科技大数据

BIG DATA

因你而改变

Big Data of Science and Technology: the Driving Force to Make Great Change

戴国强 赵志耘 / 主编

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

科技大数据

BIG DATA 因你而改变

Big Data of Science and Technology: the Driving Force to Make Great Change

戴国强　赵志耘 / 主编

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS
·北京·

图书在版编目（CIP）数据

科技大数据：因你而改变 / 戴国强，赵志耘主编. —北京：科学技术文献出版社，2018.8

ISBN 978-7-5189-4725-6

Ⅰ.①科… Ⅱ.①戴… ②赵… Ⅲ.①科学技术—数据处理—研究 Ⅳ.① G203

中国版本图书馆 CIP 数据核字（2018）第 174970 号

科技大数据：因你而改变

策划编辑：李 蕊 责任编辑：李 晴 张 红 杨瑞萍 责任校对：文 浩 责任出版：张志平

出 版 者 科学技术文献出版社

地 址 北京市复兴路15号 邮编 100038

编 务 部 （010）58882938，58882087（传真）

发 行 部 （010）58882868，58882870（传真）

邮 购 部 （010）58882873

官方网址 www.stdp.com.cn

发 行 者 科学技术文献出版社发行 全国各地新华书店经销

印 刷 者 北京时尚印佳彩色印刷有限公司

版 次 2018 年 8 月第 1 版 2018 年 8 月第 1 次印刷

开 本 710×1000 1/16

字 数 362千

印 张 28.5

书 号 ISBN 978-7-5189-4725-6

定 价 128.00元

《科技大数据：因你而改变》

编写组

主　　编　戴国强　赵志耘

编写人员　戴国强　赵志耘　袁　伟　张英杰　吴　思
常　春　刘　伟　张闪闪　杨代庆　王立学
雷　雪　陈　亮　贾　佳　张均胜　杨　岩
张兆锋　石崇德　何彦青　望俊成　王　政
张玄玄　张爱霞　赵　辉　邢晓昭　许　燕
汪芸辉　张　静　宋培彦　刘　蔚　张志娟
徐　峰　王　勇　李维波　傅俊英　郑　佳
雷孝平　张海超　曹　燕　付鑫金

序　言

钱塘江潮般的大数据浪潮汹涌来袭，紧随其后的却是天际边渐入眼帘的一幅泼墨山水，精彩的画面也徐徐铺展开来。

大数据，因为万事万物皆可量化而精彩。

当我们踏入21世纪，扑面而来的社会变革、生产力的急速发展、新技术的迅速产业化，都不及大数据年均50%的增长速度那么突出。在信息化1.0时代，微电子技术的快速发展实现了模拟声音和图像的数字化；而在信息化2.0时代，方兴未艾的新一代信息技术革命则正在加速撬动全球数字化、数据化进程。越来越多原先无法测量的东西能被度量，原先无法跟踪的行为可以被跟踪，原先无法记录的活动被数字化记录。互联网的搜索流水线正源源不断地生产着巨量数据，而物联网又悄然将量化的触角进一步延展到可以追索的极限时间和广袤空间，万物皆可量化的理想之光正在逐步照进现实。

大数据，因为万众徜徉数据海洋而精彩。

受交通大数据分享极大地便利和丰富了大众出行的启发，兼有数据生产者和数据使用者双重身份的普罗大众，对开放共享大数据之意义的认识更加深刻，对开发利用大数据之价值的需求更加迫切，对享受进而陶醉于大数据之便利的实践更加频繁。在这种较为宽松适宜的社会氛围下，数据资源就像深埋于地底、彼此隔绝的万千暗河，快速涌流出地面聚成数据之河，汇成数据海洋，又在旺盛需求的光热作用下升华成诸如政府数据云、消费数据云、医疗数据云、科学数据云、资源数据云等类型多样、伸手可及的共享数据彩云，数据资源的这一华丽转身为全社会

谋事创业提供了丰沛甘霖，孕育了无限可能。数据共享应用又源源不断创造出的新数据资源，并不断注入到数据海洋，进一步实现从 Share（享有）向 Partake（分享）【Part（参与）+Take(获取)】的转变，最终打造出一个可持续的数据循环生态系统。

大数据，因为催生创新雨后春笋而精彩。

由于数字化、网络化、智能化、泛在化和可视化等软硬件技术的持续进步，大数据资源类型不断拓展、存储中心分布更加灵活、数据传输能力越来越大，越来越快，结构化和非结构化数据加快融合统一，开发利用数据的资源和技术约束不断缩小，让创新者将更多精力和智慧投入多维度、宽视角、广领域的数据分析利用和价值创造中。创新越来越多地与大数据紧紧捆绑在一起，全社会研究大数据、挖掘大数据、存储大数据、利用大数据、服务大数据已成蔚然之势，并在科学研究、技术开发、市场应用、产业发展等创新全链条上蓬勃开展。基于大数据的创新范式让创新思维变得更加宽阔，让创新舞台变得空前宽广，让创新的道路变得更加多元。

大数据，因为颠覆传统生活模式而精彩。

曾几何时，随时找到便捷的出行工具，随手携带手机买遍天下，随心所欲地与远方亲朋好友“面对面”聊天，随地买到去往天涯海角的火车票，还是国人心中遥不可及的梦想。如今，这些梦想不仅成真，而且还在不断地颠覆着我们吃、穿、行、游、购、娱等生活的传统模式。在它们鲜亮的外表背后，大数据与大数据技术发挥着默默无闻却又居功至伟的作用。同时，企业、个人思维模式也在发生着巨大改变，消费行为、日常行为等从不可预测而变成大概率可预测，信息、产品获取变得更具个人定制化特征，因果思维更多地被相关性思维所取代。因为大数据传统生活模式、思维模式正在被快速颠覆，我们对世界的认知也得以不断延伸。

大数据的精彩还远未充分呈现，其精彩程度甚至远远超出我们目前

的想象，尚隐藏在我们未掌握、未探寻到的所在。问题的关键在于是不是需要我们继续深入挖掘？这还得从当下的世界发展困境说起。在全球化深入发展的今天，相当多的领域都已处于充分竞争状态，传统发展模式下的生产力潜力几乎已发挥殆尽，未来发展之路仍“浮云遮望眼”。尽管“创新者生，守旧者亡”“唯改革者进，唯创新者强，唯改革创新者胜”的道理早已为世人所认同，但谁来驱动创新呢？面对新一轮科技革命和产业变革加快涌起的浪潮，看似“条条大路通罗马”，其实因“九曲回肠”而难以“望尽天涯路”！未来潜于史，知古而鉴今。回顾历次工业革命的历程不难发现，每次生产力的巨大跃升，都是由于适应和满足了在当时即将到来的巨大市场需求变化，无论是以珍妮纺织机为代表的标准化制造和机械化、以福特汽车工厂为代表的流水线规模化制造和电气化，还是以工业控制和全球化分工生产为代表的模块化制造和信息化，莫不如此。

未来已走来，唯变为不变。当前，第四次工业革命晨曦乍现，以人工智能、量子信息、移动通信、物联网、大数据与云计算为代表的新一代信息经济，以基因编辑、生物合成、精准医疗、再生医学为代表的生物经济，以 3D 打印、自主控制系统、机器人、碳纳米管、石墨烯为代表的未来制造经济纷至沓来，新技术、新产品、新服务令人眼花缭乱，新产业、新模式和新业态更是层出不穷，不断冲击着我们对发展的认知。面对战略性新技术引擎“群雄并起”的局面，哪些“新动能”真正能挑起第四次工业革命的大梁呢？若能准确把握未来社会的主流市场需求，问题就能迎刃而解。

未来之动力，大数据为钥。当前，我们的社会正阔步迈向未来。其中，发展动力向更多靠内在动力转型，生产商品向定制化、服务化、绿色化升级，生产模式向网络化、智能化跨越，生产组织向扁平化、虚拟化、融合化推进，创新模式向以大数据驱动为核心的第四范式更新，发展空间加速向虚拟疆域和现实疆域并举转变。这些转变中蕴藏着的精致需求都直接或间接与以大数据为中心的新一代信息技术有着密切关系，

为此，习近平总书记将大数据谓之信息化发展的新阶段，大数据是把握并开启未来社会发展的一把钥匙。

相对于变化，找到“不变”更为关键，大数据则使其成为可能。第四次工业革命仍将依靠技术跃迁，人仍将作为支撑“核心”，而支撑引导智能化生产的是丰富多样、可共享和利用的“大数据”，它将成为第四次工业革命的标志。把大数据的潜在能力有效转化为产业需求的数据，赋能于所有参与生产活动、生产组织的新型生产力参与者，我们才有可能在全球新一轮变革与发展的竞争中继续高歌猛奏。

大数据时代，世界既变得更加透明，也越来越神秘。如果数据本身有误或不够全面，即使经过再严谨的逻辑推演、再繁复的模型计算、再高深的知识挖掘，也很难得出一个正确的结论，陷入“Garbage in, Garbage out”的循环。由于大数据的强大赋能，任何决策失误都将带来指数型风险。

以生物医药领域的热点——脑疾病治疗为例。随着人类寿命的延长，越来越多的脑疾病严重影响了人们的生活质量并给他们的家庭带来极大负担。尽管生命科学家和医学专家们通过有氧训练或者认知训练刺激神经元的新生来预防阿尔茨海默病，通过药物来刺激神经元生成从而改善抑郁症状，通过新生神经元修补神经损伤……但长期的努力却并没有取得预期的效果。

问题也因此而产生，以上治疗是否都与新生神经元有关？没有取得进展的科学问题能否归纳为“成人大脑是否存在新生神经元”？

关于这一问题的研究，2018 年两家著名刊物——《细胞—干细胞》第四期杂志发表的博尔德里尼团队研究结论与 3 月《自然》在线发表刊载的索雷尔斯团队论文观点却截然相反。索雷尔斯团队研究发现胎儿和婴儿的大脑中有大量神经元母细胞和新生神经元，但这些细胞的数量自一岁开始急剧减少，至 13 岁仍能观察到新生神经元，而所有成年个体的脑标本却未发现一例新生神经元。博尔德里尼团队的研究则发现，海马

区域的神经元母细胞数量随年龄的增长越来越少，由这些母细胞分化出的神经元却并没有随年龄增长而减少。在研究样本中，即使是老年人，在死亡时海马区域仍然有数以千计的新生神经元。

尽管两个研究团队都以人脑作为研究对象，都是对不同成熟阶段的神经元进行免疫荧光染色并用 DCX 和 PSA-NCAM 标记新生神经元，然而得到的结论却完全相反！

是博尔德里尼团队标记新生神经元时发生污染？还是索雷尔斯团队由于使用传统化学手段保存样本，技术上落后于博尔德里尼团队的大脑急速冰冻法？是 DCX 的表达并不恒定，对处理时间过于敏感？还是某个团队没有对海马区域进行完整检查？如果以上皆非，还有何原因导致两项研究结果迥异？其重要判别点又是什么？

如果以上治疗与新生神经元无关的话，那又是什么影响着治疗的有效性？已有的研究数据积累量显然无法回答这一问题，犹如盲人摸象，由于每个人只摸到了一小部分，导致对这个问题还难以达到认知上的统一。进一步设想一下，如果索雷尔斯团队的研究结果反映了真实情况，即新生神经元在成年人中几乎不存在，那之前的治疗研究、巨量投入、学科建设、人才培养等，是不是都有点南辕北辙？

以上讨论表明，数据样本量不足已成为制约科学研究、技术开发和创新的重大瓶颈因素，从思维、哲学和方法论层面思考科学技术发展的基本问题无可避免。事实上，科学家和工程师们执着探寻的因果关系与相关关系的研究对象，早已由原来的人类世界与物理世界的两极相互作用，扩大到人类世界、物理世界与网络世界的三极交互作用，影响经济和社会发展的因素显著增加。在这种情况下，如果不能首先探寻清楚基本问题，就无法看清事物发展脉络并总结其发展规律，更不要奢谈开展预测和科学决策。而大数据深度分析无疑为研究基本问题并提供解决方案等后续工作提供了一个“金刚钻”和“导航灯”，例如，通过更全面的分析而摸清上述不同研究产生矛盾结果的原因所在，并据此提出新的

研究思路。

如果说过去20年是DNA时代，也就是基于数据（Data）、网络（Net）和自动化（Automatic）的组合，通过发挥“基因”的遗传效应实现全球的高速发展，那么未来将升级到RNA时代，不仅继续发挥遗传效应，而且还将在适者生存和不确定性的前提下复制优秀品质并引导良性变异，推动经济社会继续保持协调、高速、有序发展。这里的RNA实质上概括了大数据时代的三大变化特征。

大数据“赋能”数据资源（Resource）。在大数据时代，数据快速从成果化向资源化、资产化、资本化三位一体转变，即通过深入挖掘使用大数据的潜在价值，实现大数据的保值、增值、赋值能力，即为数据“赋能”。

大数据向正新信息科技转变（Next Generation of Information Science）。在摩尔定律达到极限的今天（晶体管价格在连续50年下降后逆向上涨，若非市场操作，则可判断“摩尔定律”极限将至），传统信息技术所支撑的巨量存储和计算模式将被迫提升到科学存储和智能计算的新篇章，并据此将会源源不断地产生新的科学发现、技术创造、产品创新。

大数据与人工智能（AI）紧密结合。今天人类社会面对的挑战是必须能够完成传统机器和人类自身不愿做、不敢做和做不到的事情。资源和环境的约束要求我们必须在冗余数据和极小数据两种极限情况下也能做出正确处理，这意味着我们必须适应未来的极简需求，呈现于外部的极简，蕴藏于极其复杂的处理与计算，这绝非人类智慧的核心优势所在。多维度关联数据、不断优化的系统算法及不断升级的计算机软硬件将帮助我们更加深入地认识世界、改造世界。本书力求通过知识组织和数据挖掘等工具方法将纷繁的数据进行全面梳理和组织加工，为科学寻踪和发现提供条件；通过结构和非结构化数据存储的科学方法，在大数据的丛林中依靠可靠的路径和定位知识关联；通过机器智能帮助用户按语义提取和处理数据；利用大数据开发出最新的翻译软件，使我们获取

最新的信息更加便利；利用机器智能解决信息过载问题，实现“人找信息”到“信息找人”的巨大转变；利用大数据公共服务平台及后台技术支撑手段，提供从政府服务、产业服务、企业服务到个人服务的完整服务谱系。数据处理、数据流动、关联与分析、结果动态展示……我们希望借助这些科学的方法和手段展现大数据的潜在精彩。

在过去40年里，中国科学技术信息研究所围绕大数据持续耕耘不辍，本书内容就是长期努力的重要呈现，是对全所大数据研究实践工作的阶段性总结。写作本书的更重要目的是力图增进读者对大数据技术的发展和应用有更全面的了解，对依托大数据开展研究和创新创业的重要性有更深的理解。由于大数据所涉及领域众多，技术本身仍在迅猛发展，我们的研究也在不断深化中，不当之处敬请读者指正。

在第四次工业“革命”中人是最为关键的要素，未来跨越式发展要依靠大数据，更要依靠人类智慧！让我们回归“人定胜天”（人自身安定胜过老天的帮助）的本意，认识并按照规律踏实勤勉工作而不要奢求老天的帮助。大数据的精彩需要我们共同去了解、理解、发现、实现。

因为，有你、有我，大数据才会真的精彩！

戴国强

2018年9月28日

于中国科学技术信息研究所

目　录