数据科学与大数据技术系列

大数据导论

——大数据思维与创新应用

何　明◎等 编著

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

数据科学与大数据技术系列

大数据导论

——大数据思维与创新应用

何　明　何红悦　禹明刚
周　波　牛彦杰　余永佳　编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

当前，大数据思维作为一种前瞻性的思维模式，在政府决策、商业规划和科学研究等领域正发挥着重大作用。大数据已成为重要的战略性资源，受到政府部门、各行业企业及研究机构的重视和关注。本书主要研究大数据思维，探索创新应用，从大数据时代、大数据战略、大数据思维、大数据产业、大数据技术、各行业大数据应用，以及大数据未来发展趋势等多维度、多层次、多领域全面展开诠释。为方便读者使用，本书配备了电子课件，读者可登录华信教育资源网 www.hxedu.com.cn 免费下载。

本书适合大数据爱好者、大数据从业者和政府机关相关人员阅读，也可作为相关行业和学术领域研究者的参考书，以及大学相关课程教材。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

图书在版编目(CIP)数据

大数据导论：大数据思维与创新应用 / 何明等编著. — 北京：电子工业出版社，2020.1

ISBN 978-7-121-35941-5

Ⅰ. ①大… Ⅱ. ①何… Ⅲ. ①数据处理－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2019)第 011663 号

策划编辑：秦淑灵　杜　军

责任编辑：苏颖杰

印　　刷：三河市华成印务有限公司

装　　订：三河市华成印务有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱　　邮编：100036

开　　本：720×1000　1/16　　印张：14　　字数：355 千字

版　　次：2020 年 1 月第 1 版

印　　次：2020 年 1 月第 1 次印刷

定　　价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：qinshl@phei.com.cn。

此为试读，需要完整PDF请访问：www.ertongbook.com

序

《大数据导论——大数据思维与创新应用》是中国人民解放军陆军工程大学青年学者何明教授及其科研团队的又一佳作，也可看作3年前出版的《互联网+思维与创新》一书的姊妹篇。正如作者所言："大数据好比价值密度低的'贫矿'，大数据应用好比'沙海淘金''大海捞针'，其间充满了不确定性和偶然性。"因此，大数据思维的基本出发点是"变废为宝"，从海量的、看似无用的数据中发现潜在的利用价值。与传统的小数据相比，大数据来源广泛、获取容易，但对其进行挖掘利用要困难得多。在信息社会中，数据被视为与物质、能量同等重要的社会资源，大数据是一种稀释的资源，不同数据均弥足珍贵，只是在价值的显现程度上有差异。因此，我们不能对大数据视而不见或毫不可惜地丢弃海量数据。大数据思维有助于拓展我们对数据价值的认识，更重要的是启示我们要善于发现大数据、关注大数据、管好大数据。

该书多次强调，与传统的数据分析相比，数据挖掘得到的是关联关系而不是因果关系。许多看似毫不相关的事实，其背后隐藏着千丝万缕的联系。从哲学意义上讲，大数据分析是用宏观整体思维替代抽样统计思维，是用有偏差的数据分析替代精确的数值计算，是用定量的计算思维替代定性的理性思维。用相关性改变人们长期以来对因果关系的偏爱，是认识论的一次深刻转型。通过大数据可获得万物间相互联系的特殊规律，这些规律有一定预见能力，丰富了人们的知识，但大数据的不足之处是缺乏演绎能力，人们只能知其然而不知其所以然。经过实践的检验，这些规律或许被认为是客观规律，或许需要二次解读和理性分析。总之，数据挖掘已成为科学研究的第四范式，是对试验观察、理论推导、模拟仿真等方法的补充。但我们不能满足于关联规律的发现，只有揭示了数据内在的因果关系，才能更深入地理解和科学地运用这些客观规律。

该书专辟一章论述大数据技术。大数据技术本身不是一门学科，而是一种方法，它与云计算、机器学习等新技术密切相关。面对海量异构、动态变

化、质量低劣的数据，传统的数据处理方法难以为继，而新的处理分析技术还不够成熟。与国外相比，我们在大数据技术方面还有一定差距，但也有相对优势，比如有广泛的大数据资源，网民的数量位居世界之首，有的省市成立了“大数据发展局”“大数据管理局”，许多智慧城市建设将大数据应用作为亮点……我们有理由相信，在技术、产业的相互促进下，我国的大数据应用必将后来居上。

该书虽冠名“大数据”，但在介绍典型产业的创新应用时，也包含了小数据的运用。平心而论，两种数据之间并无严格的界限，况且在发展数字化、信息化的道路上，小数据的共享、挖掘、安全等问题还没有得到很好的解决，大数据又提出了新的挑战。为此，不少学者呼吁，在数据资源利用上，不能抓“大”放“小”、盲目跟风，对大数据的创新应用期望值不宜过高，更不能减少对小数据应用的研究。

该书内容深入浅出，并配有大量的应用案例，可作为规划、管理人员理解大数据的入门指南，也可作为大数据教学、科研人员的参考资料。随着我国信息化建设的深入和普及，我们相信将会有新的素材、新的案例不断补充进来，使该书内容更加翔实。在此，谨祝愿我国大数据应用之树枝繁叶茂，祝愿我国大数据产业发展日新月异。

中国工程院院士　戴浩

2019 年 11 月

前　言

随着人工智能、5G 及区块链技术的发展，大数据进入了深度发展时期，在政府服务、工业生产、科学研究等领域得到了空前应用，已成为事关国家经济社会发展的战略性资源。我国对运用大数据加强社会各领域建设极其重视，“一带一路”“京津冀协同发展”“军民融合”等战略与大数据紧密相关，各级政府也陆续成立了大数据管理机构。党的十九大报告指出，要推动互联网、大数据、人工智能和实体经济深度融合；《2019 年国务院政府工作报告》中指出，要深化大数据、人工智能等研发应用。因此，大数据对社会各行各业的支撑作用和影响会继续加强。

从哲学层面看，大数据思维是一种全新的思维模式。传统的自然思维模式诞生于依赖小数据和精确性的时代，看重精确性和因果关系，是信息缺乏的产物。大数据思维模式主要侧重考虑数据的整体性和相关性，一开始会与人类直觉相矛盾，但接受数据的不精确和不完美，反而使人类能够更好地预测未来和理解世界，帮助人类进一步接近事实的真相。

本书立足于当前大数据在各行各业的发展现状，根据理论创新与实践应用相结合的原则，较全面地介绍了大数据时代、战略、思维、产业和技术，并结合当前国家省市机构体制改革背景，选取市场监管、综合交通、农业农村、政务服务、公共安全、医疗健康等行业的创新应用，阐述了如何运用大数据更好地履行政府职能和提升企业效益。本书内容包括 10 章：第 1 章拥抱大数据时代；第 2 章概览大数据战略；第 3 章从哲学、运营、理政、创新等角度剖析大数据思维；第 4 章跟踪大数据产业进展；第 5 章介绍大数据技术；第 6 章至第 9 章分别分析市场监管大数据、综合交通大数据、农业农村大数据及其他行业大数据应用案例；第 10 章展望大数据的未来。

全书内容经过多次讨论和修改才得以定稿，力求能够系统梳理国内外大数据相关成果，创新大数据思维，并做到逻辑严谨、文字顺畅、深入浅出，以期为大数据从业人员、研究人员和政府决策人员提供借鉴和启发。尽管本

书编写时投入了大量的资源和精力，但书中仍难免存在错误和疏漏之处，敬请广大读者批评指正。

感谢江苏省社会公共安全应急管控与指挥工程技术研究中心、江苏省社会公共安全科技协同创新中心和江苏省应急处置工程研究中心为本书编写提供案例支持。本书的出版得到国家重点研发计划 2018YFC0806900，国家自然科学基金(青年)71901217，中国博士后科学基金资助项目 2018M633757，江苏省重点研发计划 BE2015728、BE2016904、BE2017616、BE2018754、BE2019762 等项目的支持。

感谢李功淼、张玉恒、肖毅、徐兵、张乔、王文、刘叶芳、仇功达、杨壹、许元云、张斌、顾凌枫、杨铖和刘祖均等人为本书所做的工作。特别感谢我的博士后导师戴浩院士，他以严谨的学术态度认真审阅了书稿，并对书稿提出了细致且有针对性的修改意见，使本书增色不少。

何　明

2019 年 11 月

目　录

第 1 章　大数据时代——日新月异 …… 1
1.1　大数据的崛起 …… 1
1.1.1　数据大爆炸 …… 1
1.1.2　洞悉大数据 …… 2
1.1.3　小数据与大数据 …… 6
1.2　大数据的成长 …… 8
1.2.1　互联网技术推动了大数据的泛在化 …… 8
1.2.2　存储技术支撑了大数据的大容量化 …… 8
1.2.3　计算能力加速了大数据的实时化 …… 8
1.3　挑战与机遇 …… 9
1.3.1　数据的挑战与机遇 …… 9
1.3.2　技术的挑战与机遇 …… 10
1.3.3　用户的挑战与机遇 …… 10
第 2 章　大数据战略——高瞻远瞩 …… 12
2.1　国外战略 …… 12
2.1.1　美国 …… 13
2.1.2　欧盟 …… 14
2.2　国内战略 …… 15
2.2.1　历史机遇 …… 15
2.2.2　发展规划 …… 17
2.2.3　战略蓝图 …… 19
2.3　大事记 …… 21
2.3.1　学术界大事记 …… 21
2.3.2　产业界大事记 …… 23

第 3 章　大数据思维——革故鼎新 …… 25
3.1　哲学思维 …… 25
3.1.1　总体思维 …… 26
3.1.2　相关思维 …… 27
3.1.3　容错思维 …… 27
3.2　运营思维 …… 28
3.2.1　数据收集思维 …… 28
3.2.2　数据管理思维 …… 29
3.2.3　数据应用思维 …… 31
3.2.4　数据价值思维 …… 33
3.2.5　数据事实思维 …… 36
3.3　理政思维 …… 37
3.3.1　高效决策思维 …… 37
3.3.2　阳光理政思维 …… 37
3.3.3　数据赋能思维 …… 38
3.4　创新思维 …… 39
3.4.1　跨界思维 …… 39
3.4.2　智能思维 …… 40
3.4.3　赋能思维 …… 40
第 4 章　大数据产业——风生水起 …… 42
4.1　大数据产业概述 …… 42
4.1.1　发展阶段及市场规模 …… 43
4.1.2　产业链与商业模式 …… 47
4.1.3　产业应用领域 …… 53
4.2　国外大数据产业 …… 54
4.2.1　美国 …… 55
4.2.2　日本 …… 55
4.2.3　欧盟 …… 56
4.3　国内大数据产业 …… 57

4.3.1 产业现状 …… 57
4.3.2 存在问题 …… 60
4.3.3 努力方向 …… 62
4.4 实体经济+大数据 …… 63
第 5 章 大数据技术——神兵利器 …… 65
5.1 大数据技术概述 …… 65
5.2 大数据处理框架 …… 67
5.2.1 Hadoop …… 67
5.2.2 Storm …… 68
5.2.3 Spark …… 70
5.3 数据采集与清洗 …… 70
5.3.1 数据采集 …… 71
5.3.2 数据清洗 …… 73
5.4 数据存储与管理 …… 75
5.4.1 分布式文件系统 …… 75
5.4.2 NoSQL …… 76
5.4.3 多维索引技术 …… 78
5.5 数据挖掘与分析 …… 79
5.5.1 数据挖掘的过程 …… 80
5.5.2 新型数据挖掘技术 …… 82
5.5.3 相似性连接融合技术 …… 84
5.5.4 面向领域的预测分析技术 …… 85
5.5.5 深度学习技术 …… 90
5.6 数据可视化 …… 91
5.6.1 文本可视化 …… 91
5.6.2 网络可视化 …… 92
5.6.3 时空数据可视化 …… 93
5.6.4 多维数据可视化 …… 94
5.7 大数据安全 …… 95

此为试读，需要完整PDF请访问：www.ertongbook.com

5.7.1 大数据安全技术体系 …… 95
5.7.2 大数据平台安全技术 …… 96
5.7.3 数据安全技术 …… 97
5.7.4 隐私保护技术 …… 98

第 6 章 市场监管大数据——明察秋毫 …… 100
6.1 市场监管现状分析 …… 100
6.1.1 市场监管的内涵及其现代化 …… 100
6.1.2 大数据对市场监管的作用 …… 104
6.1.3 市场监管大数据总体需求分析 …… 105
6.2 市场监管大数据的发展 …… 106
6.2.1 国外市场监管大数据的发展 …… 107
6.2.2 国内市场监管大数据的发展 …… 108
6.3 市场监管大数据体系 …… 112
6.3.1 系统体系 …… 113
6.3.2 共性支撑体系 …… 116
6.3.3 应用服务体系 …… 121
6.3.4 安全体系 …… 123
6.3.5 管理保障体系 …… 127

第 7 章 综合交通大数据——四通八达 …… 131
7.1 交通行业需求与发展现状 …… 131
7.1.1 交通行业需求 …… 131
7.1.2 交通大数据应用发展现状 …… 132
7.2 交通大数据技术 …… 133
7.2.1 大数据生命周期 …… 133
7.2.2 数据采集技术 …… 134
7.2.3 数据存储技术 …… 134
7.2.4 数据挖掘与分析技术 …… 135
7.3 交通大数据综合应用 …… 138
7.3.1 大数据平台 …… 138

7.3.2 大数据交通管理 …… 139
7.3.3 大数据便民服务 …… 143
7.4 交通大数据面临的问题与挑战 …… 144
7.4.1 交通中的自动驾驶 …… 144
7.4.2 数据可视化 …… 145
7.4.3 数据安全 …… 146
第 8 章 农业农村大数据——强本节用 …… 148
8.1 农业农村现代化的新机遇 …… 148
8.1.1 大数据为农业农村发展指明了新方向 …… 148
8.1.2 互联网为农业信息铺设了“高速路” …… 149
8.1.3 物联网为农业感知延伸了“触角” …… 150
8.1.4 线上平台为农业销售拓展了“渠道” …… 151
8.2 农业农村大数据的发展 …… 151
8.2.1 国外农业农村大数据的发展 …… 151
8.2.2 国内农业农村大数据的发展 …… 153
8.3 农业农村大数据应用 …… 158
第 9 章 其他行业大数据——百花齐放 …… 161
9.1 政务大数据 …… 161
9.1.1 数字时代的管理模式 …… 161
9.1.2 国内外现状 …… 162
9.1.3 问题与思考 …… 163
9.2 公共安全大数据 …… 165
9.2.1 警务大数据 …… 165
9.2.2 消防大数据 …… 166
9.2.3 反恐大数据 …… 167
9.3 健康医疗大数据 …… 169
9.3.1 健康医疗大数据概述 …… 169
9.3.2 健康医疗大数据的特点 …… 170
9.3.3 健康医疗大数据的应用 …… 171

9.4 粮食物资大数据 …… 175
9.4.1 大数据对粮食物资行业的影响 …… 175
9.4.2 粮食物资大数据的国内外现状 …… 176
9.4.3 粮食物资大数据的发展趋势 …… 178
9.5 智慧营区大数据 …… 181
9.5.1 智慧营区大数据体系架构 …… 181
9.5.2 智慧营区大数据的特点 …… 184
9.5.3 智慧营区大数据的应用 …… 185
9.5.4 智慧营区大数据的发展趋势 …… 187
第 10 章 大数据的未来——缤彩纷呈 …… 189
10.1 科技发展趋势 …… 189
10.1.1 大数据驱动新一代人工智能 …… 189
10.1.2 科技改变生活 …… 190
10.2 大数据产业发展趋势 …… 191
10.2.1 市场需求 …… 191
10.2.2 发展趋势 …… 192
10.3 经济发展趋势 …… 195
10.3.1 全球趋势 …… 195
10.3.2 我国趋势 …… 197
10.4 未来已来，将至已至 …… 199
参考文献 …… 201

第1章

大数据时代——日新月异

随着信息科技的不断发展，信息的获取、存储、处理和传递越来越普及、越来越快捷，产生的数据也越来越庞大、越来越重要，一个崭新的时代正悄然来临。世界正从信息时代迈向大数据时代，数据挖掘与分析等大数据技术所展现的巨大价值，正激发大众对大数据孜孜不倦的探索。

1.1 大数据的崛起

1.1.1 数据大爆炸

大数据时代赋予了人们理解摩尔定律的新视角，摩尔定律引发的技术演进正催生海量数据的涌现。电子领域的摩尔定律指出，IC 上可容纳的晶体管数目大约每两年增加一倍。与此极为相似的是，大数据时代数据生成量每两年增加一倍。物联网、云计算、人工智能等新技术能够帮助人们以前所未有的速度和精度采集、分析、存储和处理数据。从人类出现到 2000 年，人类所产生的各类数据约有 5TB(T 为计量单位，1TB 约等于 1000 亿 B)。截至 2011 年，全球产生和复制的数据已达到 1.8ZB(ZB 为计量单位，1ZB 约等于 10 亿 TB)，到 2020 年总量将达到 44ZB，其中我国数据量将达到 7.9ZB，约占全球数据总量的 18%。人类社会已经真正进入数据爆炸的时代，每时每刻都有数以千万计的数据产生。

人类一方面遨游在信息的世界里，享受着信息发展带来的福利；另一方面也不得不忍受信息大爆炸带来的困扰：过多的无关信息侵占着视觉和听觉

渠道，消耗着精力，而查找自己需要的信息，又要花费大量的时间和精力。在这个“数据大爆炸”的时代背景下，通过大数据技术，对收集到的信息进行严密而富有逻辑的整理、分析、关联，发掘出具有价值和意义的信息，就显得特别的重要。例如，音乐平台可以根据听众的听歌习惯和风格推荐个性化的歌单，新闻媒体软件可以推送读者感兴趣的新闻广告，电商平台可以根据消费者的购物记录推荐相同款式和风格的衣服，等等。

1.1.2 洞悉大数据

1. 大数据内涵

(1)对大数据定义的理解

对于大数据(Big data)，迄今没有公认的定义，通常指大量数据的集合，其数据量大到目前主流的分析方法和软件工具在合理时间内无法进行有效的获取、管理、处理，但这些信息又迫切需要整理成能够帮助企业或政府部门提供决策的有效信息。按数据对象不同，大数据可分为实体数据集合和虚拟数据集合。政府部门及企业掌握的实体数据库为实体数据集合，而微博、百度、谷歌、微信等互联网上的信息为虚拟数据集合。

(2)大数据的“三元世界”

从宏观世界角度来讲，大数据是衔接物理世界、信息空间和人类社会三元世界的纽带。物理世界通过互联网、物联网等信息技术有了在信息空间中的大数据投影，而人类社会则借助人机界面、脑机界面、移动互联网等手段在信息空间中产生自己的大数据映像。融合了三元世界的大数据具有规模大、关系复杂、状态演变等显著特征。

(3)大数据的“六度空间”

名为 Six Degrees of Separation 的数学领域猜想可以翻译为“六度分隔理论”或“小世界理论”。该理论指出：你和任何陌生人之间所间隔的人不会超过 5 个。也就是说，最多通过 5 个中间人，你就能够认识某个陌生人。

大数据与六度分隔理论的完美结合，可以成为社交媒体、商业模式、网络社会的理论基础。在社交媒体中，六度分隔理论和微信、微博、QQ 等社交软件强化了人类的社交需求，只要信息媒介传播速度足够快、人群数量足

够多，世界上的任何人就都可以迅速建立联系，产生交流。在商业模式中，运用六度分隔理论可以进一步增强精准营销的效果，通过大数据的抓取和分析，以及人工智能的筛选和匹配，最后对特定人群投放特定广告和推介。

另外值得一提的是影响力权值。虽然根据六度分隔理论，任何两个陌生人想要相互认识最多不超过 5 个中间人，但这 5 个中间人之间的联系有强有弱，即前一个人对后一个人的影响力有强有弱。换言之，就是每个中间人都有一个影响力权值，权值越大，向后一个中间人传递信息的效率和能力就越强。因此，关键不在于你认识多少人，而在于你认识哪些人。

2. 大数据特点

大数据具有如下四个特点，如图 1-1 所示。

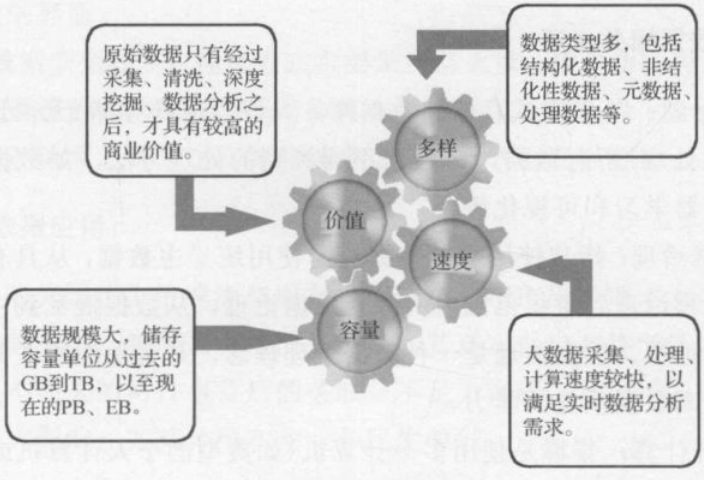

图 1-1　大数据特点示意图

一是数据规模大(Volume)。这个“大”源于广泛采集、多处存储和大量计算。普通的计算机储存容量以 GB、TB 为单位表示，而大数据则以 PB(1000TB)、EB(100 万 TB)为单位表示。

二是数据类型多(Variety)。大数据既包括地理位置信息、数据库、表格等结构化数据，也包括文本、图像、音视频等非结构化数据。不同的数据类型需要不同的处理程序和算法，所以大数据对数据的处理方法和技术也有更高的要求。

三是价值密度低(Value)。决策者要获得必需的信息，就得对大量的数据进行处理。现在通用的做法是通过使用强大的机器算法进行数据挖掘，进而获得与逻辑业务相吻合的结果，可以理解为在无边沙漠中用筛子淘取金沙，其价值密度可想而知。

四是处理速度快(Velocity)。大数据需要处理的数据有的是爆发式产生的，如大型强子对撞机工作时每秒产生 PB 级数据；有的虽然是流水式产生的，但由于用户数量众多，短时间内产生的数据量，如网站点击流、系统日志、GPS 数据等依然庞大。为了满足实时性要求，数据的处理速度必须快，过时的数据价值会贬值。例如，2011 年 3 月 11 日，日本大地震发生后，美国国家海洋和大气管理局在震后 9 分钟就推测可能发生海啸，但 9 分钟的计算延时对于瞬间被海啸吞没的生命来说还是太长了。

3. 大数据相关术语

① 数据湖：是集中式存储的数据库，允许以原样存储(无须预先对数据进行结构化处理)所有数据，并运用不同类型的处理方法，如数据挖掘、实时分析、机器学习和可视化等。

② 数据治理：指从使用零散数据变为使用统一主数据，从具有很少或没有组织和流程治理到组织范围内的综合数据治理，从数据混乱到主数据条理清晰的处理过程。数据治理是一种数据管理理念，是确保组织在其数据生命周期中存在高数据质量的能力。

③ 集群计算：集群是使用多个计算机(如典型的个人计算机或工作站)、多个存储设备冗余互联，组成对用户来说单一的、有高可用性的系统。集群计算用于实现负载均衡、并行计算等。

④ 黑暗数据：指被用户收集和处理但又不用于任何有意义用途的数据，可能永远被埋没和隐藏，因此称为“黑暗”数据，其可能是社交网络信息流、呼叫中心日志、会议笔记等。有学者估计企业 60%～90%的数据都可能是黑暗数据。

⑤ 大数据采集与预处理技术：数据的采集是进行数据分析和应用的前提。数据采集的方法手段比较多样，可通过互联网收集、数据库复制、数据

此为试读，需要完整PDF请访问：www.ertongbook.com