

北京市科学技术协会科普创作出版资金资助

从技术到应用

数据之道

邬贺铨
主编



中国科学技术出版社
CHINA SCIENCE AND TECHNOLOGY PRESS

数据之道

从技术到应用

邬贺铨 主编

中国科学技术出版社
· 北 京 ·

图书在版编目 (CIP) 数据

数据之道：从技术到应用 / 邬贺铨主编. —北京：
中国科学技术出版社，2019.8

ISBN 978-7-5046-8284-0

I. ①数… II. ①邬… III. ①数据处理—研究 IV.
① TP274

中国版本图书馆 CIP 数据核字 (2019) 第 085116 号

| | |
|------|---------|
| 策划编辑 | 郑洪炜 李 洁 |
| 责任编辑 | 李 洁 |
| 装帧设计 | 中文天地 |
| 责任校对 | 焦 宁 |
| 责任印制 | 马宇晨 |

| | |
|------|---|
| 出 版 | 中国科学技术出版社 |
| 发 行 | 中国科学技术出版社有限公司发行部 |
| 地 址 | 北京市海淀区中关村南大街 16 号 |
| 邮 编 | 100081 |
| 发行电话 | 010-62173865 |
| 传 真 | 010-62173081 |
| 投稿电话 | 010-63581070 |
| 网 址 | http://www.cspbooks.com.cn |

| | |
|-----|-----------------------------------|
| 开 本 | 880mm × 1230mm 1/32 |
| 字 数 | 155 千字 |
| 印 张 | 8.75 |
| 印 数 | 1—5000 册 |
| 版 次 | 2019 年 8 月第 1 版 |
| 印 次 | 2019 年 8 月第 1 次印刷 |
| 印 刷 | 北京利丰雅高长城印刷有限公司 |
| 书 号 | ISBN 978-7-5046-8284-0 / TP · 413 |
| 定 价 | 80.00 元 |

(凡购买本社图书，如有缺页、倒页、脱页者，本社发行部负责调换)

此为试读，需要完整PDF请访问：www.ertongbook.com

本书编委会

主 编 邬贺铨
主 任 郑伟民 刘 鹏
成 员 张 燕 梁 南 武郑浩

目录

第一章 初识大数据 / 001

什么是大数据 / 003

大数据的源头活水 / 013

大数据的六大特征 / 019

从物联网到人工智能 / 022

第二章 大数据的“道”与“术” / 031

雏形：科学研究“第四范式” / 033

破壳：麦肯锡预言“大数据”到来 / 038

兴起：云计算为大数据创造技术

前提 / 042

前沿：从野蛮生长到博弈时代 / 045

从“数据”到“智慧” / 049

“博阅”与“深思” / 055

赋大数据以大智慧 / 057

第三章 大数据处理系统 / 061

大数据处理基本概念 / 063

大数据处理过程 / 066

大数据处理模式 / 070

典型的大数据处理系统 / 080

大数据存储 / 091

第四章 大数据管理 / 099

数据管理的发展历程 / 101

大数据管理方法 / 104

数据集成 / 110

大数据隐私管理 / 113

第五章 大数据分析 / 119

数据分析流程 / 121

关键的数据分析方法 / 125

大数据可视化 / 130

第六章 大数据赋能智慧政务 / 145

智慧政务由大数据开启 / 147

如何开展智慧政务建设 / 149

智慧政务应用方向 / 153

大数据与政府治理 / 158

未来发展路径 / 169

第七章 大数据促进经济增长 / 173

精准招商 定向引资 / 175

产业转型 智能制造 / 180

数据引流 服务升级 / 184

风险控制 智慧金融 / 189

精准扶贫 个性定制 / 196

第八章 大数据服务民生福祉 / 203

大数据守护下的社会治安 / 205

融入人工智能的智慧交通 / 214

网格化监测的大数据环保 / 217

大数据与人工智能支撑下的智慧
医疗 / 224

大数据与教育 / 228

第九章 看未来，夺先机 / 237

中国大数据崛起之路 / 239

大数据开辟国家治理新路径 / 242

突破大数据发展瓶颈的方向 / 243

大数据与 AI 将融为一体 / 247

万物互联迈进潘多拉星球时代 / 249

推动大数据技术产业创新发展 / 253

大数据人才培养 / 257

参考文献 / 259

第一章

初识大数据

如果评选当今最为火热的十大技术，大数据无疑是强有力的“种子选手”。大数据已经渗透到各行各业，无人驾驶、智能交通、智慧医疗等新兴技术在一定意义上都依托于“大数据”这一信息资产，越来越多的领域将处于这场“数据风暴”之中。虽然人人都在议论大数据，但是大多数人对其“来龙去脉”却知之甚少。

什么是大数据

大数据不仅是一门技术，同时也成为一种商业模式，并正在构建一种新的生态，慢慢地改变着我们的生活。对于大数据的概念众说纷纭，很多人认为“大数据”的内涵其实就在于“数据大”。为此，明确大数据的概念和特点，成了当务之急。

大数据的定义

对于大数据，全球权威的IT研究与顾问咨询公司高德纳（Gartner）曾这样描述，“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

麦肯锡全球研究所（MGI）在《大数据：创新、竞争和生产力的下一个前沿》中给出的定义则是：大数据是指其大小超出了典型数据库软件的采集、储存、管理和分析等能力的数据集。自此，大数据从经济角度得到了全球的广泛关注。

我国信息学者涂子沛先生将大数据等同于传统的小数据（源于测量）+ 现代的大记录（源于记录）。其中，记录来源于图片、音频、视频等，而随着记录的逐渐增多，大数据也越来越“大”。

2015年8月31日，国务院发布《促进大数据发展行动纲要》指出：“大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。”

由中国信息通信研究院编写的《大数据白皮书（2016年）》提出：“大数据是新资源、新技术和新理念的混合体。从资源视角看，大数据是新资源，体现了一种全新的资源观；从技术视角看，大数据代表了新一代数据管理与分析技术；从理念视角看，大数据打开了一种全新的思维

角度。”

至今，学界对“大数据”一词仍难以给出精确衡量的技术概念定义。既然目前对大数据并没有统一的释义，我们不妨将其理解为一种资源、一种工具、一种思考和认识世界的理念。

用大数据说话

在生活中，“扑面而来”的大数据，令人不知所措。大数据作为一种技术、工具、方法，对现代社会生活的影响和冲击日益凸显，在某些领域甚至是革命性与颠覆式的。纵观人类科技发展史，似乎没有哪一次科技革命像大数据这样，从酝酿萌动到蔓延爆发，仅仅经历短短数年的时间。

用数据说话。在大数据时代，“万物皆数”，注重“量化一切”。人类生活在一个海量、动态、多样的数据世界中，数据无处不在、无时不有、无人不用，数据就像阳光、空气、水一样常见，好比放大镜、望远镜、显微镜那般重要。相较于人们过去的“凭经验办事”，如今我们必须学会“用数据说话”。

让数据发声。在包罗万象的数据海洋中，经过“打捞”不难发现，在数据中藏着诸多熠熠生辉的珍宝，此前看似毫不相关的多件事物，通过“挑拣”，也能发现事物

之间隐藏的相互关系，在帮助人们认知事物和把握局势的同时，进一步预测未来，这正是大数据的潜力与价值所在。

在探索数据价值的过程中，我们已经不仅局限于寻求问题背后的因果关系，而是将范围进一步扩大，对普遍联系的各种事物进行一一审视。在这个过程中，相关关系成为探索的重点，“是什么”在一定程度上比“为什么”更重要。正如著名大数据专家维克托·迈尔-舍恩伯格（Viktor Mayer-Schönberger）所言，“要相关，不要因果”成为大数据时代的一个显著特征。

对于相关关系，生活中的很多实例都可以帮助理解。比如，在商业场景中，通过分析可知，大多数顾客在购买牛排的同时，也会顺便购买一些胡椒粉，因为两者是餐桌上的常见搭配，而商场通过将售卖两种物品的货架摆放在一起，并提供搭配销售的优惠券，将大幅提高销售收益。

对于牛排与胡椒粉，大家比较容易将它们联想在一起，但是啤酒和尿布这两种似乎毫无关联的物品，却常同时出现在美国沃尔玛超市顾客的购物篮里。

经过调查分析才知道，在有婴儿的美国家庭中，母亲通常在家照顾孩子，而由父亲去购买尿布。在购买过程中，父亲常常在买尿布的同时，也为自己购买几瓶啤酒，久而

久之就总是出现啤酒与尿布同篮的场景。因此，发现这一规律的沃尔玛超市将啤酒与尿布摆放在相邻货架上，以提高销售收入。

除了能用于商业场景，大数据还可以提前预测流感疫情。一般在流感肆虐前，在网上搜索相关生病症状的人会增加。为此，2008年谷歌推出了“谷歌流感趋势(GFT)”，GFT根据汇总的谷歌搜索数据，近乎实时地对全球当前的流行疫情进行估测。2009年，谷歌又通过疫情预测，准确预测了H₁N₁在美国的传播，这就是相关关系的巨大力量。

“用数据说话”“让数据发声”，已成为人类认知世界的一种全新方法。世界是物质的，物质是数据的，数据正在重新定义世界的物质本源，并赋予“实事求是”新的时代内涵。我们必须善于用数据说话、用数据决策、用数据管理、用数据生活。

大数据作为一种新兴的生产要素、企业资本、社会财富，可谓取之不尽、用之不竭，而且能够重复使用、循环利用。可以说，大数据是一个信息和知识的富矿，蕴藏着无限的商机与巨大的收益，只要去深度分析和挖掘，总会有意想不到的收获。谷歌、亚马逊、Facebook、阿里巴

巴、腾讯、京东等领军企业的成功实践和辉煌业绩，就是最生动、最有力的例证。

“得数据者得天下”，除了商机与收益，大数据同时也是“未来的石油”，将成为社会创新发展的动力源泉。大数据正在推动科学研究范式、产业发展模式、社会组织形式、国家治理方式的转型与变革。“数据可以治国，还可以强国”，大数据在中国大有可为。中国是一个人口大国、制造业大国、互联网大国，这些都是最活跃的数据产生主体。根据国际数据资讯公司（International Data Corporation, IDC）预计，数字宇宙规模将在2020年达到40ZB，而中国将产生占全球21%的数据。令人可喜的是，我国已就大数据做出战略部署，制定了发展规划和行动纲要，我们可以和发达国家在同一起跑线上赛跑，并可能实现弯道超越。

借用维克托·迈尔-舍恩伯格、肯尼恩·库克耶的警示：对于大数据时代，如果你是一个人，你拒绝的话，可能失去生命；如果是一个国家的话，可能会失去这个国家的未来、失去一代人的未来。

中国发展大数据的战略

基于大数据日益成为生产资料与价值资产，加快大数

据部署，深化大数据应用，已经成为抢占“数据革命”先机的国家大事。正如习近平总书记强调的“机会稍纵即逝，抓住了就是机遇，抓不住就是挑战”。为了抓住这一革命契机，我国早已进行了政策规划，将大数据发展上升为国家战略。

2012年7月，国务院印发《“十二五”国家战略性新兴产业发展规划》，明确提出支持海量数据存储、处理技术的研发和产业化。2013年1月，工信部、国家发改委、国土资源部等五部委联合发布《关于数据中心建设布局的指导意见》，对未来中国数据中心发展指明了方向，对数据中心建设布局提供了保障措施。

2015年8月，国务院印发《促进大数据发展行动纲要》，明确提出建设数据强国。2016年3月，《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》明确实施国家大数据战略。2017年，工信部印发《大数据产业发展规划（2016—2020年）》，最终将发展目标定位为：“到2020年，技术先进、应用繁荣、保障有力的大数据产业体系基本形成”，以此构建我国大数据产业的顶层架构设计框架。

目前，在技术优势方面，我国已经具备发展大数据的

技术与产业基础，在全球十大互联网企业中，中国占据四席，特别是在智慧物流、移动支付等垂直应用领域，即便是大数据核心技术比较领先的美国，也逊色于中国。

同时，由于起步较晚，中国大数据发展的局限也不容忽视，与发达国家相比，中国在新型计算平台、分布式计算架构、大数据处理、分析和呈现等相关核心技术方面与国外相比仍存在差距。

国外发展大数据的战略

从2008年《自然》杂志将“大数据”一词带入大众视野，直至其在各行各业发挥重要作用，不过10年时间。在这10年里，综览世界各国，大多处于大数据发展的初级阶段。即便是美国、日本、欧盟这些发达国家和地区，在大数据这一新兴技术的发展中也几乎处于同一起点。

在这次大数据浪潮中，美国是较早做出反应的国家。2009年，美国为了加快实现公共部门的开放共享，倾力建设了Data.gov这一门户网站，整合了包括财政、金融、医疗、科教、交通、能源等在内的50个部门的数据，并通过OGPL平台的建设和完善，制定数据交换、用户交流、数据资源管理等制度，进一步完善平台功能。

2012年3月，美国政府宣布实施《大数据研究与发