· 专家面对面 ·　DOI：10.3969/j.issn.1672-9455.2018.22.001

大数据时代必知

廖　生[1]，应斌武[2]，关　明[3]，张　本[4]

（1.四川大学华西口腔医院/华西口腔医学院信息中心，成都 610041；2.四川大学华西医院实验医学科/四川大学华西临床医学院医学检验系，成都 610041；3.复旦大学附属华山医院中心实验室，上海 200040；4.中国人民解放军陆军军医大学第一附属医院流行病学中心，重庆 400038）

中图法分类号：R-05　　**文献标志码**：A　　**文章编号**：1672-9455(2018)22-3325-04

近年来，随着计算机及信息技术的飞速发展和行业普及应用，行业应用系统规模迅速增大，行业应用产生的数据呈爆炸式增长。西方发达国家已经从科技战略层面上提出一系列的大数据研究计划来推动大数据技术的探索、研究及应用。大数据带来巨大挑战的同时也带来了巨大的商业机遇，不断产生的海量大数据包含很多小数据不具备的潜在深度价值，数据分析挖掘能够对行业、企业带来巨大的商业价值，对实现各种高附加值的服务打下坚实的理论基础及创新应用方法，能够进一步提升行业、企业的经济效益及社会效益。

来源于各个不同领域的大数据尽管代表着不同的事物，隐含着不同的信息，但是都具备 4 个重要特征，简称 4V 特征，即大容量（volume）、快速更新（velocity）、多类型（variety）和高价值（value）。所谓大容量，虽然没有一个绝对的标准，但与传统数据量相对一般都在 TB 级别以上。Velocity 是指数据产生和更新的速度非常快。Variety 是指数据种类繁多，除了文字信息以外，还包括各种图像、图形、视频及声音等多媒体数据。Value 是指大数据中隐藏了极高价值的信息，这些信息需要通过机器学习与数据挖掘才能转化为人类所能理解的知识。

大数据在医疗行业的应用能够显著提高医疗效率和医疗效果，能够更加合理调配医疗资源。医院在建立大数据仓库及云计算服务的平台上，通过对现有模型的优化及开发新模型，能够使自身优势更加充分发挥，使各种资源进行统一协调，提高效率，有效减少宝贵的医疗资源浪费。在大数据的支撑下，一些新型数据分析模型的运用，如：Hadoop、机器学习（machine learning，ML）、自然语言处理（natural language processing，NLP）等将对医疗行业带来质的飞跃，从而引导医疗行业在诸多方面发生深刻变革。大数据已经逐渐深入人类的日常生活，医疗行业的从业者有必要掌握大数据的知识及方法，并加以合理运用，方能将其作用发挥出来，持续改进医疗的研究及实践，提升人类健康水平。

笔者在此邀请了另外 3 位专家从不同角度对大数据关键问题进行解读。

1　如何看待大数据（big data）时代到来

廖　生

廖生：技术日新月异的今天，人们看到 7 nm 芯片已经诞生，5G 通讯[1]触手可及，随着数据爆炸式增长，人类科学已经突破了试验归纳、模型推演、仿真模拟，随之演进到科学第四范式（paradigm），即数据密集型科学发现（data-intensive scientific discovery），也就是常说的“大数据”[2]。从根本上说，大数据使我们能够收集和分析我们正在产生的大量数据。大数据背后的理念是我们在互联网上做的每件事都留下了数

作者简介：廖生，医学博士，副教授，研究生导师，美国哈佛大学高级访问学者，博士后。长期从事口腔基础医学、临床医学及医疗信息化相关专业的教学、科研及管理工作，结合交叉学科优势在相关专业发表 SCI 论文 20 余篇，负责国家自然科学基金面上项目 2 项，青年基金 1 项，教育部博士点基金 1 项，科技部国家重点实验室开放基金 1 项，四川大学青年基金 1 项，作为主研参与多项国家及省部级基金研究及国际合作项目。现任四川大学华西口腔医院/医学院信息中心主任，中国卫生信息学会理事，国家卫健委信息中心专家，中华口腔医学会口腔信息管理专委会副主任委员，中华口腔医学会信息技术专委会理事，四川省卫生信息学会理事。

作者简介：应斌武，男，四川大学华西医院实验医学科主任/四川大学华西临床医学院医学检验系主任。

作者简介：关明，男，复旦大学附属华山医院中心实验室主任，华山医院检验医学科副主任。

作者简介：张本，男，中国人民解放军陆军军医大学第一附属医院流行病学中心主任。

字足迹。这种分析不断增长的数据量的能力正在彻底改变人们理解世界及其中的一切方式。有了这么多信息，大数据的整合在人们的日常生活中是不可避免的。大数据是如此之大，它会时刻影响我们的工作、锻炼和购物方式，而这仅仅是个开始。

应斌武

应斌武：大数据可能看不见，摸不着，却在潜移默化地改变人类的生产、生活。随着人们对数据的不断积累，对数据的分析，将无数杂乱的数据精确分类后使其服务于个体，大幅度提高了人们交流和工作效率，发挥着重大的商业效能。数据显示，1990－2000 年，《大英百科全书》(Encyclopedia Britannica)用户数量流失十分巨大。与之相反的是“维基百科”(Wikipedia)用户数量在 2000－2010 年迅猛增长。虽然《大英百科全书》与“维基百科”提供的服务相差无几，但维基百科对大数据进行分析后能够将内容精准地推送给用户，使得“维基百科”在与《大英百科全书》的竞争中占尽先机。同样，科学家试图通过对海量的基因信息进行分析解读人类生命所有的秘密。随着人类基因组计划(human genome project)的顺利完成，二代测序成本大幅度降低，并且在可以预见的将来，基因测序的费用还会有很大的下降空间，测序及分析技术的发明、发展及成本的快速降低，在人类历史上第一次给这一期望提供了可行的工具[3]。

关　明

关明：大数据不仅仅记录了人们的日常行踪，更包括了驾车习惯、阅读习惯、阅读爱好、购物喜好等。试想当人从一出生，每天的状态就以数字的形式记录下来，这就十分利于以后对他进行全方位的分析、教育及诊治。商业可以通过这些数据分析出他的喜好，当他进行购物时，销售就能更精确地向他推荐谷歌或者苹果的产品，商店也可以分析出他愿意购买电脑还是手机。十几年前苹果手机的出现使得人们的日常生活渐渐被数据化。到今天，智能手机的普及使得大数据的搜集成为现实。现在收集数据不仅仅是将用户的数据传输到服务器上，还能够及时通过大数据得出的结果反馈给用户。即使是一块小小的智能手表，能采集的人体数据也是海量的，并能与用户实时交互。大数据给人们带来的便利正在逐渐显现，大数据可以勾勒出人的全方位轮廓，甚至可以预测出未来会发生的事情。

张　本

张本：大数据应用是人类社会信息化发展的一个重要阶段，其应用给经济和社会生活带来了一系列变化。随着大数据与各行各业融合发展，智能化的综合网络将遍布社会各个角落，信息技术也将改变人们的学习方式、工作方式和娱乐方式。一大批新的就业形态和就业方式将被催生，商业交易方式、政府管理模式、社会管理结构也会发生深刻变化。同时，大数据也让人们对于未来充满无限期待。

2　如何看待结构化数据(structured data)和非结构化数据(unstructured data)的价值

廖生：一提到数据，大部分人会联想到具有行列格式的结构化数据，而非结构化数据则是数据结构不规则或者不完整，没有预定义的数据模型，不方便用二维逻辑来表现的数据，因此通常不适合主流关系数据库，通常包括报表、影像、扫描文档、Web 页面及多媒体音频和视频信息等诸多在医疗行业大量存在的数据格式。近年来，由于用于存储和管理此类数据的新平台的出现，它在信息系统中越来越普遍，并且能够被各种商业智能和分析应用程序组织使用。使用传统的分析技术难以发掘非结构化数据中的信息和知识，但非结构化数据却蕴藏着更大的宝藏。

应斌武：纵观全球，95％的信息都是非结构化的，就算是人们试图将非结构化数据结构化，结构化的数据也不会超过 10％。非结构化数据几乎是结构化数据的 10 倍之多[4]。过去的世界信息是不对等的，但是现在，世界各处所能获取的信息都是平等的。当你

传递一个信息时，会同时产生大量的非结构化数据。此前，人们难以开发一个能准确翻译出一篇文章的软件，而谷歌在整合了大量的非结构化数据后，能够根据语言、语境等数据准确地将文章翻译成不同的语言。Facebook 也将众多无组织的信息整合到一起，使人们不用再为处理众多杂乱的信息而抓狂。现在的手机软件，他们不需要时时刻刻监控你的位置，只需要收集你在什么地点的使用次数多就可以为你制订特别的服务。在大数据时代，廉价的非结构化数据对于人们来说是提升生产力的巨大机会。

关明：在医疗行业，约 80％的医疗数据是自由文本构成的非结构化数据，其中不仅包括大量的描述性文本，也包括包含非统一文字的表格字段。必须通过医学自然语言处理技术，将非结构化医疗数据转化为适合计算机分析的结构化数据是医疗大数据分析的基础，对医疗非结构化数据的深究必将极大促进医疗行业的发展。

张本：随着越来越多的非结构化沟通渠道和来源（电子邮件、社交等），人类的沟通方式发生了巨大变化。需要处理所有这些非结构化通信以及由此产生的非结构化数据和信息。各类机构需要知道如何将这些信息收集在一起并根据其中蕴含的知识进行发展。此外，在这些非结构化数据中隐藏着大量其他的重要信息。人工智能消除了对规则的需求，只有人工智能可以处理这些数据，因为非结构化数据根本无法通过传统方法进行处理。人工智能利用其“智能”超越信息本身，能够实现信息自动处置，学习人类行为，还可以进行预测。

3　大数据如何在精准医疗（precision medicine）领域发挥作用

廖生：25％的美国人都会服用名叫舒喘灵（albuterol）的药物。但是在服用这种药物时，传统医疗并没有考虑到患者的年龄、生活环境、呼吸空气的情况、家里是否有宠物。不同患者都以同样的方式服用药物，使我们并不能知道这是否对缓解患者的病情及时、有效。同样，在治疗糖尿病、癌症和其他疾病时也会面临同样的情况。传统医疗的逻辑是依靠患者的症状、化验结果做出诊疗决策。精准医疗是医生利用患者全方位各种维度的健康及诊疗信息为患者制订独一无二的治疗方案，并且因为基因在人个体之间所具有的相似性，为一个患者所制订的精准治疗方案也可以运用于成千上万符合相同数据模型的患者。通过对患者基因的分析，可以发掘癌症的发生、发展过程，并制造出新的药物，一个很好的例子就是格列卫（gleevec）[5]。美国哥伦比亚大学的科学家也运用大数据，对美国食品药品监督管理局的药物反应数据，同时结合了一组来自 30 万患者、160 万的心电图数据，发现了 8 组药物混合服用可能与长期性的心律异常相关[6]。其中最显著的罗氏芬（rocephi）和兰索拉唑（prevacid），分别单独服用并未发现风险，但是当这两种药物同时服用时，服用者会出现心律异常，严重时甚至会导致服用者死亡。科学家们希望通过数据发掘和大数据分析，尽可能全面研究不同药物组合可能带来的用药不良反应。十年前，人们对自己的所有 DNA 序列进行测序需要 1 亿美元，五年前这个费用需要 10 000 美元，而现在只需要 1 500 美元。随着人们对自己身体健康的重视程度不断地增强，对医疗健康的投入在不断地增加，而精准医疗的作用就是提高有限医疗资源的运转效率，提高医疗质量，降低个体及社会医疗成本。

应斌武：精准医疗依靠对患者众多数据的收集、联通、分析。通过对患者的数据分析、样本化验、身体检查、进食情况的了解、常用药的调查、基因分析，制订精确的治疗计划，将每个患者作为特殊的个体对待，能够使患者在最短的时间获得最显著的治疗效果。作为朝阳产业的精准医疗，目前处于产业生长的萌芽期，其发展壮大需要经历基础科学技术创新、科技成果转化，以及医学临床应用，完成产业裂变，从而逐步构建起融合研发、生产、治疗的精准医疗的完整产业链。

关明：在医学史中，疾病预防和治疗都是基于“标准化患者”的预期结果。传统做法是将同类疾病的患者的数据汇集在一起用于统计分析，汇总其结果形成临床指南并告知数十亿患者的健康和疾病管理。这种方法在特定的历史时期，在一定程度上促进了医学发展，但它忽略了重要的个体差异，这可能导致预后不确定性的发生。精准医学旨在为个体患者量身定制临床治疗计划，目标是在正确的时间向正确的患者提供正确的诊疗。组学技术的最新进展为临床医生提供了更全面的患者信息。测序和相关数据存储成本的降低，以及有效数据分析方法的开发，使得大规模生物医学数据分析研究成为前所未有的可能。这些进步可以提高复杂疾病诊断的准确性，患者可以从靶向治疗中受益，使疾病的诊治关口提前。尽管如

此,大数据在精准医疗中的应用仍然存在许多挑战。用于数据存储,数据库管理和计算分析的传统方法不足以满足每年生成的数 PB(1 PB 等于 1 000 TB,等于 1 000 000 GB)的生物医学数据。此外,随着数据集变得更大和更多样化,需要先进的分布式文件存储和计算方法来处理数据。

张本:肿瘤的高度复杂性意味着使用大数据的方法与用于某些其他类型疾病的方法明显不同。每位肿瘤患者可以有数千个维度参数,但只有少数相似的患者能够被汇总研究,因为肿瘤都是独一无二的。目前,科学家主要在使用这些数据在 3 个层面上分析研究肿瘤:(1)细胞层面,寻找个体肿瘤细胞数据的模式,以发现相关遗传标志物。寻找共同特征可以帮助我们更好地预测个体肿瘤如何发生、发展,以及哪种药物治疗可能最有效;(2)患者层面,患者的病史和 DNA 数据可用于根据肿瘤、基因以及治疗对类似疾病模式和遗传学患者的影响,辅助确定最佳治疗组合;(3)人口层面,可以分析更广泛的人口数据,根据患者不同的生活方式、地理位置和癌症类型为其提供治疗策略。这些不同方法的最终目标是让肿瘤学家能够为每位患者提供特定癌细胞的定制药物进行治疗,并限制严重不良反应的风险。

4 大数据应用中需要特别注意的问题

廖生:采用所有数据而不是样本数据,能够大幅度提高大数据解决问题的能力。大数据已经能够在一定程度上辅助医生利用全方位的患者健康及疾病数据做决策,在经验的基础上产生新的发现,新的治疗计划,这样对于病情的诊断、治疗及预后判断会更加准确和及时,且数据的可用性带来了对未知的健康问题相关因素的关注。大数据不能替代抽样调查。目前的大数据技术远不能达到“普查”,即不能简单地将目前的大数据做“总体”来使用。绝大多数的大数据与传统的经过科学实验统计得到的真实数据是不一样的。诸多情况下,经过简单计算的大数据并不如经过精心设计、复杂计算的实验而取得的小数据可靠,尤其对于那些概念相对复杂、涉及面较为广泛的不易聚类的命题,比如在某个时间段很多人搜索“流行感冒”,可能只是上映了一场关于流行感冒的电影,不一定是流感突然爆发。

应斌武:由于数据量大幅增加,不同行业系统根据不同时代需求所建设的系统的采集手段、数据标准多种多样,在每个时间平面上难免会出现一些不准确的数据混入大数据库,且因数据的流动性难免需要相互融合、交叉变异,形成新的数据结构,充斥噪声的大数据难以满足精准应用要求,数据的质量亟待提高。

关明:由于目前大数据理论过于依靠数据的大量汇集,那么一旦数据本身出现问题,在只问有什么,不问为什么的模式下,就很可能出现颇具危险性的或者不合时宜的大数据,即因为数据本身携带的问题,而导致不该出现的错误预测和决策。况且,医学处于随时变化的环境,只看其果,不究其因,着实危险。

张本:大数据时代,隐私的泄露极易发生,且不同于一般行业的数据,医疗数据具有其特殊的敏感性和重要性。医疗数据的来源和范围具有多样性的特征,包括病历信息、医疗保险信息、健康日志、基因遗传、医学实验、科研数据等。个人的医疗数据关系到个人的隐私保护,医疗实验数据、科研数据不仅关系到数据主体的隐私、行业的发展,甚至关系到国家安全。

参考文献

[1] RUSSELL C L. 5 G wireless telecommunications expansion:Public health and environmental implications[J]. Environ Res,2018(165):484-495.

[2] HEY T. The Fourth Paradigm - Data-Intensive Scientific Discovery[M]// E-Science and Information Management. Springer Berlin Heidelberg,2012:1.

[3] ANON. International consortium completes human genome project[J]. Pharmacogenomics,2003,4(3):241.

[4] GANDOMI A,HAIDER M. Beyond the hype:Big data concepts,methods,and analytics[J]. International Journal of Information Management,2015,35(2):137-144.

[5] ANON. FDA approves Novartis′ Gleevec[J]. Expert Rev Anticancer Ther,2001,1(1):3.

[6] LORBERBAUM T,SAMPSON K J,CHANG J B,et al. Coupling data mining and laboratory experiments to discover drug interactions causing QT prolongation[J]. J Am Coll Cardiol,2016,68(16):1756-1764.

(收稿日期:2018-06-27 修回日期:2018-08-28)