BIG DATA DEMYSTIFIED

How to use big data, data science and AI to make better business decisions and gain competitive advantage

大数据实战

大数据、数据科学和人工智能在商务决策中的应用

大卫·斯蒂芬森（David Stephenson）◎ 著
邵真 ◎ 译

中国人民大学出版社

BIG DATA DEMYSTIFIED

How to use big data, data science and AI to make better business decisions and gain competitive advantage

大数据实战

藏书章

大数据、数据科学和人工智能在商务决策中的应用

大卫·斯蒂芬森（David Stephenson）◎ 著
邵真 ◎ 译

中国人民大学出版社
• 北京 •

图书在版编目(CIP)数据

大数据实战/（美）大卫·斯蒂芬森（David Stephenson）著；邵真译. —北京：中国人民大学出版社，2019.9
书名原文：Big Data Demystified
ISBN 978-7-300-27192-7

Ⅰ.①大… Ⅱ.①大… ②邵… Ⅲ.①数据处理-研究 Ⅳ.①TP274

中国版本图书馆CIP数据核字（2019）第158121号

大数据实战
——大数据、数据科学和人工智能在商务决策中的应用
大卫·斯蒂芬森 著
邵 真 译
Dashuju Shizhan

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010-62511242（总编室）		010-62511770（质管部）
	010-82501766（邮购部）		010-62514148（门市部）
	010-62515195（发行公司）		010-62515275（盗版举报）
网　　址	http://www.crup.com.cn		
经　　销	新华书店		
印　　刷	天津中印联印务有限公司		
规　　格	170 mm×230 mm　16开本	**版　　次**	2019年9月第1版
印　　张	15插页1	**印　　次**	2019年9月第1次印刷
字　　数	150 000	**定　　价**	49.00元

版权所有　侵权必究　印装差错　负责调换

译者序

在大数据时代，企业不仅关注大数据技术本身，而且关注大数据技术与业务运营和业务战略的关联。管理者经常面临的问题是，哪些大数据技术更适合企业？大数据技术是否有助于企业提高业务运营效率，更快更好地做出业务决策？企业应该如何有效地应用大数据技术来支持企业的业务战略并获得长期的竞争优势？

上述问题在大卫·斯蒂芬森（David Stephenson）所写的《大数据实战》这本书中得到了很好的解答。大卫·斯蒂芬森是数据科学和大数据分析领域的专家，在宾夕法尼亚大学沃顿商学院讲授商务管理和数据分析课程，并担任过多家国际知名企业的专业咨询顾问，在数据科学领域具有丰富的研究和实践经验。

本书是大卫·斯蒂芬森结合自己近 20 年的行业咨询经验所写

的，包括两大部分。

第一部分关注大数据技术本身，共分为五章：第 1 章介绍了大数据的起源和基本概念，让读者了解大数据的重要性及其发展趋势；第 2 章以 IBM 所开发的深蓝计算机打败国际象棋世界冠军以及 AlphaGo 打败围棋世界冠军作为开篇案例，引出了人工智能和机器学习的概念，阐述了人工智能的起源及其发展；第 3 章和第 4 章通过塔吉特超市和《华盛顿邮报》的真实案例，阐述了大数据在精准预测、营销和客户关系管理方面所扮演的重要角色；第 5 章介绍了大数据生态系统的概念，并结合云计算模式的概念，阐述了大数据分析中硬件资源和软件资源开源的重要性。

第二部分介绍了大数据生态系统与企业运营和业务战略的关联，共分为六章：第 6 章和第 7 章从企业客户、产品性能、竞争对手和外部环境四个方面阐述了大数据技术如何帮助企业更好地制定并调整业务战略；第 8 章和第 9 章关注大数据分析技术的实施，阐述了企业如何有效地选择数据分析模型和分析方法、硬件资源和软件资源来支持业务战略，以获取竞争优势；第 10 章关注大数据团队的组建，介绍了平台工程师、数据工程师、业务分析师和算法专家等与大数据分析相关的职位及其工作职责；第 11 章关注大数据治理，阐述了大数据隐私准则、数据保护和规范遵守等相关的概念；第 12 章通过 IBM 开发的人工智能应用程序在医疗诊断中失败的案例，阐述了传统企业向数据驱动型企业转型时需要注意的事项。

本书的内容与大数据技术的发展紧密结合，各章内容中融入了

很多大数据分析的经典案例，能够让读者更好地了解大数据的相关知识，理解大数据生态系统及其与企业运营战略的关联。本书内容通俗易懂，适合对大数据及其应用感兴趣的读者。

本书由哈尔滨工业大学管理学院的邵真组织翻译并负责全文的统稿工作。在翻译的过程中，得到了课题组学生的支持，其中张心翼参与翻译了第 1～2 章，潘正远参与翻译了第 3～5 章，赵睿参与翻译了第 6～7 章，黄启亮参与翻译了第 8～9 章，张宸靓参与翻译了第 10 章，王倩参与翻译了第 11～12 章。在此向各位同学表示感谢！

由于时间紧迫和学识所限，本书的翻译可能存在不当和疏漏之处，恳请广大读者批评指正。

邵　真

前　言

你可能经常听到大数据这个词，但你真的知道大数据究竟是什么吗？大数据为什么如此重要？大数据能否对你的组织造成影响，从而带来改进和竞争优势？是否存在这种可能——不使用大数据会让你在竞争中处于劣势？

本书的目的在于解析“大数据”这一名词，同时向你提供用数据科学和机器学习来充分利用这些数据的实践方法。

大数据是一类新的数据，具有以下特点：数据量大，并且数据量还在持续、迅速地增长，同时，其数据结构并不符合传统的数据结构。“大”这个字眼是一种轻描淡写的说法，它并不能充分地说明实际情况的复杂程度。我们所处理的数据不仅仅是比传统的数据量大，更是与传统的数据有本质上的区别，就好比一辆摩托车并不是

大一点的自行车，一片海洋也并非大一点的游泳池。大数据带来了新的挑战，创造了新的机会，模糊了传统的竞争界限，因而需要新的方法来帮助我们从数据中获取有形的价值。海量数据与为处理此类数据而开发的技术结合，提供了对大规模数据的洞察角度，由此掀起了一股机器学习的浪潮，产生了由计算机驱动汽车的无人驾驶系统、比医生更为精确的心脏病预测系统，以及比人类更精通复杂游戏（如围棋[1]）的计算机系统。

为什么大数据是一个规则的改变者？正如我们将看到的，通过大数据，我们可以获得对数据更深层次的洞察力，从而理解促进消费者购买的因素以及生产线效率降低的原因。大数据不仅可以让商家实时地为全球数以百万计的用户提供高度个性化的体验，而且能够为诸如癌症研究、航天、粒子物理学等领域提供同时分析十亿量级数据的计算能力。大数据还提供了数据和计算资源，使得人工智能重新崛起，其中最具代表性的就是引领全球的基于深度学习[2]的技术。

没有局限于数据本身，在过去的二十年里，研究者和工程师们还开发出了硬件和软件结合的一整套生态系统，来收集、存储、处理和分析这些丰富的数据。本书将这些硬件和软件工具统称为大数据生态系统。这一生态系统能够帮助我们从大数据中挖掘出巨大的价值并将其应用于商业、科学和健康领域。想要利用大数据，你需要将大数据生态系统中的各个部分整合在一起，并选择出适合你的应用场景的最佳解决方案。你还需要为这些数据提供合适的分析方

法，众所周知，数据科学由此产生。

总的来说，大数据远不止简单的数据和技术。大数据应用于商业、科学和社会等领域，对你所从事的工作产生了巨大的影响。你的决策远不止购买一项技术。在本书中，我将会总结与大数据相关的工具、应用和处理方法，分析如何从多种形式的现代数据中获得价值。

大部分组织将大数据视为它们数字化转型的重要部分。许多成功的组织在运用大数据和包括深度学习在内的数据科学等方面已经做得很好了。研究表明，大数据的运用和收益增长（超过50%的收益增长）之间有着很强的联系。企业利用数据科学技术实现关键绩效指标（key performance indicators，KPI）[3] 10%～20%的增长是很常见的。

对于那些尚未开始利用大数据和数据科学的组织来说，最大的障碍就是不知道大数据应用所获得的收益是否值得付出与努力。我将在本书中阐明大数据应用所带来的益处，并通过案例来说明其中的价值和风险。

在本书的第二部分，我将描述在组织内确立一项数据战略和完成数据项目的实践步骤。我将讨论如何招募到合适的人，并创建一个收集和使用数据的计划。我也会讨论关于数据科学和大数据工具应用的具体领域。此外，我将对如何雇用合适的人来执行这些计划给出几点建议。

章节概述

第一部分　大数据揭秘

第 1 章　大数据的故事

本章将介绍大数据是如何发展成一个现象、大数据如何在过去的短短几年里变成一个如此重要的话题、大数据从哪里来、谁在使用大数据以及驱动人们使用大数据的原因是什么、大数据如何在今天实现了过去不可能完成的事情。

第 2 章　人工智能、机器学习和大数据

本章将介绍人工智能（AI）发展的一段简史，阐述人工智能是如何与机器学习联系在一起的，介绍神经网络和深度学习人工智能的应用及其如何与大数据产生联系，并对人工智能领域的工作者提出一些建议。

第 3 章　为什么大数据有用

本章将阐述我们的数据使用方式和思维方式是如何改变的，大数据如何创造新的机遇并改变现有的分析技术，通过成功的故事和案例来阐述数据驱动的含义。

第 4 章　大数据分析的应用案例

本章将阐述 20 个大数据分析和数据科学的商业应用案例，重点阐述如何应用大数据改变现有的数据分析方式。

第 5 章　理解大数据生态系统

本章将阐述关于大数据的主要概念，如开源代码、分布式计算和云计算。

第二部分　将大数据生态系统应用到组织中

第 6 章　大数据如何指导组织战略

本章将阐述如何应用大数据从客户、产品性能、竞争者和其他外部因素的视角来指导组织战略。

第 7 章　形成大数据和数据科学的战略

本章将提供分步指导：根据业务目标和利益相关者的建议进行数据规划，组建项目团队，确定最相关的数据分析项目，执行项目直至完成。

第 8 章　实施数据科学——分析、算法和机器学习

本章将阐述几种主要的数据分析方法、如何选择模型和数据库，并阐述使用敏捷的数据分析方法来实现商业价值的重要性。

第 9 章　选择技术

本章将阐述如何为大数据解决方案选择合适的技术、需要做出哪些决策、如何通过可利用的资源来实现决策。

第 10 章　组建团队

本章将阐述大数据和数据科学项目中需要的关键角色、如何在雇用和外包间做出权衡。

第 11 章　数据治理与法律遵从

本章将阐述隐私原则、数据保护、监管合法性和数据治理，并从法律、声誉和内部视角来分析其影响；讨论关于个人身份信息（PII）、链接攻击和欧洲的新隐私法规（GDPR）；通过案例分析因使用数据不当而陷入困境的公司。

第 12 章　在组织中成功部署大数据

本章将结合一个备受瞩目的项目失败案例，阐述在组织中成功部署大数据的最佳实践，并在如何使组织转型为数据驱动、如何在组织中部署数据分析人员、如何有效地使用资源以整合数据方面给出建议。

注释

[1] 围棋：中国的二人棋盘游戏。目标是用己方棋子围出最多的区域。

[2] 深度学习：利用数百层以上的人工神经网络的人工智能。

[3] 关键绩效指标（key performance indicators，KPI）：通常用于组织中评价绩效的量化标准，用于制定目标和衡量进展。

目 录

CONTENTS

第一部分 大数据揭秘

第 1 章 大数据的故事 3

到了 21 世纪初，是什么发生了改变 4

数据为什么变得这么多 5

产生数字化数据设备的广泛应用 6

正在迅速下降的磁盘存储成本 12

RAM 成本的直线下降 14

处理能力成本的直线下降 15

为什么大数据成为如此热门的话题 15

成功的大数据先驱者 18

开源软件为软件开发人员提供了竞争环境 18
云计算让启动和扩展计划变得更加容易 20
小贴士 21
问题 21
第 2 章 人工智能、机器学习和大数据 24
什么是人工智能和机器学习 25
人工智能的起源 26
为何近来人工智能再次兴起 27
人工神经网络和深度学习 28
人工智能如何帮助分析大数据 31
一些谨慎的话 31
小贴士 34
问题 34
第 3 章 为什么大数据有用 36
全新的数据使用方式 36
一种新的数据思维方式 37
遵循数据驱动的方法 38
更强的数据工具 41
小贴士 48
问题 48
第 4 章 大数据分析的应用案例 50
A/B 测试 50

推荐引擎/下一个最佳购物建议 52
预测：需求和收入 55
节省 IT 成本 56
市场营销 57
社交媒体 58
定价 59
客户维系/客户忠诚度 59
购物车弃置管理（实时） 60
转化率优化 60
商品定制化（实时） 61
重新定位（实时） 61
诈骗检测（实时） 61
减少客户流失 62
预测维护 63
供应链管理 64
顾客终身价值 64
线索评分 65
人力资源 66
情绪分析 67
小贴士 68
问题 68

第 5 章　理解大数据生态系统　70

什么让数据变“大”　70

分布式数据存储　73

分布式计算　74

快数据/流数据　75

雾计算/边缘计算　76

开源软件　77

许可　79

代码分发　80

开源的好处　80

大数据中的开源软件　81

云计算　82

小贴士　83

问题　84

第二部分　将大数据生态系统应用到组织中

第 6 章　大数据如何指导组织战略　89

你的客户　89

获取数据　90

使用数据　97

你的竞争者　102

外部的因素　104

你的产品 105
小贴士 107
问题 107
第 7 章 形成大数据和数据科学的战略 109
项目团队 110
启动会议 119
启动输出 121
范围界定阶段 123
小贴士 126
问题 127
第 8 章 实施数据科学——分析、算法和机器学习 128
四种分析方法 128
模型、算法和黑箱 135
人工智能和机器学习 140
分析软件 143
分析工具 148
敏捷分析 149
小贴士 151
问题 152
第 9 章 选择技术 154
交付给最终用户 159
选择技术时需要考虑的方面 160