卫生政策研究进展

Progress in Health Policy Research

健康医疗大数据专刊
2017 年　第 7 期
（总第 82 期）

上海市卫生和健康发展研究中心　　2017 年 11 月 10 日

编者按　近年来，随着健康医疗信息化的广泛推进，健康医疗大数据的应用范围正在快速扩展，构建基于健康医疗大数据的卫生决策机制并付诸应用实践，对全面提升健康医疗领域的治理能力和水平发挥着重要作用。本期专刊以健康医疗大数据为主题，收录了《健康医疗大数据概念及政策现状》《健康医疗大数据应用领域探索与思考》《健康医疗大数据在卫生决策中的应用研究》《基于健康医疗大数据的卫生决策机制构建研究》4 篇文章；同时，为了更为系统和深入地了解健康医疗大数据发展应用及大数据技术，本刊还转载了《健康医疗大数据发展应用的思考》《大数据技术的应用现状与展望》2 篇文章。谨供领导和同志们参阅。

卫生政策研究进展

2008 年 11 月创刊
第 10 卷 第 7 期（总第 82 期）
2017 年 11 月 10 日
（内部交流）

主管
上海市卫生和计划生育委员会

主办
上海市卫生和健康发展研究中心
（上海市医学科学技术情报研究所）

编辑出版
《卫生政策研究进展》编辑部
上海市北京西路 1477 号 807 室
邮编：200040
电话：021-22121872
传真：021-22121623
E-mail：phpr@shdrc.org
网　址：www.shdrc.org

上海市连续性内部资料准印证
（K）第 0649 号

目　　次

专题研究

健康医疗大数据概念及政策现状

王贤吉　杨山石　宋捷　信虹云　王力男　何江江　金春林

【摘　要】随着医疗行业技术和卫生信息化的蓬勃发展，医疗卫生领域也迎来了大数据时代，健康医疗大数据的应用将对普通民众的生活和政府政策的制定产生深刻的影响。本研究通过对健康医疗大数据相关概念的界定和对国内外健康医疗大数据相关政策的梳理总结，分析我国健康医疗大数据现存的挑战并提出相应的政策建议，以促进健康医疗大数据的充分利用。

【关键词】健康医疗大数据；政策；建议

随着云计算、物联网等新兴技术的推广和应用，各种不同类型的数据不断急速剧增。海量数据的出现、对数据挖掘分析的需要，都暗示着大数据时代已经到来。医疗行业早就面临着海量数据及非结构化数据的双重挑战，随着医疗行业技术和卫生信息化的蓬勃发展，医疗卫生领域也迎来了大数据时代。健康医疗大数据的应用可以帮助改善对于公众健康的监控、更好地对医疗服务进行定价、提高研发效率、支持临床决策，并对普通民众的生活和政府政策的制定产生深刻的影响。本文通过对健康医疗大数据政策的梳理总结，分析健康医疗大数据现存的挑战，以期提出相应的政策建议，促进健康医疗大数据的充分利用。

一、健康医疗大数据的相关概念界定

（一）大数据与健康医疗大数据的定义

大数据（big data），或称巨量资料，是需要新处理模式才能具有更强

第一作者：王贤吉，男，助理研究员
通讯作者：金春林，男，研究员，上海市卫生和健康发展研究中心主任，上海市医学科学技术情报研究所所长
作者单位：上海市卫生和健康发展研究中心，上海 200040；上海市医学科学技术情报研究所，上海 200031
基金项目：第四轮上海市公共卫生体系建设三年行动计划（2015—2017）项目（GWIV-33）

的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产，需要强大（如云计算等）的手段来处理[1]。从概念内涵来看，大数据具有新的特点，可概括为“5V”，即 Volume（数据规模大）、Velocity（处理速度快）、Variety（数据类型多）、Value（价值密度低）、Veracity（数据准确性低）[2]。从概念外延来看，大数据具有三个层面的含义：**一是**指一种新型的研究方法或发现新知识的分析技术；**二是**指由数据驱动的新型决策机制。大数据不仅仅是一种技术，更是一种理念创新和模式创新；**三是**指数字时代下的新型治理模式。大数据不仅仅是政府管理的一种新手段或新工具，它还将带来政府管理改革的一个全新阶段。

健康医疗大数据特指在医疗领域内产生的大数据，是随着医疗卫生信息化的发展，医疗数据类型和规模正急速地增长，在允许的时间里，无法用常规软件对医疗数据进行获取、管理和整合成为能够使用的有用信息[3]。目前国家层面还没有相关文件对健康医疗大数据给出明确界定，地方层面仅福州市近期发布的《福州市健康医疗大数据资源管理暂行办法》（榕政综〔2017〕122 号）明确了健康医疗大数据范围，即所有与医疗卫生和生命健康活动相关的数据集合，覆盖全员人口和全生命周期、涉及国家公共卫生安全和生物信息安全的大量数据[4]。近年来我国学者也多从应用视角研究健康医疗大数据涵义，国家卫生计生委统计信息中心主任孟群认为信息网络技术与传统的公共卫生、医疗服务、医疗保障、药品管理、计划生育、综合管理等 6 大业务的深度融合，以及互联网 + 健康医疗、精准医学、人工智能等新兴领域，所产生的大量与健康相关的大数据，都可以称为健康医疗大数据[5]。代涛将健康医疗大数据分为健康医疗服务数据、生物医学数据、医疗保险数据、医药研发与管理数据、公共卫生数据、行为与情绪数据、统计相关数据、人口管理数据以及与人类健康相关的环境数据[6]。许培海、黄匡时认为健康医疗大数据泛指

与健康和生命有关的所有数据，即个人从出生至死亡的一系列生命过程所产生的数据，并构建了包含临床数据、非临床数据（公共卫生数据为主）及个体生命周期数据在内的三维空间模型[7]。

健康医疗大数据的来源主要包括以下四个方面[8]：（1）制药企业/生命科学；（2）临床医疗/实验室数据；（3）费用报销/利用率；（4）健康管理/社交网络。目前，健康医疗大数据分析和应用主要集中在“临床医疗业务”、“医疗支付”、“药物研发”、“新型商业模式”、“公众健康监控”等五大领域[9]。

（二）健康医疗大数据的相关利益方及主体

健康医疗大数据的相关利益方包括政府、医院、企业、医务人员、患者和普通大众。健康医疗大数据可以为医务人员提供包括临床辅助决策、治疗方法与疗效比较、不良反应与差错分析提醒等服务；健康医疗大数据可以为患者和普通大众提供自我健康管理、健康预测与预警等服务[10]；药企可以通过健康医疗大数据的提取和分析，评价新药的安全性、有效性、潜在的副作用，提高研发效果；医院可以通过对医院感染数据的全面分析，提高医院感染管理防控效能，维护患者健康[11]；政府通过健康医疗大数据建设可以提高政府决策的科学性，提高政府制定政策和决策效率，也可以更有效应对突发事件，提升政府危机管理的能力和水平，大力提高政府的公信力。其中政府作为数据的最大需求者和提供者，大部分的健康医疗大数据都掌握在政府手中，政府是健康医疗大数据的主体。因此，政府和相关政策在健康医疗大数据建设中的作用尤为重要。

二、国内外健康医疗大数据相关政策

（一）国外

从目前健康医疗大数据的整体形势来看，国外尤其是美国、英国、日本等发达国家在推进健康医疗大数据的发展和应用方面起步早、发展

成熟，各国也有其不同的特色和亮点。如美国对精准医疗的推动、英国对个性化医疗的积极探索、日本对大数据的高效利用、美国和澳大利亚提及的建立起信息共享和交互的技术标准或公共政策，注重健康信息隐私和安全、新加坡促进健康医疗大数据在养老方面的发展、韩国的生物医学银行概念以及法国对相关人才发展的重视。除此之外，发达国家大多重视大数据基础设施的建设和智能技术的开发，鼓励政府等公共部门提升大数据分析能力，以灵活利用医疗数据，改进健康管理和疾病预防，详见表 1。

表 1　国外健康医疗大数据相关政策及内容

国家	年份	政策	内容
美国	2012	《大数据研究与发展倡议》	美国的人口、医疗等公共部门通过挖掘网络、搜索引擎等积累的海量数据，实现了对人口流动、传染病蔓延等情况的实时分析与检测。
	2015	《美国联邦政府医疗信息化战略规划（2015—2020）》	在保护健康信息隐私和安全的前提下，建立起信息共享和交互的技术标准，加强公众、医疗机构和公共卫生机构快速查找、获取电子健康信息的能力。
	2015	《“精准医学”计划》	拟投资 2.5 亿美元启动精准医疗计划（Precision Medicine Initiative）。根据该计划，美国将搜集 100 万人的个人健康信息以及测得他们的基因组序列，以实现短期目标：鉴定新的癌症亚型；与药厂等私人部门合作测试精准疗法的临床效果；拓展对癌症疗法的认识（抗药性、鸡尾酒疗法、肿瘤复发等）。长期目标：将精准医疗逐步覆盖到其他健康和疾病相关的所有领域。
日本	2012	《活跃 ICT 日本战略》	ICT 即信息与通讯技术，主要关注大数据，并将其作为 2013 年六个主要任务之一，聚焦大数据应用所需的社会化媒体等智能技术开发，以及在新医疗技术开发、缓解交通拥堵等公共领域的应用。
	2014	《创建最尖端 IT 国家宣言更新》	鼓励各方在医疗健康大数据平台上，灵活利用医疗数据，改进健康管理和疾病预防，建立健康长寿型社会。

（续表 1）

国家	年份	政策	内容
新加坡	2012	《健康医疗 2020 总体规划》	发展数字医疗技术，保持卫生系统高效并应对人口老龄化。
澳大利亚	2013	《公共服务大数据战略》	推动公共部门利用大数据分析进行服务改革，制定更好的公共政策，保护公民隐私。
英国	2013	《癌症患者数据库》	英格兰医疗保健当局宣布，英格兰将建立世界最大的癌症患者数据库，为个人化的癌症治疗提供基础。这个数据库将保存和整理英国每年 35 万新确诊的肿瘤病例的全部数据。建立这个数据库的目的是在此基础上推动“个人化医疗”，针对每位患者的癌症类别和具体情况对症下药。
法国	2013	《法国政府大数据五项支持计划》	促进本国大数据发展，推动经济社会发展，重点发展领域：人才培养、交通、医疗卫生。
韩国	2015	《健康医疗行业提升计划》	强化已有健康行业的基础设施，打造一个连接不同医学基础设施的平台，包括：生物医学银行和健康医疗大数据。

（二）国内

近年来，国内政府对健康医疗大数据的应用价值愈发重视，2015 年起国家密集出台了一系列政策文件，主要提及了以下几个方面的内容：（1）建立医疗网络信息平台，加强区域医疗卫生服务资源整合，推动电子健康档案和电子病历数据整合共享；（2）推动基因技术和精准医疗技术发展；（3）发展预防、治疗、康复、养老等的一体化健康服务新模式；（4）产学研联合的大数据发展机制；（5）推进智慧健康医疗；（6）支持有关企事业单位、第三方机构开展医疗健康大数据创新应用研究；（7）加强健康医疗数据安全保障和患者隐私保护，详见表 2。

表 2　我国健康医疗大数据相关政策及内容

年份	政策	内容
2015	《国务院关于积极推进“互联网 +”行动的指导意见》（国发〔2015〕40 号）	发展基于互联网的医疗卫生服务，支持第三方机构构建医学影像、健康档案、检验报告、电子病历等医疗信息共享服务平台，逐步建立跨医院的医疗数据共享交换标准体系。积极在线预约诊疗、候诊提醒、划价缴费、诊疗报告查询、药品配送等便捷服务。引导远程医疗服务。鼓励互联网企业与医疗机构合作建立医疗网络信息平台，加强区域医疗卫生服务资源整合，充分利用互联网、大数据等手段，提高重大疾病和突发公共卫生事件防控能力。积极探索互联网延伸医嘱、电子处方等网络医疗健康服务应用。发展基因检测、疾病预防等健康服务模式。
2015	《国务院关于印发促进大数据发展行动纲要的通知》（国发〔2015〕50 号）	构建电子健康档案、电子病历数据库，建设覆盖公共卫生、医疗服务、医疗保障、药品供应、计划生育和综合管理业务的医疗健康管理和服务大数据应用体系。探索预约挂号、分级诊疗、远程医疗、检查检验结果共享、防治结合、医养结合、健康咨询等服务，优化形成规范、共享、互信的诊疗流程。鼓励和规范有关企事业单位开展医疗健康大数据创新应用研究，构建综合健康服务应用。
2016	《国家发改委办公厅关于组织实施促进大数据发展重大工程的通知》（发改办高技〔2016〕42 号）	在健康医疗、社保就业、教育文化、交通旅游等领域，推动传统公共服务数据与互联网、移动互联网、移动穿戴设备数据的汇聚整合，鼓励社会机构开展应用研究，开发便民服务应用，优化公共资源配置，提升公共服务水平。
2016	《中华人民共和国国民经济和社会发展第十三个五年规划纲要》	提升健康信息服务和大数据应用能力，发展远程医疗和智慧医疗。
2016	《国家创新驱动发展战略纲要》	促进组学和健康医疗大数据研究，发展精准医学。开发数字化医疗、远程医疗技术，推进预防、医疗、康复、保健、养老等社会服务网络化、定制化，发展一体化健康服务新模式。
2016	《国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见》（国办发〔2016〕47 号）	部署通过“互联网 + 健康医疗”探索服务新模式、培育发展新业态，努力建设人民满意的医疗卫生事业，为打造健康中国提供有力支撑。

（续表 2）

年份	政策	内容
2016	《国家信息化发展战略纲要》	推进智慧健康医疗，完善人口健康信息服务体系，推进全国电子健康档案和电子病历数据整合共享，实施健康医疗信息惠民行动，促进和规范健康医疗大数据应用发展。探索建立市场化远程医疗服务模式、运营机制和管理机制。加强区域公共卫生服务资源整合，探索医疗联合体等新型服务模式。
2016	《“十三五”国家信息化规划》	推进健康医疗临床和科研大数据应用，推进基因芯片和测序技术在遗传性疾病诊断、癌症早期诊断和疾病预防检测中的应用，推动精准医疗技术发展。推进公共卫生大数据应用，全面提升公共卫生监测评估和决策管理能力。
2016	《国务院关于印发“十三五”卫生与健康规划的通知》（国发〔2016〕77 号）	全面深化健康医疗大数据应用。推进健康医疗行业治理、临床和科研、公共卫生大数据应用，加强健康医疗数据安全保障和患者隐私保护，积极应用物联网技术、可穿戴设备等，探索健康服务新模式，发展智慧健康医疗，强化预防、治疗、康复的精细服务和居民连续的健康信息管理业务协同。
2017	《智慧健康养老产业发展行动计划（2017—2020 年）》	利用物联网、云计算、大数据、智能硬件等新一代信息技术产品，能够实现个人、家庭、社区、机构与健康养老资源的有效对接和优化配置，提升健康养老服务质量效率水平。

三、健康医疗大数据相关挑战及政策建议

健康医疗大数据时代已经到来。目前，国务院、国家卫生计生委等部门已经出台了一系列相关政策文件，以充分激发健康医疗大数据的创新活力，探索数据新应用，培育发展新业态。然而，医疗领域本身具有的独特性和复杂性，使得大数据技术在医疗领域的应用推广仍存在诸多挑战，诸如医疗数据的部门间共享仍存在制度壁垒，隐私数据的泄露与不当应用的风险较大，相关人才的缺乏等等。为了保障健康医疗大数据在采集、存储、分析、应用、管理等各环节的规范性，促进其在我国的良性发展，急需制定相关政策进行引导和监督。

（一）完善顶层设计，建设共享开放的健康医疗大数据应用基础体系

政府重视数据的跨部门、跨区域共享与开放。但是，不同系统或不

同平台之间标准并不统一，可能缺乏统一的数据和传输标准，不同层级部门之间数据平台建设水平也参差不齐，这都给数据的共享和开放带来了问题。另外，建设全国健康医疗数据资源集成和共享平台涉及多方监管部门和参与主体，实施起来存在较大难度。建议进一步完善顶层设计，使得建设共享开放的健康医疗大数据应用基础体系有迹可循。整合利用电子政务外网、网络运营商网络、政务云资源等现有设施资源，推动政府健康医疗信息系统与居民电子健康档案、电子病历三大数据库互联互通，促进各级各类医疗卫生机构应用信息系统数据采集、集成共享和业务协同，推广结构性数据、统一数据采集标准、强化数据共享共用，加快建设形成覆盖省、市、县的健康医疗业务专网，建成统一权威、互通共享的全民健康信息综合管理平台。

（二）隐私信息保护尚缺乏，需加强立法和完善技术规范

大数据在给社会带来巨大效益的同时也暴露出许多关于数据安全和隐私的问题。以患者隐私为例，隐私泄露途径更加多样化：医学研究的参与、信息发布、数据挖掘、数据库的安全性、医院管理不当和医护人员的影响（外包）、数据共享等，这对信息及数据安全保证提出了更高要求。对于隐私信息保护应该分辨出哪些医疗数据属于隐私数据，哪些医疗数据可以共享和利用，明确隐私保护对象。目前国外对于隐私信息的保护主要通过立法和制定安全管理规范实现，国外对医疗健康信息隐私的立法保护主要采取两种模式：一种模式是在基础隐私保护法律框架下将医疗健康信息从个人隐私信息中划归出来单独立法并制定执行标准施以保护，另一种模式是将医疗健康隐私信息纳入个人信息、敏感信息施以综合保护。我国可借鉴国外模式，加强数据安全方面的立法，同时制定《医疗大数据安全管理办法》，加强数据质量管理，确保健康医疗数据的合法、可用，探索制定健康医疗隐私信息开放办法，完善数据访

问控制系统和应用管理规范，尽可能保护数据隐私[12]。

（三）制定健康医疗信息化人才发展计划，加强医学信息学学科建设

由于医疗行业本身的专业性，健康医疗大数据同时也具备复杂性的特点，这就对医疗信息工作人员提出了更高的要求：既要熟悉医疗健康和经营管理知识，又要精通数据分析和信息利用相关知识。相关人才的缺失限制了健康医疗大数据的挖掘与利用。建议加强健康医疗信息化复合型人才队伍建设。建立和完善医学信息学人才培养体系，鼓励支持高等院校、健康医疗机构、行业协会、健康医疗信息企业合作建立教育实践、实训基地，开展继续教育和培训，畅通医学信息人才职业发展通道，推进健康医疗大数据应用发展的人才技术交流与合作。

参考文献

[1] 杨旭.数据科学导论[M].北京理工大学出版社, 2014.

[2] 于施洋,王建冬,童楠楠.国内外政务大数据应用发展述评:方向与问题[J].电子政务,2016,01:2-10.

[3] 李永欢.广东省医疗大数据建设中的政府作用研究[D].广西师范大学, 2015.

[4] 福州市人民政府.福州市人民政府关于印发《福州市健康医疗大数据资源管理暂行办法》的通知（榕政综〔2017〕122号）[Z].2017-04-21.

[5] 孟群, 胡建平, 董方杰,等.我国健康医疗大数据资源目录体系建设研究[J]. 中国卫生信息管理杂志,2017,14(3):387-391.

[6] 代涛.健康医疗大数据发展应用的思考[J].医学信息学杂志,2016, 37(2):1-8.

[7] 许培海, 黄匡时.我国健康医疗大数据的现状、问题及对策[J].中国数字医学,2017,12(5):24-26.

[8] Zhang Z, Zhou Y, Shou-Hong D U, et al. Medical Big Data and the Facing Opportunities and Challenges[J]. Journal of Medical Informatics, 2014.

[9] 蔡佳慧,张涛,宗文红.医疗大数据面临的挑战及思考[J].中国卫生信息管理杂志,2013(4):292-295.

[10] 李国杰.大数据研究的科学价值[J].中国计算机学会通讯,2012,8(9):8-15.

[11] 汪鹏,吴昊,罗阳,等.医疗大数据应用需求分析与平台建设构想[J].中国医院管理,2015,35(6):40-42.

[12] 孟群,毕丹,张一鸣,等.健康医疗大数据的发展现状与应用模式研究[J].中国卫生信息管理杂志,2016,13(6):547-552.

（责任编辑：唐密）

健康医疗大数据应用领域探索与思考

李艳[1]　杨山石[2]　王贤吉[2]　王力男[2]　何江江[2]　金春林[2]

【摘　要】健康医疗大数据已成为国家重要的基础性战略资源。推动健康医疗大数据的应用，建设健康大数据中心、区域中心、应用发展中心已成为目前我国健康医疗大数据产业发展的重心。本文尝试梳理健康医疗大数据可能的应用领域，以期为健康医疗大数据的实际应用提供参考。

【关键词】健康医疗大数据；应用领域；思考

近年来，大数据引起了产业界、科技界和政府部门的高度关注。2008 年《Nature》出版专刊“Big Data”，2011 年《Science》也推出关于数据处理的专刊“Dealing with data”，2012 年 3 月奥巴马宣布美国政府投资 2 亿美元启动“大数据研究和发展计划”。美国政府认为大数据是“未来的新石油”，必将给未来的科技与经济发展带来深远的影响，并将“大数据研究”上升为国家意志[1]。医疗行业与人们健康息息相关的特殊性决定了健康医疗大数据的重要性，加快推动大数据与医疗健康行业的结合将为健康医疗服务带来深刻变革。2016 年《国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见》（国办发〔2016〕47 号），该政策的颁布奠定了健康医疗大数据作为国家基础性战略资源的重要地位[2]。本文拟阐述当前健康医疗大数据的应用领域及面临的挑战和风险，以期为我国健康医疗大数据发展提供参考。

第一作者：李艳，女，硕士在读
通讯作者：金春林，男，研究员，上海市卫生和健康发展研究中心主任，上海市医学科学技术情报研究所所长
作者单位：1. 华中科技大学同济医学院医药卫生管理学院，武汉 430030
2. 上海市卫生和健康发展研究中心，上海 200040；上海市医学科学技术情报研究所，上海 200031
基金项目：第四轮上海市公共卫生体系建设三年行动计划（2015—2017）项目（GWIV-33）

一、健康医疗大数据的应用

我国卫生统计信息发展为大数据技术在医疗卫生领域中的发展和应用提供了广阔空间。在技术层面:传统基于数据仓库的商业智能(Business Intelligence，简称 BI）和数据挖掘应用可以迁移到大数据环境，结合非结构化数据的分析、挖掘，以及传感器技术产生的大量实时监测数据的分析；在业务层面：涵盖面向医生的临床辅助决策和科研，面向管理者的管理决策辅助、行业监管、绩效考核，面向居民的健康监测，面向药品研发的统计学分析就诊行为分析等方面都大有可为[3]。

2014 年，国家卫生计生委发布了国家卫生、计生资源整合顶层设计规划——“4631-2 工程”*。基于“4631-2”框架中的 6 大业务应用，以及健康医疗大数据在实际中的应用，本文将健康医疗大数据的应用领域归纳为公共卫生、医疗服务、医疗保障、药品管理和综合管理等方面。

（一）公共卫生

1. 疾病预防

在疾病预防方面，大数据可以使研究者比之前任何时候都更加了解健康及其影响因素。据估计，只有 10% ～ 15% 的健康影响因素已被医疗服务提供者所测定，剩下的 85% ～ 90% 的影响因素，包括健康行为、遗传因素、自然和社会经济环境等均未被测定[1]。

随着互联网、物联网、医疗卫生信息系统及相关信息系统等普遍使用，可以系统全面地收集健康危险因素数据，包括环境因素（利用 GIS 系统采集大气、土壤、水文等数据），生物因素（包括致病性微生物、细菌、病毒、真菌等的监测数据），经济社会因素（分析经济收入、营养条件、

* 注：其中，“4”代表 4 级卫生信息平台，分别是：国家级人口健康管理平台，省级人口健康信息平台、地市级人口健康区域信息平台及区县级人口健康区域信息平台；“6”代表 6 项业务应用，分别是：公共卫生、医疗服务、医疗保障、药品管理、计划生育、综合管理;“3”代表 3 个基础数据库，分别是:电子健康档案数据库、电子病历数据库和全员人口个案数据库；“1”代表 1 个融合网络，即人口健康统一网络；最后一个“2”是人口健康信息标准体系和信息安全防护体系。

人口迁移、城镇化、教育就业等因素数据），个人行为和心理因素，医疗卫生服务因素，以及人类生物遗传因素等。利用大数据技术对健康危险因素进行比对关联分析，针对不同区域、人群进行评估和遴选健康相关危险因素及制作健康监测评估图谱和知识库，促进居民健康水平的提高[3]。Asthmapolis公司研发了一种追踪器，能够记录哮喘患者吸入器的使用，将信息传输到中央数据库，可以了解个人、团体和人群的流行趋势。这些数据还会与疾病预防控制中心的哮喘危险因素数据结合，用以帮助患者制定个性化的预防、治疗计划[4]。

2. 疫情检测

大数据技术也可以用于疫情监测。疾病防控是否及时，发生疫情时能否第一时间做出反应并制定防治方案，是疫情控制的关键。利用大数据实时监控疾病的症状和源头并与病原数据库里的病毒进行特征比对分析，寻找与疾病症状吻合的病原，可确定疾病类型，提前预测疫情暴发，为疫情控制争取时间；同时，大数据实时监控也可及时发现匹配不到的未知或新型病原，针对新病原尽早开始疫苗研发[5]。谷歌公司对流感的准确预测就是运用大数据技术的成功案例。谷歌公司把美国人在网络上频繁使用的检索关键字，与美国疾控中心流感传播时期的数据进行比较，通过大数据分析辨别出人们是否感染了流感[6]。

（二）医疗服务

1. 疾病诊疗

通过大数据挖掘分析建立临床决策支持系统和用药、医嘱的自动报错系统，可以有效地减少医疗差错。利用大数据全面分析患者特征数据和诊疗数据，比较多种干预措施的有效性，可以找到针对特定患者的最佳治疗途径[7]；对大量电子病历中的数字化信息进行分析处理，发现最有效的临床路径以及单病种等[7]。安大略理工大学的卡罗琳·麦格雷戈

博士和其研究队伍与 IBM 合作，采用软件来监测处理即时的患者信息，实施对早产儿的病情诊断，在明显感染症状出现的 24 小时之前，系统就能监测到早产儿身体发出的感染信号 [8]。

2. 循证医学

大数据对循证医学也有着巨大的作用，利用大数据技术分析个人数据集，可以为循证医学提供最坚实的证据，发现小样本无法发现的细微差别，为医生提供最新的证据，指导临床实践 [1]。纪念斯隆—凯特琳癌症中心和沃森超级计算机合作，利用 60 万份医疗证据，150 万条患者记录以及肿瘤研究领域中 42 种医疗杂志和临床试验的 200 万页文本数据，沃森可以现场为医疗工作者提供治疗建议，与沃森合作的护士，约 90% 采纳了其建议 [9]。

3. 个性化医疗 [10]

个性化医疗服务的最大特点是在个人实施健康管理基础上，通过对个人健康危险因素进行全面评估，制定具有差异化的健康促进计划。MapReduce 和 Hadoop 分布式系统应用于临床大数据处理和分析，给疾病诊断和个性化治疗开辟了新的途径，被认为是当前医学界的重大进展 [11]。如在肿瘤个性化治疗方面，美国临床肿瘤学会的“肿瘤学快速研究系统的多阶段计划（Cancer Lin Q）”使用可获得的开源和专有软件，对 1 万例乳腺癌患者的电子病历进行 22 项专项评估。评估完成后，通过浏览和检索病历、产生假设、评价质量，为临床医生实时提供循证信息和治疗进展，并确定临床试验的参与资格，针对不同患者进行差异化治疗 [12]。

（三）医疗保障

1. 降低医疗成本

利用医疗保险大数据，建立定价环节的自动化系统模型，可以改

进费用补偿方式和降低医疗成本；通过对大量数据分析可以确定病人健康保险优惠计划的补偿额度，能更加有效地利用医疗资源，改进医疗成本管理；利用软件识别出高度使用医疗保险患者，分析某个社区或卫生系统的医疗成本趋势，可以帮助医疗服务提供者针对某类患者或某类疾病状态制定成本控制策略。美国采用医药福利管理模式（即第三方医保支付审核和控费）来降低医疗成本[13]。北京东软望海科技有限公司将DRGs与医疗成本数据有效结合，实现了科室成本核算和项目成本核算、病种成本核算，促进了医院精细化运营管理[24]。

2. 防范医保欺诈

大数据分析可以帮助医保机构找出一些典型的理赔费用风险问题，例如分解住院、不合理医疗检查项目或者不合理高值医用耗材、诊断和处方药品指征不匹配、药品剂量超标等，并通过大数据分析和机器审核，可以快速筛选出存在欺诈风险的索赔请求，有效降低欺诈成功率。成都市就利用大数据建立了智能辅助审核系统对住院治疗费用单据进行审核，能筛出疑似过度医疗行为，初步判断治疗、用药是否符合“临床规则”[8]。Heetal等分析澳大利亚医保数据，有效发现了参保人员的医疗欺诈行为[15]。

（四）药品管理

1. 医药研发和评价

大数据可以应用于药品研发的每一个阶段。药品研发前，利用大数据对患者乃至大众的行为和情绪进行测量，挖掘患者症状特点、行为习惯、喜好等，找到符合患者症状的药品和服务，针对性地调整和优化药品[16]；研发成功后，通过大数据分析公众疾病药品需求趋势，利用大数据确定最优的投入产出比，从而实现最优资源组合和节约成本。药品上市前，通过大数据扩大样本数和采样分布范围，分析药物副作用以及药

品不良反应可以克服传统临床试验和副作用报告分析中样本数小、采样分布受限等因素的影响，使结果更具有说服力，有利于缩短药品上市时间，降低企业成本；药品上市后，通过整合上市后各研究阶段可获得的所有数据，全面把握上市药品的安全性、有效性和经济性，为临床合理用药提供更有价值的参考。医药公司还可以通过大数据技术优化物流信息平台，提高管理效率[17]。美国食品与药品管理局计划使用大数据方法，通过综合和追踪包括科研人员、消费者、出版物、制造、广告、药房和促销支出等多种来源的数据，调查药品核准标示外的市场销售[18]。

2. 药物使用监测

根据现有的电子病历以及医院信息系统数据，利用大数据技术对海量的临床诊疗数据进行分析，可对药物使用效果以及某种药物剂量或多种药物不同剂量比例与实际治疗效果之间的关系进行分析。加利福尼亚州的 Kaiser Permanente 机构将临床数据与成本数据联系起来作为数据集，分析发现了药物不良反应，随后抗关节炎药（Vioxx）从市场中撤出。南通大学附属医院利用 Hadoop 技术挖掘所定义的关键信息数据，并对此数据进行评估，从而得到关于药品使用剂量关系规则，为病人创建合理的服药剂量数据[20]。

（五）综合管理

1. 医院管理

在医疗行业和机构的管理领域，可以通过各种统计和分析，让管理者从多个角度全局性地掌握医疗机构运营的总体情况，找到医院医疗质量管理不足的环节和医疗资源分配不合理的地方；对医疗质量和效益指标进行精确计算，监控医疗行为过程中的各环节，提高过程质量管理、监控，实现终末的质量评价；也可进行医生绩效分析、成本核算和控制、供应链分析、市场数据挖掘等，为管理者进行科学化、合理化的决策提

供强有力的数据支持，从而提升医疗机构的运营管理水平。马萨诸塞州蓝十字和蓝盾医疗保险（Blue Cross Blue Shield，简称 BCBSMA）将大数据分析嵌入到业务流程，帮助业务决策者深入了解医院运营情况[21]。

2. 医院评价

综合运用健康医疗大数据资源和信息技术手段，可健全医院评价体系，完善现代医院管理制度。根据疾病疑难危重的分级指数进行分级，同级别病种比对包括治疗方式、平均住院日、住院费用、出院后是否再次入院、入院死亡率等，并将这几项内容综合比对医师医疗服务能力水平；而不同级别病种数量比对体现的是专科医疗服务能力水平。很多医院目前已经开展了基于 DRGs 的医疗服务绩效分析和基于病种对比的医疗服务能力分析，评价体系充分运用了大数据理念[22]。

二、风险与挑战

国内外趋向成熟的大数据技术研究推动了卫生统计信息步入更快的发展阶段，实现“数据＋环境（产生数据的环境及其条件因素）→信息＋规律（信息变化的规律性、学习效应及其总结）→知识＋思想（利用知识的逻辑框架及其知识库建立）→智慧”这样一个螺旋式学习提升和价值发现过程[3]。但是，在实践过程中，仍不可避免地面临一些风险和挑战。

（一）健康大数据使用中的安全、保密、共享、开放等医学伦理学问题

健康大数据不可避免地涉及人群的隐私信息，包括身体现况、健康史、个人信息，甚至基因、蛋白数据等，如若泄露，极可能会使患者的日常生活遭到难以预料的侵扰[23]。如由于缺乏对病人隐私保护的足够关注，英国医疗健康大数据旗舰平台 care.data 被迫停摆[24]。若将数据加入到大数据库之前，通过电脑程序将能够被识别的患者个人信息从医疗

记录中去除，理论上讲可以克服这个问题。但由于缺乏个体的识别信息，其他数据将无法和研究样本整合，难以证实因果行为和健康状况的关系，不能进行某类人群大范围的研究。通过特殊处理（如去识别化、数字身份加密等）可以较好地解决此问题，但仍绕不开信息识别，而且去识别化本身也需要处理可识别的信息，可能造成患者健康信息在不知情、未授权的情况下被他人盗用。

此外，健康大数据的收集、存储、维护及使用方面，不仅涉及个人隐私问题，还牵涉公众利益甚至国家安全。《国务院关于印发促进大数据发展行动纲要的通知》(国发〔2015〕50号)[25]中反复提及共享和开放的战略，强调由政府主导共享和开放数据，降低公众获取和利用政府数据资源的难度及成本，为公共卫生健康大数据研究铺平道路。与此同时，大数据意味着大责任、大伦理，任何单位或个人使用健康大数据时均应该严格申请审查并备案，在法律允许的框架内使用相关数据，承担风险责任[26]。健康大数据的使用过程既要破除壁垒，让信息互联互通；又要充分隐私保密，杜绝隐患。

（二）突破大数据的关键技术

突破大数据的关键技术，推动其在公共卫生中的应用半结构化和非结构化数据量呈几何级数增长，传统的分析技术面临着较大的冲击和挑战。数据的广泛存在性使得数据越来越多地以不同的形式散布于不同的系统和平台之中[27]。健康医疗大数据除了大数据所具有的数据规模大、数据类型多、处理速度快、价值密度低、数据准确性低（简称“5V”）等特征外，还具有多态性、不完整性、时间性及冗余性等特征，为了便于进行分析，需要解决数据的多源异构性、数据的质量问题、数据的整合等。特别需要指出的是，在大数据时代虽然允许不精确的出现，但最基本、最重要的任务还是应该尽可能减少错误，保障质量。除上述技术

挑战外，还有数据信息孤岛问题普遍存在，标准化难以实施等技术和非技术困难尚未得到有效彻底地解决。

（三）甄别健康大数据使用中的“误差”，提高精度

大数据也会产生“大错误（Big Error）”[28]，流感在2013年最先袭击美国且造成十分严重的危害。当时科学家们先利用大数据技术，之后又采用传统的公共卫生监测方法分析流感的影响程度并进行估计，结果显示前者对流感的高峰期影响水平明显高估。“大数据”可以作为有效的工具来评估疾病负担和传播，Google流感趋势（Google Flu Trends，GFT）结合疾控中心的数据网络可以大幅提高预测性能，在流感传播和流行期间这一改进的模型可以更准确预测未来一周的感染情况[29]。同样，卫生服务人员需要认识到垃圾数据的存在以及有责任维护数据的完整性和准确性[10]。健康大数据使用者应认识到大数据本身不可能替代其他数据；虽然其弥补了很多以前数据的缺陷，但只是弥补性而不是取代性的功能，在疾病与健康预测方面甄别健康大数据的“误差”尤为重要。

三、小结

信息技术与经济社会的交汇引发了数据迅猛增长，大数据已经成为推动经济转型发展的新动力、重塑国家竞争优势的新机遇和提升政府治理能力的新途径。健康医疗大数据是国家重要的基础性战略资源，在当前健康医疗大数据产业大力发展的关键阶段，必须正确引导和规范健康医疗大数据资源的共享与应用，努力克服健康医疗大数据所面临的风险，将挑战转化为机遇，逐步实现健康医疗大数据的融合共享、开放应用。

参考文献

[1] 王潇,张爱迪,严谨.大数据在医疗卫生中的应用前景[J].中国全科医学, 2015,18(1):113-115.

[2] 国务院办公厅.国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见（国办发〔2016〕47号）[Z]. 2017-06-24.

[3] 周光华,辛英,张雅洁,等.医疗卫生领域大数据应用探讨[J].中国卫生信息管理杂志, 2013,10(4):296-300.

[4] Groves P,KayyaliB,Knott D, et al. The“big data” revolution in healthcare. McKinsey Quarterly [2013-1-15]. http://www.payerfusion.com/wp_content/uploads/2014/02/The big data revolution in healthcare _1.pdf.

[5] 钟敏. 浅析大数据在疾病防控中的应用[J]. 海峡预防医学杂志, 2017,23(4):88-90.

[6] 张怡,李柯. 大数据在医疗领域各角色中的应用[J].电脑知识与技术:学术交流, 2015,22(21):10-11.

[7] 王帅,沈明辉,冯昌琪,等. 基于医院大数据的基层医疗机构诊疗决策支持模式[J]. 中华医学图书情报杂志, 2015,24(4):66-99.

[8] 周雪晴, 罗亚玲. 信息化建设中医疗大数据现状[J].中华医学图书情报杂志, 2015,24(11):48-51.

[9] SchonbergerMV,Cukier K.大数据时代［M］.盛杨燕,周涛,译.浙江:浙江人民出版社, 2013:9.

[10] Ward JC. Oncology Reimbursement in the Era of Personalized Medicine and Big Data[J]. Journal of Oncology Practice, 2014,10(2):83-86.

[11] Mohammed EA, Far BH, Naugler C. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends[J].Bio Data Mining, 2014,7(1):22.

[12] Sledge Jr GW, Miller RS, Hauser R. Cancer Lin Q and the future of cancer care [C].American Society of Clinical Oncology educational book. American Society of Clinical Oncology, 2012:430-434.

[13] 高红玉, 周利生. 信息经济时代药品福利管理商业模式创新[J]. 中国药物与临床, 2013, 13(11):1505-1507.

[14] 毕丹, 董可男, 薛鲁宁,等. 健康医疗大数据产业分析与前景展望[J]. 大数据时代, 2017,2(4):6-20.

[15] He H, Wang J, Graco W, et al. Application of neural networks to detection of medical fraud[J]. Expert Systems with Applications, 1997, 13(4):329-336.

[16] 迟梦雅. 大数据时代医药企业面临的机遇和挑战[J].商业经济, 2014,33(11):43-44.

[17] 代涛.健康医疗大数据发展应用的思考[J].医学信息学杂志, 2016,37(2):1-8.

[18] SzlezakN,EversM,WangJ,et al. The role of big data and advanced analytics in drug discovery, development, and commercialization [J].ClinPharmacol Ther, 2014,95(5):492-495.

[19] IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/ wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/.

[20] 顾洁莹, 陈昊. 基于数据挖掘技术的药物剂量控制系统[J].中国数字医学, 2016,11(4):61-63.

[21] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential[J]. Health information science and systems, 2014, 2(1): 3.

[22] 戴明锋, 孟群. 医疗健康大数据挖掘和分析面临的机遇与挑战[J] .中国卫生信息管理杂志, 2017,14(2):126-130.

[23] Bellazzi R. Big data and biomedical informatics:a challenging opportunity[J].Yearb Med Inform, 2014,9(1):8-13.

[24] 洪延青, 何延哲. 英国健康医疗大数据平台care.data为何停摆?[J].中国经济周刊, 2016(29):77-79.

[25] 国务院关于印发促进大数据发展行动纲要的通知(国发〔2015〕50号)[EB/OL]. [2017/10/26]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

[26] Dove ES,Özdemir V. What role for law, human rights, and bioethics in an age of big data, consortia science, and consortia ethics? The Importance of Trustworthiness[J]. Laws, 2015,4(3):515-540.

[27] Khan N, YaqoobI,HashemIAT,et al. Big data: survey, technologies, opportunities, and challenges[J]. Scientific World J, 2014,2014:712826.

[28] Khoury M J, Ioannidis J P A. Big data meets public health[J]. New Zealand Medical Journal, 2014, 93(676):1054-1055.

[29] Davidson MW, Haim DA, Radin JM. Using networks to combine “Big Data” and traditional surveillance to improve influenza predictions[J]. Sci Rep, 2015(5):8154.

（责任编辑：丛鹂萱）

健康医疗大数据在卫生决策中的应用研究

杨燕　杨山石　王贤吉　何江江　王力男　金春林

【摘　要】当前医改已进入攻坚期和深水区，健康医疗大数据与深化医药卫生体制改革、促进卫生计生事业科学发展密切相关，要确保科学决策、正确决策，必须要完善决策智力支持系统，构建大数据集成平台，为决策咨询论证、调查评估等方面提供有力支撑。本文通过梳理国内外健康医疗大数据在卫生决策领域中的实践经验，提出大数据在医疗卫生决策中的应用可能性和挑战。

【关键词】健康医疗大数据；医改；数据集成平台；决策机制

2012 年美国发布《大数据研究和发展倡议》，揭开政府主导大数据应用的序幕 [1]。近年来，我国卫生信息化建设步伐加快，信息化为管理和决策服务的效果逐步显现。利用先进的信息技术，可以实时生成汇总数据，实现对卫生工作的实时监督、动态管理、科学决策，更好地向居民提供各类卫生服务，提高居民健康水平。

一、健康医疗大数据与卫生决策

当前，医疗卫生体系发展正处于大数据背景下，数据作为原材料，只有在特定研究背景下，经过加工处理，才能产生有用的信息。如何将医疗卫生领域中的海量信息收集、分析汇总，最终形成可以指导决策的证据，值得关注。通过大数据分析发现医疗质量不足和医疗资源分配不合理的地方，从而为卫生政策决策提供科学依据，更有效地解决卫生领域发展中的各种问题，达到优化医疗卫生行业资源配置，解决医疗资源

第一作者：杨燕，女，研究实习员
通讯作者：金春林，男，研究员，上海市卫生和健康发展研究中心主任，上海市医学科学技术情报研究所所长
作者单位：上海市卫生和健康发展研究中心，上海 200040；上海市医学科学技术情报研究所，上海 200031
基金项目：第四轮上海市公共卫生体系建设三年行动计划（2015—2017）项目（GWIV-33）

有限性的目的。

二、国外健康卫生大数据决策应用实践

（一）加拿大大数据卫生决策应用发展经验

加拿大卫生决策支持系统基于建立覆盖全体居民的“电子健康档案解决方案”（Electronic Health Record Solutions, EHRS），并实现数据的共享与可交互性[3]。早在2001年，加拿大就成立了名为Infoway的机构，该机构旨在促进和加速整个国家范围内电子健康档案的研发。Infoway集合大量的人力、物力、财力，耗时8年，制定出加拿大电子健康记录蓝图（EHRS blueprint），该蓝图被誉为迄今为止最完善的国家级EHRS和区域卫生信息网络建设规划与实施文档，用以指导和规划全加拿大电子健康系统的建立和实施。Inforway还计划建立全国性的电子健康档案系统、药品信息系统、实验室信息系统、影像系统、公共卫生信息系统和远程医疗系统，实现医疗机构和其他机构之间的互联互通，并于2020年实现所有加拿大人拥有可交互的电子病历EHRS[4]。

以Panorama为例[5]，2003年SARS大规模爆发，促使加拿大政府开始思考以往公共卫生事件监管系统的弊端，开始全面建设旨在解决信息技术鸿沟的公共卫生监测IT运用系统（Public Health Surveillance IT Application,Panorama）。Panorama能收集、共享和分析广泛的卫生信息，不仅对辅助卫生保健人员管理公共卫生事件起到至关重要的作用，还能帮助公共卫生人员有效管理疾病个案与疾病暴发、免疫和疫苗存储[6]。

（二）美国大数据卫生决策支持系统发展经验

目前，美国卫生决策支持系统主要应用于医疗服务、公共卫生突发应急事件管理、医院管理、医疗保险管理等领域。美国典型的卫生决策支持系统包括美国疾病监测报告系统（National Electronic Disease Surveillance System, NEDSS）、美国医生医嘱录入系统（Computerized

Physician Order Entry，CPOE）等[7]。

NEDSS[8]建设目标是在美国联邦政府、州和地方统一数据和信息系统标准的基础上，执行公共卫生概念数据模型（PHCDM）或者HL7参考信息模型（RIM），通过整合美国联邦政府、州和地方数据，建立一个有效、完整、能互操作的信息系统，实时收集和分析疾病数据，监测并评估疾病发展趋势，确定公共卫生突发事件，指导疾病的预防、控制和治疗。

CPOE[7]通过药物处方信息的电子化录入、消费者历次用药记录的自动整合和个人电子处方数据的汇总，为医生提供药物相互作用和不良反应的信息，帮助医务人员做出正确的药物指导，促进临床合理用药，减少不良事件发生；同时提供重复用药检验，与医疗保险用药目录匹配和基本政策补贴计算，规范医疗行为，有效降低医疗费用。

（三）英国大数据卫生决策支持系统发展经验

2013年，英国启动了健康医疗大数据平台care.data。care.data集中了最详尽的医疗数据，包含了全英国的家庭医生（General Practitioner, GP）和医院记录的病历数据，以及社会服务信息。不仅可以用于“直接医疗”（direct care）的目的，还可以通过数据的统一归口、共享、分析，更好地认识患者及疾病，研发药物和治疗方式；认识公共卫生和疾病的发展规律；保障每一个患者都能获得个性化、高质量的服务，还可以用于比较不同区域资源状况和医疗质量，使有限资源的利用达到最大化[9]。

目前，英国电子病历系统由国家医疗概要记录、本地医疗详细记录以及为管理、研究和其他“次要”目的提供汇总数据辅助使用服务三部分组成[10]。患者的概要记录全部存储于全国性中心数据库骨架。通过中心网络传输，不仅可以实现电子病历在不同医院之间的共享，通过对全国范围电子病历的大数据进行分析和数据挖掘，可以辅助于公共卫生决

策，如发现地区疾病谱、流行病扩散趋势[11]。

三、对我国大数据卫生决策应用发展的启示

我国目前尚处于卫生决策大数据应用探索的初期，通过总结国外实践经验，以期为我国未来卫生大数据决策未来应用与发展，提供参考。

（一）卫生信息化建设起基础性作用

加拿大政府设立国家层面决策支持系统的建设者 Infoway, 负责领导全国医疗信息化建设[12]。Infoway 的核心任务不仅包括进行全国范围的项目投资，还具有指导、监督、确定发展方向、制定标准等作用 , 它作为战略投资者，既避免了政府机构的低效，又起到宏观指导、整合各区域资源的引导和协调作用。美国主要卫生决策支持系统的快速发展也得益于近年来计算机和网络通信技术的不断进步，尤其是在数据仓库、数据挖掘、人工智能等方面取得的突破[13]。发展卫生信息化的发展水平在某种程度上决定着卫生决策支持系统的发展水平，卫生信息化建设为决策支持系统建设提供了网络设施、数据来源以及广泛的区域间系统互联和信息交换。

我国尚未有类似统筹负责国家层面的信息机构。2016 年 6 月 21 日，《国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见》（国办发〔2016〕47 号）（以下简称《意见》），该《意见》明确健康医疗大数据是国家重要的基础性战略资源，需要不断夯实健康医疗大数据应用基础，加快建设统一权威、互联互通的人口健康信息平台，推动实现政府决策依托平台建设。2017 年 4 月—7 月，中国健康医疗大数据产业发展集团公司、中国健康医疗大数据科技发展集团公司和中国健康医疗大数据股份有限公司相继筹建，推动《意见》落到实处，我国健康医疗大数据建设正在不断推进。

（二）政府发挥主导作用

其次，加拿大政府始终发挥主导作用、高度重视基础设施建设是加拿大大数据决策支持系统建设取得很大进展的关键因素之一[14]。国家是卫生数据最大的拥有者，对卫生决策支持系统的投入是卫生大数据决策应用的主要驱动力，因此，加大政府的引导、整合、投入和监管，尤其是基础设施和数据库的建立和维护，应该成为我国卫生大数据决策应用发展的重要方向。近年来，我国颁布了《促进大数据发展行动纲要》、《全国医疗卫生服务体系规划纲要（2015—2020 年）》等，提出要推动大数据的应用，提高医疗卫生服务能力和管理水平。目前我国卫生信息化基础相对薄弱、人口数量大、医疗机构众多，要在政府的引导下，调动多方力量整合现有数据，形成统一结构的数据库，同时注重统筹规划，不断完善卫生管理机制、运行机制和相关的政策制度，为卫生决策支持系统研制和应用提高适宜的外部环境，促进我国决策支持系统有序发展。

（三）技术和人才是关键

随着大数据时代的带来，半结构化和非结构化的真实世界数据迅猛增长，传统分析方法已经无法满足需求[15]。大数据的技术革新体现在各个环节[16]，在数据采集、预处理和存储方面：通过从传感器网络、社交媒体等数据源中获取结构化、半结构化和非结构化数据，并将数据主体进行预处理与存储，在这一环节，大量数据被集成和变换，成为“合格”的数据，从而被进一步分析与利用；在数据分析方面：必须对经过预处理之后的数据精心分析与挖掘，目前主要分析方法包括并行计算、实时计算与流式计算等。大数据的发展也对人才提出了更高的要求，不仅需要数据库管理人员、统计分析人员、大数据新技术等应用人才，也需要管理人员能够正确认识健康医疗大数据应用与卫生业务及管理需要之间的关系，并能够科学决策[17]。

（四）数据安全性风险

隐私性是医学信息的重要特征，医疗卫生信息中存在大量隐私信息。利用医疗卫生数据不可避免涉及患者隐私问题，在大数据时代，云端可以每时每刻对用户的信息进行采集，使每一个用户成为“透明人”，大数据背景下面临的用户隐私保护、数据内容可信验证、访问控等安全挑战更为严峻。大数据应用不断发展使得信息使用更加便捷、范围更广，但也增加了保护患者隐私的难度[18]。因此一方面需要通过采取技术手段来限制用户对医疗信息资源的权限管理[16]，例如：数据发布匿名保护技术、社交网络匿名保护技术、数据追溯技术和数据中心体系结构建设等；同时还完善隐私保护和侵犯追责等方面法律规章，使得医疗卫生数据在合法范围内使用。

参考文献

[1] Tom K. Big data is a big deal[EB/OL].http:// www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

[2] 孟群, 毕丹, 张一鸣,等. 健康医疗大数据的发展现状与应用模式研究[J]. 中国卫生信息管理杂志, 2016, 13(6):547-552.

[3] 田晨. 大数据背景下循证决策的应用研究[D].湖南大学,2016.

[4] 杨帆. 保健对象电子健康档案研究与数据库实现[D]. 第三军医大学, 2009.

[5] Frisch L E, Borycki E M, Capron A, et al. Public Health Informatics in Canada[M]// Public Health Informatics and Information Systems. Springer London, 2014.

[6] 胡红濮, 代涛, 高星. 加拿大卫生决策支持系统的发展与启示[C]// 中国医学科学院/北京协和医学院医学信息研究所/图书馆2011年学术年会. 2012.

[7] 代涛. 卫生决策支持系统发展的国际经验[J]. 中国循证医学杂志, 2012, 12(3):247-250.

[8] Ward J, Hildebrandt C, Patel A. NEDSS Base System (NBS): Electronic Data Exchange and Workflow Decision Support:[J]. Online Journal of Public Health Informatics, 2017, 9(1):211.

[9] 洪延青, 何延哲. 英国健康医疗大数据平台care.data为何停摆?[J]. 中国经济周刊, 2016(29):77-79.

[10] 陈荃, 代涛, 李新伟. 英国卫生决策支持系统发展与启示[J]. 中国循证医学杂志, 2012, 12(4):371-373.

[11] British Telecommunications plc. Connecting health care. 2011-01-09. Available at: http:// www.N3.nhs.uk.

[12] Lau F. Extending the infoway benefits evaluation framework for health information systems.[J].

Studies in Health Technology & Informatics, 2009, 143:406-13.

[13] 李新伟, 代涛, 胡红濮. 美国卫生决策支持系统建设与应用[J]. 中国循证医学杂志, 2012, 12(3):251-255.

[14] 孙敬水. 加拿大电子政府发展的经验及其对我国的启示[J]. 科技管理研究, 2004, 24(1):25-28.

[15] 潘惊萍, 张子武, 段占祺,等. 医疗卫生大数据应用探索[J]. 中国卫生信息管理杂志, 2016, 13(4):420-424.

[16] 廖建新. 大数据技术的应用现状与展望[J]. 电信科学, 2015, 31(7):1-12.

[17] 李岳峰, 周光华, 孟群. 我国医改卫生统计与信息化人才培训框架设计与思考[J]. 中国卫生信息管理杂志, 2013, 10(2):120-124.

[18] 周永军, 洪梅. 医院数字化建设的伦理要求[J]. 中国医学伦理学, 2010, 23(3):109-111.

（责任编辑：甘银艳）

基于健康医疗大数据的卫生决策机制构建研究

杨山石　王贤吉　杨燕　何江江　王力男　金春林

【摘　要】对医疗卫生行业而言，健康医疗大数据的应用正在快速发展，并已逐步显现出开创卫生决策新纪元的潜力。本文在前期研究的基础上，结合公共决策的发展演变，总结提炼出基于健康医疗大数据的卫生决策机制，并结合具体实践过程中可能面临的挑战，给出了推进健康医疗大数据在卫生决策中应用的建议。

【关键词】健康医疗；大数据；卫生；决策

大数据作为政府治理现代化的一种技术支撑，能够有效帮助政府机构优化公共决策。对医疗卫生行业而言，构建基于健康医疗大数据的卫生决策并付诸应用实践可以通过辅助各级卫生部门决策者进行科学决策，实现合理配置医疗资源、监测防控疾病、提高医疗服务质量、有效控制医疗费用不合理增长等目标。

一、公共决策的发展演变

卫生决策作为公共决策的一种，具有公共决策的共性，公共决策的发展演变历程中通用的模式模型对于构建基于健康医疗大数据的卫生决策机制具有重要的指导作用。

（一）传统的公共决策模式

传统的公共决策模式大致分为三种[1]：一是直接依靠决策者所具有的分析问题和理性判断能力进行决策，但决策结果容易受到个人主观见

第一作者：杨山石，女，科技管理工程师
通讯作者：金春林，男，研究员，上海市卫生和健康发展研究中心主任，上海市医学科学技术情报研究所所长
作者单位：上海市卫生和健康发展研究中心，上海 200040；上海市医学科学技术情报研究所，上海 200031
基金项目：第四轮上海市公共卫生体系建设三年行动计划（2015—2017）项目（GWIV-33）

解的误导；二是通过相关利益者分析进行决策，决策参与者代表不同利益，通过求同存异，形成最终决策共识，但决策过程往往效率低下；三是通过数据抽样，运用统计学方法得出结论，决策有规范的研究过程，包括样本量确定、研究设计、现场调查、数据分析等，但样本的代表性存疑，统计推断的外部效度有待检验。

（二）“DIKW 金字塔”模型

随着计算机和网络技术快速发展，“数据”获得成本的下降，决策者寻求基于数据的决策。拉塞尔 • 阿克夫（RussellAckoff）提出了“数据 - 信息 - 知识 - 智慧”（Data-Information-Knowledge-Wisdom）“DIKW 金字塔”模型（见图 1）[2]。

在此模型中，数据位于底部，是所收集研究对象的有关资料，信息是对收集资料进行整合分析得到的结果，知识则是对信息进行加工转换得到的产品，智慧是依据知识实施公共决策。

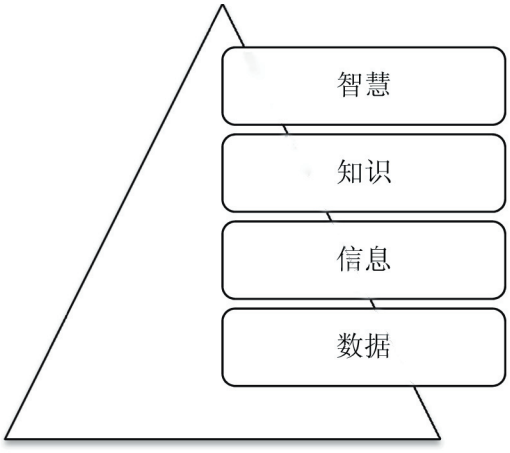


图 1 “DIKW 金字塔”模型

（三）“神经系统”模型

“DIKW 金字塔”模型较好地阐释了数据驱动决策的理念。由于大数据具有的“全样本分析”和“相关性分析”等特性，牛正光等人提出了基于大数据的“数据 - 智慧”（Data-Wisdom）的“神经系统”模型（见

图 2）[3],并认为“DIKW 金字塔”模型将向“DW 神经系统”模型转变。

“DW 神经系统”模型中，决策分为五步：第一步确定待解决的问题和决策目标;第二步围绕问题目标，广泛收集数据，并建立大数据库；第三步进行数据分析，实现隐形数据显性化转变；第四步，数据分析结束后，向决策者汇报信息，解释结果，依据决策目标的满足程度决定是否需要重新整合、分析数据；第五步是决策。

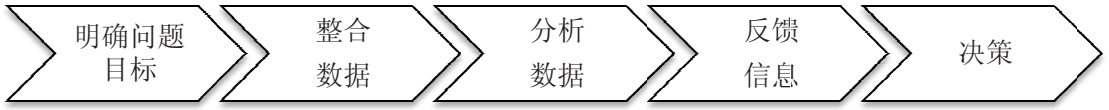


图 2 “DW 神经系统型”决策模式

（四）智慧公共决策

胡税根等提出了智慧公共决策 [4]。智慧公共决策与“DW 神经系统”模型内容相似，皆以大数据系统为支撑，决策者依靠相关数据分析所得到的信息和证据制定决策。从本质上来讲，智慧公共决策是以大数据驱动为核心，以新一代信息技术为支撑，以公共利益最大化为目标，具有全面感知、客观透明、实时连续等特征的一种全新的公共决策，既是对已往公共决策思想的继承，又是基于大数据技术实现的模式创新和突破。

从决策主体来看，更加关注多元参与，决策的关键不再是传统权威，而是网络化多元共同体的共同经验、学习过程和话语赋权。

从决策过程来看，可以运用分布式文件系统、分布式数据库、批处理技术，以及开源实现平台等最新的云计算技术，智能化分析数据之间及数据与环境之间的广泛联系，实时连续地为决策制定提供辅助。

从决策手段来看，通过明确问题所在，利用数据寻找解决途径，并不断监测决策的效果，从而决定下一步的行动和措施。

从决策目标来看，决策借助大数据技术优势，利用传感器、射频识别、

数据检索分类工具、条形码等方法，结合互联网、物联网、等技术全面感知社会事项及公众所需，采取网络地图、标签云、历史流图等最新的大数据可视化技术把握过去、现在与未来的发展规律和历史逻辑，展示决策者行为的全过程。

二、基于健康医疗大数据的卫生决策机制构建

与教育、交通等其他公共部门相比、卫生领域具有自身的特点，因此，公共大数据决策在卫生领域的具体应用还需要结合卫生事业的特性。本文在前期研究的基础上，结合公共决策的发展演变，总结提炼出基于健康医疗大数据的卫生决策机制（见图 3）。

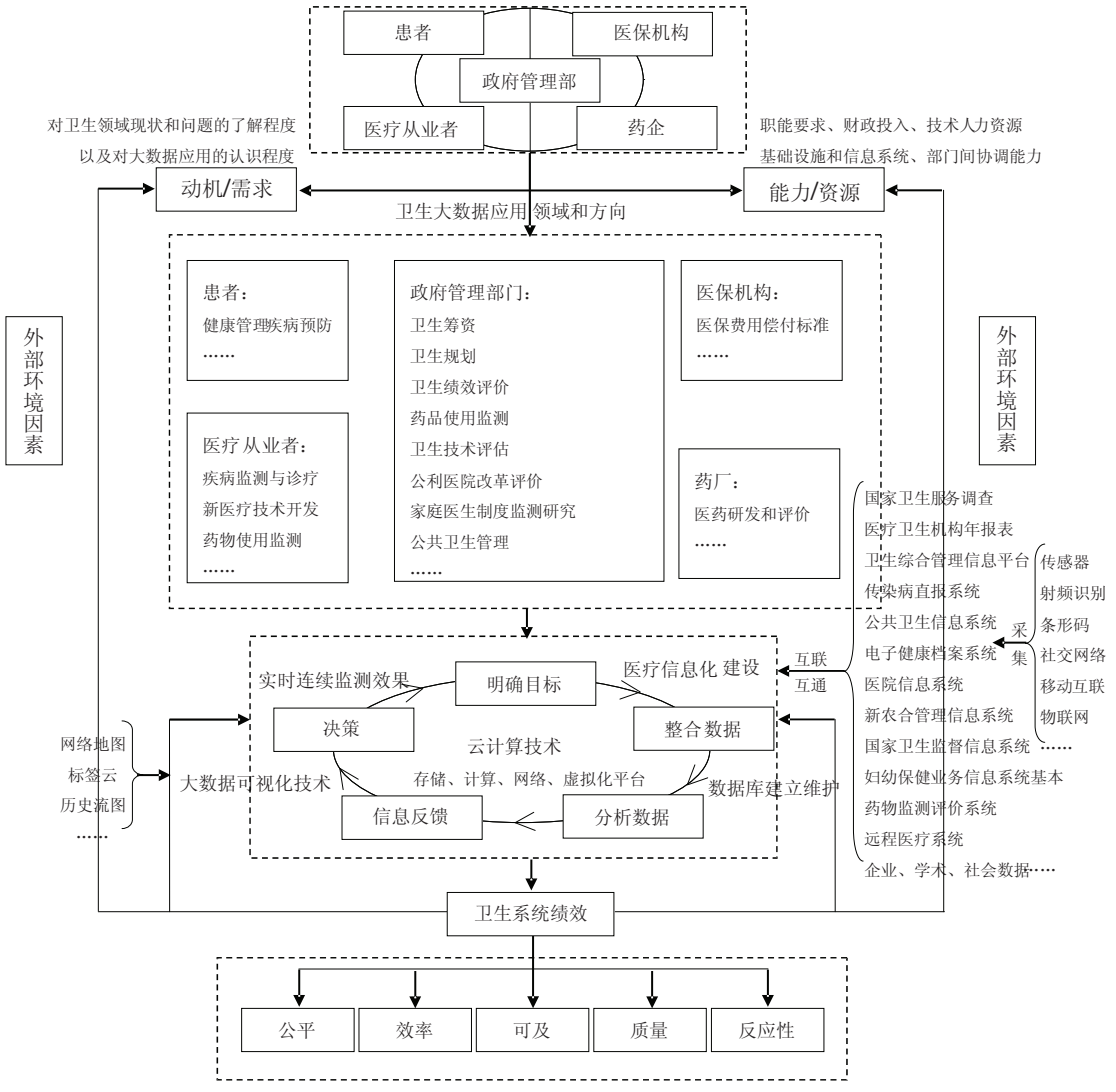


图 3　基于健康医疗大数据的卫生决策机制

（一）动机和能力驱动

动机和能力是驱动公共部门大数据应用的关键因素[5]，因为公共部门只有意愿和能力兼具，才能使大数据应用发挥效力。

结合卫生领域特征，卫生部门主要负责人对卫生领域现状和问题的了解程度以及对大数据应用的认识程度是影响卫生决策大数据应用的关键因素；而各卫生部门职能要求、财政投入、技术人力资源、信息化基础设施和管理系统、多部门之间的协调能力，则是支撑卫生决策大数据应用的关键因素。

（二）利益相关者影响

各大医疗产业核心人群的需求不容忽视。患者渴望加强健康管理和疾病预防，且可以用合理的医疗花费，在任何医疗场所都能得到正确的治疗等；医疗从业者进行疾病监测与诊疗、药物使用监测、新医疗技术开发等，并期望获得与其劳动能力相匹配的薪酬等；医保机构希望加强合理利用卫生服务和药物，优化医保费用偿付标准，减少保险支付的需求等；药企通过创新研制，向市场提供安全有效的药物以获得利润报酬等。

政府管理者在借助大数据对这些因素加以研究的基础上，可进行卫生筹资、卫生规划、卫生绩效评价、卫生技术评估、公立医院改革评价、家庭医生制度监测、公共卫生管理等方面的研究和应用。

（三）数据来源与利用

政府不再是数据的唯一或主要来源，卫生决策中的重要数据源包括：药厂、患者、医疗从业者、医保机构、政府管理机关[6]。健康医疗大数据不仅来源于各级各类卫生信息系统，包括国家卫生服务调查、医疗卫生机构报表、传染病直报系统、公共卫生信息系统、电子健康档案系统、医院信息系统、新农合管理信息系统、国家卫生监督信息系统、妇幼保

健业务信息系统、基本药物监测评价系统、远程医疗系统等；还来源于传感器、射频识别、条形码、社交网络、移动互联、物联网等，非结构化数据与结构化数据并存。

如何在 PB 级数量单位的数据中获取有效信息，政府应加强与社会机构或企业合作，来加快健康医疗大数据应用。围绕需要解决的问题和目标，借助云计算技术，通过医疗信息化建设实现各类信息源的互联互通，建立并维护数据库，进行数据的整合与分析，并将分析的结果以网络地图、标签云、历史流图等大数据可视化形式展现出来，最终辅助医疗卫生行业政府管理部门进行决策。

（四）决策效果的评价

政府不仅在决策之前可以掌握更充分的信息，而且可以在政策实施过程中通过社交媒体实时、在线收集公众意见，实现对公众关注的问题以及对政策的反应的监测并及时调整相应政策。大数据基于数据驱动的认知能够洞察传统方式未能识别的模式和相关关系，并且可以建立预测模型，对政策实施后的效应进行仿真模拟[7]。

大数据在卫生决策中的应用究竟会对卫生系统绩效产生怎样的影响？可以结合卫生系统绩效的五项具体产出：公平、质量、效率、可及和反应性加以分析。通过实证研究，分析大数据应用的价值创造和对为卫生系统绩效的影响，并以此作为未来卫生大数据决策广泛应用的依据。

三、思考与建议

健康医疗大数据为卫生决策乃至整个医疗卫生事业带来了发展机遇。但是，与其他领域一样，健康医疗大数据在卫生决策中的应用也存在着诸多挑战。只有迎接这些挑战，才能将健康医疗大数据辅助卫生决策的价值充分发挥出来。

当前，我国的健康医疗数据资源汇集程度和健康医疗大数据应用于

卫生决策的水平与发达国家相比仍有一定的差距，国内外在行政体制和法律制度等方面也有较大的差异。我们不能简单地照搬欧美国家的一些做法，而是要在汲取借鉴的同时立足国情，不断提高数据资源的质量，有序推进健康医疗大数据在卫生决策领域的示范性应用。

（一）提升健康医疗大数据采集能力

采集大数据，需要有安全、高效、可控的基础网络和各种业务数据库作为保障。各级政府要加大财政投入以及信息化基础设施和管理系统建设力度，明确卫生决策和健康医疗大数据资源管理的综合协调部门，以卫生决策需求为导向制定健康医疗大数据资源规划，建立公共基础数据资源的标准，完善数据资源采集、共享、利用和保密等相关制度，完善卫生信息资源目录体系，扩大健康医疗大数据的采集和交换共享范围。包括对跨部门的相关数据进行整合、通过多种途径获取政府内外部相关数据，构建横向到边、纵向到底的健康医疗信息网络，形成国家健康医疗大数据应用体系。

同时需要注重非结构化数据的汇集，推进可穿戴设备、人工智能、健康电子产品、健康医疗移动应用等产生的数据资源以及网路舆情数据等规范接入共享平台。

（二）实现健康医疗大数据比对更新

通过对大数据的跨部门协同管理，将分散在政府各个机构的相关数据关联起来，能够有效地减少公众重复提交数据和政府重复输入数据，有利于改善服务体验并降低行政成本[7]。目前我国卫生部门的数据采集仍以各自采集为主，纵向采集但缺乏横向比对，给数据质量带来了很大的问题，主要体现在记录遗漏、数据项错误、信息过时等。

如人口信息采集，是基于部门直接连通到社区的业务系统。各部门在采集人口信息时，都侧重于自身所服务的人群，因而各部门人口采集

的口径都不尽相同，与部门直接服务对象相关的数据比较准确（如民政的低保户数据、计生的育龄妇女数据），但与其业务不直接相关的人口数据质量相对较低。造成人口数据不够准确、不够全面的主要原因是部门间人口数据口径不一，部门间、系统间数据共享不够且缺乏数据比对机制，一线人员输入、核对数据的责任心不强且缺乏评价手段，未形成数据更新的机制等等。

因此，应在完善数据采集机制的基础上，建立数据更新策略，通过条块结合比对、第三方验证的方法，即部门权威数据与社区网格化核实相结合，同时引入数据涉及到的个人和企业、其它部门的相关数据加以校验，以提高数据的完整性、准确性和及时性。

（三）加强健康医疗大数据挖掘应用

在提高卫生大数据数量和质量的同时，从技术上加强对健康医疗大数据的深度挖掘，将数据的关联性和潜在价值性挖掘出来。为此，在人才培育方面，要强化医学信息学学科建设和人才融合培育，打造高层次、复合型的核心技术研发人才和科研团队，建立多层次多类型的健康医疗大数据应用人才培养培训体系，推进政府、高等院校、科研院所、医疗机构、企业人才共育模式，促进健康医疗大数据人才队伍建设[8]。

探索大数据在卫生决策领域的应用可遵循“效益优先、突出重点、先易后难、逐步推进”的原则，综合考虑数据的可得性、数据分析的复杂度和可能的经济和社会效益等因素，确定卫生大数据应用于决策的重点和时序。

卫生大数据应用不限于政府自身的应用，政府还可以将非涉密的卫生大数据逐步对社会开放，构建政、产、学、研合作共赢的大数据采集应用平台，提高整个社会对卫生信息资源的开发利用水平，这样一方面可以丰富应用，优化公众体验；另一方面也有利于卫生大数据产业发展。

参考文献

[1] 张红凤，韩琭，闫韶华. 转型期公共决策模式路径优化：从传统模式向动态协同模式的转变[C]. 泰山学术论坛—市场的决定性作用与公共政策创新专题暨中国经济理论与管理前沿论坛，2014.

[2] Rowley J. The wisdom hierarchy: representations of the DIKW hierarchy[J]. Journal of Information Science, 2010, 33(2):163-180.

[3] 牛正光，奉公. 基于大数据的公共决策模式创新[J]. 中州学刊，2016，232(4)：7-11.

[4] 胡税根，单立栋，徐靖芮. 基于大数据的智慧公共决策特征研究[J]. 浙江大学学报(人文社会科学版)，2015，45(3)：5-15.

[5] 马亮. 公共部门大数据应用的动机、能力与绩效：理论述评与研究展望[J]. 电子政务，2016(04)：62-74.

[6] 田晨. 大数据背景下循证决策的应用研究[D].湖南：湖南大学，2016.

[7] 何军. 大数据对我国电子政务的影响及对策研究[J]. 发展研究，2014，02:56-60.

[8] 金兴，王咏红. 健康医疗大数据的应用与发展[J].中国卫生信息管理杂志，2016，13(2)：187-190.

（责任编辑：卢伟霞）

健康医疗大数据发展应用的思考

代涛

【摘　要】 随着计算机科学和信息技术的快速发展，健康医疗信息化得到广泛应用，在医疗服务、健康保健和卫生管理过程中产生海量数据集，形成健康医疗大数据。健康医疗大数据的发展与应用对提升医药卫生服务水平、促进健康产业发展等方面发挥着重要作用，许多国家对此已经形成共识，一些发达国家已将其作为国家重大战略并付诸实践。本文在简要介绍大数据概况的基础上，重点阐述健康医疗大数据的主要内容、应用领域、面临挑战和发展趋势等内容。

【关键词】 大数据；健康医疗；应用领域；挑战；趋势

随着物联网、移动互联网、云计算等新兴信息技术的迅猛发展和普及应用，行业应用系统规模迅速扩大，产生的数据呈现前所未有的爆发式增长态势[1]。国际数据公司（International Data Corporation, IDC）的Digital Universe研究报告指出：2013年全球数据总量超过4.4ZB，并且每两年将翻一番，预计2020年将达到44ZB[2]。数据的复杂性也急剧增长，数据多样性、变化快、低价值密度等复杂特征日益显著[1-2]。其复杂性对传统的计算技术和信息技术提出严峻挑战，“大数据”一词已不仅在于数据规模的定义，更代表着大数据处理所需的新技术方法，以及大数据分析和应用所带来的新发明、新服务和新机遇[3-4]。

大数据可发挥其全样本、深入关联、注重相关性等优势，解决以往存在的“信息碎片化”、“盲人摸象”等问题，提升人们的洞察力和统筹

第一作者：代涛，男，研究员，博士
作者单位：中国医学科学院医学信息研究所，北京 100020
注：本文转载自《医学信息学杂志》2016年第37卷第2期。

规划能力[5]。越来越多的证据证明，只要拥有足够长的连续历史数据，足够强大的计算分析工具，就可以根据过去和现在预测未来。大数据分析挖掘将能为行业/企业带来巨大的商业价值，为衣食住行、健康、社交、信用等与生活息息相关领域提供高附加值的增值服务，进一步提升行业/企业的经济、社会效益[6-9]。

未来，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、控制和运用也将成为国家间和企业间新的争夺焦点[10-11]。联合国于2012年发布关于大数据政务的白皮书《大数据促进发展：挑战与机遇》[12]，以推动各国政府机构、重大行业对大数据技术的研究和应用。美英等欧美发达国家从国家科技战略层面提出了一系列的大数据技术研发计划[13]，如美国发布《大数据研究和发展计划》[14]，推出“数据—知识—行动”计划[15]，启动“数据开放行动”[16]。欧盟正在力推《数据价值链战略计划》[17]，英国发布《英国数据能力发展战略规划》[18]，加拿大发布《健康大数据分析白皮书》[19]，澳大利亚发布《公共服务大数据战略》[20]，日本发布《创建最尖端IT国家宣言》[21]，韩国提出“大数据中心战略”[22]。中国在2015年相继出台《推进“互联网+”行动指导意见》[23]和《促进大数据发展行动纲要》[24]，系统部署大数据发展工作。

一、健康医疗大数据的主要内容与应用领域

（一）主要内容

随着健康医疗信息化的广泛应用，在医疗服务、健康保健和卫生管理过程中产生海量数据集，形成健康医疗大数据[25-26]。健康医疗大数据通常可以划分为以下几个方面：以电子健康档案、电子病历、医学影像、检验检查等为主的健康医疗服务数据；基因序列、蛋白质组等生物医学数据；新型农村合作医疗、城镇职工基本医疗保险、城镇居民基本医疗

保险等医疗保险数据；药物临床试验、药物筛查、基本药物集中采购、医疗机构药品与疫苗电子监管等医药研发与管理数据；疾病监测、突发公共卫生事件监测、传染病报告等公共卫生数据；患者行为表现、保健品购买记录、健身信息等行为与情绪数据；卫生资源与医疗服务调查、计划生育统计等统计数据；居民婚姻、家庭、计划生育登记等人口管理数据；与人类健康密切相关的空气污染物和气候状况等环境数据[27-28]。

（二）主要应用领域

健康医疗大数据将为临床诊疗、药物研发、卫生监测、公众健康、政策制定和执行等带来创造性变化，全面提升健康医疗领域的治理能力和水平，创造极大的价值[27]。据麦肯锡预测，如果有效利用健康医疗大数据，每年可为美国带来3000多亿美元的价值[6]。

1. 为临床诊疗管理与决策提供支持

通过效果比较研究，精准分析包括患者体征、费用和疗效等数据在内的大型数据集，可帮助医生确定最有效和最具有成本效益的治疗方法。利用临床决策支持系统可有效拓宽临床医生的知识，减少人为疏忽，帮助医生提高工作效率和诊疗质量。通过集成分析诊疗操作与绩效数据集，创建可视化流程图和绩效图，识别医疗过程中的异常，为业务流程优化提供依据。

2. 为药物研发提供支持

通过分析临床试验注册数据与电子健康档案，优化临床试验设计，招募适宜的临床试验参与者。通过分析临床试验数据和电子病历，辅助药物效用分析与合理用药，降低耐药性、药物相互作用等带来的影响。通过及时收集药物不良反应报告数据，加强药物不良反应监测、评价与预防。通过分析疾病患病率与发展趋势，模拟市场需求与费用，预测新药研发的临床结果，帮助确定新药研发投资策略和资源配置。

3. 为公共卫生监测提供支持

大数据相关技术的应用可扩大卫生监测的范围，从以部分案例为对象的抽样方式扩大到全样本数据，从而提高对疾病传播形势判断的及时性和准确性。将人口统计学信息、各种来源的疾病与危险因素数据整合起来，进行实时分析，可提高对公共卫生事件的辨别、处理和反应速度并能够实现全过程跟踪和处理，有效调度各种资源，对危机事件做出快速反应和有效决策。

4. 为公众健康管理提供帮助

通过可穿戴医疗设备等收集个人健康数据，辅助健康管理，提高健康水平。为医患沟通提供有效途径，医生可根据患者发送的健康数据，及时采取干预措施或提出诊疗建议。集成分析个体的体征、诊疗、行为等数据，预测个体的疾病易感性、药物敏感性等，进而实现对个体疾病的早发现、早治疗、个性化用药和个性化护理等。

5. 为医药卫生政策制定和执行监管提供科学依据

整合与挖掘不同层级、不同业务领域的健康医疗数据以及网络舆情信息，有助于综合分析医疗服务供需双方特点，服务提供与利用情况及其影响因素，人群和个体健康状况及其影响因素，预测未来需求与供方发展趋势，发现疾病危险因素，为医疗资源配置、医疗保障制度设计、人群和个体健康促进、人口宏观决策等提供科学依据。通过集成各级人口健康部门与医疗服务机构数据，识别并对比分析关键绩效指标，快速了解各地政策执行情况，及时发现问题，防范风险。

二、以开放促进大数据的应用创新

（一）全球开放数据概况

大数据应用的基础是数据足量全面。自 2009 年开始，为推动政府数据开放与共享，促进社会应用创新，美英等发达国家纷纷开展数据开

放运动，其后许多发展中国家也相继开展[29-30]。2011 年 9 月美国、英国、巴西、墨西哥、印度尼西亚、挪威、菲律宾、南非 8 个国家联合签署《开放政府声明》，成立开放政府联盟；经过 4 年多的发展，截至 2016 年 1 月其成员国发展为 69 个[31]。2013 年 6 月美、英、法、德、意、加、日、俄 8 国领导人在 G8 峰会上签署《开放数据宪章》，明确数据开放 5 大原则、14 个重点领域和 3 项共同行动计划[32]。美、英等国均在国家层面设立统一的数据开放门户 data. gov.[国别]，并分级分类发布一定数量的数据集。

万维网基金会和开放数据研究院于 2013 年和 2014 年连续两次从准备度（readiness）、实施度（implementation）和影响力（impact）3 个方面，对全球近 80 个国家和地区的开放数据情况进行评估并发布评估报告《开放数据晴雨表》[33-34]。其中，2014 年全球开放数据排名前 10 的国家和地区为英国、美国、瑞典、法国、新西兰、荷兰、加拿大、挪威、丹麦、澳大利亚，见表 1，中国排名第 46[34]。

表 1　2014 年度“开放数据晴雨表”评出的开放数据排名前 10 的国家和地区[34]

排名	国家 / 地区	总体得分	准备度（%）	实施度（%）	影响力（%）
1	英国	100.00	98	100	100
2	美国	92.66	96	88	100
3	瑞典	83.70	100	76	88
4	法国	80.21	91	75	84
4	新西兰	80.01	81	88	55
6	荷兰	75.79	95	76	57
7	加拿大	74.52	90	75	58
7	挪威	74.59	88	73	64
9	丹麦	70.13	94	54	95
10	澳大利亚	68.33	92	69	43

（二）健康医疗数据开放与应用情况

健康医疗作为重要的民生领域，美英等国家均将其作为优先开展数据开放的领域。截至2016年1月美国国家数据开放平台data.gov上共发布192119个数据集；其中健康医疗类数据集1701个，以医疗保险与补助相关数据居多（32.67%），其次为人口统计类数据（15.41%）、疾病控制与公共卫生相关数据（10.55%）、卫生管理与质量监测相关数据（10.17%），再次为疾病治疗相关数据（7.18%），卫生费用相关数据（6.96%）。英国国家数据开放平台data.gov.uk上共发布16332个数据集；其中健康医疗类数据集1613个，以卫生管理与质量监测类数据（37.47%）和人口统计数据（32.15%）居多，两者共占69.62%，其次为卫生费用相关数据（9.55%）、疾病控制与公共卫生数据（6.35%）等。

纽约大学管理实验室（GovLab）全面调研了美国企业利用开放数据创造新的商业、产品与服务的情况，研究发现美国政府开放的健康医疗数据已在80余家企业中得到有价值的应用，企业主营业务涉及健康医疗、科学研究、研究与咨询、数据与技术、生活方式与客户服务、食品、保险、教育、法律、金融与投资、地理空间等多种领域[35]。实现增值利用的数据具有以下共同特点：符合国家数据再利用许可；为开放格式的结构化数据；持续更新；数据粒度较为精细[36]。

三、健康医疗大数据发展应用面临的主要挑战

（一）健康医疗大数据发展应用的体制机制缺失

大数据的融合应用、共享协作等体制机制不健全。面对来自不同机构、采取不同格式、遵循不同标准的多源数据，如何实现数据、技术与应用的有机融合，仍存在诸多障碍。多学科、产学研、跨机构的合作机制缺失，面临数据融合共享渠道不畅、产业自主创新实力不强、运行机制不顺、政策法规缺位等瓶颈问题。

（二）基础设施的能力和质量仍需提高

随着健康医疗大数据的飞速增长，对基础设施的能力和质量提出更高要求。一方面要处理不同设备和应用系统所产生和收集的呈指数增长的数据，另一方面要利用适当的管理模式将信息化基础设施打造成持久的研究与应用平台，确保连续性并实现跨领域合作[11]。数据量增加、跨地区跨国界计算、协同应用等在传输速度、可靠性和服务质量等方面提出了更高的要求。同时数据的时效性和折旧性需求并存，需要具备更先进的计算能力和更高容量的吞吐能力。

（三）健康医疗大数据发展应用的关键技术需要新的突破

首先，当前的标准和技术难以满足健康医疗大数据整合应用的要求，缺少统一的标准、固定的描述格式和表示方法等，不同层次结构化、半结构化与非结构化数据的集成融合困难[37]。其次，软硬件协同与数据处理的时效性局限。目前，分布式系统的一致性、可用性和分区容错性3者不可兼得，难以解决医疗卫生数据采集、处理的实时性以及动态索引、先验知识缺乏等难点问题；硬件异构要求软件适应不同机器多核CPU的并行处理机制；大部分能量损耗于大规模集群的闲置节点上。

（四）数据管理面临质量、保存、整合等诸多挑战

首先是数据质量问题，人类基因组学、健康行为、公共卫生检测等相关数据规模、产生速度和复杂度的增加使得各种类型的误差和错误更容易被引入系统，分布式数据清洗、质量检测、修复等挑战性问题突出。其次是数据保存的问题，各种存储技术缺乏统一的标准从而难以兼容，导致大量数据的丢失，对数据在新旧系统之间的迁移提出了巨大挑战。此外，数据整合度欠缺，数据尚未完整嵌入到业务流程和组织管理实践中，如患者监控数据尚未整合到临床诊疗中，临床数据尚未整合至公共卫生服务和重大疾病、传染病监测中等。

（五）安全与隐私保护措施欠缺

安全隐私保护薄弱影响数据的共享范围，健康医疗大数据涉及患者的隐私、医疗机构 / 企业的安全或者其他特殊要求，存在较为严重的安全隐患。基因组学的发展和研究活动规则的改变，使得隐私的泄露几乎不可避免。传统数据库通过基于数据粒度的安全性控制实现安全隐私保护，但是大数据的操作还比较欠缺有效的安全保护措施。

（六）复合型人才严重缺乏

推动健康医疗大数据应用发展亟需大批的复合型人才。据麦肯锡预测，即使在美国这样的信息技术强国，其相关人才缺口也将于 2018 年达到 14 ～ 19 万 [6]。目前，世界上仅有少数公司掌握大数据分析核心技术，全世界范围内都亟需数据解释人员，利用信息技术将数据处理后的可视化结果展现给决策者，将大数据分析的结果转化为政策，直接为医疗服务、管理、决策提供支撑。因此，急需推进政府、高等院校、科研院所、医疗卫生机构、企业等人才共育模式的建立。

四、健康医疗大数据的发展趋势

（一）健康医疗大数据将更快发展和更广泛应用

随着大数据采集、存储、组织、整合、挖掘、协同与互操作等技术的快速发展，健康医疗大数据的应用将更为广泛。主要方向有：基于多感知器和智能终端的健康医疗数据采集，基于云平台的分布式存储与并行计算，动态大数据的实时处理及非结构化数据处理，基于领域本体的数据标注与语义提取，多元异构数据的深度整合，海量动态数据的学习、推理、预测与知识发现等。这些新理论与新技术的突破，将为健康医疗大数据驱动的创新应用提供更加强有力的支撑。

（二）健康医疗大数据驱动临床决策支持和精准医学研究

针对健康威胁大、发病率高、诊疗费用高、改进实效好的肿瘤、心

脑血管疾病和老年慢性病等疾病，建设专病临床医学数据中心，同时利用基因芯片与基因测序技术，获得海量个体的基因组、蛋白质组、代谢组数据，应用大数据分析挖掘技术开展疾病发生发展机理、早期诊断、疗效比较研究，发现疾病治疗相关的靶标，从而提高其预防和诊疗水平已成迫切需要，将成为临床决策支持和精准医疗研究的重要领域。分布式存储与并行计算架构、异构数据整合与挖掘等技术将在基因组学、转录组学、蛋白质组学、代谢组学、表型组学等生物医学大数据研究中发挥重要作用。

（三）电子健康档案向着精细化、智能化和便捷化的方向发展

汇聚个人全面健康信息，建立覆盖全体居民的电子健康档案云平台，让每位公民拥有一份标准化的电子健康档案，并及时方便地获取健康医疗数据。电子健康档案云平台的建设将有助于推动在线病情跟踪与咨询，减少重复检查带来的时间和经济负担。基于电子健康档案开发疫苗接种提醒、处方遵从性提醒、药物相互作用提醒等功能，将有助于实现集预防、治疗、康复和健康管理于一体的个人全生命周期的健康管理。同时，可通过电子健康档案分析全人群健康状况、发病和患病情况，及时获取异常公共卫生事件，提高公共卫生监控的覆盖面和响应速度。

（四）互联网环境下更有助于实现个性化与社会化的健康管理

各类传感器、可穿戴设备、智能手机的迅速发展和应用，使得移动医疗能够真正连接用户与服务，借助互联网将优质医疗资源带到患者身边，使得居家养老、居家护理、慢病管理等健康服务更加便捷化与个性化，促使健康服务模式由治疗向预防和保健转变，催生健康服务新业态。基于社交网络的患者交流与医患沟通将更加普遍，医疗机构更多地借助社交网络平台、移动 APP 等与患者沟通，主动收集患者需求并推送合适的健康医疗服务。同时，并行计算、高维分析、自我量化算法等大数

据处理技术的广泛应用，将提升面向心脑血管、糖尿病等慢性病患者的个性化健康服务的质量与效率。

（五）健康医疗大数据更加注重开放共享与隐私保护

随着大数据的应用价值逐步显现，部分国家着力推进政府数据开放共享以促进社会应用创新。健康医疗作为重要的民生领域，美英等国均将其作为优先开展数据开放的领域。数据开放所带来的一个全新挑战是对个人隐私与数据安全的威胁，在开放共享的同时应强化健康医疗信息安全的技术支撑。一要加强健康医疗行业网络信息安全等级保护、网络信任体系建设，提高信息安全监测、预警和应对能力；二要建立信息安全认证审查机制、数据安全和个人隐私影响评估体系，将信息安全流程化制度化；三要从技术上采取数据封装、数据分离、去除个人标识信息等措施保护个人隐私。

五、思考与建议

健康医疗大数据的发展与应用将推动健康医疗模式的革命性变化，有助于扩大健康医疗资源供给、降低医疗费用、提升医疗服务质量和效率，进而对我国经济、社会、科技和人民生产生活等产生重大而深远的影响。为进一步推动健康医疗大数据的发展应用，要重点做好以下工作：一要尽快制定促进健康医疗大数据发展应用的政策措施，推动建立基于互联网、云服务的健康医疗服务新模式，构建健康医疗信息共享服务平台，推动健康医疗大数据的开放共享和深化应用。二要加快推动相关技术研发和标准规范建设，构建大数据采集、存储、组织、整合、挖掘、协同、互操作和安全保护技术体系。三要促进技术、方法、数据与决策的多维融合，加快专病临床医学数据示范中心建设，推动多来源、多类型、多层面数据的融合应用，促进健康医疗服务的个性化与精细化。四要加强复合型人才培养和开发，构建适应大数据环境下“产—学—研”相结

合的人才培养机制，造就一批高层次的人才队伍。

参考文献

[1] Tony Hey,Stewart Tansley,Kristin Tolle (著）,潘教峰,张晓林等（译）.第四范式:数据密集型科学发现 [M].北京:科学出版社,2012.

[2] Vernon Turner,John F. Gantz, David Reinsel, et al. The Digital Universe of Opportunities:rich data and the increas¬ing value of the internet of things [R/OL]. IDC Analyze Future, 2014. [2014-05-10]. http://idcdocserv.com/1678.

[3] 中国计算机学会大数据专家委员会.中国大数据技术与产业发展白皮书（2013) [R/0L]. [2014-05-10]. http://www.ccf.org.cn/sites/ccf/ccfziliao.jsp? contentld = 2774793649105.

[4] 黄宜华.深入理解大数据:大数据处理与编程实践[M].北京:机械工业出版社,2014.

[5] 黄明达.21世纪人类大健康产业时代的机遇与挑战[EB/0L].[2015-10-09]. http://finance,china.com/fin/x^201404/14/5888837.html.

[6] James Manyika,Michael Chui,Brad Brown,et al.Big da¬ta:the next frontier for innovation,competition, and pro¬ductivity [R]. McKinsey Global Institute，McKinsey & Company. 2011.

[7] 维克托•迈尔-舍恩伯格,肯尼思•库克耶（著）,周涛等（译).大数据时代:生活、工作与思想的大变革[M],杭州:浙江人民出版社,2013.

[8] 埃里克•托普（著）,张南等（译）.颠覆医疗:大数据时代的个人健康革命[M].北京:电子工业出版社,2014.

[9] Momica Bulger, Greg Taylor, Ralph Schroeder. Data-Driven Business Models: challenges and opportunities of big data [R/OL]. [2015-05-10]. http://www.nemode.ac.uk/wp-content/uploads/2014/09/nemode _ business_ models_ for_ bigdata-2014-oxford, pdf.

[10] 李国杰.大数据研究的科学价值[J].中国计算机学会通讯,2012, 8 (9):8-15.

[11] 李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域--大数据的研究现状与科学思考[J].中国科学院院刊,2012,27(6):647-657.

[12] UN Global Pulse. Big Data for Development:challenges Opportunities [R/OL]. [2012-05-30]. http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopme-nt-UNGlobalPulseJune2012. pdf.

[13] 张勇进,王璟璇.主要发达国家大数据政策比较研究[J].信息化研究,2014,(19):1-14.

[14] Mhyeon Gutmann. Big Data R&D Initiative [EB/OL].[2012-08-16]. http://www.digitalpreservation.gov/meetings/documents/ndiippl2/Day%202/BigData _ Gutmann_ DPI2. pdf.

[15] Executive Office of the President of United States. Fact Sheet:data to knowledge to action:new announcements[R]. [2014-02-10]. https://www.whitehouse.gov/sites/default/files/microsites/ostp/Data2Action% 20A-nnouncements.pdf.

[16] Executive Office of the President of United States. Big Data:seizing opportunities preserving values[R/OL]. [2014-08-16]. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may—1_2014.pdf.

[17] 宇燕.欧盟大数据发展战略[EB/OL]. [2014-12-10]. http://www.mofcom.gov.cn/article/i/jyjl/m/

201412/20141200826137.shtml.

[18] HM Government. Seizing the Data Opportunity:a strategy for UK data capability [EB/OL]. [2013-11-06]. https://www.gov.uk/govemment/uploads/system/uploads/attachment—data/flle/254136/bis-13-1250-strategy-for-uk-data-capability- v4.pdf.

[19] Canada Health Infoway. Big Data Analytics in Health White Paper [R/OL]. [2014-03-12].https://www.info-way-inforoute.ca/en/component/edocman/resources/technical-documents/emerging-technology/1246-big-data-analytics-in-health-white-paper-full-report.

[20] Commonwealth of Australia. Australian Public Service:bet¬ter practice guide for big data [R/OL]. [2015-02-16]. http://www.finance.gov.au/sites/default/files/APS-Better-Practice-Guide-for-Big-Data.pdf.

[21] 蒙遗善.日本:用大数据创建最尖端IT国家[EB/ OL].[2014-01-14]. http://www.china-cloud,com/dashujuzhongguo/disanqi/2014/0114/22689.html.

[22] 顾洪文.大数据国家档案之韩国:大数据从基础设施起步[EB/OL]. [2014-01-13]. http://www.china-cloud,com/dashujuzhongguo/disanqi/2014/0110/22652.html.

[23] 国务院.关于积极推进“互联网+”行动的指导意见[国发〔2015〕40号][Z]. 2015.

[24] 国务院.关于印发促进大数据发展行动纲要的通知[国发〔2015〕50号][Z]. 2015.

[25] Rosemary Wyber,Samuel Vaillancourt,William Perry, et al. Big Data in Global Health;improving health in low-and middle-income countries [J]. Bulletin of the World Health Organization,2015.93(3):203-208.

[26] Alex Pentland,Todd G. Reid,Tracy Heibeck. Big Data and Health:revolutionizing medicine and public health[J]. WISH Big Data and Health Report,2013.

[27] 代涛.健康领域如何掘金大数据[N].健康报,2015,928:(6).

[28] 俞国培,包小源,黄新霆,等.医疗健康大数据的种类、性质及有关问题[J].医学信息学杂志,2014,(6):9-12.

[29] Noor Huijboom,Tijs Van den Broek. Open Data:an inter¬national comparison of strategies [J]. European Journal of ePractice,2011,12 (1):4-16.

[30] Open Knowledge. Global Open Data Index: tracking the state of government open data[EB/OL]. [2015-09-09]. http://index,okfn.org.

[31] Open Government Partnership. What is the Open Govern¬ment Partnership?[EB/OL]. [2016-01-31]. http:// www.opengovpartnership.org/.

[32] Cabinet Office of the United Kingdom. Open Data Charter [EB/OL].[2013-06-18]. https://www.gov.uk/gov-emment/publications/open-data-charter.

[33] Tim Davies. Open Data Barometer:2013 global report[R/OL]. [2014-01-04]. http://www.opendatabarome-ter.org.

[34] The World Wide Web Foundation. Open Data Barometer Global Report:second edition[R/OL]. [2015-02-01]. http://www.opendatabarometer.org.

[35] GOVLAB. Open Data 500 U.S. Open Data Compass:what types of companies use which agencies,

data?[EB/OL].[2016-02-22]. http://www.opendata500.com/us.

[36] 李姣,郭海红,郭珉江,代涛.美英政府开放健康医疗数据的主题分布与开放程度量化研究[J].图书情报工作,2015,59(20):132-137.

[37] 张振,周毅,杜守洪.医疗大数据及其面临的机遇与挑战[J].医学信息学杂志,2014,35 (6): 2-8.

（责任编辑：康乐妮）

大数据技术的应用现状与展望

廖建新

【摘　要】本文梳理了大数据研究的4项关键技术：“数据的采集、预处理与存储”、“数据的分析与挖掘”、“数据的隐私保护”、“数据中心体系结构”，挖掘和展示了国内外大数据研究的热点，以期对该领域的研究有一个系统而全面的认识。从处理平台、分析产品、标准化3个方面介绍了大数据技术应用的现状，并列举了大数据现有的产品应用及各行业应用案例。最后，从大数据的分析策略、深度学习、隐私保护和数据质量几个方面揭示了大数据研究存在的挑战和机遇，以期为后续大数据技术应用的发展提供参考。

【关键词】大数据；数据分析；数据挖掘；非结构化数据；隐私保护；并行计算；社交网络

自2012年兴起的“大数据”潮流，让“big data”这个IT圈子里的名词一下风靡了各个行业。虽然大数据的重要性得到了大家的一致认同，但是对大数据的理解却众说纷纭，以至于除了IT以及相关行业以外，各种如餐饮业、房地产业、金融业等行业都声称自己采用了“大数据”技术，并迫不及待地宣布了自己的“大数据”战略。就在这种全民热炒的时代，更需要科技工作者保持冷静的头脑。大数据是一个抽象的概念，除去数据量庞大这一特征，大数据还有一些其他的特征，这些特征决定

第一作者：廖建新，北京邮电大学网络与交换技术国家重点实验室网络智能研究中心主任、二级教授、博士生导师
作者单位：北京邮电大学网络与交换技术国家重点实验室，北京 100876
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目(No.2013CB329102)，国家自然科学基金资助项目（No.61471063，No.61372120, No.61271019, No.61101119, No.61121001），教育部科学技术研究重点（重大）项目(No.MCM20l30310)，北京高等学校青年英才计划基金资助项目(No,YETPO473)，高等学校博士学科点专项科研基金资助课题（NO.20100005 U0008)
注：本文转载自《电信科学》2015年第7期。

了大数据与“海量数据”和“非常大的数据”这些概念之间的不同。

高德纳分析员 Doug Laney 曾于 2001 年在一次演讲中指出，数据增长有 3 个方向的挑战：数量（volume），即数据多少；速度（velocity），即资料输人、输出的速度；种类（variety），即多样性；这 3 方面的特征即大数据最先提出的 3V 模型。2011 年，在国际数据公司（IDC）发布的报告中[1]，大数据被定义为：“大数据技术描述了新一代的技术和架构体系，通过高速采集、发现或分析，提取各种各样的大量数据的经济价值。”大数据的特点可以总结为 4 个 V，即 Volume（体量浩大）、Variety（模态繁多）、Velocity（生成快速）和 Value（价值巨大但密度很低）3 这种 4V 定义得到了更广泛的认同，这种定义指出了大数据最为核心的问题，就是如何从规模巨大、种类繁多、生成快速的数据集中挖掘价值。在那以后，业界对大数据的解读越来越全面，相继把大数据的基本特征扩展到了 5V、7V、甚至 11V 特征，扩充了 Veracity（真实性）、Validity（有效性）、Variability（易变性）、Viability（存活性）、Volatility（波动性）、Visibility（可见性）、Visualization（可视性）等新维度。

除了上述主流的定义，还有人使用 3S 或者 3I 描述大数据的特征。3S 指的是：大小（size）、速度（speed）和结构（structure）；3I 指的是：定义不明确的（ill-defined）、令人生畏的（intimidating）、即时的（immediate）。因此，为了保证大数据的可控性，需要缩短数据搜集到获得数据洞察之间的时间，使得大数据成为真正的即时大数据。就大数据究竟该如何定义，工业界和学术界还有不少讨论[2-4]。但是，大数据的关键并不在于如何定义或如何界定，而是如何提取数据的价值，如何利用数据，如何将“一堆数据”变为“大数据”。

随着数据为王的大数据时代的到来，产业界需求与关注点发生了重大转变：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关注大数据的处理能力。大数据处理的兴起也改变了云计算的发展方向，使其进入以分析即服务（analytics-as-a-service，AaaS）为主要标志的 Cloud2.0 时代。大数据还引起了科技界对科学研究方法论的重新审视，正在引发科学研究思维与方法的一场革命。Schnberger 对大数据引发的思维变革进行了总结。在大数据时代，通过挖掘和分析处理，大数据可以为人的决策带来参考答案，但是并不能取代人的思考。正是人的思维，促使众多利用大数据的应用出现，因此，大数据更像是人的大脑功能的延伸和扩展，而不是大脑的替代品。如果能有效地组织和使用大数据，将对社会经济和科学研究发展产生巨大的推动作用，同时也孕育着前所未有的机遇。

本文将在系统分析大数据研究关键技术的基础上，梳理国内外大数据现有的技术产品，总结分析企业大数据的策略和商用应用案例。最后，在分析大数据时代面临挑战的基础上，尝试展望大数据研究可能存在的机遇，提出正确应对大数据挑战的观点，以期对该领域的研究有一个系统而全面的认识，为后续大数据研究方向的发展提供参考。

一、大数据的关键技术

大数据技术，从本质上来说就是从类型各异、内容庞大的数据中快速获得有价值信息的技术。目前，随着大数据领域被广泛关注，大量新的技术已经开始涌现出来，而这些技术将成为或者已经成为大数据采集、存储、分析、表现的重要工具。

大数据处理的关键技术主要包括：数据采集、数据预处理（数据清理、数据集成、数据变换等）、海量数据存储、数据分析及挖掘、数据的呈现与应用（数据可视化、数据安全与隐私等）。图 1 展现了如何将大量的数据最终转化成为有价值信息的一般步骤，基本囊括了大数据领域的关键技术。正如图 1 所示，数据经过一系列的加工和处理，最终以

有价值的信息的形式到达用户手中。需要特别注意的是，在数据分析中，云技术与传统方法之间进行联合，使得一些传统的数据分析方法能够成功地运用到大数据的范畴中来。

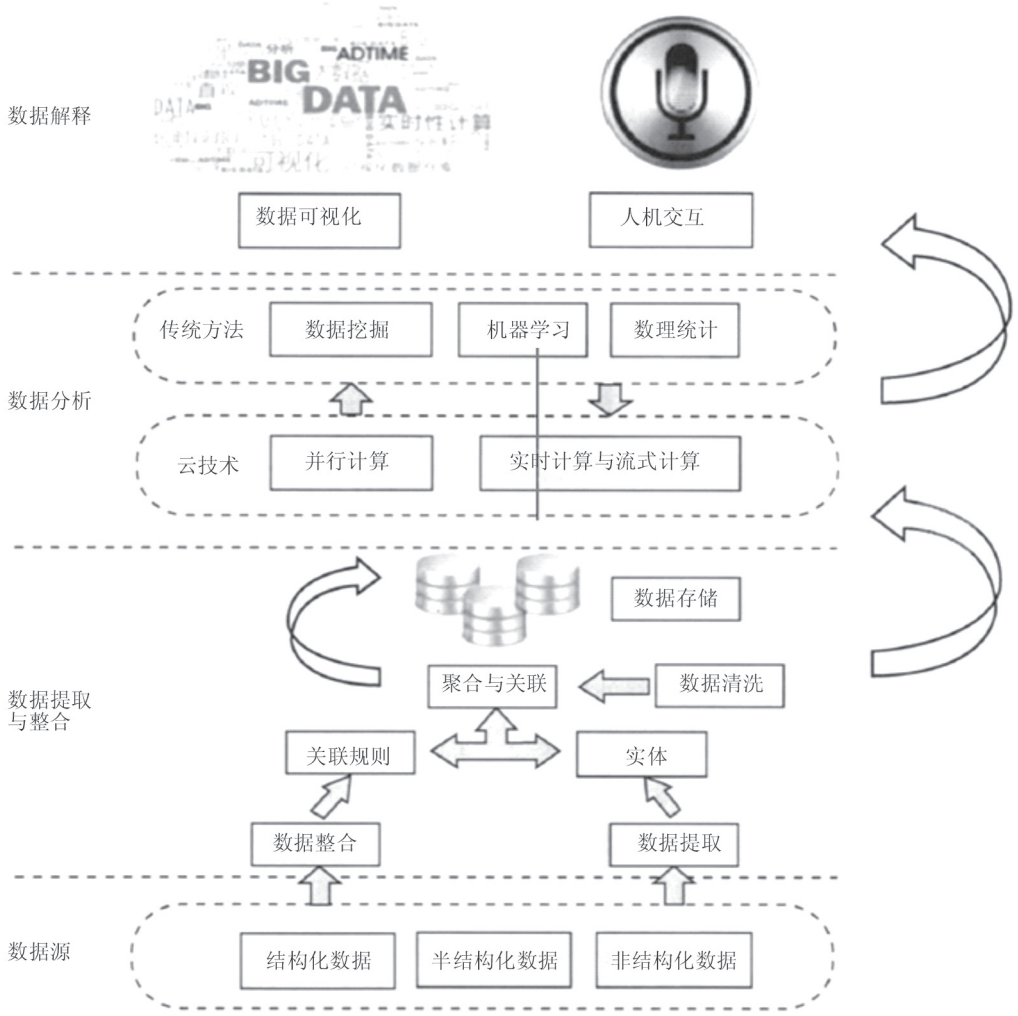


图 1　大数据领域关键技术

下面将对大数据领域中的核心技术进行简要的介绍与阐述。

（一）数据的采集、预处理与存储

大数据技术的第一步也是关键一步便是数据的采集。从传感器网络、社交媒体等数据源中获取结构化、半结构化和非结构化数据，并将数据主体进行预处理与存储是大数据环境下处理与分析数据的基础。大量数据在这一环节中被清理、集成和变换，成为“合格”的数据，从而被进一步分析与利用。

1. 数据的采集技术

在大数据这一范畴下，数据是指通过传感器网络、无线射频数据、社交网络数据及移动互联网数据等方式获得的结构化、半结构化 (或称为弱结构化) 及非结构化的海量数据，是大数据分析挖掘的根本。

大数据采集体系如图 2 所示。由图 2 可知 , 一般可将大数据采集体系分为智能感知层和基础支撑层。智能感知层主要包括数据传感网络、无线射频网络、智能识别网络 (二维码) 及资源接人系统，实现对非结构化、半结构化、结构化的海量数据的智能化识别、定位、接入、传输、监控、初步处理和管理等。而基础支撑层主要提供大数据服务平台所需的物理介质，如数据库资源、物理传输资源、物联网资源等。

图 2　大数据采集体系

2. 数据预处理技术

数据预处理技术主要包括数据清理、数据集成以及数据变换 [11]。数据清理可以去掉噪声数据以及异常数据，纠正数据中的不一致。而数据集成可以将来自不同数据源的数据合并成一致的数据存储，如数据仓库。数据变换可以改进涉及距离度量的挖掘算法的精度和有效性，将不同度

量下的数据归一化使得数据的应用比较有意义。这些数据预处理技术在数据分析之前使用，可以大大提高数据分析的质量，提高分析的速度与准确性。

3. 海量数据存储技术

在大数据的环境下，为保证高可用、高可靠和经济性，往往采用分布式存储的方式来存储数据，采用冗余存储的方式来保证存储数据的可靠性，即为同一份数据存储多个副本。海量存储的关键技术包括并行存储体系架构、高性能对象存储技术、并行 I/O 访问技术、海量存储系统高可用技术、嵌入式 64 bit 存储操作系统、数据保护与安全体系、绿色存储等。

广泛适用的分布式文件存储系统的设计思想不同于传统的文件系统，这类系统往往是针对大规模数据处理而特殊设计的[13-14]。它们虽然运行在一些非常廉价的普通硬件之上，但是却可以提供容错的功能，从而给用户提供总体上性能较高的服务。一个分布式集群一般由一个主服务器和大量的块服务器构成，许多用户可以同时访问[15]主服务器包含了所有的元数据，包括名字空间、访问控制信息、从文件到块的映射以及块的当前位置。主服务器还控制系统活动范围，定期通过心跳消息与每一个块服务器通信，并收集它们的状态信息。

（二）数据的分析与挖掘

为了实现抽取大数据中价值的信息，必须对经过预处理之后的数据进行分析与挖掘。目前针对海量数据的分析方法主要有并行计算、实时计算与流式计算。

1. 并行计算

并行计算是指同时使用多个计算资源完成运算。其基本思想是将问题进行分解，由若干个独立的处理器完成各自的任务，以达到协同

处理的目的[16]。在大数据时代下，串行的处理方式难以满足人们的需求，现在主要采用并行计算的方式。目前在大数据环境下所提出的并行计算往往是指任务级别的并行计算，而指令或进程级别的并行计算模型往往具有更强大的处理大规模数据的能力，只是现有的 GPU（graphics processing unit，图像处理器）编程模型还不够完善。大量挖掘算法开始针对并行架构进行调整，使其能够利用并行的优势，对大规模的数据进行更好的处理。LuoD 等提出的非平凡策略，可以将一系列的数据挖掘问题进行并行处理，其中包括支持向量机（support vectormachine，SVM）、非负最小二乘问题等。由此得到的算法，可以通过目前流行的并行处理结构进行实现，数据分析效果获得了显著的提升[17]。Gao F 等提出了一种新的近似算法，将基于核函数的数据挖掘算法的应用范围推广到了海量数据集上。该算法大幅度降低了计算核矩阵时的计算开销和内存使用，但却没有对结果的精确度产生很大的影响。此外，并行结构下该算法的实现也被提出，真实数据集上的测试结果显示，提出的算法可以大量节省时间和空间上的消耗 Shim K 在并行结构的框架下，讨论了如何将传统的数理统计方法和并行结构相结合，以便进行大数据分析[19]。

2. 实时计算与流式计算

传统的数据流处理[20-21]受到数据采集速度和内存容量等因素的限制，往往只能处理小规模的数据流。但随着数据采集技术和数据传输技术的进一步发展，使得短时间内积累大量的历史数据成为可能。与此同时，目前大数据环境下对数据流处理的要求不断提升，使得历史数据规模的增加也成为必然。

目前关于数据流的处理研究主要可以分为两类：集中式和分布式。在集中式环境下，数据流的计算受到存储资源，特别是内存容量的限制，

主要通过概要数据、准入控制和 QoS 降价等方法，以牺牲服务质量为代价最终实现伸缩性。而在分布式环境下，针对由多个算子组成的数据流处理网络，主要是通过平衡在多个节点上的算子分布来最终实现伸缩性。

实时计算同样为实时数据的挖掘奠定了基础，此时数据挖掘有了特定的对象——数据流。为了有效地处理数据流 , 需要建立新的数据结构，并将新的数据结构与传统的挖掘算法相结合 [22]。因为并不存在无限大的空间存储数据流，所以需要在正确性和存储空间之间进行平衡。常用的数据结构和挖掘技术有滑动窗口、多分辨率方法、梗概、随机算法等。

3. 深度学习技术

深度学习概念的出现，源自于复杂数据结构处理以及复杂特征提取任务之中遇到的与人工智能相关的问题，这些问题普遍需要对高阶抽象概念进行表述，具有非线性、语意性等特征。深度学习网络结构通常由多层非线性运算网络组成，每一层的输出作为下一层的输入，能够从海量数据中提取并学习到有效的复杂特征，进而用于数据的检索、分类、回归等问题之中 [24]。

深度学习概念起源于人工神经网络的研究，而人工神经网络结构在复杂高维数据处理问题上有许多不足，并且有着泛化能力差、训练速度慢等问题。2006 年 Hinton 等 [25] 提出的用于训练深信度网络（deep belief network，DBN）的无监督学习算法，将之前用于训练网络的全局优化算法拆分成若干个子任务逐次执行，在保证算法性能的前提下极大地提高了网络训练效率。在此之后，大量研究工作被引人，深度学习理论被不断丰富，将深度学习技术应用于信号处理、图像识别、语音处理等领域的尝试与研究也日益增加。

深度学习结构试图找到数据内部的结构特征，发觉数据间的真实关联规则。在处理实际任务的过程中，数据的表现形式、关系模式是多种

多样的；与之对应，深度学习结构也发展出多种结构，以应对不同场景下的数据处理需求。目前有卷积神经网络（convolutional neural network，CNN）[26]DBN 两种主流的深度学习结构。在自然语言处理和信息检索领域，已经有了大量的 DBN 应用案例 [27]。

（三）数据的隐私保护

在大数据时代，云端可以每时每刻对用户的信息进行采集，使每一个用户成为“透明人”，因此当前亟需针对大数据面临的用户隐私保护、数据内容可信验证、访问控制等安全挑战，提出相应的解决方案。

1. 数据发布匿名保护技术

对于大数据中的结构化数据（或称关系数据）而言，数据发布匿名保护是实现其隐私保护的核心关键技术与基本手段。以典型的 A 匿名方案为例，通过元组泛化、抑制等数据处理，将准标识符分组。每个分组中的准标识符相同且至少包含个元组，因而每个元组至少与 k-1 个其他元组不可区分。实现方法包括基于裁剪算法的方案以及基于数据置换的方案等。在大数据场景中，数据发布匿名保护问题较之更为复杂：攻击者可以从多种渠道获得数据，而不仅仅是同一发布源。

2. 社交网络匿名保护技术

由于社交网络具有图结构特征，其匿名保护技术与结构化数据有很大不同。

社交网络中的典型匿名保护需求为：用户标识匿名与属性匿名（又称点匿名），在数据发布时隐藏了用户的标识与属性信息；用户间关系匿名（又称边匿名），在数据发布时隐藏用户间的关系。目前的边匿名方案有基于边的随机增删方案、基于节点聚集的匿名方案、基于基因算法的实现方案、基于模拟退火算法的实现方案以及先填充再分割超级节点的方案等。

3. 数据溯源技术

数据溯源（data provenance）也被译成“数据世系”。其基本出发点是帮助人们确定数据仓库中各项数据的来源，例如了解它们是由哪些表中的哪些数据项运算而成的，据此可以方便地验算结果的正确性，或者以极小的代价进行数据更新。数据溯源的基本方法是标记法，如通过对数据进行标记来记录数据在数据仓库中的查询与传播历史。后来概念进一步被细化为 why 和 where 两类，分别侧重数据的计算方法和出处。除数据库以外，数据溯源还包括 XML 数据、流数据与不确定数据的溯源技术。数据溯源技术也可用于文件的溯源与恢复。例如，通过扩展 Linux 内核与文件系统，创建一个数据起源存储原型系统，可以自动搜集起源数据。未来数据溯源技术将在信息安全领域发挥重要作用。

（四）数据中心体系结构

网络数据中心作为大数据服务的天然载体，是实现大数据分析的基础设施。传统数据中心的拓扑采用树型分层结构，但性能不佳且可扩展性差。为此，Al-Fares 等提出了一种基于“胖树”的类树型结构拓扑，其主要目的是在网络端节点实现更高的聚合带宽。通过增加一定的布线复杂度来连接形成一个胖树型网络，端交换机用来连接服务器。在全负载最坏的情况下，这个结构仍然可以实现约 87% 的聚合带宽。此外，美国加州大学的 Guo 等人将并行计算的一些思想引人数据中心结构的设计中，提出了 DCell[28]。DCell 是一种递归构建的数据中心，它使用性能强大的商业级 PC 和低端的交换机，拓展性相对于节点的度具有双倍指数的增长关系。清华大学的李丹提出了一种使用双网卡 PC 机和低端交换机来搭建数据中心的方案 FiConn[29]。在 FiConn 这一拓扑中，链路被分为若干级别，通过一种自适应的机制来实现平衡负载和提高网络吞吐量的目的。

二、大数据技术应用现状

大数据已经渗透到每一个行业和业务领域，逐渐成为重要的生产因素，对大数据的分析处理正是业界的流行趋势。

（一）大数据处理平台

大数据时代产生了许多更为复杂的新数据类型，企业需要更加强大的数据处理平台来分析这些数据。国外的如 Google、Amazon、IBM、Microsoft 等企业，都已经意识到大数据对企业长远发展的重要意义，推出了多种有关大数据的应用平台，非常有名是 Spark 平台和 Oracle 大数据机等知名项目，为企业的产品规划、决策及战略发展提供了重要的数据支撑 [30]。

1. 大数据计算框架

目前，基于分布式处理技术的开源软件框架有 Apache Hadoop，Storm 和 Spark 等。Hadoop 能够利用简单的模型并行处理分布在各个集群计算机中的数据，只是难以满足大数据的实时处理和挖掘的需求。Storm 是基于拓扑的实时流处理计算逻辑，由 Zookeeper 连接 Nimbus 和 Supervisor，并协调分布式环境中的同步、命名等问题。Spark 是一种基于内存的流式实时数据处理平台，如图 3 所示。Spark 处理某个时间段窗口内的事件流，将用户定义的一系列 RDD 转化成 DAG 图，然后 DAG scheduler 把 DAG 再转化成 taskset（作业集），这样 taskset 就可以向集群申请计算资源，集群把这个 taskset 部署到 worker 中进行运算。开发者的任务是定义一些 RDD，并定义相应的转化动作，最后将这一系列的 RDD 投放到 Spark 的集群中运行。

比较典型的分布式流计算框架还有 IBM 的 InfoSphere Streams，它可以根据从多样的海量数据中提取出的自相关数据，实现对来自相关数据的信息和知识的积极分析和管理。Apache Drill 可以用于解决大数据

集的互动分析问题，可以处理分散在上千台服务器中的数据，它的主要优势在于可以在较小的时延内快速响应交互式查询。

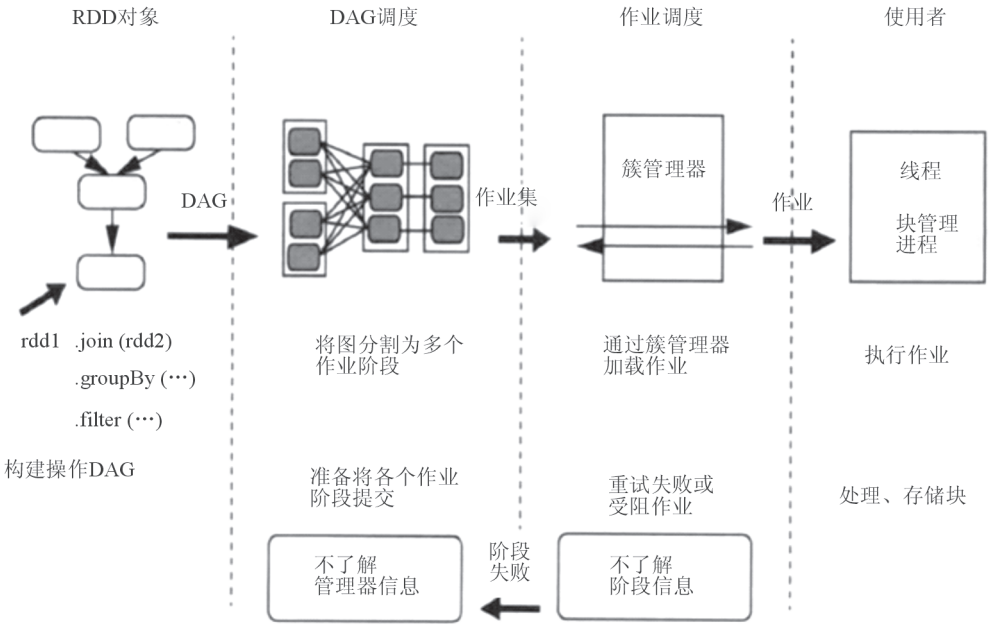


图 3　Spark 框架

2. 大数据存储框架

对于大数据应用，NoSQL 数据库被认为比 SQL 更适合处理和捕捉海量数据。NoSQL 数据库是一种分布式横向扩展技术，它使用了分布式节点集来提供高度弹性扩展功能，让用户可以通过添加节点来动态处理负载。MongoDB 是一个介于关系数据库和非关系数据库之间的产品，支持的数据结构非常松散，因此可以存储比较复杂的数据类型。Neo4j 是一个高性能的 NoSQL 图形数据库，它将结构化数据存储在网络上而不是表中，Neo4j 具有嵌入式、高性能、轻量级等优势，在处理复杂的网络数据时，具有很好的性能表现。Oracle NoSQL Database 是来自 Oracle 的键值 NoSQL 数据库，具备数据备份和分布式键值存储系统。设计时考虑到了高扩展性和高可用性，并可部署于多个互相复制的节点上，以便进行快速故障切换及负载均衡。Apache HBase 源于 Google

的 Bigtable，是一个分布式面向列存储的 NoSQL 数据库。在 Hadoop 和 HDFS 之上提供了像 Bigtable 一样的功能。HBase 快照添加了类似“过程协作”的新功能，用于在线快照、写时备份、恢复和克隆。

除此之外，Greenplum 的统一分析平台结合 Greenplum DB 和 Greenplum Hadoop 为企业构建高效处理结构化、半结构化、非结构化数据的大数据分析平台。GoodData 提供的是基于云的数据分析服务，向 SaaS 提供商提供技术集成服务，使得这些 SaaS 提供商可以向最终用户提供诸如仪表盘、报表等功能。

3. 大数据云平台

国内外的众多技术公司将大数据处理技术融人云平台[31]。Amazon EMR（elastic Map Reduce）是亚马逊（Amazon）提供的大数据分析云服务，使用 Hadoop 开源框架将数据分布在 Amazon EC2 实例集群中并进行处理，分析用户提交的海量数据分析作业。BigQuery 是谷歌（Google）基于 Dremel 和 Hadoop 集群所提供的交互式大数据分析云服务 BigQuery 允许企业用户上传大数据集到谷歌的存储器中分析数据，同时可以建立应用程序共享服务。国内的引跑科技推出的大数据平台 EngineOne，是大数据处理技术和云平台的结合。EngineOne 云操作系统提供了构建和部署云应用程序所需的全部工具和 API，能让用户在基础设施上弹性部署并运行应用程序。

华存数据、曙光、Oracle 等知名国内外公司纷纷推出了性能卓越的大数据一体机。软件方面应用相应的数据库系统进行高级分析，而硬件方面一般都是云服务器、集群等作为软件的支持。华为公司还推出了名为 FusionCube 的融合一体机，把计算、存储、网络以及虚拟化平台有机融合在统一的架构之下。

（二）大数据分析产品

面对如此庞大而复杂的数据，需要专门设计的硬件和软件工具进行专业处理。该类型的处理工具分为两种：数据可视化产品可将纷繁的数据呈现在用户面前，并对数据进行统计整理，方便用户了解当前数据分布情况；数据分析类产品则侧重于数据的分析和挖掘以及为企业用户提供比较完整的解决方案。

1. 可视化产品

这类软件可以将复杂的大规模数据可视化，允许用户通过操作界面方便地与数据进行互动，并对其执行进一步的高级分析，以便获得更深入的业务洞察，更好地帮助用户进行企业决策。借助这类软件，用户可以从原始信息出发，最终得到隐含在结构化与非结构化数据中的重要信息。

IBM 的 Cognos Insight 商业智能解决方案可以向业务线使用人员提供丰富多变的可视化界面。它通过所有的系统和资料资源，提供了无缝密合的报表、分析、记分卡、仪表盘等解决方案。Google 等诸多公司也推出了适用于其他行业的数据可视化工具。

2. 分析类产品

Google Analytics（谷歌分析）是 Google 公司推出的网站分析工具，用户把网站信息提交给 Google Analytics，Google Analytics 自动给出分析结果及丰富的图表式报告。Google Analytics 也可分析社交广告、移动广告、搜索广告的投放效果，使广告的作用极大地发挥。百度统计比 Google Analytics 更加本土化，更加适合中国用户。百度统计提供独立的 IP 地址数，分析报告的时效性更强，而且将百度指数和热门搜索词功能整合其中，把相关的信息串在一起，帮助分析流量原因。

对于多数企业来说，仅仅依靠网站分析是远远不能满足需求的，需

要功能更强的大数据分析工具来辅助企业决策。阿里云开放数据处理服务（open data processing service，ODPS），就是构建在大规模分布式计算系统上的，针对海量数据的处理工具，适用于海量数据统计、数据模型、数据挖掘、数据商业智能等诸多互联网应用。RadiumOne 通过分析 Facebook 中用户的社交网络行为，预测用户会对哪类产品感兴趣。而 Mintigo 推出的搜索引擎用来挖掘潜在用户，它首先对营销需求、目标行业以及用户人群进行分析。Connotate 的数据采集和分析服务适合企业使用，以指导他们更好地做出决策。Luminoso 利用 Connotate 在社交网站中发现消费者对新产品的喜好，以帮助他们做出更好的决策和提高。

（三）大数据的标准化

在大数据标准化方面，目前最为实质性的是国际标准化组织（International Organization Standardization，ISO）/ 国际电工委员会（International Electrotechnical Commission，IEC）第一联合技术委员会（Joint Technical Committee1，JTC1）成立了大数据研究组，由美国国家标准技术研究所（National Institute of Standards and Technology，NIST）牵头，NIST 系统地开展了大数据架构、数据、安全需求等方面的研究。由于大数据作为一项新兴技术，国内外相关标准的研制还处于起步阶段[32]，简要介绍如下。

1.ISO/IEC JTC1 SC32

持续致力于研制信息系统环境内及之间的数据管理和交换标准，为跨行业领域协调数据管理能力提供技术性支持。SC32 下设 4 个工作组和几个研究组：WG1 电子业务；WG2 元数据；WG3 数据库语言；WG4SQL 多媒体和应用包。JTC1/SC32 现有的标准制定和研究工作为大数据的发展提供了良好基础。

2.IS0/IEC JTC1 SG2

负责大数据国际标准化的大数据研究组，有利于统筹开展大数据的标准化工作。调研各个标准化组织在大数据领域的关键技术、参考模型以及用例等标准基础；确定大数据领域应用需要的术语与定义；评估分析当前大数据标准的具体需求，提出 ISO/IEC JTC1 大数据标准的优先顺序；向 2014 年 ISO/IEC JTC1 全会提交大数据建议的技术报告和其他研究成果。

3.NIST

力争形成达成共识的定义、术语、安全参考体系结构和技术路线图，提出数据分析技术应满足的互操作性、可移植性、可用性和扩展性需求，安全有效地支持大数据应用的技术基础设施。

4. 全国信息技术标准化技术委员会（以下简称“全国信标委”）

全国信标委 (TC28) 负责开展国内数据标准化工作，在元数据、数据库、数据建模、数据交换与管理等领域推动相关标准的研制与应用，为提升跨行业领域数据管理能力提供标准化支持。

（四）大数据产品应用

互联网行业大数据应用的起步较早，其数据类型丰富、范围广，与各垂直行业不断进行着深度的融合目前，互联网行业的大数据主要应用在社交网络、B2C 业务、精准营销、广告监测等方面，经典案例有百度、新浪微博、阿里巴巴等。

百度以传统的搜索引擎为基础，进一步推出了特色鲜明的百度大数据引擎。该引擎包括开放云、数据工厂和百度大脑 3 个部分。此前这些技术被应用在语音、图像和文本识别以及自然语言和语义理解方面，现在这些技术被用来对大数据进行智能化的处理、分析和学习，甚至组建了拥有 $2x10^n$ 个参数的深度神经网络 3 基于图片识别技术，百度还新推

出了拍照搜索功能，命名为百度识图。在使用手机拍摄照片后，搜索引擎能自动识别出用户拍摄的对象，并返回该对象的相关信息。对于搜索未知事物时经常遇到的词不达意的困惑，拍照的方式将更加直观、全面、易用。

新浪微博通过各种渠道聚集了海量数据信息，已沉积了将近 $5x10^n$ 个人与人之间的联系图、人与物之间的联系图谱。经过研讨这些数据可以更精准地为每个用户推送更准确的广告 [35]。“带着微博去旅行”就是新浪微博推出的一项微博旅游品牌活动，那么针对到达这些地区的旅行用户，可以及时推送当地的旅行信息、抢手景点以及电话费用等信息。移动端产品“Page 页面”是聚合了用户兴趣爱好社交关系数据的综合展示页面，连接人、物和兴趣，将微博的弱社交网络转换为真正的社会网络。

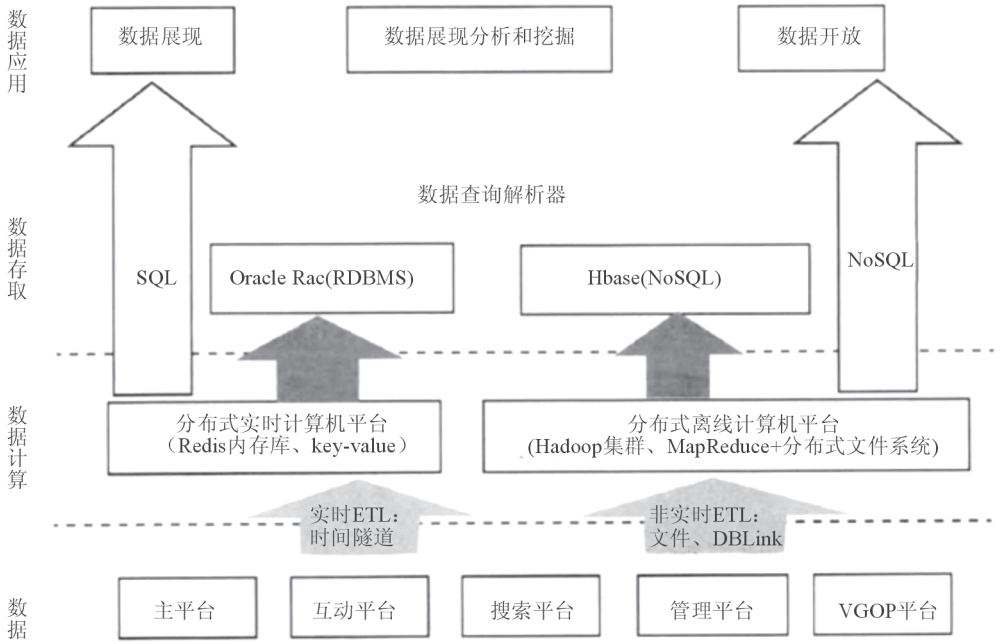
为了使淘宝商家能够实现数据化运营，阿里巴巴的数据魔方应运而生。根据用户的互联网痕迹进行渠道营销效果优化，从而调整营销资源在各个渠道的投放。例如，数据魔方可以根据不同时段的网站访问情况，分析顾客的需求和地理位置，实现按时按地点地精准地投放广告。还可以结合历史数据，给商家提供比较合理的价格参考。

（五）大数据行业应用

大数据在现今的公共管理、零售、互联网、电信、金融等众多行业快速渗透 [35]，影响力迅速扩大，已经成为重要的生产因素。目前各行业纷纷加快大数据的应用规模，重构未来的核心竞争力。对于很多行业而言，如何有效利用这些海量数据正成为赢得竞争的至关重要的因素之一。

对运营商来说，可利用数据与网络资源优势，挖掘数据的商业智能应用，为运营商创造更大的价值，这对于运营商转型升级具有重大的战略意义为了应对大数据时代的挑战，中国移动率先提出了大数据超细分微营销精服务的理念，将助力其全面提升精细化运营水平，改善用户体

验。借助大数据分析识别用户特征，细分用户，进而指导公司制定合适的营销策略。其中，手机阅读业务需要数据分析与挖掘支持的解决方案，为精细化运营提供科学依据。杭州东信北邮信息技术有限公司为此研究、设计和开发了一整套大数据系统框架，具体架构如图 4 所示。特别是在 ETL、分布式计算以及智能推荐引擎方面分别设计和开发了一套完整的产品，有效地完成了新用户的挖掘和各种用户群体的分析，并针对不同情况给出了详细的营销方案，实现了精细化的营销策略。



三、大数据存在的挑战与机遇

目前对大数据的探讨还处于起步阶段，一些有价值的研究开始尝试全局性地整理大数据生态环境之中各部分、各模块之间的关系大数据带来的描述与存储的挑战、分析与理解的挑战、挖掘与预测的挑战都有别于传统数据，除去开发新型数据库形式、提高数据挖掘算法性能、增强并行数据处理能力的研究与尝试，还需要对大数据有全局性的认识，通过分析大数据相关技术的发展情况与不足，找到大数据发展过程中的痛

点，以应对大数据发展中的机遇和挑战。结合大数据关键技术，从如下几个角度对大数据技术存在的机遇和挑战进行分析和预测。

（一）大数据的新型分析策略

大数据时代的数据分析需求，不仅体现在对海量数据的处理能力上，同时也体现在对新的知识类型的兴趣上。进人大数据时代以来，数据查找、筛选甚至回归、关联规则分析等传统数据分析目标已经渐渐无法令人们满足，从而有了新的数据分析的目的，这就必然促使有针对性地开发设计新的数据分析策略。数据是对现实事物的采样而非完整描述，这是大数据时代的一个崭新视角。在这个思想下，大数据处理就变成了大规模的常规数据处理，而除了依赖针对大数据的规模性数据处理算法之外，又多了依据数据采样进行数据分析的一条新路。以伯克利大学的BlinkDB^ 数据库系统为例，其运算速度如图 5 所示。在执行大规模数据查找的任务时，系统将数据库视为一个大样本集，逐次从数据库中以不同规模进行采样，使得通过采样结果得到的查询信息不断逼近实际结果，这一策略使得其在保证相同误差水平的条件下取得比 Hive 快 200 倍的查找速度。在社交网络应用方面，采样的策略同样有非常广泛的应用空间。往往不需要了解一个人的全部操作和发言，而只需要随机地从这个人的全部操作之中取规模很小的一个样本集，就可以得到与原有策略准确度不相上下的用户行为分析结果。

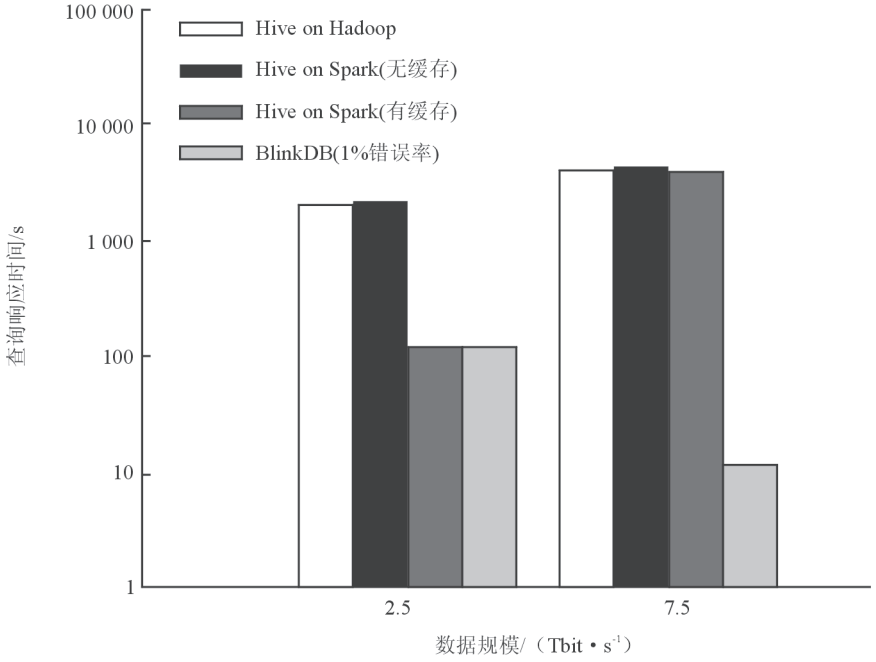


图 5　BlinkDB 的运算速度

（二）大数据的深度学习问题

在近几十年的研究中，研究人员普遍发现，基于人工神经网络结构的深度学习技术具有很好的学习能力、应用空间。然而目前深度学习领域的研究仍然存在一些有待解决的问题，这些问题又集中体现在如下 3 个方面。

1. 理论分析

深度学习技术的成型依赖于对网络训练方法理论的创新，而现有深度网络所存在的诸多问题，也需要理论方面的突破性进展才能得到真正的解决。深度模型之所以存在训练方面的困难，一个主要的原因在于深度网络使用了非线性神经元结构，而非线性系统虽然具有更好的拟合目标函数的能力，却较难应用于基于梯度的算法进行优化训练。如果能够找到更好的 RBM 求解策略，或是发现新的可以替代 RBM 的结构，都将对深度学习技术的研究与应用产生举足轻重的影响。

2. 数据特征提取

现阶段的一些结合 RBM 与 HMM、CRF 等传统特征提取算法的资

料还基本处于将二者机械叠加的水平：由于 DBN 结构本身就是多层简单结构的叠加，这种叠加虽然具有形式简明、训练策略明朗的优势，却也牺牲了很多扩展性与针对性。发掘深度结构的潜在优势，找到更好的方法，建立适合不同数据特点的深度学习结构，将成为下一阶段深度学习领域研究的重点。

3. 训练与优化求解技术

基于梯度的训练算法对随机初始化的深度神经网络结构收效较差，容易陷入局部最优而导致性能下降，这一现象引出了一系列半监督训练策略、回火策略等算法及变形，依然需要在训练算法、并行性优化等方面大幅度提升现有技术水平。

（三）大数据的隐私保护问题

大数据时代的隐私问题与以往常规数据时代的数据隐私问题相比，具有很多新的特征，其中较为主要的包括数据收集过程、数据分析过程中带来的隐私问题。

随着个人电子终端的传感器化、用户行为记录的细粒度化、信息收集的隐秘化等现象的出现，数据收集过程逐渐呈现出一种以往完全无法想象的状态，即由用户主动提供数据向数据收集者主动收集数据变化、用户知情向用户不知情变化、数据实际使用方式模糊化 [40]。用户往往在毫不知情的前提下，个人数据、操作记录等信息就已经被隐秘地收集了，而实际上用户有权力知晓这些数据是否被共享、误用、恶意传播、销毁。数据收集过程中的隐私保护问题有别于一般的数据安全问题，较难通过技术方式进行解决，主要依赖相关政策法规的约束。

现有的大数据隐私研究除去继承常规数据的隐私处理方法之外，还使用了诸如分级访问控制策略 [41] 使用添加数据机制防范背景知识分析攻击 [42] 等策略。基于“知情与同意”策略的位置隐私保护策略在大数

据时代已经不足以满足隐私保护的实际需求，基于启发式隐私度量、概率推测和隐私信息检索的大数据隐私保护机制的相关研究，仍有大量的工作需要进行。

（四）大数据的数据质量与可用性问题

在解决大数据存储、传输、处理的相关问题，建立数据中心基础架构、实现大数据标准化规范的过程中，数据质量与可用性始终是无法忽视的问题[43-44]。大数据之中蕴藏着巨大的潜在价值，要想确保能够从大数据中取得正确的信息，进而以之辅助决策、实现相关应用，关键的前提是保证数据质量。然而在实际情况中，能够取得的数据往往是有很多冗余、噪点、误差的质量不高的数据，这对大数据的知识发现、应用都有负面的影响。确保数据可用性同样是一项十分困难的任务，大数据在一致性、精确性、完整性、时效性、实体统一性这 5 个数据可用性维度上都存在着一定的可用性难题。

大数据可用性面临的主要挑战包括如下几个方面。

1. 探索从物理信息系统等多数据源有效地获取高质量大数据的理论和方法，研究高效数据过滤方法，建立多模态大数据融合计算的理论和算法，实现高质量数据的获取和精准整合，继而发现数据演变规律。

2. 大数据可用性面临很多数据恢复、定量评估、质量融合管理、数据演化方面的新问题，应当结合大数据处理的新方法、新技术，有针对性地对大数据可用性进行定性定量的分析，设计并建立大数据可用性的理论模型、形式化系统和评估算法。

3. 弱可用数据指的是无法彻底修复的数据，在大数据环境下，弱可用数据的规模较以往大幅度增加。除进一步研究数据错误自动检测与修复的理论和技术之外，对近似计算理论和技术的研究、对弱可用数据的知识校验和纠偏、对弱可用数据的知识演变机理方面的研究，也是解决弱可用数据问题的重要思路。

四、结束语

在大数据时代，数据的收集、获取和分析都更加快捷，这些海量的数据将对人们的思考方式产生深远的影响。纵观人类社会的发展史，人的需求及意愿始终是推动科技进步的源动力。随着物联网的兴起，移动感知技术的发展，数据采集技术的进步，人不仅是大数据的使用者和消费者，还是生产者和参与者。基于大数据的社会关系感知、众包、社交网络大数据分析等与人的活动密切相关的应用，大数据将在各行业产生重大的社会价值，也必将引起社会活动的巨大发展和变革。未来还需要进一步加强大数据标准化顶层设计，推动开放数据集建设，建立健全的大数据标准体系，重点突破一批涉及大数据发展的基础性、方法性、公共性标准的研制，为大数据发展和应用夯实基础。

参考文献

[1] Gantz J, Reinsel D. Extracting Value from Chaos. IDC iView Report, 2011

[2] Schnberger V M, Cukier K.大数据时代：生活、工作与思维的大变革.盛杨燕，周涛译，杭州：浙江人民出版社，2013

[3] Team 0 R. Big Data Now: Current Perspectives from O’Reilly Radar. Sebastopol: O’Reilly Media, 2011

[4] Grobelnik M, Big data tutorial, http://videolectures.net/ eswc2012grobelnik big data/[2016-03-10], 2012

[5] 张引，陈敏，廖小飞.大数据应用的现状与展望.计算机研究与发展，2013, 50(S2)

[6] Binzenhofer A，Tutschku K’ Graben B A D, ei a/. A P2P-based framework for distributed network management. Lecture Notes in Computer Science, 2006(3883): 198-210

[7] Tutschku K, Chevul S, Binzenhfer A, Schmid M, et al. A seJf-organizing concept for distributed end-to-end quality monitoring. University of Wurzburg Institute, Wurzburg, Germany, 2006

[8] 李强，王宏，王乐春.基于P2P的分布式网络管理模型研究. 计算机-丁程，2006, 32(13)： 150-152

[9] Karagiannis T, Papagiannaki K，Faloutsos M. Blinc: multilevel traffic classificaion in the dark. Proceedings of the 2005 Conference on Applications，Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, Pennsylvania, USA，2005

[10] Karagiannis T, Roido A, Aloutsos M, el al. Transport layer identiiication of P2P traffic. Proceedings

of the 2004 ACM SIGCOMM Internet Measurement Conference, Taormina, Italy, 2004

[11] 鲍静，范生万.基于数据挖掘的图书数据预处理.大学图书情报学刊，2008, 26(2): 31-33

[12] 刘云霞.数据归约的统计方法研究及应用（博士学位论文）. 厦门大学，2007

[13] 刘鹏.云计算技术基础.北京：电子工业出版社，2011

[14] 王庆波，金滢，何乐.虚拟化与云计算.北京：电子工业出版社，2010

[15] 王鹏.云计算的关键技术与应用实例.北京：人民邮电出版社，2010

[16] Agrawal D, Bernstein P, Bertino E, et al. Challenges and opportunities with big data. Challenges and Opportunities with Big Data-ResearcGate, 2012, 6(12): 2032-2033

[17] Luo D, Ding C, Huang H. Parallelization with multiplicative algorithms for big data mining. Proceedings of IEEE 12th International Conference on Data Mining, Brussels, Belgium, 2012： 489-498

[18] Gao F, Abd-Almageed W, Hefeeda M. Distributed approximate spectral clustering for large-scale datasets. Proceedings of the 21st International ACM Symposium on High-Performance Parallel and Distributed Computing, Delft, the Netherlands, 2012: 223-234

[19] Shim K. MapKeduce algorithms for big data analysis, and storage of big data. Proceedings of the VLDB Endowment, Istanbul, Turkey, 2012: 2016-2017

[20] Abadi D J, Ahmad Y, Balazinska M，et al. The design of the borealis stream processing engine. Proceedings of 2nd Biennial Conference on Innovative Data Systems Research， Asilomar, CA，USA, 2005

[21] Motwani R, Widom J, Arasu A, et al. Query processing, resource management, and approximation in a data stream management system. Proceedings of the 2003 CIDR Conference, Asilomar, CA, USA, 2003

[22] Han J, Kamber M, Pei J, et al. Data Mining: Concepts and Techniques- Translated by Fan M, Meng X F. Beijing: China Machine Press, 2005

[23] Al-Fhres M，Loukissas A, Vahdat A. A scalable, commodity data center network architecture. Proceedings of ACM SIGCOMM, Seattle, WA, USA, 2008

[24] Mitchell B, Sheppard J. Deep structure learning: beyond connectionist approaches. Proceedings of the 11th International Conference on Machine Learning and Applications (ICMLA), Boca Raton, Florida, USA, 2012: 162-167

[25] Hinton G，Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006，18(7): 1527-1554

[26] FuKusHIMA K. Neocognitron ： a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 1980，36(4): 193-202

[27] Ranzato M，Susskind J，Mnih V, et aL. On deep generative models with applications to recognition. Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 2011: 2857-2864

[28] Guo C, Wu H, Tan K, et aL. DCell: a scalable and fault-tolerant network structure for data center.

Proceedings of ACM SIGCOMM, Seattle, WA, USA, 2008

[29] Li D, Guo C, Wu H, et aL. FiConn: using backup port for server interconnection in data centers. Proceedings of IEEE INFOCOM, Rio de Janeiro, Brazil, 2009

[30] 胡雄伟，张宝林，李抵飞.大数据研究与应用综述.标准科学，2013(10):18-21

[31] 方巍，文学志，潘吴斌等.云计算：概念、技术及应用研究综述. 南京信息工程大学学报：自然科学版，2012，4(4): 351-361

[32] Chandarana P, Vijayalakshmi M. Big data analytics frameworks. Proceedings of 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), Mumbai, India, 2014: 430-434

[33] 中国电子技术标准化研究院.大数据标准化白皮书，2014 China Electronics Standardization Institute. Big Data Standardization White Paper,2014

[34] 宫夏屹，李伯虎，柴旭东等.大数据平台技术综述.系统仿真学报，2014，26(3): 489-496

[35] 黄哲学，陈小军，李俊杰等.面向服务的大数据分析平台解决方案.科技促进发展，2014,10(1)：52-59

[36] 袁玮.云计算在电信行业经营分析系统中对大数据的处理探析.硅谷，2014(6)

[37] Demchenko Y, De Laat C, Membrey P. Defining architecture components of the big data ecosystem. Proceedings of 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, USA, 2014: 104-112

[38] Zeng D, Lusch R. Big data analytics: perspective shifting fromtransactions to ecosystems. IEEE Intelligent Systems, 2013, 28(2):2-5

[39] Agarwal S, Mozafari B, Panda A, et al. BlinkDB: queries with bounded errors and bounded response times on very large data. Proceedings of the 8th ACM European Conference on Computer Systems, Prague, Czech Republic, 2013: 29-42

[40] 孟小峰.位置大数据隐私保护研究综述.软件学报，2014，25(4): 693-712

[41] Cheng Y, Park J, Sandhu R. Preserving user privacy from third-party applications in online social networks. Proceedings of the 22nd International Conference on World Wide Web Companion, New York, USA, 2013:723-728

[42] Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms. Proceedings of the 41st Annual ACM Symposium on Theory of Computing, Bethesda, Maryland, USA, 2009: 351-360

[43] 李默涵，李建中，高宏.数据时效性判定问题的求解算法.计算机学报，2012，35(11): 2348-2360

[44] 刘波，耿寅融.数据质量检测规则挖掘方法.模式识别与人工智能，2012，25(5): 835-844

（责任编辑：牛静雅）

研究　传播　交流　影响

Research Dissemination Communication Impact

上海市卫生和健康发展研究中心

Shanghai Health Development Research Center (SHDRC)

中国　上海

Shanghai China