

数据科学: 问题导向的交叉学科创新

在大数据应用技术需求的推动下, 大数据科学研究和应用已经成为科技领域中的热点. 同时人们开始从这些工程实践中提炼出一些共性的思路、方法和工具, 开启了一门新科学——数据科学.

从思想方法上, 数据科学研究继承了统计学的一些思想, 例如在大量数据上做统计性的搜索、比较、聚类或分类等分析归纳, 其结论是一种相关性, 而并非一定是某种因果关系. 虽然都依赖大量的计算, 但数据科学与计算机模拟不同, 并非是基于一个已知的数学模型, 而是用大量数据的相关性取代了因果关系和严格的理论和模型, 并基于这些相关性获得新的“知识”. 早在 2007 年, 著名计算机科学家吉姆·格雷就指出“数据密集型科学”已经成为继实验、理论、计算模拟之后的第 4 个科学研究范式. 吉姆·格雷博士当时指出, 发展数据科学的主要阻力在于工具的缺乏, 以及开发这些工具的巨大成本^[1]. 2010 年, *Science* 杂志在对 1700 多位科学家的调研后验证了这一观点: 技术的缺乏影响了科学家们分析、保存以及分享数据^[2].

在筹备本期专辑过程中, 《科学通报》编辑部向国内 2500 名科学家发送了问卷, 重复了这一调研, 回收 276 份. 虽然调研的样本范围不同, 没有直接的可比性, 但是结果也揭示了数据科学总体的一些发展趋势. 表 1 展示了主要数据的对比. 我们欣喜地看到, 83% 的样本都表示听说过大数据. 而且与之前的调研相比, 科学家们使用了更大的数据集, 更多的人开始使用公共数据集. 然而, 只有 28% 的样本认为自己体验过大数据技术, 同时经费和技术问题仍是影响数据科学发展的重要因素. 这些变化一方面是国际科学界加大了对数据科学的投入, 同时也因为大数据和云计算技术本身的发展, 一定程度上降低了数据处理的成本和技术门槛. 我们预期影响数据科学发展的客观因素会很快淡化, 而科学家的科研理念和专长等主观因素将会起到更为关键性的作用.

数据科学发展的第一推动力是人才. 数据科学是一个交叉学科的领域. 美国加州大学伯克利分校统计系的郁彬教授提出, 一个合格的数据科学家应具备的基本素质和技能可用 SDC³ 来概括, 即统计(statistics)、领域知识(domain knowledge)、计算技术(computing)、合作能力(collaboration)以及与非本领域的人沟通的能力(communication)(<http://statistics.berkeley.edu/~binyu/ps/papers2014/IMS-pres-address14-yu.pdf>). 这对科学家来说是非常高的要求. 更为棘手的是, 目前数据科学家在评价体系和职业发展前景、经费获取方面都存在相当的障碍. 2013 年 11 月美国摩尔基金会和斯隆基金会共同资助了 3780 万美元的“数据科学学院”项目, 其重要目标之一, 是加强跨学科数据科学人才的培养, 以及探索适合数据科学家的职业发展规划(http://www.moore.org/newsroom/in-the-news/2013/11/12/%20bold_new_partnership_launches_to_harness_potential_of_data_scientists_and_big_data).

本专辑在组稿过程中采用了交叉学科的思路, 邀请了一批不同领域的科学家, 就数据科学在自然科学、工程、社会科学、经济等各领域的应用, 以及统计、计算等数据科学的核心技术, 进行了报道. 虽然各领域对于

表 1 《科学通报》与 *Science* 对科学家使用大数据状况的调研结果对比

Table 1 How do scientists use big data? –Survey results by *Chinese Science Bulletin* and *Science*

调研问题(答案: 是或否)	<i>Science</i> (2010 年) ^[2]	《科学通报》(2014 年)
使用超过 100 GB 数据集	20%	30%
需要特殊的数据分析	84%	96%
所需的数据分析因为技术原因无法解决	27%	33%
缺少数据分析的经费	80%	65%
使用过公共数据集	44%	73%
与他人成功分享过已发表的原始数据	74%	75%

“大数据”的理解千差万别，数据的形式更是多样，然而，我们可以看到，各个领域大数据在面临的问题、需要的系统、方法和工具上都是相通的。例如，对于存储容量、带宽和经济性的要求、数据处理系统的弹性、灵活性和易用性的要求，以及对于跨国家、跨研究组数据共享的需求，成为了大数据对计算机系统领域需求的热点。从计算工具上看，语义分析、复杂网络分析、不同模态数据特征的融合、多媒体数据的特征抽取等，都是数据科学应用，尤其是社会科学领域应用的热点。如何能更好地提炼、抽象这些方法和工具，将之扩展到更多的领域，让更多的科学家和工程技术人员一起探讨在科研中使用大数据的共性问题，同时在科研中培养出进行跨领域数据科学的人才，可能是数据科学资金、技术和人才匮乏问题的合理解决方法。

正如本期中《大数据背景下的心理学：中国心理学的学科体系重构及特征》一文^[3]所提出的，“发现问题、提出问题、解决需求的价值在研究层面上远远大于精致计算、细致操作的价值”。我们希望本期专辑能起到抛砖引玉的作用，让不同学科背景的研究者围绕数据科学这一题目，进行跨学科的深入探讨，拓展思路，提出并解决对科学发展、对国计民生至关重要的问题，为促进我国数据科学的进一步发展做出贡献。

- 1 Gray J, Hey T, Tansley S, et al. Jim Gray on eScience: A Transformed Scientific Method. Washington: Microsoft Press, 2007
- 2 Science Staff. Introduction – Challenges and opportunities. Science, 2011, 331: 1515–1516
- 3 Yu F, Peng K P, Zheng X J. Big data and psychology in China (in Chinese). Chin Sci Bull, 2015, 60: 520–533 [喻丰, 彭凯平, 郑先隽. 大数据背景下的心理学：中国心理学的学科体系重构及特征. 科学通报, 2015, 60: 520–533]

张云泉^①, 徐葳^②, 龙桂鲁^③

① 中国科学院计算技术研究所, 计算机体系结构国家重点实验室, 北京 100190;

② 清华大学交叉信息研究院, 北京 100084;

③ 清华大学物理系, 清华信息科学技术国家实验室(筹), 北京 100084