大数据研究的科学价值

李国杰
中国科学院计算技术研究所

关键词：大数据　数据科学　第四范式

近年来，“大数据”已经成为科技界和企业界关注的热点。2012年3月，美国奥巴马政府宣布投资2亿美元启动“大数据研究和发展计划”，这是继1993年美国宣布“信息高速公路”计划后的又一次重大科技发展部署。美国政府认为大数据是“未来的新石油”，将“大数据研究”上升为国家意志，对未来的科技与经济发展必将带来深远影响。一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制也将成为国家间和企业间新的争夺焦点。

与大数据的经济价值相比，大数据研究的科学价值似乎还没有引起足够的重视。本文试图对基于大数据的科学研究（包括自然科学、工程科学和社会科学）谈几点粗浅的认识，希望引起有关领域科技人员的争鸣。

推动大数据的动力主要是企业经济效益

数据是与自然资源、人力资源一样重要的战略资源，隐含巨大的经济价值，已引起科技界和和企业界的高度重视。如果有效地组织和使用大数据，将对经济发展产生巨大的推动作用，孕育出前所未有的机遇。奥莱利（O’Reilly）公司断言：“数据是下一个‘Intel inside’，未来属于将数据转换成产品的公司和人们。”

基因组学、蛋白组学、天体物理学和脑科学等都是以数据为中心的学科。这些领域的基础研究产生的数据越来越多，例如，用电子显微镜重建大脑中的突触网络，1立方毫米大脑的图像数据就超过1PB。但是，近年来大数据的飙升主要还是来自人们的日常生活，特别是互联网公司的服务。据IDC公司统计，2011年全球被创建和被复制的数据总量为1.8ZB（10^{21}），其中75%来自于个人（主要是图片、视频和音乐），远远超过人类有史以来所有印刷材料的数据总量（200PB）。谷歌公司通过大规模集群和MapReduce软件，每个月处理的数据量超过400PB；百度每天大约要处理几十PB数据；Facebook注册用户超过10亿，每月上传的照片超过10亿张，每天生成300TB以上的日志数据；淘宝网会员超过3.7亿，在线商品超过8.8亿，每天交易数千万笔，产生约20TB数据；雅虎的总存储容量超过100PB。传感网和物联网的蓬勃发展是大数据的又一推动力，各个城市的视频监控每时每刻都在采集巨量的流媒体数据。工业设备的监控也是大数据的重要来源。例如，劳斯莱斯公司对全世界数以万计的飞机引擎进行实时监控，每年传送PB量级的数据。

数据为王的大数据时代已经到来，战略需求也发生了重大转变：企业关注的重点转向数据，计算机行业正在转变为真正的信息行业，从追求计算速度转变为大数据处理能力，软件也将从编程为主转变为以数据为中心。采用大数据处理方法，生物制药、新材料研制生产的流程会发生革命性的变化，可以通过数据处理能力极高的计算机并行处理，同时进行大批量的仿真、比较和筛选，大大提高科研

和生产效率。数据已成为矿物和化学元素一样的原始材料，未来可能形成“数据探矿”、“数据化学”等新学科和新工艺模式。大数据处理的兴起也将改变云计算的发展方向，云计算正在进入以“分析即服务”（analysis as a service，AaaS）为主要标志的Cloud 2.0时代。

IBM、Oracle、微软、谷歌、亚马逊、Facebook等跨国巨头是发展大数据处理技术的主要推动者。自2005年以来，IBM投资160亿美元进行了30次与大数据有关的收购，促使其业绩稳定高速增长。2012年，IBM股价突破200美元大关，3年之内翻了3倍。华尔街早就开始招聘精通数据分析的天文学家和理论数学家来设计金融产品。IBM现在是全球数学博士的最大雇主，数学家正在将其数据分析的才能应用于石油勘探、医疗健康等各个领域。eBay通过数据挖掘可以精确计算出广告中的每一个关键字为公司带来的回报。通过对广告投放的优化，2007年以来eBay产品销售的广告费降低了99%，而顶级卖家占总销售额的百分比却上升至32%。目前推动大数据研究的动力主要是企业经济效益，巨大的经济利益驱使大企业不断扩大数据处理规模。

科技界要应对大数据带来的技术挑战

大数据研究的热潮激励基础研究的科研人员开始考虑“数据科学”问题。但必须指出，目前大数据的工程技术研究已走在科学研究的前面。当前的局面是各个学科的科学家都以自己为主处理本领域的海量数据，信息领域的科学家只能起到助手的作用。也就是说，各领域的科学问题还掌握在各学科的科学家手里，计算机科学家所提炼出的具有共性的大数据科学问题并不多。当技术上解决不了的问题越来越多时，就会逐步凝练出具有共性的科学挑战问题。在条件还不成熟的时候，计算机科学家应虚心地甘当一段时期的“助手”，虚心与各应用领域的科研人员合作，努力解决各领域大数据处理提出的技术挑战问题。对于网络大数据方面，计算机学者的主动性可能会较早发挥出来。

美国政府六个部门启动的大数据研究计划中，除了国家科学基金会的研究内容提到要“形成一个包括数学、统计基础和计算机算法的独特学科”外，绝大多数研究项目都是应对大数据带来的技术挑战，重视的是数据工程而不是数据科学，主要考虑大数据分析算法和系统的效率。例如，国防部高级研究计划局（DARPA）的大数据研究项目包括：多尺度异常检测项目，旨在解决大规模数据集的异常检测和特征化；网络内部威胁计划，旨在通过分析传感器和其他来源的信息，进行网络威胁和非常规战争行为的自动识别；Machine Reading项目，旨在实现人工智能的应用和发展学习系统，对自然文本进行知识插入。能源部（DOE）的大数据研究项目包括：机器学习、数据流的实时分析、非线性随机的数据缩减技术和可扩展的统计分析技术，其中，生物和环境研究计划的目标是大气辐射测量等气候研究设施，系统生物学知识库项目是对微生物、植物等生物群落功能的数据驱动的预测。国家人文基金会（NEH）项目包括：分析大数据的变化对人文社会科学的影响，如数字化的书籍和报纸数据库，从网络搜索，传感器和手机记录交易数据。国家科学基金会（NSF）的大数据项目的重点也是围绕突破关键技术，包括：从大量、多样、分散和异构的数据集中提取有用信息的核心技术；开发一种以统一的理论框架为原则的统计方法和可伸缩的网络模型算法，以区别适合随机性网络的方法。

现有的数据中心技术很难满足大数据的需求，需要考虑对整个IT架构进行革命性的重构。存储能力的增长远远赶不上数据的增长，设计最合理的分层存储架构已成为信息系统的关键。数据的移动已成为信息系统最大的开销，目前传送大数据最便宜的方式是通过飞机或地面交通工具运送磁盘而不是网络通信。信息系统需要从数据围着处理器转改为处理能力围着数据转，将计算用于数据，而不是将数据用于计算。大数据也导致高可扩展性成为信息系统最本质的需求，并发执行（同时执行的线程）的规模从现在的千万量级提高10亿级以上。

在应对处理大数据的各种技术挑战中，以下几个问题值得高度重视：

高效处理非结构化和半结构化数据 目前采集到的数据85%以上是非结构化和半结构化数据，传统的关系数据库无法胜任这些数据的处理，因为关系数据库系统的出发点是追求高度的数据一致性和容错性。根据CAP理论（consistency, availability, tolerance to network partitions），在分布式系统中，一致性、可用性和分区容错性三者不可兼得，因而并行关系数据库必然无法获得较强的扩展性和良好的系统可用性。系统的高扩展性是大数据分析最重要的需求，必须寻找高扩展性的数据分析技术。以MapReduce和Hadoop为代表的非关系数据分析技术，凭借其适合非结构数据处理、大规模并行处理、简单易用等突出优势，在互联网信息搜索和其他大数据分析领域取得了重大进展，已成为大数据分析的主流技术。尽管如此，MapReduce和Hadoop在应用性能等方面仍存在不少问题，还需要研究开发更有效、更实用的大数据分析和管理技术。

新的数据表示方法 目前表示数据的方法，不一定能直观地展现出数据本身的意义。要想有效利用数据并挖掘其中的知识，必须找到最合适的数据表示方法。若在一种不合适的数据表示中寻找大数据的固定模式、因果关系和关联时，可能会落入固有的偏见之中。数据表示方法和最初的数据填写者有着密切关系。如果原始数据有必要的标识，就会大大减轻事后数据识别和分类的困难。但为标识数据给用户增添麻烦往往得不到用户认可。研究既有效又简易的数据表示方法是处理网络大数据必须解决的技术难题之一。

数据融合 数据不整合就发挥不出大数据的重大价值。网上数据尤其是流媒体数据的泛滥与数据格式种类太多有关。大数据面临的一个重要问题是个人、企业和政府机构的各种数据和信息能否方便地融合。如同人类有许多种自然语言一样，作为信息空间（cyberspace）中唯一客观存在的数据难免有多种格式。但为了扫清网络大数据处理的障碍，应研究推广不与平台绑定的数据格式。大数据已成为联系人类社会、物理世界和信息空间的纽带，需要通过统一的数据格式构建融合人、机、物三元世界的统一的信息系统。

数据的去冗余和高效率低成本的数据存储 数据中有大量的冗余，消除冗余是降低开销的重要途径。大数据的存储方式不仅影响效率也影响成本，需要研究高效率低成本的数据存储方式。需要研究多源多模态数据的高质量获取与整合的理论和技术、错误自动检测与修复的理论和技术、低质量数据上的近似计算的理论和算法等。

适合不同行业的大数据挖掘分析工具和开发环境 不同行业需要不同的大数据分析工具和开发环境，应鼓励计算机算法研究人员与各领域的科研人员密切合作，在分析工具和开发环境上创新。当前跨领域跨行业的数据共享仍存在大量壁垒，海量数据的收集，特别是关联领域数据的同时收集还存在很大挑战。只有进行跨领域的数据分析，才有可能形成真正的知识和智能，产生更大的价值。

大幅度降低数据处理、存储和通信能耗的新技术 大数据的处理、存储和通信都将消耗大量的能源，研究创新的数据处理和传送的节能技术是重要的研究方向。

“数据科学”研究的对象是什么?

计算机科学是关于算法的科学，数据科学是关于数据的科学。从事数据科学研究的学者更关注数据的科学价值，试图把数据当成一个“自然体”来研究，提出所谓“数据界”（data nature）的概念，颇有把计算机科学划归为自然科学的倾向。但脱离各个领域的“物理世界”，作为客观事物间接存在形式的“数据界”究竟有什么共性问题目前还不清楚。物理世界在信息空间中有其数据映像，目前一些学者正在研究的数据界的规律其本质可能是物理世界的规律（还需要在物理世界中测试验证）。除去各个领域（天文、物理、生物、社会等）的规律，作为映像的“数据界”还有其独特的共同规律吗？这是一个值得深思的问题。

任何领域的研究，若要成为一门科学，研究的内容一定是研究共性的问题。针对非常狭窄的领域的某个具体问题，主要依靠该问题涉及的特殊条件和专门知识做数据挖掘，不大可能使大数据分析成为一门科学。数据研究能成为一门科学的前提是，在一个领域发现的数据相互关系和规律具有可推广到其他领域的普适性。事实上，过去的研究已发现，不同领域的数据分析方法和结果存在一定程度的普适性。IBM的经验表明：电网数据分析的算法也可应用于供水和交通管理上。抽象出一个领域的共性科学问题往往需要较长的时间，提炼“数据界”的共性科学问题还需要一段时间的实践积累。计算机界的学者至少在未来5~10年内，还需要多花一些精力协助其他领域的学者解决大数据带来的技术挑战问题。通过分层次的不断抽象，大数据的共性科学问题才会逐步清晰明朗。技术上解决不了的问题积累到相当的程度，科学问题就会浮现出来。

当前数据科学的目标还不十分明确，但与其他学科一样，科学研究的道路常常是先做“白盒研究”，知识积累多了就有可能抽象出通用性较强的“黑盒模型”和普适规律。数据库理论是一个很好的例子。在经历了层次数据库、网状数据库多年实践以后，柯德（Codd）发现了数据库应用的共性规律，建立了有坚实理论基础的关系模型。之前人们也一直在问数据库可不可能有共性的理论。现在我要做的事就是提出像关系数据库这样的理论来指导海量非结构化万维网（Web）数据的处理。

信息技术的发展使我们逐步进入“人—机—物”融合的三元世界，未来的世界可以做到“机中有人，人中有机，物中有机，机中有物”。所谓“机”就是联系人类社会（包括个人身体和大脑）与物理世界的信息空间，其最基本的构成元素是不同于原子和神经元的比特（bit）。物理空间和人类社会（包括人的大脑）都有共性的科学问题和规律，与这两者有密切联系的信息空间会不会有不同的共性科学问题？从“人—机—物”三元世界的角度来探讨数据科学的共性问题，也许是一个可以尝试的突破口。

目前，大数据的研究主要是将其作为一种研究方法或一种发现新知识的工具，而不是把数据本身当成研究目标。作为一种研究方法，它与数据挖掘、统计分析、搜索等人工智能方法有密切联系，但也应该有不同于统计学和人工智能的本质内涵。大数据研究是一种交叉科学研究，如何体现其交叉学科的特点需要认真思考。

在传统数据挖掘研究中，急用先研的短期行为较多，多数是为了某个具体问题研究应用技术，统一的理论还有待完善。传统的数据挖掘技术在数据维度和规模增大时，所需资源指数级增加，应对PB级以上的大数据还需要研究新的方法。统计学的目标是从各种类型的数据中提取有价值的信息，给人后见之明（hindsight）或预见（foresight），但一般不强调对事物的洞察力（insight），不强调因果逻辑。需要将其他方法和统计方法结合起来，采用多元化的方法来建立综合性模型。传统人工智能（AI）（如机器学习）可以接受$N\log N$甚至N^3级复杂度的算法，但面对PB级以上的海量数据，$N\log N$甚至线性复杂性的算法都难以接受，处理大数据需要更简单有效的人工智能算法和新的问题求解方法。

数据背后的共性问题——关系网络

观察各种复杂系统得到的大数据，直接反映的往往是一个个孤立的数据和分散的链接，但这些反映相互关系的链接整合起来就是一个网络。例如，基因数据构成基因网络，脑科学实验数据形成神经网络，万维网数据反映出社会网络。数据的共性、网络的整体特征隐藏在数据网络中，大数据往往以复杂关联的数据网络这样一种独特的形式存在，因此要理解大数据就要对大数据后面的网络进行深入分析。网络有不少参数和性质，如平均路径长度、度分布、聚集系数、核数和介数等，这些性质和参数也许能刻画大数据背后的网络的共性。因此，大数据面临的科学问题本质上可能就是网络科学问

题，复杂网络分析应该是数据科学的重要基石。

目前，研究万维网数据的学者以复杂网络上的数据（信息）传播机理、搜索、聚类、同步和控制作为主要研究方向。图1是1999年画出的万维网分布，由此导出无尺度网络（scale free network）。最新的研究成果表明，随机的无尺度网络不是一般的“小世界”，而是“超小世界（ultrasmall world），网络规模为N的最短路径的平均长度不是一般小世界的ln*N*，而是lnln*N*。网络数据科学应发现网络数据与信息产生、传播、影响背后的社会学、心理学、经济学的机理以及网络信息涌现的内在机制，同时利用这些机理研究互联网对政治、经济、文化、教育和科研的影响。

过去几个世纪主宰科学研究的方法一直是“还原论”，将世界万物不断分解到最小的单元。通过解构复杂系统，还原论带给我们单个节点和链接的理论。作为一种科研范式，还原论已经快走到尽头。尽管对单个人、单个基因以及单个原子等了解得越来越多，但我们对整个社会、整个生命系统、物质系统的理解并没有增加很多，有时可能距离理解系统的真谛更远了。例如，以解剖学为基础的现代医学离真正了解人体活动机理还有很大距离，据统计，医生对病因的判断有一半是错误的。而网络理论则反其道而行之，通过组装这些节点和链接，帮助我们重新看到整体。基于大数据对复杂系统进行整体性的研究，也许将为研究复杂系统提供新的途径。从这种意义上看，“网络数据科学”是从整体上研究复杂系统的一门科学。

发现无尺度网络的阿尔伯特（Albert-László Barabási）教授在2012年1月的《自然物理学》（*Nature Physics*）上发表一篇重要文章“The Network Takeover”。文章认为：20世纪是量子力学的世纪，从电子学到天文物理学，从核能到量子计算，都离不开量子力学；而到了21世纪，网络理论正在成为量子力学可尊敬的后继，正在构建一个新的理论和算法框架。

因果关系与相互关系

与传统的逻辑推理研究不同，大数据研究是对数量巨大的数据做统计性的搜索、比较、聚类和分类等分析归纳，因此继承了统计科学的一些特点。统计学关注数据的相关性或称关联性，所谓“相关性”是指两个或两个以上变量的取值之间存在某种规律性。“相关分析”的目的是找出数据集里隐藏的相互关系网（关联网），一般用支持度、可信度和兴趣度等参数反映相关性。两个数据A和B有相关性，只反映A和B在取值时相互有影响，并不能告诉我们有A就一定有B，或者反过来有B就一定有A。

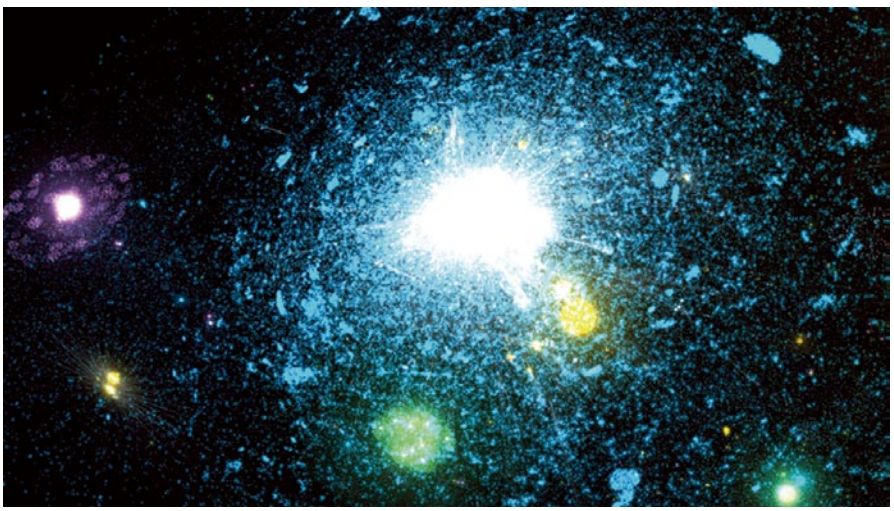

图1　万维网分布

严格来讲，统计学无法检验逻辑上的因果关系。例如，根据统计结果：可以说“吸烟的人群肺癌发病率会比不吸烟的人群高几倍”，但统计结果无法得出“吸烟致癌”的逻辑结论。我国概率统计领域的奠基人之一陈希孺院士生前常用这个例子来说明统计学的特点。他说：假如有这样一种基

因，它同时导致两件事情，一是使人喜欢抽烟，二是使这个人更容易得肺癌。这种假设也能解释上述统计结果，而在这种假设中，这个基因和肺癌就是因果关系，而吸烟和肺癌则是有相关性。统计学的相关性有时可能会产生把结果当成原因的错觉。例如，统计结果表明：下雨之前常见到燕子低飞，从时间先后看两者的关系可能得出燕子低飞是下雨的原因，而事实上，将要下雨才是燕子低飞的原因。

也许正是因为统计方法不致力于寻找真正的原因，才促使数据挖掘和大数据技术在商业领域广泛流行。企业的目标是多赚钱，只要从数据挖掘中发现某种措施与增加企业利润有较强的相关性，采取这种措施就是了，不必深究为什么能增加利润，更不必发现其背后的内在规律和模型。谷歌广告获得巨额收入经常被引用作为大数据相关分析的成功案例，美国《连线》（*Wires*）杂志主编克里斯·安德森（Chris Anderson）在他的著名文章“The End of Theory”的结尾发问：“现在是时候问这一句了：科学能从谷歌那儿学到什么？”

因果关系的研究曾经引发了科学体系的建立，近代科学体系获得的成就已经证明，科学是研究因果关系最重要的手段。对于相关性研究是可以替代因果分析的科学新发展还只是因果分析的补充，不同的学者有完全不同的看法。我们都是从做平面几何证明题开始进入科学的大花园的，脑子里固有的逻辑思维模式少不了因果分析，判断是否是真理也习惯看充分必要条件，对于大数据的关联分析蕴含的科学意义往往理解不深。对于简单的封闭系统，基于小数据的因果分析容易做到，当年开普勒发现行星三大定律，牛顿发现力学三大定律都是基于小数据。与距离平方成反比的万有引力定律可以从开普勒三定律通过逻辑演绎得到，并没有采用大量的数据统计得到1.999次方或2.001次方的精确结果。但对于开放复杂的巨系统，传统的因果分析难以奏效，原因在于系统中各个组成部分之间相互有影响，可能互为因果，因果关系隐藏在整个系统之中。现在的“因”可能是过去的“果”，此处的“果”也可能是别处的“因”，因果关系本质上是一种相互纠缠的相关性。在物理学的基本粒子理论中，颇受重视的欧几里德量子引力学（霍金所倡导的理论）本身并不包括因果律。因此，对于大数据的关联分析是不是“知其然而不知其所以然”，其中可能包含深奥的哲理，不能贸然下结论。

社会科学的大数据

根据数据的来源，大数据可以粗略地分成两大类：一类来自物理世界，另一类来自人类社会。前者多半是科学实验数据或传感数据，后者与人的活动有关系，特别是与互联网有关。这两类数据的处理方式和目标差别较大，不能照搬处理科学实验数据的方法来处理万维网数据。

科学实验是科技人员设计的，如何采集数据、处理数据事先都已经想好了，不管是检索还是模式识别，都有一定的科学规律可循。美国的大数据研究计划中专门列出寻找希格斯粒子（被称为“上帝粒子”）的大型强子对撞机（LHC）实验。这是一个典型的基于大数据的科学实验，至少要在1万亿个事例中才可能找出一个希格斯粒子。2012年7月4日，欧洲核子研究中心（CERN）宣布发现新的玻色子，标准差为4.9，被认为可能是希格斯玻色子（承认是希格斯玻色子粒子需要5个标准差，即99.99943%的可能性是对的）。设计这一实验的激动人心之处在于，不论找到还是没有找到希格斯粒子，都是物理学的重大突破。从这一实验可以看出，科学实验的大数据处理是整个实验的一个预定步骤，发现有价值的信息往往在预料之中。

万维网上的信息（比如微博）是千千万万的人随机产生的，从事社会科学研究的学者要从这些看似杂乱无章的数据中寻找有价值的蛛丝马迹。网络大数据有许多不同于自然科学数据的特点，包括多源异构、交互性、时效性、社会性、突发性和高噪声等，不但非结构化数据多，而且数据的实时性强，大量数据都是随机动态产生。科学数据的采集一般代价较高，大型强子对撞机实验设备花了几十亿美元。因此对采集哪些数据做过精心安排。而网

络数据的采集相对成本较低，网上许多数据是重复的或者没有价值，价值密度很低。一般而言，社会科学的大数据分析，特别是根据万维网数据做经济形势、安全形势和社会群体事件的预测，比科学实验的数据分析更困难。

未来的任务主要不是获取越来越多的数据，而是数据的去冗分类、去粗取精，从数据中挖掘知识。几百年来，科学研究一直在做“从薄到厚”的事情，把“小数据”变成“大数据”，现在要做的事情是“从厚到薄”，要把大数据变成小数据。要在不明显增加采集成本的条件下尽可能提高数据的质量。要研究如何科学合理地抽样采集数据，减少不必要的数据采集。两三岁的小孩学习识别动物和汽车等，往往几十张样本图片就足够了，研究清楚人类为什么具有小数据学习能力，对开展大数据分析研究具有深远的指导意义。

近十年来增长最快的数据是网络上传播的各种非结构化或半结构化的数据。网络数据的背后是相互联系的各种人群，网络大数据的处理能力直接关系到国家的信息空间安全和社会稳定。从心理学、经济学、信息科学等不同学科领域共同探讨网络数据的产生、扩散、涌现的基本规律，是建立安全和谐的网络环境的重大战略需求，是促使国家长治久安的大事。我国拥有世界上最多的网民和最大的访问量，在网络大数据分析方面已经有较强的基础，有可能做出世界领先的原始创新成果，应加大网络大数据分析方面的研究力度。

数据处理的复杂性

计算复杂性是计算机科学的基本问题。对于科学计算，我们主要考虑时间复杂性和空间复杂性。对于大数据处理，除了时间和空间复杂性外，可能需要考虑解决一个问题需要多大的数据量，暂且称为“数据量复杂性”。数据量复杂性和空间复杂性不是一个概念，空间复杂性要考虑计算过程中产生的空间需求。

设想有人采集完全随机地抛掷硬币的正反面数据，得到极长的“0”和“1”数字序列，通过统计可计算出现正面的比例。可以肯定，收集的数据越多，其结果与0.5的误差越小，这是一个无限渐进的过程。基于唯象假设的数据处理常出现这类增量式进步，数据多一点，结果就好一点点。这类问题的科学价值可能不大。反过来，可能有些问题的数据处理像个无底洞，无论多少数据都不可能解决问题。这种问题有些类似NP（non-deterministic polynomial，非确定性多项式）问题。我们需要建立一种理论，对求解一个问题达到某种满意程度（对于判定问题是指有多大把握说“是”或“否”，对于优化问题是指接近最优解的程度）需要多大规模的数据量给出理论上的判断。当然，目前还有很多问题还没有定义清楚，例如，对于网络搜索之类的问题，如何定义问题规模和数据规模等。

对从事大数据研究的学者而言，最有意思的问题应该是，解决一个问题的数据规模有一个阈值。数据少于这个阈值，问题解决不了；达到这个阈值，就可以解决以前解决不了的大问题；而数据规模超过这个阈值，对解决问题也没有更多的帮助。我把这类问题称为“预言性数据分析问题”，即在做大数据处理之前，我们可以预言，当数据量到达多大规模时，该问题的解可以达到何种满意程度。

与社会科学有关的大数据问题（比如舆情分析、情感分析等）遇到许多过去没有考虑过的理论问题，目前的研究才刚刚开始，迫切需要计算机学者与社会科学领域的学者密切合作，共同开拓新的疆域。借助大数据的推力，社会科学将脱下“准科学”的外衣，真正迈进科学的殿堂。

科研第四范式是思维方式的大变化

2007年，已故的图灵奖得主吉姆·格雷（Jim Gray）在他最后一次演讲中描绘了数据密集型科研“第四范式”（the fourth paradigm）的愿景（图2是微软公司出版的纪念吉姆·格雷的关于第四范式的专著）。将大数据科研从第三范式（计算机模拟）

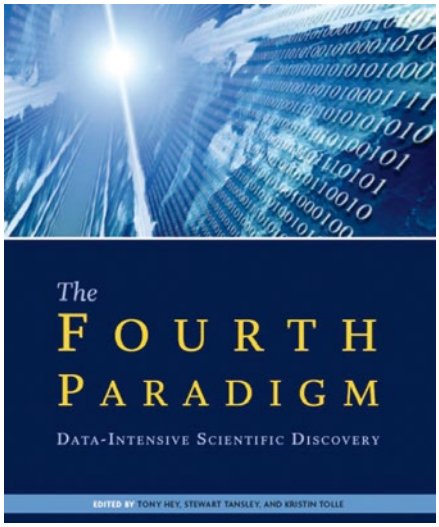

图2 《第四范式》封面

中分离出来单独作为一种科研范式，是因为其研究方式不同于基于数学模型的传统研究方式。谷歌公司的研究部主任彼得·诺维格（Peter Norvig）的一句名言可以概括两者的区别："所有的模型都是错误的，进一步说，没有模型你也可以成功"（All models are wrong, and increasingly you can succeed without them）。PB级数据使我们可以做到没有模型和假设就可以分析数据。将数据丢进巨大的计算机机群中，只要有相互关系的数据，统计分析算法就可以发现过去的科学方法发现不了的新模式、新知识甚至新规律。实际上，谷歌的广告优化配置、战胜人类的IBM沃森问答系统都是这么实现的，这就是"第四范式"的魅力！

美国《连线》杂志主编克里斯·安德森2008年曾发出"理论已终结"的惊人断言："数据洪流使（传统）科学方法变得过时"（The data deluge makes the scientific method obsolete）。他指出获得海量数据和处理这些数据的统计工具的可能性提供了理解世界的一条完整的新途径。PB（Petabytes）让我们说：相互关系已经足够（Correlation is enough）。我们可以停止寻找模型，相互关系取代了因果关系，没有具有一致性的模型、统一的理论和任何机械式的说明，科学也可以进步。

克里斯·安德森的极端看法并没有得到科学界的普遍认同，数据量的增加能否引起科研方法本质性的改变仍然是一个值得探讨的问题。对研究领域的深刻理解（如空气动力学方程用于风洞实验）和数据量的积累应该是一个迭代累进的过程。没有科学假设和模型就能发现新知识，这究竟有多大的普适性也需要实践来检验。我们需要思考：这类问题有多大的普遍性？这种优势是数据量特别大带来的还是问题本身有这种特性？所谓从数据中获取知识要不要人的参与，人在机器自动学习和运行中应该扮演什么角色？也许有些领域可以先用第四范式，等领域知识逐步丰富了再过渡到第三范式。我想，不管对"科研第四范式"的理解有多深，我们得承认：科研第四范式不仅是科研方式的转变，也是人们思维方式的大变化。

我对大数据的理解很肤浅，希望以上的"抛砖"能引来晶莹的"美玉"。■

李国杰

CCF名誉理事长，CCF会士，本刊主编。中国科学院计算技术研究所首席科学家，中国工程院院士。