让大数据分析更可信

王腾蛟，李喜莲
北京大学信息科学技术学院，北京 100871

摘要
大数据在学术界和产业界的各个领域正扮演着愈加重要的角色，但同时，大数据是否可信，引发了无数研究者的广泛关注和激烈讨论。从大数据名称的历史演变、大数据应用的案例分析以及大数据工程的角度探索大数据的可信程度，并由此总结出保证大数据分析正确性需要解决的3个挑战：正确选择数据源、科学抽样有代表性和有价值的数据、严谨完备的大数据工程分析方法。

关键词
大数据；数据仓库；海量数据；大数据工程

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017017

Making big data analysis more credible

WANG Tengjiao, LI Xilian
School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China

Abstract
Big data is playing an increasingly important role in various areas of academia and industry. However, whether big data can be trusted has caused widespread concern and intense discussion among countless researchers. The credibility of big data from the historical evolution of big data names, case studies of big data applications and big data engineering was explored, and thus the three challenges needed to be addressed to ensure the correctness of big data analysis were concluded: the right choice of data source, the scientific sampling of representative and valuable data, the rigorous and complete big data engineering analysis method.

Key words
big data, data warehouse, massive data, big data engineering

1 引言

《Nature》大数据专栏曾经论述数据在数学、物理、工程、社会及经济等领域扮演着愈加重要的角色，《光明日报》报道，以“社交网络分析”“观点挖掘”等大数据技术为基础的人文社会科学研究方法逐渐被采纳，并成为重要的发展方向，人文社会科学的大数据研究必然会成为人文社会科学的主流领域之一。因此，学术界大数据已经成为众多学科的一种科研方法，工业界依据大数据的分析结果进行决策支持的方法已被广泛应用于各个行业。大数据是否可信，引发了人们的激烈讨论。

本文主要从大数据的历史演变、大数据分析的可信性以及大数据工程的角度，分析大数据面临的机遇和挑战以及如何应用科学合理的工程方法保证大数据分析的正确性。

2 大数据名称的历史演变

从数据管理系统诞生以来，人们对如何管理和分析“大”数据的期望不断提高，同时也在努力探索数据管理和数据分析的新技术与新方法。从数据库领域的研究者来看，当前大数据概念的诞生，有其历史必然性，并有一个清晰的演变轨迹：从超大规模数据库（very large data base，VLDB）的发展到海量数据（massive data）的兴起，再到当前大数据（big data）的热潮。

2.1 20世纪70—80年代：VLDB的发展

20世纪70—80年代，人们用VLDB描述管理大规模数据的技术。

这一时期的数据管理技术有了突破性的进展，即诞生了关系数据库的概念，关系数据库管理系统在产业界快速发展。1970年，“关系数据库之父”Codd E F博士在期刊《Communications of the ACM》上发表了一篇名为《A relational model of data for large shared data banks》的论文，提出关系模型的概念，奠定了关系模型的理论基础。其后涌现了很多关系型数据库系统，其代表产品有Oracle、IBM公司的DB2、微软公司的MS SQL Server以及Informix等。有了关系数据库管理系统的支撑，数据管理的规模显著提高，人们管理数据的能力也大大增强。

同时，大量的研究学者和学术会议聚焦于探讨新的应用领域面临的数据管理问题和技术挑战。1975年，第一届VLDB学术会议在美国马萨诸塞州的费雷明汉召开。经过几十年的发展，现在VLDB学术会议已成为数据库领域中最主要、规模最大的国际学术会议之一。

2.2 20世纪90年代—21世纪初：海量数据的兴起

20世纪90年代—21世纪初，人们开始用海量数据来描述管理和分析大规模数据的技术。

这一时期的数据管理技术有了另一项突破性的进展：“数据仓库”和“数据挖掘”技术从概念走向应用。20世纪90年代初，美国著名信息工程专家Inmon W H在其里程碑式的著作《Building the data warehouse》提出“数据仓库”的概念，认为数据仓库是一个面向主题、集成、相对稳定、反映历史变化的数据集合，是用于决策支持的知识管理。数据仓库与数据库不同，其重点不是数据的简单堆积，而是

从容量庞大型数据库中抽取数据，并将其清理、转换为新的存储格式和集中统一、随时可用的知识和信息。“数据挖掘”也是随“数据仓库”产生的一个新概念，是一个自动或半自动化地从大量数据中发现有效、有意义、潜在有用、易于理解的数据模式的复杂过程。

20世纪90年代后期，全球的数据资源高速增长，主要体现在两个方面：一是互联网及其应用的快速发展产生了海量数据；二是各个行业大批成熟的业务信息系统投入运行，并积累了海量数据。此时人们认识到，数据是一种宝贵的资源，但还没有充分发挥作用。

在学术界，1991年、1993年、1994年相继举行了数据库中知识发现（knowledge discovery in database，KDD）专题讨论会，并于1995年由专题讨论会发展成为国际会议年会。1999年，国家重点基础研究发展计划（“973”计划）中设立了“网络环境下海量信息的组织与处理的理论与方法研究”项目，探索并解决网络环境下拥有海量信息、运行海量进程的服务软件系统的设计、实现和维护中的重大问题。

2.3 当前：大数据的热潮

当前全球的大数据热潮，无论深度还是广度都超过了VLDB和海量数据时代。

但是大数据热潮也带来了巨大的争议。一种观点认为大数据技术推动了新的应用；另一种观点则认为，当前的大数据缺少革命性的技术进展，是广泛的应用需求再次激发了对数据管理和分析技术的探索。然而，全社会毕竟开始认识到数据的重要性，计算机科学家毕竟Lee T B曾经强调数据的价值：数据的价值将使得比系统本身更持久（data is a precious thing and will last longer than the systems themselves）。全球大数据市场规模如**图1**所示。

大数据的出现吸引了政府和科学领域的研究热潮，有支持者也有反对者。

在政府与国际关系方面，2012年3月29日奥巴马政府公布了“大数据研究和发展计划”，旨在帮助美国提升从海量和复杂的数据中获取知识的能力，将大数据的研究上升为国家意志。

在科学研究方面，2008年9月《Nature》杂志出版了一期大数据专刊，论述数据在数学、物理、工程及社会、经济等领域扮演的愈加重要的角色。2011年2月，《Science》也推出了一期关于数据处理的专刊，从互联网技术、互联网经济学、超级计算、环境科学等多个方面介绍了海量数据带来的技术挑战。2012年11月7日，微软发布《第四范式：数据密集型科学发现》，大数据科研方法从之前的实验归纳型科研、模型推演型科研、计算模拟型科研推进到第四范式——数据密集型科研。

反对的言论也有很多，比如以机器学习领域著名学者Jordan M I教授为代表的反对者认为，“大数据的冬天”将要来临。

仔细分析反对者的态度，并没有否定大数据的发展趋势。反对者认为，海量数

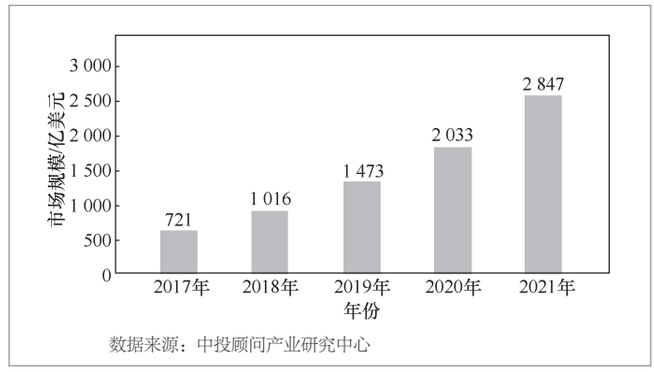

图1 全球大数据市场规模

据确实可以经过分析得到一定的结论，但是面对同样的数据，不同的方法可能得到的结论是矛盾的。并且大数据从“定义”到“4V（volume，variety，velocity，value）”，仅仅强调了数据的特点，并没有强调新的规范高效的数据管理技术，才引发了大家在大数据核心技术方面的争议。

3 大数据是否可信

在大数据的热潮中，大数据已经成为学术界众多学科的一种科研方法，在产业界，依据大数据的分析结果进行决策支持的方法已被广泛应用于各个行业。但是人们注意到，大数据有成功亦有失败的案例，在这个背景下，大数据是否可信，引发了人们的激烈讨论。

大数据在政治[1]、经济[2]和文化[3]领域获得了一些成功应用，如股票预测、房价涨跌预测、政治情感预测等。罗伯特·邦德等人在《Nature》杂志上发文比较了网络社会网络①和面对面社会网络②影响政治行为的路径。他们在2010年美国国会大选期间对6 100万个Facebook用户实施了一项发送政治动员消息的随机控制实验，研究发现，这些消息直接影响了数百万人的政治的自我表达、信息寻找和现实生活中的投票行为[1]。《Science》上另外一篇文章谈到“用大数据分析世界文化中心的兴衰”，使用谷歌Freebase收集和分析15万人出生和死亡的时间、地点数据。他们利用这些计算机领域看似很少的数据，惊奇地发现“历史事件和人口迁移变化之间的联系，而这些历史事件通常鲜为人知，人们通常忽略了他们的历史意义”[3]。大数据的这些成功应用促使大家注重用大数据的方法解决问题，并且大数据技术正在不断改善着人们的研究视野、研究方法和研究工具。

3.1 重新审视“谷歌预测流行病错误”

2008年，谷歌公司拟定了一系列高科技卫生计划，并投资1 500万美元预防下一代全球性流行病。用各地区网民搜索“流行病”关键词（比如温度计、流感症状、胸闷等）的数目，预测流行病爆发的时间、地点。谷歌公司第一次开始预测流感就取得了很好的效果，比美国疾病预防控制中心（Centers for Disease Control and Prevention，CDC）提前两个星期预测到了流感的爆发。但美国流感季节的到来似乎让谷歌公司的算法有些措手不及，它估计的流感峰值几乎是CDC的两倍，有些地方的数据甚至显示出更大的差距[4]。

《Nature》和《Science》上也抨击了谷歌公司的这项计划，“这些系统反应快，覆盖人群范围广，但是也有着很多弊端”。这在当时被认为是“big data”的“big error”，反对方认为出现这一失误的原因在于，仅仅按照点击频率预测流行病的工程方法是有问题的，因为关键词点击和现实理解有差距，点击并不代表患病。

其他研究者则把目光投向了Twitter。Twitter可能是除社交媒体外，可供公众选择的一个获取信息的最大渠道。约翰斯•霍普金斯大学的团队最近已经对几千条与流感相关的Twitter信息进行了分析，以便梳理出哪些信息是由真正患有流感的民众发布，哪些仅仅只是在发布流感新闻的链接。

从《Nature》和《Science》对谷歌公司预测流行病错误的批评可以看出，拥有了数据之后，应该从不同角度加以审视，确保对数据的来源、验证方法、控制手段以及是否有脏数据等问题都了

① 互联网下的社交网络，如Facebook或者Twitter

② 真实的社交网络

如指掌，才能开展下一步的数据分析。Twitter的结果进一步证明了数据源固然重要，但是选取和抽样合理的样本、采用符合实际的分析方法，才能得出合理的结论。否则即使拥有再大的数据集，脱离场景来谈大数据的应用，长久而言也会收获甚微。

3.2 研究现状

最近几年大数据的失败案例屡见不鲜，导致错误的原因也有很多。2015年，Gartner公司分析师Svetlana Sicular对于大数据项目的失败，全面总结出7个主要原因，并划分3个大类：战略、技术、分析。战略方面的失败提到选择了错误的应用场景；技术层面的失败包括无法应对料想外的问题和缺乏大数据分析技能，表明目前存在的大数据技术手段都不太成熟，没有标准的处理流程；数据分析本身的失败包括，对数据过于自信（如谷歌公司的流感预测），提出了错误的问题，应用了错误的模型。虽然目前大数据结合了众多学科研究的方法，比如数据科学的兴起，与神经网络、深度学习、语义计算以及人工智能等其他相关技术的结合，逐渐成为大数据分析领域的热点。但缺乏完整规范的理论体系和技术体系作为支撑，导致部分尝试以失败告终。

那么大数据能否作为一种科研方法以及是否可信？从以上案例分析总结3个方面：正确选择数据源、科学抽样有代表性和有价值的数据、严谨完备的大数据工程分析方法，在某种程度上是决定大数据成功与否的关键。但是目前还是存在很多由于错误数据源导致的失败案例，而且也没有形成科学的抽样方法，值得一提的是，近两年已经在全球范围内兴起大数据的标准化工作。

3.3 数据源的选择

哈佛商学院客座教授托马斯·达文波特（Davenport T H）写了一本名为《Keeping up with the quants》的新书，其中认为，管理大数据项目的一个重要部分是要问正确的问题：如何定义问题、需要哪些数据、来自哪里等。说明了大数据研究中非常重要的一点是数据源的选择。

近两年来，大数据被应用到天气预报、影视制片、交通、商场等与人们息息相关的行业。不同的行业必须选择领域相关、主题相关的数据，要先明确期望的目标，再确定正确的数据源，最后才可能利用大数据分析技术得到期望的结果。《纸牌屋》选择用户的电影评分数据，基于对观众偏好的分析，预测、设计观众喜欢的剧情，找观众喜爱的演员出演相关的角色，甚至可以预测票房。京东集团选择每天积累的上亿用户购买行为数据，分析出消费者的个人兴趣，以数据驱动算法，从而实现精准推荐。数据源的正确选择是充分利用大数据时代带来的巨大价值的首要条件。

3.4 大数据的抽样

有一种观点认为，大数据得到的是数据的全集，不再需要传统的抽样方法。然而这种观点也遭到了很多的质疑，甚至是批评。

社会调查是典型的数据抽样方法，通常被认为是用代表性样本测量和分析公众意见中有关社会和政治问题的科学方法。尽管如此，社会调查还是受到其样本量小、更新速度慢、无响应率高以及昂贵成本的限制[5,6]。近些年来，网络空间大数据成为获取和理解公众舆论的新资源，因为它的数据量大、更新速度快，并且容易获

得[7,8]。例如，这些网络空间的意见被用来预测票房[9]和政治选举[10]。然而，一些科学家持有相反的观点，例如有些专家认为，互联网舆情系统得到的结果往往是偏激的，数据虽然大，但代表的更多的是一类偏激人群的观点。他们怀疑，由于网络空间和现实世界之间的舆论差异，社会调查不能被大数据取代，除非这种差异可以量化和科学地消除[11]。

在大数据时代，如何利用大数据解放社会调查的限制，又能将科学合理的抽样方法应用于大数据分析，或者构建一套针对海量数据抽样的科学方法，是大数据分析的关键，也是一个值得重视的新的研究方向。

3.5 大数据工程

软件工程是当前软件危机诞生时产生的一项技术，用于保证大规模软件开发的质量。

大数据分析显然是一个巨大的工程，如何保证这个工程正确实施，也是一个值得关注的研究方向。如果能够像软件工程方法一样，利用完备的理论体系、技术体系和测试体系，也许大数据的一些失败可以避免，并且能够得到更广泛的应用。

总之，科学严谨的大数据研究需要大家沉下心做工程分析，重视采用科学的工程方法，评估从数据收集到应用的每一个技术环节。比如借鉴“软件工程”的概念，建立、完善和补充大数据从获取、整理、存储、索引到查询与检索、分析与挖掘、加工与展现的整个生命周期的基础理论、技术方法以及系统。这两年，在全球范围内兴起了对大数据的标准化工作，为国内加快大数据标准化研究提供了很好的机会。2014年6月，中国电子技术标准化研究院编制的大数据标准化白皮书详细描述了大数据标准化要完成的详细工作，包括基础标准、数据处理、数据安全、数据质量、产品和平台、应用和服务等各个方面，说明业界已经开始重视大数据作为一种科研方法的规范化。期待未来的大数据工程可以保证大数据分析的正确性。

4 结束语

本文从3个方面分析了大数据面临的机遇和挑战，即大数据的历史演变、大数据分析的可信性以及大数据工程。

大数据概念的诞生，与其历史演变轨迹密不可分：从VLDB到海量数据，再到当前的大数据，每一个阶段都吸引了学术界的广泛关注，并促进了产业界大规模数据管理和分析技术的迅猛发展。

大数据时代带来的巨大效益显而易见，同时大数据时代的不确定性和失败的前车之鉴也给很多领域带来了挑战。由于缺乏科学合理的理论体系和技术体系，众多专家质疑大数据是否可信。保证大数据分析的正确性，需要应对3个重要挑战，即选择正确的数据源，抽样有代表性和有价值的数据，再加上完备的大数据工程分析方法，期待未来的大数据工程可以保证大数据分析的正确性。

参考文献:

[1] BOND R M, FARISS C J, JONES J J, et al. A 61-million-person experiment in social influence and political mobilization[J]. Nature, 2012, 489(7415): 295-298.

[2] EINAV L, LEVIN J. Economics in the age of big data[J]. Science, 2014, 346(6210): 1243089.

[3] SCHICH M, SONG C, AHN Y Y, et al. A

network framework of cultural history[J]. Science, 2014, 345(6196): 558-562.
[4] KHOURY M J, IOANNIDIS J P A. Big data meets public health[J]. Science, 2014, 346(6213): 1054-1055.
[5] CURTIN R, PRESSER S, SINGER E. Changes in telephone survey nonresponse over the past quarter century[J]. Public Opinion Quarterly, 2005, 69(1): 87-98.
[6] WEISBERG H F. The total survey error approach: a guide to the new science of survey research[M]. Chicago: University of Chicago Press, 2005.
[7] CROSAS M, KING G, HONAKER J, et al. Automating open science for big data[J]. The Annals of the American Academy of Political and Social Science, 2015, 659(1): 260-273.
[8] KING G. Ensuring the data-rich future of the social sciences[J]. Science, 2011, 331(6018): 719-721.
[9] ASUR S, HUBERMAN B A. Predicting the future with social media[C]//2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Aug 31-Sept 3, 2010, Toronto, Canada. New Jersey: IEEE Press, 2010: 492-499.
[10] CONNOR B O, BALASUBRAMANYAN R, ROUTLEDGE B R, et al. From tweets to polls: linking text sentiment to public opinion time series[C]//4th Int'l AAAI Conference on Weblogs and Social Media, May 23-26, 2010, Washington, DC, USA. [S.l.: s.n.], 2010.
[11] GOLDER S A, MACY M W. Digital footprints: opportunities and challenges for online social research[J]. Annual Review of Sociology, 2014, 40(1): 129-152.

作者简介

王腾蛟(1973-),男,北京大学信息科学技术学院教授、博士生导师,北京大学文理大数据研究中心常务副主任,主要研究方向为数据库管理系统、互联网数据分析、数据仓库与数据挖掘等。

李喜莲(1992-),女,北京大学信息科学技术学院硕士生,主要研究方向为大数据、机器学习和数据挖掘等。

收稿日期: 2017-01-06
基金项目: 国家自然科学基金资助项目(No. 61572043);国家重点研发计划基金资助项目(No. 2016YFB1000704)
Foundation Items: The National Natural Science Foundation of China (No. 61572043), The National Key Research and Development Program (No. 2016YFB1000704)