大数据时代思维方式变革的哲学意蕴

宋海龙
（解放军信息工程大学 理学院，郑州 450001）

摘　要：大数据时代思维方式变革呈现出追求全样本、接纳混乱性、关注相关关系等特征。从哲学层面分析，全样本体现的是开放系统的理念，肯定了事物作为系统与其环境之间存在的物质、能量和信息的交流，强调了事物自身演化发展的可能性。混乱性与大数据相伴生，接受混乱性是挖掘数据中隐含的潜在价值、对事物的演化发展做出精确预测的基本途径。相关关系是大数据时代统计因果关系的体现，这是由全样本系统、混乱性数据自身的非定域性及其与数据采集、分析过程的不可分离性所决定的，是在技术层面据以预测事物演化发展的前提。大数据时代思维方式变革的哲学意蕴还体现在科学研究范式的转化、人生态度的转变等方面。

关键词：大数据时代；全样本；相关关系；思维方式；哲学意蕴
中图分类号：B81　　**文献标志码：**A　　**文章编号：**1002－7408(2014)－05－0088－03

大数据(Big Data)正在以前所未有的速度颠覆着人们探索世界的方法。[1]所谓大数据，是指数据规模巨大、类型繁多、更新速度极快的数据库。[2]其要义有二：一是指数据量大到无法在一定时间内用常规软件工具对其内容进行提取、管理、分析、处理和应用的数据集合；二是指提取、管理、分析、处理和应用这些数据需要全新的技术体系。维克托·迈尔·舍恩伯格在《大数据时代》一书中指出，大数据时代处理数据理念上应有三大转变：要全体不要抽样；要效率不要绝对精确；要相关不要因果。[3]这些理念上的变化正在引起人们认识世界和改造世界的思维方式发生深刻变革。对大数据时代的数据理念进行理性分析，探析大数据时代思维方式变革的哲学意蕴，是顺应大数据时代大势，树立大数据思维，应对大数据时代挑战的基础工作。

一、小样本与全样本

维克托·迈尔·舍恩伯格指出，在小数据时代，人们针对随机性采样进行样本分析，这样的采样分析的精确性随着采样随机性的增加而大幅提升，且采样分析的精确性与样本数量的增加关系不大。在大数据时代，采用的是全数据模式，样本等于总体，[3]36－37大数据时代的样本分析需要全新工具——大数据科学。小数据时代的样本经典统计学分析能够更快更容易地发现问题，但不能预见事先未考虑到的问题；大数据时代的样本分析具有更开阔的视野，全样本是一座等待开采的金矿，具有发现问题的无限可能性。

数据科学与经典统计学具有本质的区别：第一，数据规模不同。大数据要分析的是与某事物相关联的所有数据，而不仅仅是传统意义上的样本数据。后者的使命是用尽可能少的数据来印证尽可能重大的发现，这已经成为我们的思维定式。第二，动静标准不同。在大数据时代，数据不再是静止、陈旧的，而是开放、动态的。小数据时代的经典统计学最基本的要求是数据方向单一、精确无误，大数据科学处理的数据容许不精确甚至错误、繁杂甚至混乱。经典统计学所使用的数据功能单一，用毕即弃。大数据时代的数据信息具有无限的、潜在的使用价值，可以永远贮存、反复使用。第三，数据收集形式不同。小数据时代，经典统计依赖于数据的随机采样，数据来源渠道单一、范围小，且无法显示细节信息。大数据时代，人们尽可能多渠道、多领域、多方式收集数据，接受其间混杂的错误和凌乱。第四，哲学关系不同。经典统计学关注的是因果关系，大数据转变了人们的思想，不再探求难以捉摸的因果关系，转而关注事物的相关关系。[4]

从科学层面看，小样本和全样本的区别仅仅在于信息科学的发展所提供的样本数据量的变化、样本分析工具的变化。而从哲学层面看，小样本和全样本的区别不仅在于样本数据量大小的不同，而且在于研究事物的思维方法的哲学基础不同。小样本遵循的是一种传统、封闭、静态地看待事物的理念，这种理念隔离了研究对象与其他事物、与环境之间的联系，斩断了研究对象自身的动态发展过程，终结了研究对象在发展过程中重塑自身新质、重建与其他事物之间联系的可能性，显露出明显的机械论哲学思维向唯物辩证法过渡的痕迹。全样本遵循的是现代、开放、动态地看待事物的理念，这种理念本质上是将研究对象视为一个开放的系统，肯定了研究对象与其所处环境之间存在的物质、能量和信息交流，强调了系统自身演化发展的可能性。维克托·迈尔·舍恩伯格曾列举事例：将一个在社区内有很多联系关系的人从社区中剔除掉，这个关系网会变得没那么高效但却不会解体；但如果将一个与所在社区之外的很多人有着连接关系的人从这个关系网中剔除，整个关系网很快就会破碎成很多小块。[3]43由此可见，环境对于系统存在及演化的重要意义。

系统遵循层次性原理，根据需要，我们可以灵活调整研究对象作为系统的规模范围，将其作为母系统或子系统

基金项目：国家社科基金军事学项目“适应大规模联合作战要求的军事信息人才培养问题研究”(13GJ003－068)资助。
作者简介：宋海龙(1964－)，男，河南方城人，解放军信息工程大学理学院教授，博士，研究方向：科学史、科学技术哲学。

看待；系统遵循相干性原理，其内部各要素之间发生着非线性相互作用，在系统演化过程中产生的扰动因素作为微“涨落”或者被系统的自稳机制所消除，或者被系统的非线性作用机制放大成巨“涨落”，进而引起系统朝着新的方向演化。这样，全样本就从根本上确保了研究对象的完整性，而不是肢解研究对象，仅仅抽取其某个方面、某个片段进行研究；确保了样本分析的客观性，而不是先入为主地按照事先设定的预案来“绑架”研究对象，为进一步揭示其本质和规律、做出新发现创造了条件。

二、精确性与混乱性

大数据时代，混乱是数据规模扩大的逻辑前提和必须付出的代价。维克托·迈尔·舍恩伯格指出，执迷于精确性是信息缺乏时代和模拟时代的产物。只有5%的数据是结构化且能适用于传统数据库的。如果不接受混乱，剩下95%的非结构化数据都无法使用，只有接受不精确性，我们才能打开一扇从未涉足的世界的窗户。其实，数据混乱并不可怕，精确的数据也未必能够保证获得令人满意的结果，这里的关键不在于数据的精确与否，而在于数据量的大小。在大数据时代，要想获得大规模数据带来的好处，混乱应该是一种标准途径，而不应该是竭力避免的。[3]60

在这里，我们可以看到维克托·迈尔·舍恩伯格面对精确和混乱时的矛盾情结。一方面，他认为因数据测量的不确定性所导致的混乱是不可避免的，量子力学中测不准关系原理就说明了这一点。“20世纪20年代，量子力学的发现永远粉碎了‘测量臻于至善’的幻梦。然而，在物理学这个小圈子之外的一些测量工程师和科学家仍沉湎在完美测量的梦中。”[3]47另一方面，他又认为数据的错误与混乱是我们在描摹现实时采取的一种不得已而为之的折衷方案。错误性不是大数据本身所固有的。它只是我们用来测量、记录和交流数据的工具的一个缺陷。如果说哪天数据变得完美无缺了，不精确的问题也就不复存在了。错误并不是大数据固有的特性，而是一个亟需我们去处理的现实问题，并且有可能长期存在。[3]56

从哲学的层面看，精确性与混乱性(或称为模糊性)是一对范畴，具有辩证统一的关系。[5]首先，精确性与混杂性都有一定的适用范围，不能混淆，这是由主体的主观目的性与客体的客观规定性所决定的。任何被称为精确的东西，在更高的意义上或更大的范围内又是混乱的，只有在一定条件下，精确性才具有绝对的意义。在小数据时代，主体认识目的的确定性容许通过对认识对象简单化的处理得到精确性的数据信息；在大数据时代，主体认识目的具有多元性、开放性和变动性，认识对象作为系统自身具有复杂性，决定了其数据信息的混乱性。其次，精确与混乱相互包含、相互转化，一方的存在和发展要以另一方的存在和发展为条件。精确是以否定混乱为前提，精确的目的是消除混乱；接受混乱，是为了得到精确。在大数据时代，数据科学的功能即是通过对表面上看似混乱的数据进行分析，从而得到精确的结论，以对事物发展做出正确的预测。

数据的混乱性，是大数据时代我们必须接受的通向精确性的惟一途径。只有接受了数据的混乱性，我们才有可能通过大数据科技手段，挖掘出数据中隐含的潜在价值。

三、因果关系与相关关系

维克托·迈尔·舍恩伯格认为，在大数据时代，知道是什么就够了，没有必要知道为什么。人们开始注重相关关系，而不再像小数据时代那样一定要追寻因果关系。因果关系往往来自经验，来自于经验中的直觉、信念，经不起实证的检验。寻找因果关系是现代社会的一神论，大数据推翻了这个论断。但我们又陷入了一个历史的困境，那就是我们活在一个“上帝已死”的时代。用相关关系取代因果关系，就能取得一石二鸟的功效：“既不损坏建立在因果推理基础之上的社会繁荣和人类前行的基石，又取得实际的进步”。[3]23相关关系的核心是量化两个数据值之间的数理关系。相关关系强是指当一个数据增加时，另一个数据很有可能也随之增加；相关关系弱是指当一个数据增加时，另一个数据值几乎不会发生变化。例如，沃尔玛公司发现，飓风用品与蛋挞食品的销售量之间存在相关关系，于是将蛋挞与飓风用品摆在一起销售，取得不错效果。再如，亚马逊公司曾聘用专业书评家来引导消费者购书，但专业书评家团队的业绩相较于只关注相关关系而不注重理性分析的软件分析要差很多，后来该部门被裁撤。[3]70-71大数据时代的相关关系分析，是克服因果探寻传统思维模式和特定领域里的固有偏见、深刻洞悉数据中潜藏的奥秘以进行科学预测的有效途径。大数据时代将要释放出的巨大价值使得我们选择大数据的理念和方法不再是一种权衡，而是通往未来的必然改变。[3]94

毫无疑问，就技术操作的层面上讲，在大数据时代，寻求数据之间的相关关系而不理会因果关系就能够对事物发展做出科学预测。如在2009年，谷歌公司的工程师们通过对海量相关数据的建模、分析，在流感爆发几周前，早于卫生组织而正确预测出了甲型H1N1流感传播的途径、时间和区域。[3]3-4

但如果从哲学层面进行分析，就存在以下几个问题：第一，世界上存在不存在因果关系；第二，因果关系与相关关系之间有无联系；第三，如何理解大数据时代的因果关系。关于本体论意义上的因果关系的探寻，是一个古老的哲学命题。英国近代著名哲学家休谟(David Hume，1711-1776年)认为，人的一切认识都来自感觉经验，在感觉之外，不管是物质实体，还是精神实体，经验都不能告诉我们它们是否存在。同样，被认为存在引起和被引起关系的因果关系也不能由经验证明。“任何物象都不能借它所呈现于感官前的各种性质，把产生它的原因揭露出，或把由它所生的结果揭露出来。”[6]因此，因果关系只是人们在认识世界时为求方便而做出的人为假定。显然，维克托·迈尔·舍恩伯格在《大数据时代》中对于因果关系的态度是受到了休谟的影响，但他又未明确否认因果关系的存在，他只是认为，在大数据时代，相关关系比因果关系更加重要。我们认为，存在因果关系是事物演化发展的逻辑条件，也是我们认识世界本质的逻辑前提；揭示因果关系是自然科学的中心任务，也是大数据时代隐藏在相关关系背后支配事物发展变化的决定力量。

关于因果关系与相关关系之间的关系问题，是科学与技术关系在大数据时代背景之下的一种折射。科学是探究因果关系即因果律的学问，而技术是解决问题的方法、技巧，两者关注的焦点存在差异，但两者并非对立的关系，如同技术解决“怎么做”、科学回答“为什么”一样，相关关系可以在实践中引导我们“怎么做”，而因果关系可以回答我们“为什么”这样做。其实，正如维克托·迈尔·舍恩伯格在《大数据时代》一书中一直努力想讲清为什么在大数据时代我们思维方式会出现种种新变革一样，即使在大数据时代，我们大家也一直没有放弃过对因果关系的追寻，这是由我们的思维本性所决定的，我们的思维智慧和习惯不允许我们仅仅止步于相关关系，而一定要挖掘出其背后隐藏的因果关系。进一步说，在我们进行数据分析之前，我们也一定在头脑中存在着关于因果判断的各种猜测，尽管这些猜测可能并未影响到数据分析的结果。

大数据时代的因果关系是一种什么样类型的因果关系？笔者认为，大数据时代人们之所以更加关注相关关系，是因为建立在经典统计学基础之上的经典因果关系(或经典因果律)在此已不适用，取而代之的或许应该是建立在量子统计学基础之上的统计因果关系(或统计因果律)，但这一点还没有被大家所认识，没有引起大家的足够重视。讨论这一问题，必须从20世纪上半叶著名物理学家爱因斯坦与玻尔关于量子力学领域中因果律的争论说起。爱因斯坦坚持的经典因果关系是微观粒子数据对测量仪器和测量过程具有独立性的拉普拉斯因果论，玻尔坚持的统计因果关系是微观粒子数据对测量仪器和测量过程具有相关性的非拉普拉斯因果论。大数据时代的全样本，因其作为系统的层次性、开放性和动态性，就其物理学特征而言，失去了空间中存在的定域性和与测量仪器的可分离性，而具备了量子力学中所描述的微观客体所具有存在的非定域性和与测量技术手段的不可分离性。大数据时代数据分析的结果与使用的分析工具、分析的过程之间均存在着相关关系。因此，大数据时代的因果关系不再适用于经典的拉普拉斯决定论，而适用于统计决定论。[7]

四、结语

大数据时代思维方式变革的三个典型特征：追求全样本而非小样本、混乱性而非精确性、相关关系而非因果关系，三者之间具有辩证统一的关系。全样本就必须接纳混乱的数据；大数据之大足以弥补数据混乱对样本分析结果引起的负面影响；对于全样本的分析处理，遵循的是统计决定论的因果关系，而不是经典拉普拉斯决定论的因果关系，在操作层面上直接表现出来的是相关关系。

大数据对思维方式变革的影响还远不止此。如科学研究范式从几千年前的经验科学过渡到几百年前的理论科学，再过渡到几十年前的计算科学，最后过渡到今天大数据时代的数据管理和软件分析。[8]又如，大数据时代永垂不朽变得容易，想很快地被人遗忘成为奢望。从结绳记事，到纸质印刷，再到电子传播，每一次媒介技术的进步，都极大拓展了人类的“记忆”能力。维克托·迈尔·舍恩伯格书中解释了“遗忘”如何因为数字技术和互联网的发展从常态变成例外：“人类对完整记忆的需求一直在持续上升，这让如今的世界已经被设置为记忆模式。”现在不是我们不想“遗忘”，而是我们无法“被遗忘”，大数据时代，是一个“记忆”高度完善的时代，也是一个很难“遗忘”的时代。[9]再如，大数据时代只关注数据和相关关系的做法，正在导致丧失生活意义的严重后果。巴拉巴西在《爆发》中介绍了一个虚拟网站，用户通过在搜索框中输入自己的名字就能知悉自己任何时间任何地点的监控录像信息。这种系统的建立在理论和技术方面上已不成问题。可以设想：一个人将其位置数据、财产信息都传到网上，但是关于这个人你一无所知，因为没有任何关于他性格、喜好等个性化信息，这是一个“什么都有，但什么都缺”的典型案例。[10]

正如维克托·迈尔·舍恩伯格指出的那样，面对大数据，人类社会曾沿袭多年的思维方式正在发生着变革：小数据时代推崇的抽样分析、精确性、因果律，在大数据时代均受到了挑战。我们需要树立大数据思维的概念，需要用新的信息分析框架来解读大数据。“人类灵感产生的各种火花第一次可以通过大数据多方面多层次爆发出来，这将是个美丽的新世界——人类的创造力可以在大数据中充分得到精彩的发现！”[11]

参考文献：

[1]赵国栋，易欢欢，糜万军. 大数据时代的历史机遇[M]. 北京：清华大学出版社，2013：5.

[2]孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，(1).

[3]舍恩伯格·W·M，库克耶·K. 大数据时代[M]. 盛杨燕，周涛，译. 杭州：浙江人民出版社，2013：17－18.

[4]薛红文. 大数据与统计学[N]. 现代物流报，2013－09－16.

[5]辛晓晖. 模糊性和精确性应作为一对范畴纳入哲学教科书[J]. 福建论坛(文史哲版)，1986，(2).

[6]休谟. 人类理解研究[M]. 关文运，译. 北京：商务印书馆，1957：28.

[7]黄继. 爱因斯坦与玻尔物理实在观的比较分析[J]. 南京航空航天大学学报(社会科学版)，2004，(3).

[8]邓仲华，李志芳. 科学研究范式的演化——大数据时代的科学研究第四范式[J]. 情报资料工作，2013，(4).

[9]张超. 大数据时代：当“被遗忘的权利”成为一个问题[J]. 中国图书评论，2013，(8).

[10]刘德寰，李雪莲. 大数据的风险和现存问题[J]. 广告大观(理论版)，2013，(3).

[11]孙黎. 大数据的想象力[J]. IT经理世界，2013，(21).

【责任编辑：孙　巍】