是任何数据，

不管大数据，还是传统数据，都具有的属性。只不过，传统数据是围绕特定意图并根据集中设计而形成的，价值密度很高；而大数据是用户自发形成的，比较散乱，价值密度低，追寻其意义和逻辑的工作也就更复杂。

其次，大数据具有强烈的技术属性。一方面，大数据的产生和形成与以互联网为代表的信息技术的迅猛发展有关；另一方面，数据的收集和处理也离不开信息技术。可以说，正是信息技术的无远弗届和强大处理能力，成就了大数据之大。离开信息技术，不仅没有物理意义上的大数据，也不会有逻辑意义上的大数据。传统数据的搜集和处理也会运用技术，但这些技术多是模块化、标准化和单机版的，易学易用，而大数据收集和处理涉及的技术就要复杂得多。

第三，大数据具有强烈的社会属性。大数据有两个基本来源[2]：一个是物理世界，比如对气象、设施、机械等运作状况的监测结果，另一个便是人类社会。社会研究主要涉及第二种来源的大数据。与传统数据的形成是一个高度控制性的过程不同，大数据的形成是一个高度开放性的过程。原因在于，大数据是特定人群范围在特定时间内活动的实时印迹和同步记录。这意味着，民众在数据形成中的角色由以往的被动变成了主动（包括自动）。在此过程中，参与的主体、过程和结果均不受研究者选择和控制。可以说，正是民众广泛而主动地参与数据形成，才成就了大数据之大。民众在数据形成过程中的广泛参与性，就是这里所说的大数据的社会属性。

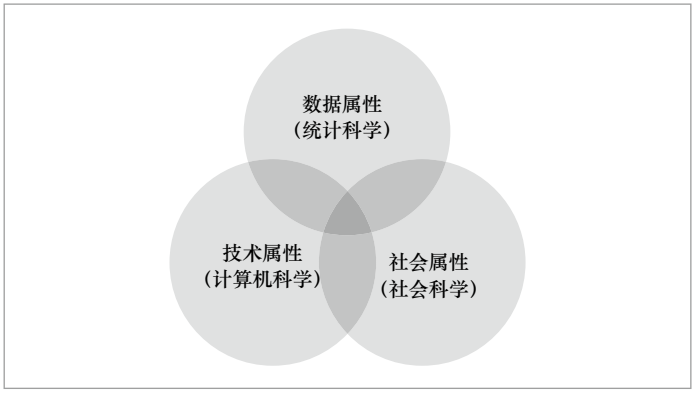

图1 大数据的三重属性与相关学科

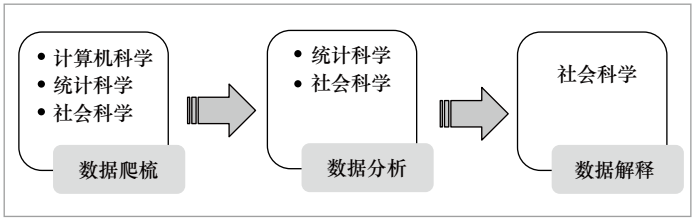

图2 大数据研究的基本流程与不同学科的角色

既然大数据同时具有上述三重属性，那么，如**图1**所示，任何关于大数据的分析和应用就必须同时处理这三重属性，方能修得正果。而这需要3个学科，即统计科学、计算机科学和社会科学的通力合作。其中，统计科学侧重应对数据属性，计算机科学侧重处理技术属性，社会科学则侧重探寻社会属性。

那么，3个学科究竟应该怎样分工和合作呢？这要从大数据社会研究的过程说起。基于大数据的社会研究大体可以划分为3个阶段：数据爬梳、数据分析和数据解释。如**图2**所示，随着阶段的变化，3个学科所扮演的角色及相互关系也会发生变化。

首先来看第一阶段，数据爬梳。该阶段的中心任务是实现数据形态从杂乱数据（messy data）向主题数据（thematic data）、从物理数据（physical data）向逻辑数据（logic data）的转变。具体来说是两个内容：一是数据的抽取，即根据特定的研究目的，从海量、多变而杂乱的数据足迹中把与研究主题相关的数据识别出来；二是根据分析的需要，把抽取出来的数据重新分类和赋值，实现数据的结构化。巧妇难为无米之炊，只有形成符合相应逻辑和格式要求的数据，后续分析和解释才有

2016014-8

米下锅。很显然，计算机科学是完成该任务的主角。原因很简单：以大数据之海量、多变和杂乱，传统的数据处理软件根本无法应付，必须运用深度学习、社会计算、知识计算等专门技术[3]。而这些技术之复杂和更新速度之快，不是其他学科的学者短时间能够掌握的，即使能够掌握也很不符合效率原则，因此必须有计算机科学的人加入。

但这并不是说社会科学和统计科学在这一阶段不重要，事实正好相反。如上所述，数据爬梳的核心任务是实现杂乱数据向主题数据、物理数据向逻辑数据的转变。这意味着，主题和逻辑的确定非常关键，否则数据的抽取和结构化就没有方向。而主题和逻辑来自对社会的洞察。这就需要社会科学。迈尔-舍恩伯格等人在《大数据时代》一书中提倡“让数据自己说话”，这个说法是站不住的。数据自己不可能说话，而只有经过理论指导的爬梳之后才能说话。没有爬梳，数据就是一团乱麻，不能说话；即使说话，也是胡话。而要爬梳，就离不开理论的指导。

当然，社会科学对主题和逻辑的确定并非一蹴而就，也需要不断地探索。所谓探索，就是在理论构想与数据事实之间来回折中，最后选择一个最佳方案。在此过程中，必然进行一些初步的、探索性的统计分析，因此，在这一阶段，统计科学的介入也是必不可少的。

数据爬梳一旦完成，就进入第二个阶段——数据分析，即挖掘数据之间的逻辑关系。这自然要用到统计工具，但模型的建立、参数的选择等，都离不开社会理论的指导。这已经是社会研究的常识，不赘述。由于爬梳好的数据已经按照一定主题和逻辑实现了结构化，因此可以用传统的社会统计软件进行分析，计算机科学相应就退出了这一阶段的工作。

接下来是第三个阶段——数据解释，即从当下数据之间已知的逻辑关系出发，推断更有一般性的规律，揭示更有一般性的原理。这个过程主要靠社会科学的理论思辨发挥作用，故连统计科学也退出舞台。

综上所述，社会科学是唯一贯穿3个阶段的学科。但这并不是说社会科学具有高于其他两个学科的特殊地位，毋宁说，社会理论对于大数据研究非常重要。这一点是连迈尔-舍恩伯格和库克耶都不否认的。在《大数据时代》中，他们一方面声称要终结因果分析，以便“让数据自己说话”，但另一方面也承认，因果关系的终结并不等于理论的终结，“理论的终结”的说法是荒谬的[1]：“大数据时代绝对不是一个理论消亡的时代，相反地，理论贯穿于大数据分析的方方面面。”

然而，当前的大数据研究，特别是国内的大数据研究，颇有些迷信“让数据自己说话”，忽视甚至轻视社会理论的倾向较为严重。事实上，即使是持这种态度的研究，也不是完全没有理论的指引，只是这些“理论”多属非专业学者对社会的直觉，不够系统和严密；或者不了解相关领域的理论进展，偶然发现一个理论便如获至宝，然后不顾适用条件地大用特用。社会科学的加入有利于改变这种凭感觉进行数据爬梳的状态。在大数据研究的草莽时代，凭直觉进行相关研究也许能在短期内取得一些甚至很“惊艳”的成绩，但从长期来看是没有竞争力的，是不可持续的。毕竟大数据具有强烈的社会属性，而术业有专攻，社会也不是凭直觉或所谓“智慧”就能参透的。

计算机科学家格雷曾在2007年提出大数据是“第四研究范式”的观点[4]。根据该观点，人类历史上先后有实验、理论推演、电脑仿真3种科学发现范式。而现在

人类能够采集和处理的数据是如此之多和大，以致研究者能够直接依靠现实的数据进行科学探索和发现。这就是所谓第四范式，即“数据密集型的科学发现（data-intensive scientific discovery）”。该观点虽然突出大数据在科学探索过程中的驱动作用，但并不否认理论的指导意义。第四范式的精髓并不是用大数据完全代替前三代范式中的实验、理论和模拟，而是在新的基础上将实验、理论、模拟与数据统一起来。第四范式中的“格雷法则”正是理论发挥引领作用的体现。

6 当前大数据社会研究面临的主要难题

大数据的三重属性决定了基于大数据的社会研究需要信息技术、统计分析和社会思想3种力量，从而需要计算机、统计学和社会学3个学科的紧密合作。然而，当前大数据社会研究的主要障碍正在于这3个学科之间的合作比较困难。事实上，在小数据时代，这3个学科曾经有过很好的合作。但大数据迥异于小数据的特征，使得原来的合作方式难以为继，而新的合作方式又一时难以建立。造成这种局面的原因，可以概括为两个方面：一是技术或曰生产力方面，即每个学科在大数据时代都面临新的困境，难以充分满足彼此的要求；二是体制或曰生产关系方面，即正是在这种情况下，不同学科之间的关系需要加紧调整和磨合，但由于学科属性、学科建制和市场选择等原因，调整和磨合的过程很艰难。

在历史上，计算机科学、统计学和社会学这3个学科一直有合作。相对来说，社会学与统计学的合作更紧密，社会学借助新的统计技术和模型得以迅速发展，而社会学问题的挑战也是推动统计学不断发展的重要动力。两个学科的紧密结合，使社会统计学成为一门独立的课程和研究领域。而计算机技术也为社会统计学的形成和发展贡献良多，社会统计的软件化和SPSS、SAS、STATA等著名统计软件的推出，更是使得社会统计作为一种理论和方法达到前所未有的繁荣。然而，大数据的出现对每个学科及其相互关系都提出了挑战。

首先，对社会学来说，以前虽然也在不断地收集和分析数据，但数据爬梳的任务很少。即使有，也主要是清除其中的噪音（比如数据中的作伪和逻辑矛盾）。而且由于这些数据都是根据一定研究设计而收集的，量小、集中、形态稳定并且结构化，因此，即使是噪音的清除，也可以用标准化、模块化的统计软件进行，社会学家只要在既有的统计软件平台上编程即可实现。而到了大数据时代，数据的基本特征是海量而且价值密度低，更严重的是多源、多变、异构、杂乱，数据爬梳的重点也随之从噪音的清除转向了数据的识别（抽取）和结构化。也就是说，大数据是高度非标准化、非结构化的，标准化、模块化的统计软件显然不能胜任。由于大数据的上述特征，甚至连噪音的清除也难以通过统计软件完成。

统计软件是标准化、模块化的，相对来说易学、易用，社会学家经过努力还能够掌握和运用。而现在大数据的处理，常常要求根据研究对象的特征从头构造或调整算法甚至处理系统，这就超出了一般社会学家的能力。社会学家即使努力为之，也不符合效率原则。总之，大数据使社会学对计算机科学的依赖程度大大加深。

在大数据出现之前，社会学也有通过编码把定性数据转变成可用于统计分析的计量数据的做法。这个工作在基本原理上与大数据爬梳相似，即反复聚类。其工作

2016014-10

过程大致如下：初步理论构想→通览原始资料→初步分类并编码→通览分类结果→调整理论构想→再读原始资料→调整分类并编码→……如此循环往复，直到分类和编码达到理论要求为止。这样一个工作过程，现在虽然有Nvivo等软件的帮助而省力很多，但仍然无法应付大数据。除了大数据本身规模巨大、结构复杂等原因之外，更重要的是，以往分类和编码的对象是确定的，而大数据的一个重要特征是频繁变动，这意味着前后两次分类面对的对象很可能不同（比如试图对网络公共知识分子进行分类，前后两天抓取的网络公共知识分子在人数和构成上很可能不同），从而使前一次分类对后续的分类调整失去参考意义，通过反复聚类实现合理分类编码的期望随之落空。要适应大数据多变的特征，必须依赖计算机技术。

然而，可能让社会学失望的是，计算机对大数据的处理也不是手到擒来。其挑战主要在3个方面[3]：一是数据本身的复杂性，即数据的类型和模式多样、关联关系繁杂、质量良莠不齐，使得数据的感知、表达、理解和计算等多个环节都面临不少难题；二是计算的复杂性，即大数据多源、异构、量大、多变等特征使传统的机器学习、信息检索、数据挖掘等计算方法不能有效支持大数据的处理、分析和计算；三是系统的复杂性，目前的计算硬件和软件系统在系统架构、计算框架和处理方法上都还难以满足大数据处理的需要。由于这些原因，目前计算机科学在数据爬梳技术的精度、效率、成本和易用性等方面，都难以满足社会研究的需要。当然，除了这些技术限制之外，还有伦理、法律等方面的限制。

其次，在社会学更加依赖计算机科学的同时，计算机科学也更加依赖社会学。原因在于，计算机科学虽然在信息处理方面技术非常强悍，但与以往不同的是，大数据是人类社会活动的产物，具有强烈而且不断变动的社会属性。离开对这些社会属性的理解，再好的算法和系统都不知道派什么用场，怎么派用场。如前所述，一些计算机学者凭着对社会的直觉也可能做出不错的大数据研究，但这并非长久之计。就此而言，计算机科学对社会学的依赖也在加深。然而，社会学的思想和理论通常比较晦涩、飘忽，让计算机学者难以在这些思想理论与计算机的工作对象之间建立起映射关系。社会学如何将抽象的思想和理论变成计算机学者可以理解、可以操作的任务，还有很长的路要走。

第三，数据爬梳也需要统计学的介入，但统计学面对大数据同样有自己的难题。数据爬梳并不是一个纯粹的技术过程，而是一个数据与思想反复碰撞的过程。在此过程中，需要不断对爬梳出来的数据进行统计分析，然后相应调整理论方案或技术路线。大数据再大，终归也是一种数据，必然适用统计学。统计学对数据爬梳也很重要。然而，传统统计学是基于小数据时代的条件发展起来的，无论理论还是方法都是以样本为基础展开的。但在大数据时代，数据的属性发生了很多变化，从而对统计学的传统理论和方法构成严峻挑战。比如，大数据中不同个案的发生经常不是独立随机事件，而是相互之间存在着强烈的正反馈或负反馈（典型表现是围绕特定事件而展开的公共讨论）；大数据独特的分布特征（比如重尾分布）会导致方差、标准差等标准方法变得无效，分布理论、大数定律和中心极限定理的应用也会受到限制[5]。如此等等。

总而言之，大数据对3个学科既有的研究范式都提出了新的挑战。这些挑战，使它们一方面更加相互依赖，但另一方面也使它们比以前更加难以满足彼此的需要，以致难以走到一起，或者不欢而散。这就

更需要3个学科求同化异，以更大的耐心和毅力推进合作。但不幸的是，合作的推进又面临学科属性、学科建制和学术市场等方面的障碍。

首先，因学科属性不同，3个学科在研究活动的组织方式上存在重大差别，从而影响相互之间的合作。在3个学科中，相对而言，计算机学科的研究活动具有更强的工程性质。这表现在，它可以将研究目标分解为若干边界比较清晰的任务，然后交由不同的研究人员和组织去实施，实现分进合击。相应地，其研究活动通常采用团队作战的实验室体制。同样由于其活动的工程性质，计算机学科的研究进度相对可控制、可预测。而统计学，尤其是社会学的研究活动则具有鲜明的思想属性。思想工作是很难分解的，难以想像让甲思考A部分，乙思考B部分，然后组合起来，就形成一个思想了。因此社会研究常常以个体的形式进行，很难采取团队作战的方式。与此同时，即使个人的思考，也比较依赖灵感，进度很难控制和预测。学科属性的差异给学科之间的合作造成一定困难。

举例言之。社会学家经常在拿到数据后，一时在理论上没有思路，于是陷入沉思，很长时间没有下文。也许突然有一天，理论灵感来了，他就急不可耐地想探测一下数据，看看新的思路是否可行，如果不可行又该如何调整，如此反复。正因为如此，社会学家的研究工作常常显得大起大落，节奏很不稳定。这虽然是社会学研究活动的固有特征，但确实让其他学科很难配合，甚至引起一些误解，认为社会学家“不靠谱”。

其次，还有学科建制上的障碍。按当前体制，这3个学科往往分属不同的研究单位。组织归属不同，科研议程的设置、资源的配备、绩效的考核也就不同。以前，学科之间在建制上的分割并不构成学科合作的严重障碍。因为在那个时候，学科之间的结合通常是知识的结合，而不需要组织建制的结合，只要有那么一两个学术精英善于结合不同学科的知识，创造出若干标准化的知识模板或研究范式，其他学者只管遵循和借鉴就可以了。在此过程中，学科之间主要是在知识上打交道，无需在组织和人员上打交道，即使打交道，也无需很多，现在则不然。大数据的基本特征恰恰是高度复杂，亦即高度非标准化。这一方面意味着，学科合作已经难以通过标准化的知识模板进行，而常常需要围绕特定问题“一事一议”地、面对面地碰撞和交流，从而需要把学科合作从知识层面延伸到组织和人事层面；另一方面也意味着，学科合作涉及的知识越来越多，越来越细，越来越复杂，相应地，标准化的知识模板也越来越难以形成。这样，怎样打破学科壁垒，如何通过组织和人员的融合来实现学科之间的融合，就成为一个重大问题，目前还没有找到有效的破解之道。

最后是市场选择。在大数据开发的两种取向中，社会研究更偏于科学取向，产品质量要求高，生产周期长，生产成本高，短期内却难以见到效益，自然在市场上不讨喜，因而在研究资源的获取上受到很大限制。而3个学科中的统计学，特别是计算机科学，其工作更容易被市场接受，更容易走应用路线。这样一种局面，对3个学科能否亲密合作，把一场注定艰辛的“爱情长跑”坚持到底是一个严峻的考验。从目前来看，形势并不乐观。

7 结论与展望

随着互联网的普及和信息技术的迅速发展以及国家对大数据社会治理的力推，大数据研究也越来越热。当前大数据开发

2016014-12

中存在着科学和应用两种取向，且呈应用取向完全压倒科学取向之势，这不利于大数据研究的可持续发展。大数据兼有技术、数据、社会三重属性，要推进科学取向的大数据研究，就必须有机地结合信息技术、统计和社会思想3种力量。这内在地要求计算机科学、统计科学和社会科学3个学科摒弃门户之见，实现通力合作。大数据研究绕不过社会科学，社会科学也绕不过大数据。在当前，由于各自技术能力的局限，学科属性的差异、学科体制的障碍、市场选择的偏向，3个学科之间的合作还比较困难。

这导致目前完整意义上的大数据研究并不多。从社会科学方面来看，多是利用一些已经比较结构化的大数据展开研究[6,7]，真正自己从头采集和爬梳数据的研究非常少[8]。由于这些数据的变量比较少，变量的取值和层次以及样本的代表性等，不尽符合社会学命题的要求，以致能够进行的社会学理论推演十分有限，甚至只能做一些粗浅的、宏观层面的描述统计。而计算机科学虽然在数据爬梳方面做了很多工作，但在研究主题的凝练和对社会机制的理解方面都比较薄弱，即使拉泽尔等人著名的《计算社会科学》[9]一文亦不免此病。这是缺乏社会理论引领的结果。总的来看，要真正做出既有思想深度，又有坚实数据支撑的大数据研究，还任重而道远。

现代社会是一个复杂而多变的巨系统，社会治理不能凭感觉率性而为。顺应社会和技术形势的变化，在社会治理过程中主动利用大数据，是社会治理方略的重大进步。与自然世界的运作一样，社会运作也有自己的规律。大数据虽然看上去庞大而“全面”，但其中蕴含的社会规律并不会自然显露，同样需要经过艰苦的科学探索，这就需要积极推进科学取向的大数据社会研究。离开坚实的社会研究，所谓以大数据为基础的社会治理只会是一枕黄粱。当前，在大数据研究领域，包括对大数据社会治理的研究，广泛存在着急功近利的倾向和对应用取向的迷恋。这要求政府应在尊重应用与科学两种取向合理分工的前提下，充分发挥调节作用，把科学取向的大数据研究提上重要日程，同时加大资源投入，将大数据研究作为一个基础性和战略性项目来支持。

致谢

感谢杜小勇、周雪光、张尹霞、庄家炽以及2016年1月16日中国人民国家发展与战略研究院“大数据与社会治理”会议上各位同仁的意见和建议。

参考文献：

[1] 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[M]. 盛杨燕，周涛，译. 杭州：浙江人民出版社，2013：27-97.
MAYER-SCHÖNBERGER V, CUKIER K. Big Data: A Revolution that Will Transform How We Live, Work, and Think[M]. Translated by SHENG Y Y, ZHOU T. Hangzhou: Zhejiang People's Publishing House, 2013: 27-97.

[2] 李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊，2012，27(6)：647-657.
LI G J, CHENG X Q. Research status and scientific thinking of big data[J]. Bulletin of the Chinese Academy of Sciences, 2012, 27(6): 647-657.

[3] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014，25(9)：1889-1908.
CHENG X Q, JIN X L, WANG Y Z, et al. Survey on big data system and analytic technology[J]. Journal of Software, 2014,

25(9): 1889-1908.

[4] HEY T, TANSLEY S, TOLLE K. 第四范式:数据密集型科学发现[M]. 潘教峰, 张晓林, 译. 北京: 科学出版社, 2012.
HEY T, TANSLEY S, TOLLE K. The Fourth Paradigm: Data-Intensive Scientific Discovery[M]. Translated by PAN J F, ZHANG X L. Beijing: Science Press, 2012.

[5] 游士兵, 张佩, 姚雪梅. 大数据对统计学的挑战和机遇[J]. 珞珈管理评论, 2013(2):165-167.
YOU S B, ZHANG P, YAO X M. The challenges and opportunities that big data brings to statistics[J]. Luojia Management Review, 2013(2): 165-167.

[6] 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[J]. 社会学研究, 2015(1): 23-48.
CHEN Y S. The trajectory of sociology over two centuries: a cultural study using millions of books[J]. Sociological Studies, 2015(1): 23-48.

[7] RIJT A V D, SHOR E, WARD C, et al. Only 15 minutes? The social stratification of fame in printed media[J]. American Sociological Review, 2013, 78(2):266-289.

[8] GARY K, JENNIFER P, ROBERTS M E. Reverse-engineering censorship in China: randomized experimentation and participant observation[J]. Science, 2013, 345(6199): 1-10.

[9] LAZER D, PENTLAND A, ADAMIC L, et al. Computational social science[J]. Science, 2009, 323(5915): 721-723.

作者简介

冯仕政 (1974-), 男, 中国人民大学社会与人口学院教授、副院长, 主要研究方向为政治社会学、社会治理、社会不平等。

收稿日期:2016-01-20
基金项目:中国人民大学科研基金资助项目“当前中国网络群体性事件的形成及治理研究”(No.13XNL005)
Foundation Item: Research Funds for Central Universities, and the Research Funds of Renmin University of China “The Online Collective Action in Current China” (No.13XNL005)