

大数据时代对统计学的挑战^{*}

邱东

内容提要: 本文首先探讨了面对大数据潮流应持有的科学态度, 然后从大数据能否淹没整个世界、信息与噪声能够泾渭分明吗、统计学与数据科学究竟是什么关系、大数据潮流对统计学究竟产生了什么样的影响等四个方面论述了大数据对统计学的挑战。

关键词: 大数据; 信息; 噪声; 数据科学; 统计学

中图分类号: C829.2

文献标识码: A

文章编号: 1002-4565(2014)01-0016-07

The Challenge of Statistics in the Age of Big Data

Qiu Dong

Abstract: This paper discusses the trend to big data which is due from scholars to scientific attitude, and then discusses the challenges of big data from four aspects as following: Can big data cover the whole world? Can Information and noise be quite distinct from each other? What's relationship between statistics and data sciences? What kind of impact generated on the trend of big data?

Key words: Big Data; Information; Noise; Statistics; Data Sciences

一、除了机遇还有挑战

世界潮流, 浩浩荡荡, 不可阻挡, 国人讲究识时务者为俊杰, 信息时代, 数据爆炸。大数据大势当前, 究竟采取什么样的态度才是真正的“识时务”?

大数据时代并不会自动生成, 总是需要不断地提出和解决大数据发展所遇到的问题和矛盾, 才会有切实的进步。事物发展的不同阶段有不同的“时务”, 需要不同的应对。

2009年, 大数据成为互联网信息技术行业的流行词汇。而早在1980年, 著名未来学家A. 托夫勒

出版《第三次浪潮》, 其中已将大数据赞颂为“第三次浪潮的华彩乐章”。此间30余年, 能不能看作大数据发展的萌芽期? 多数人对数据爆炸还懵懵懂懂, 世界需要赛博世界(Cyber world)的开拓者, 需要大数据潮流的预示者, 需要导师, 需要先声夺人。

一旦人们接受大数据汹涌而来的现实, 就需要既讲机遇, 也讲挑战。我们固然仍需要启蒙, 需要科普, 需要科学理论和方法论的“二传手”, 但不需要跟风, 不需要屏蔽了部分信息的“偏息图”, 不需要抓住一点不及其余的“唯数据论”, 不需要“应运而生”的投机者。我们更需要切实有学术增加值的数

^{*} 本文为第十七次全国统计科学讨论会特邀论文。

衡等传统计算机技术与现代网络技术融合起来, 把多个计算实体整合成一个具有强大计算能力的系统, 并借助SaaS、PaaS、IaaS、MSP等商业模式把它分布到终端用户手中。云计算的核心理念就是不断提高“云”处理能力来减少用户终端的处理负担, 使用户终端简化成一个单纯的输入输出设备, 并能按需享受强大的“云”计算处理能力。可见, 统计技术与云计算技术的融合是一种优势互补, 只有这样统计

技术才能在大数据时代一展身手、有所作为, 才能真正把统计思想在数据分析中得到体现, 实现统计分析研究的目的。

数据创造统计, 流量创新分析。由于各个应用领域的不断变化, 特别是数据来源与类型的不断变化, 使得统计学还难以成为一门真正成熟的科学。因此, 在数据分析的世界里, 不断提高驾驭数据的能力是统计学发展的终身动力。

据学科的拓展,更需要批判性思维。当事物的内在矛盾尚未充分暴露时就提出预警,这是一种制衡性的存在,是构成时代“全息图”的必要条件之一。

总之,在事物的不同成长期,学者应该有不同的担当,其使命的重心要有所不同。面对任何社会潮流,学者应该努力去做一位“麦田里的守望者”。

基于以上认识,且基于已经有很多人在阐述大数据对统计的机遇,本文论述大数据对统计可能形成的挑战。

二、大数据能否淹没整个世界

(一) 互联网外还有大鱼

面对大数据的迅猛发展,有的人认为,大数据可以覆盖整个世界,万维而结,天网恢恢,疏而不漏,概莫能外。甚至有人进而产生一种臆想:即使不能全覆盖也不要紧,大数据都覆盖不了的,那就是落伍的,就不值得覆盖,无关大局,推断整体时可以放弃。

实际情况未必完全如此。比如,部分技术精英,最早使用互联网和手机的信息技术先驱者,后来却竭力躲避“技术专政”。部分政治、宗教人士,最典型的如本·拉登,倾向于远离互联网。还有部分富人为了避税、避仇等原因,也尽可能躲避互联网的覆盖。显然,这三部分人是大数据难以覆盖的,而他们的经济行为恰恰对分析社会格局非常重要,推论时不可忽视。

无论是数理统计,还是数据科学,都对其在经济特别是金融领域的应用情有独钟。那么,在大数据时代的经济领域里,所谓地下经济(未观测经济)能不能被完全取缔?甚至,现金交易和易货贸易能不能被彻底消除?果真能够“数据全覆盖”,就意味着信息技术对人类经济行为的一种根本颠覆。数据已成海量、指数型增长,我们就能做此断言吗?

大数据的倡导者通常把民主、开放和理性作为必然的前提,这个前提确实应该得到满足,然而其不同国家和地区的实现程度是大不相同的。人类社会并不会同步进入大数据时代,“整个世界可能被割裂成大数据时代、小数据时代和物数据时代”(知名IT评论人谢文语)^[1],同一个地球,却是三种时代并存。

(二) 数据再大也是相对的

无论数据形成多么迅猛,无论覆盖如何全面,无论规模怎样大,大数据集仍然存在“数据黑暗地带”或“数据阴影区域”,也就是说,大数据集仍然存在着无法周全的“信号问题”。

数据的确大到了意想不到的程度,然而“大数据之大”也是相对的。海,对于人类、对于地球而言固然大,对宇宙来说就不那么大了。即便局限于地球,所谓海量数据对所研究的问题而言,规模也仍未见得就足够大。

比如,从皮尤研究中心可以获悉,美国上网的成年人中只有16%使用推特网(Twitter),与整体人口相比,其中年轻人和城市人的比例偏多,因而对全社会状况的分析来说,这绝不是一个具有代表性的样本。推特网的数据显示,人们离家越远越快乐。或许的确有人如此,但对多数人来说,这种推论是真实的么?

有报道称,全球所有数据的90%产生于过去两年,如果这个趋势按照大数据拥趸者的估计那样持续,那么今天的大数据相对而言只是明天的小数据,我们不仅不能穷尽所有数据,而且我们对数据的掌握始终将是非常有限的。从逻辑上讲,今天的“几近全覆盖”到了明天就会大打折扣,今天因数据全面而得到的结论也许会被明天否定,所谓全覆盖之说缺乏延展性。

参考文献

- [1] Viktor Mayer-Schönberger, Kenneth Cukier. 盛杨燕等译. 大数据时代[M]. 杭州: 浙江人民出版社, 2013.
- [2] Bill Franks. 黄海等译. 驾驭大数据[M]. 北京: 人民邮电出版社, 2013.
- [3] 涂子沛. 大数据[M]. 桂林: 广西师范大学出版社, 2013.
- [4] 郭晓科. 大数据[M]. 北京: 清华大学出版社, 2013.
- [5] C. R. Rao. 统计与真理[M]. 北京: 科学出版社, 2004.
- [6] 龚耘, 彭克慧. 哲学的故事[M]. 北京: 光明出版社, 2005.
- [7] 李金昌. 统计思想研究[M]. 北京: 中国统计出版社, 2009.

- [8] 张小明. 应急科技: 大数据时代的新进展[N]. 光明日报, 2013-10-14.

作者简介

李金昌,男,49岁,浙江义乌人。浙江工商大学副校长,统计学教授,博士生导师。研究方向为经济统计、统计理论与方法、抽样技术、政府统计。

(责任编辑:程 晔)

牛津大学教授维克托·迈尔·舍恩伯格被誉为“大数据时代的预言家”,他和肯尼思·库克耶编写了《大数据时代》^[1]一书,其中明确指出“人们总是受到现有测量和认知工具的局限,我们明天使用的工具很可能比今天的强大数倍甚至上千倍,我们现在所拥有的知识较之明天就显得微不足道了。”^[1]

(三) “道魔博弈”是动态无尽的

确实,科学技术再强大,也始终处于不断改进之中,相对于所要解决的问题而言,科学技术总是不完善的。因此,海量数据的规模与人类的有效处理能力之间也将一直存在着紧张关系,当前的主流软件工具能否在合理时间内完成海量数据处理的全过程,以生成有助于各类决策的信息,将始终是挑战性的。

更为要紧的疑问是,如果现实世界能被赛博世界完全操纵,不管它多么科学、多么先驱,现实世界还会存在么?如果大数据真是“全能方法”,那人类还能进步么?还用进步吗?说到底,大数据还是要为人类服务的,就人类的工具与其工作对象的关系而言,到底是道高一尺魔高一丈,还是道高一尺魔高九寸?到底是水涨船高,还是“水涨没顶”?

微软的史密斯说,“如果给我提供所有数据,我就能拯救世界。”可谁都知道,世界还在发展之中,而数据不过是对世界运行的记录,只要世界还没有完结,就不可能提供出“所有数据”。可见,史密斯聪明地预设了一个不可能满足的前提条件,他拯救世界的能力不可证伪。史密斯显然是在效仿阿基米德,给我一个支点,我就能撬动地球。

三、信息与噪声的辩证关系

大数据发展也引发了对基本概念的重新思考。信息和数据含义不同,但二者密切相关。英文的 data,我们通常译为“数据”。有学者提出,“数据之据”表明了其内涵的质的规定性,按照这种说法,可以有“数码与数据”之别,或者说“有据之数”与“无据之数”之别。我们知道,“具象数据”肯定给出了某种信息,但抽象的数码全都是信息吗?循着这类问题思考,就涉及到了数据学科的基础理论甚至哲学层面,个人的见解可能会有很大差异。

不过笔者认为,以下关于信息与噪声的认识应该是比较容易得到认同的。

(一) 信息与噪声的“一体性”

在科学领域,很少看到所有数据都集中到一个

明确的结论上的情况。真正的数据非常噪杂。数据就在那里,信息和噪声同时空存在,统计学家纳特·西尔弗说“只要能将信号与噪声区分开来,我们就能够获得所需的任何信息。”^[2]这位预测界的“神奇小子”说的是绝对真理,但问题恰恰在于信号与噪声难以区分,二者随使用者的变化而变化。从最终用途看,大部分数据对用户而言都是噪声。

人们拥有的信息呈指数增长,而需要验证的假设也正在以同样的速度增长。亟待解决的问题及其复杂程度也正在以同样的速度增长。大量的信息成倍增加,但有用的信息却非常有限。西尔弗指出:“噪声的增长速度要比信号快得多”^[2],因此,“信号的比例正在缩小,我们需要找到更好的方法对信号和噪声进行区分。”^[2]数据科学家通常用 4V 表达大数据的特点,有学者补充提出稀疏性(Sparsity),即有价值的信息相对于数据量而言非常至少,这个补充很有见地。

没有信息,就无法辨明真相。这会让人们以为:信息越多,就越靠近真相。很多情况如此,但这不是绝对的。信息只是得到真相的必要条件,而不是充分条件。无信息则无真相,不等于有信息便有真相。有信息而无真相,其重要原因就是噪声与信息的“一体性”。1936 年,《文学文摘》对 240 万美国公民调查,得出兰登胜选的预测,而乔治·盖洛普只调查了 5000 人,却宣布罗斯福会赢。这是小数据战胜大数据的典型案列,主要原因就是《文学文摘》的大数据中噪声过多。

科学使社会变得明朗,但科学同样也使社会组织变得更加复杂。信号过多又会使意义识别工作异常困难,这些信号可能会被淹没在震耳欲聋的噪声中。在统计学中,将噪声误以为信号的行为被称为过度拟合^[2]。遗憾的是,现实中过度拟合的事例太多。

无论从静态还是从动态看,信息与噪声之间都没有绝对的界限。如此说来,大数据时代同时也就是“大噪声时代”。

(二) 信息和噪声的“对象相对性”

种种情况表明了信息和噪声的“对象相对性”。

数据库专家杰克·奥尔森指出“数据能满足其既定的用途,它才有质量。”“信息的质量不仅取决于它本身,还取决于它的用途。”^[3]

对某些人而言是信息,对另外一些人而言则可

能完全是噪声。比较典型的例子如,第二次世界大战时美军内部使用印第安语联络,实际上已经是明码呼叫,对当时的日军而言,则是不可破解的密码。

本来是有用的、待用的信息,可是接受者没有能力在适当的时间里接受和消化,甚至成为接受其他更有用信息的负担,此信息对该接受者而言实质上处于噪声状态。究竟是信息还是噪声,依接受者的数据处理能力而定。

数据中信息多少还与使用者性格密切相关。对“狐狸型决策者”而言,数据越多,其提取的有用信息越多,其决策成功的可能性越大;对“刺猬型决策者”而言,数据越多,被噪声影响的可能性就越大,其决策失败的可能性也就越大^[2]。

环境对数据有用性的确定也有相当大的影响。比如,小规模恐怖袭击在以色列被视为普通犯罪,这里的每个人对恐惧都已经麻木了,以色列真正不能容忍的是潜在的大级别恐怖袭击。在这种状态下,涉及小规模恐怖袭击的消息对以色列人来说就不是什么有用的信息,不会影响其日常生活的安排。但类似消息对外地的旅游者而言则恐怕是致命的。

(三) 信息公布的广度与其价值可能存在逆相关关系

对多数人都可轻易得到的信息,其信息价值往往不大。多数人视若无睹的,甚至视为噪声的信息,对少数人而言其价值可能最大。

在某种意义上,待用信息即噪声,因为它虽然可能成为信息资源,但其潜在状态的保留也需要消耗资源。数据的价值是或然的,数据处理的损耗却是实在的。同时,待用信息的拥有量过大,还会影响用户选择的效率,也具有负面作用。

本来信息越多越好,这是人类倡导大数据的本意。然而信息广度与其价值背反却是人类难以完全解决的一个悖境。中国社会科学院信息化研究中心秘书长姜奇平指出“大数据的取舍之道,就是把有意义的留下来,把无意义的去掉。”^[4]

但取舍绝不是这么简单。社会是多元的,意义也是多元的,甲的“有意义”,可能正是乙的“无意义”。全社会达成共识的意义少之又少,难道只保留那些普世价值么?如果鼓励开放社会,那么众人意义的叠加,数据就很可能无可删除,哪一项也“舍”不掉。争吵到最后都得“取”,信息拥堵问题恐怕还在。

(四) 部分信息的价值具有时效性

信息获取时间的先后与其价值大小可能存在正相关关系。如果比其他人延迟获取,所得信息往往会失去其应用价值。

人们决策往往不能等到掌握全部甚至大部分信息,“限时性决策”的场合非常多,只有决策时由决策者掌握的数据才是具有实效的信息,否则就是噪声。

信息的时效性和多样性也为决策者坚持预定目标提供了方便。最为典型的例子就是美国发动的伊拉克战争。西尔弗指出:从错综复杂的数据中很容易看到你想要的信息,当我们过于想要发动一场战争时,不可靠的信号来源也会被解读成可靠的。例如伊拉克战争发生前,伊拉克工程师拉菲德·阿尔贾纳比因痛恨萨达姆而撒谎,他声称自己过去服务的种子工厂是一座制造生化武器的秘密工厂,随后他承认自己编造了一个虚假信息^[2]。而美方明知其证词为噪声,也将之作为信号加以利用。反战的民众没能及时得到相关信息,而战后得到的确切信息已经于事无补。

(五) 信息获取量的边际效用递减关系

对信息的接受者而言,通常都是信息越多越好。但如果发送信息的频率过高,获取持续时间过长,信息接受者就会产生审美疲劳,所得到信息的边际效用将从大变小,甚至为负,转变为噪声。

西尔弗指出“信息的增长速度远远超过了人们处理信息和分辨信息的速度,”^[2]面对数据盛宴,人们往往无从下手,从胃口和食物的配比看严重过量。面对过量的信息,人们本能地进行筛选,按照自己的偏好,选出喜欢的,忽略其他的。对于超负荷信息,人类只能启动自身固有的减载能力。

经济学诺奖得主赫伯特·西蒙说过“显而易见信息消费了什么,这就是信息接受者的注意力。因此丰富的信息导致了注意力的匮乏。匮乏资源不是信息,而是我们关注信息的处理能力。注意力是组织活动的主要瓶颈。”^[3]

尽管信息技术有助于人们处理信息,然而多数人仍然不可能做到极致,注意力的提升毕竟是有限的,如果信息过度膨胀,就容易出现“数据涝灾”。

当然信息和噪声之间还可能存在其他关系,以上五条不过抛砖引玉。

掌握大数据,正是为了得到有益于人类的信息,

主观上还会倾向于屏蔽或许无益于人类的噪声。显然,明确信息与噪声之间的辩证关系,才可能形成较为正确的大数据观。

四、统计学与数据科学

(一) “同一”还是“之一”

有的统计学者将统计学完全等价于数据科学,认定二者是一码事,持“同一论”或“唯一论”。此观点是否成立,涉及到对学科关系的不同认识。

宣称统计就是数据科学,作此断定恐怕有自我中心之嫌:大数据时代来临,统计就是一切,现实果真这么简单吗?

笔者涉猎相关文献得到的学科关系信息是:有不少著述在讨论数据科学分支时提到了统计学,就是说,不少学者持“之一论”或“属于论”。这意味着:统计学并不等价于数据科学。甚至还有学者认为统计学与数据科学之间应该存在某种交叉关系。

常常有统计学者将统计学定义为“一门数据科学”,这个定义或许没什么错,但如果统计学确是数据科学体系中的一门分支学科,那么这个定义虽然也提供了某些学科信息,但其贡献的新信息量实在有限。如同说“北京是中国的一个城市”,标示意义就没那么大。

统计是一门数据科学,大数据时代数据科学特别重要,所以统计学特别重要。这个推断逻辑上没问题。问题在于:“统计学特别重要”不等于“统计学唯一重要”。只有证明统计学是唯一的数据科学,二者同一,才能说明大数据时代“统计学唯一重要”。

(二) 三种学科关系

第一是统计学与其他数据科学的关系。

借助于大数据的便利,我们可以很方便地找到国内外对数据学(dataology)或数据科学的种种阐述,所提出的主要学科内容有:基础理论研究、数据实验和逻辑推理方法研究、领域数据学、数据资源的开发利用方法和技术研究等,远远超出了统计学的范围。

2013年5月,“香山科学会议”也围绕数据科学和大数据进行了专题讨论,会议邀请了管理、计算机、数学、经济、生物、社会和法律等领域的专家学者,其内容远远不止于统计学。可以说,统计学家并不是这种会议的主角,这也反证了统计学不是唯一的数据科学。

有人概括出数据科学的两个主要内涵:一是研究数据本身,另一个是为自然科学和社会科学研究提供一种新方法,称为“科学研究的数据方法”。还有人提出四门数据科学:计算机科学、数理统计学、图形设计学和人机交互学。舍恩伯格和库克耶甚至将作家也列入数据科学家的结合体!

当然,数据科学还在发展的过程之中,其定义还在总结和争论之中。但至少有一点对多数学者而言是明确的:无论是从属关系还是交叉关系,统计学只是数据科学体系中的一个组成部分或构件。

第二是统计学与自己学科的基础——概率论的关系。

笔者以为,统计学无非“应用概率学”。美国国家科学基金会(NSF)专门组织世界一流专家研究撰写了国际统计学发展报告——《统计学:二十一世纪的挑战和机遇》,其中并没有否认概率论作为统计学的数理基础。我们看到,这些专家充分了解大数据的时代背景及其对统计学的影响。

第三是统计学与所应用领域专门学科的关系。

统计学对所应用领域而言,其应用范围越广,越是大数据,越说明其方法论性质,越说明其工具性和基础性。

在数据科学探讨的相关文献中,我们发现,不少学者特别强调“学科群”的观念,强调自然科学、社会科学、数据科学之间的相互交叉,强调实质性科学在领域应用中的主导性。

(三) 四类世界与四种科学

人们对世界对科学有着各种各样的分类,体现不同的视角和观察格局。应对大数据时代,本文提出四类世界和四种科学。

广义地理解,应对物质世界,我们有物理科学。应对精神世界,我们有心理科学。应对赛博世界,我们有“数理科学”。应对行为世界,我们有“事理科学”。

这些学问都是人类多维思维的科学成果,然而它们之间并不是外在的板块叠加关系,而是彼此有机渗透的交错关系。正像现实世界并不是四个(或多个)分立的世界,而是同一世界四个不同维度的展示。真要学好用好某一科学成果,不可能单取其一。

世界是多元的,这意味着世界不只是data,故而不可用data代替一切,不能过度依赖data。定性和“较质”在大数据时代仍然不可或缺。

五、大数据时代的统计重心转移

(一) “样本 = 总体”意味着什么

舍恩伯格和库克耶指出:大数据不用随机分析法(抽样调查)这样的捷径,而采用所有数据的方法^[1]。所谓“所有数据”是一种相对的说法,但在工作思路,似乎又回转向了“全面调查”,数据科学家甚至提出了“样本 = 总体”的准则。这种巨大的调整,是否意味着统计重心需要转移呢?

维克托·迈尔·舍恩伯格和肯尼思·库克耶认为,随机采样方法存在许多固有的缺陷:一是采样的随机性很难实现,二是不适合考察子类别的情况,三是调查结果缺乏延展性,四是采样忽略了细节考察,五是无法用于奇异值分析^[1]。而大数据分析则可以弥补随机采样法的上述缺陷。

统计实务是一个包含着各不同阶段的全过程,从统计设计到基础数据收集、数据处理、数据分析、统计信息发布,环环相扣。统计学要为统计实务提供理论和方法论指导,需要针对不同阶段的各自特点而进行科学研究。

如果说原来的统计实务的重心在于收集基础数据,难点在于如何获取数据,那么在大数据时代,则更在于如何选择有用数据。如果说原来重心是“做加法”——无中生有,那么现在重心则是“做减法”——“有中生用”。用贺铿教授的话说,原来统计学的特点是以小见大,现在则更在于“由繁入简”。

前面讲到信息与噪声的内在关系,对统计学也会产生重要影响。在专门的抽样调查中,统计设计充分注意到了样本的随机性,不会有那么多噪声冲击。到了大数据时代,统计处理将面临许多非随机数据,如何剔除噪声就成了更为突出的任务。

另外,全球所有数据的90%产生于过去两年,这还意味着,从时间序列的角度看,大数据“厚今薄古”,数据呈严重“偏态分布”,如何在时序数据分析中注意这一特点的影响,也应该是一个重要课题。

由大数据时代的影响,至少有两点需要引起注意:其一是统计数据处理比基础数据收集更为重要;其二是统计设计环节的格外重要,总体上需要更注重不同阶段的衔接和反馈关系。

(二) 数据的识别问题

舍恩伯格和库克耶提出大数据时代的三大趋势,其中之一就是相关分析对原来因果分析的替代。

这种所谓替代并不是全然的,在大数据时代人们并不是放弃了因果关系的分析,而是借用相关分析作为重要乃至主要途径。因果关系不过是相关关系中的一部分,分析相关关系也正是为了间接得出对因果关系的认识。

大数据时代,数据的识别问题更为重要。这意味着,现实与理论、方法间需要反复作用,实质性科学的主导更为重要。哪怕是从纯海量数据中提取信息,没有先验认识,也还需要后验认识,从数据中总结出的数量规律能否成立?这是无论如何也不能避开的问题。

就是舍恩伯格和库克耶也不同意安德森的极端观点。他们指出“大数据绝不会叫嚣‘理论已死’”,因为“大数据是在理论的基础上形成的”,无论是如何搜集数据、分析数据,还是解读研究结果,都得依赖理论。“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。”^[1]

微软研究院首席研究员、MIT公民媒体中心客座教授凯特·克劳福德指出,“数据无法自己说话,而数据集——不管它们具有什么样的规模——仍然是人类设计的产物。大数据的工具——例如Apache Hadoop软件框架——并不能使我们摆脱曲解、隔阂和错误的成见。”^[5]一些著名案例可以说明这一点。

波士顿的StreetBump应用程序比较聪明,对公路质量状况,它试图从驾驶员的智能手机上取得数据。如果驾车经过路面坑洼处,智能手机就可以灵敏地反馈所受到的震动。然而,克劳福德指出,如果仅仅依靠智能手机用户的信息,那只是一个“自我选择样本”——它必然导致非智能手机用户数据的缺失,即年老和不那么富有的市民很容易被排除在数据分析之外。即便就年轻和富有者而言,智能手机信息也存在偏差的可能。因为驾车的质量不同,防震能力不同,智能手机所感应的震动程度也会有所区别。

数据误导的另一个典型案例是,2012年“谷歌流感趋势”过高地估计了年度流感发病率。通过人们在网搜索记录来判断这些人是否患上了流感,风险很大。人们可能是为了提前得到相关信息而搜索,可能是为了亲戚朋友而搜索,一个人感冒可以有N个人搜索,也完全可能只是关心社会动向而搜索,如何区分患者搜索和非患者搜索,还需进一步甄别。

类似的事例都表明,如果依赖有缺陷的大数据,完全可能对公共决策造成相当大的影响。

(三) 虚拟信息的识别问题

除了对“数据之据”的考察外,还有虚拟信息的识别问题。例如,许多推特账号实际上是机器人自动程序或“半机器人”系统(即得到机器人程序辅助的人工控制账号),还有虚假账号。最近的估计显示,可能存在多达 2000 万个虚假账号。

由此,当我们使用网络数据分析社会状况时,首先要警觉的问题是,数据中有没有由自动化算法系统产生的?如果有的话,究竟有多少?“架势无线”的 CEO 叶忻坦言,市场中此类数据的噪声太多,会导致数据价值大大降低。以互联网营销为例,大量的“刷量”以及水军好评差评等数据已经严重干扰了数据的准确性。

此外,在技术层面如何实现“数据去重”?如何确保数据的完整性?如何剔除网络病毒的影响,如此等等,也是不可忽视的。

(四) 为什么要提防“数据独裁”

舍恩伯格和库克耶专门论述了“数据独裁”问题,其典型案例就是罗伯特·麦克纳马拉。第二次世界大战时麦克纳马拉就是“统计控制队”的一名精英,战后当他接手福特公司时,数据管理成为他执掌福特的杀手锏。然而,上有政策下有对策,工厂经理会迅速生成总裁所要的数据。比如,麦克纳马拉规定,只有在旧车型的所有零件没有存货时,才能生产新车型,于是生产线经理就把剩余的零件全部倒进河里。麦克纳马拉担任美国国防部长后,评判战争进度的方法就是看对方的死亡人数。由于长官热衷于数据,下级为了达成命令或升迁,就一层一层地将数字扩大化,只要那是上级希望听到的数字。

政策制定者一旦锁定了一个特定变量,这个变量就会逐步失去其作为经济指标的价值,这是“古德哈特定律”所揭示的^[2]。只要利用数据进行管理,就必定有人利用数据进行“反管理”。至少人们具有这样一种行为倾向:在博弈中尽可能生成符合自己主观意愿的信息。甚至,罪犯可以利用“数据印迹”嫁祸于人,藏匿自己,“魔道之争”将在一个新的平台上展开。

按照索罗斯的“自反性原则”(the principle of reflexivity)事物参与者的看法与其所处状态互相影响。问题的要害恰恰在于,参与者自身正是他们所要解决的问题的组成部分^[2]。只要涉及到人类本身,就一定存在动态博弈。大数据对博弈的支持其实是“两边下注”的,社会事务的不确定性依然存在,甚至更为复杂。由此,“数据独裁”往往适得其反。

在大数据时代,还将不可避免地遭遇隐私保护问题,由此引发道德风险的若干争议。如何把握好数据开发和隐私保护的平衡,是对人类的重大挑战,统计学者也不能置身事外。这个话题笔者将专文另述。

参考文献

- [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]. 杭州:浙江人民出版社,2013年(中文版).
- [2] 纳特·希尔弗. 信号与噪声[M]. 北京:中信出版社,2013年(中文版).
- [3] 涂子沛. 大数据[M]. 桂林:广西师范大学出版社,2012.
- [4] 维克托·迈尔·舍恩伯格. 删除——大数据取舍之道[M]. 杭州:浙江人民出版社,2013年(中文版).
- [5] 凯特·克劳福德. 对大数据的再思考[J/OL].
- [6] 艾伯特·拉斯洛·巴拉巴西. 爆发——大数据时代预见未来的新思维[M]. 北京:中国人民大学出版社,2012年(中文版).
- [7] 数据科学与大数据的科学原理及发展前景——香山科学会议第462次学术讨论会综述[J/OL].
- [8] 刘大椿,刘永谋. 思想的攻防——另类科学哲学的兴起和演化[M]. 北京:中国人民大学出版社,2010.
- [9] D. 萨尔斯伯格著,邱东等译. 女士品茶——20世纪统计怎样改变了科学[M]. 北京:中国统计出版社,2004年(中文版).
- [10] 美国国家科学基金会(NSF)著,缪柏其等译. 统计学:二十一世纪的挑战和机遇[M].
- [11] 邱东. 经济统计学学科论[M]. 北京:中国财政经济出版社,2013.
- [12] George Soros. The Soros Lectures[M]. Public affairs, 2010.

作者简介

邱东,男,1990年毕业于东北财经大学,获经济学博士学位。现为北京师范大学长江学者特聘教授,国民核算研究院学术委员会主席。研究方向为国民经济统计。

(责任编辑:陈晨)