

大数据策略

如何成功使用大数据与
10个行业案例分享

DATA DIVINATION
BIG DATA STRATEGIES

[美] Pam Baker 著
Bob Gourley 参编
于楠 译

大数据应用与技术丛书

大数据策略

如何成功使用大数据与
10 个行业案例分享

[美] Pam Baker 著
Bob Gourley 参编
于楠 译

清华大学出版社

北 京

Pam Baker

Data Divination: Big Data Strategies

Copyright © 2015 by Cengage Learning.

Original edition published by Cengage Learning. All Rights reserved. 本书原版由圣智学习出版公司出版。版权所有，盗印必究。

Tsinghua University Press is authorized by Cengage Learning to publish and distribute exclusively this simplified Chinese edition. This edition is authorized for sale in the People's Republic of China only (excluding Hong Kong, Macao SAR and Taiwan). Unauthorized export of this edition is a violation of the Copyright Act. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

本书中文简体字翻译版由圣智学习出版公司授权清华大学出版社独家出版发行。此版本仅限在中华人民共和国境内(不包括中国香港、澳门特别行政区及中国台湾)销售。未经授权的本书出口将被视为违反版权法的行为。未经出版者预先书面许可，不得以任何方式复制或发行本书的任何部分。

ISBN: 978-7-302-43902-8

北京市版权局著作权合同登记号 图字: 01-2015-2359

Cengage Learning Asia Pte. Ltd.

151 Lorong Chuan, #02-08 New Tech Park, Singapore 556741

本书封面贴有 Cengage Learning 防伪标签，无标签者不得销售。

版权所有，侵权必究。侵权举报电话：010-62782989 13701121933

图书在版编目(CIP)数据

大数据策略 如何成功使用大数据与 10 个行业案例分享/ (美)帕姆·贝克(Pam Baker) 著;
(美)鲍勃·格雷(Bob Gourley) 参编; 于楠 译. —北京: 清华大学出版社, 2016

(大数据应用与技术丛书)

书名原文: Data Divination: Big Data Strategies

ISBN 978-7-302-43902-8

I. ①大… II. ①帕… ②鲍… ③于… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 117019 号

责任编辑: 王 军 刘伟琴

封面设计: 周周设计局

版式设计: 思创景点

责任校对: 曹 阳

责任印制: 刘海龙

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座

邮 编: 100084

社 总 机: 010-62770175

邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者: 北京密云胶印厂

经 销: 全国新华书店

开 本: 185mm×260mm

印 张: 19.25

字 数: 468 千字

版 次: 2016 年 7 月第 1 版

印 次: 2016 年 7 月第 1 次印刷

印 数: 1~4000

定 价: 49.80 元

产品编号: 062275-01

译者序

2009年甲型H1N1流感爆发的几周之前，互联网巨头谷歌公司在《自然》杂志上发表的一篇论文令公共卫生官员们和计算机科学家们感到震惊。文中解释了谷歌不需要分发口腔试纸和联系医生，便能够在全美范围内预测冬季流感的传播，甚至可以具体到特定的地区和州，而这仅仅是通过观察人们在网上的搜索记录来完成的。这是当今社会所独有的一种新型能力：通过对海量数据进行分析，获得具有巨大价值的产品和服务，或者深刻的洞见。在下一轮流感来袭之前，我们将拥有一种更好的预测工具，预防流感的传播。除了公共卫生领域，大数据引起社会生活巨大变革的例子已不胜枚举。2012年《纽约时报》的一篇专栏中指出“大数据时代已经降临”，文中提到，大数据逐步渗透到人类生产生活的方方面面，对各行各业都造成影响。人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。区别于传统通过经验和直觉的方式给出判断，未来将会日益基于对海量数据的分析而做出决策，并巧妙地用来激发新产品和新型服务。我们可以预见，无论是学术界、工业界，还是政府、商界，所有领域都将面临这种进程。

无论是最早提出大数据定义的咨询界翘楚麦肯锡咨询公司，还是全球最具权威的IT研究顾问高德纳公司，在大数据技术方面均提到数据的专业化处理。这是因为我们会对大数据庞大的容量、繁多的类型、高速的数据处理与流转，以及低价值密度感到有些无所适从。只有不断提高数据的加工能力，挖掘内部的潜在联系，才能实现数据增值，并基于分析结果做出决策，最终指引新一轮经济增长和消费者盈余浪潮的到来。

作为一名著名的商业分析师、科技撰稿人，本书作者帕姆·贝克(Pam Baker)长期致力于大数据研究与分析的一线工作。大量工作经验的积累和体会，使本书与传统的工程技术类图书有着本质的区别。本书并不是为工程技术人员讲述如何配置数据服务器或优化数据存储容量等，相反，它更偏重于为企业经理或合作伙伴提供大数据方面的经验策略，帮助理解数据分析的应用问题等。

作者以亲切易懂的语言为读者深入浅出地讲述了大数据工作以及相应商业策略的制定，描述了跨产业融合的可预见持续性等，这些绝非是对以往大数据策略书籍和主流媒体讨论的知识的汇总和提炼。首先，本书从大数据策略的制定入手，提出了策略的重要性，指明“如果没有强有力的数据策略，即便存储大量的数据或使用各种大数据工具，也无法帮助我们得到最终的分析结果”。其次，以策略为中心，帕姆·贝克继续阐述了如何正确提取数据源、解决数据难题、动态调整策略、协同分析，以及处理数据隐私等热点、敏感问

题。最后，围绕大数据的关键业务问题和管理框架，作者充分发挥自己科技记者的一贯风格，汇聚了众多行业在大数据分析方面的示例，这也是本书最引人入胜的部分，其中包括政府部门、国防情报部门、安全行业、医疗保健领域、小型企事业、能源领域、交通运输业、商品零售业、银行金融业、制造业等。依托于大数据策略的使用，她还分析了各个领域的机遇与挑战，并以独到的观点解释了如何更加有效地使用大数据分析结果，使之能够促进社会经济的发展与稳定。相信这些翔实丰富的实际用例以及作者的精妙阐述可以帮助读者深刻地领会大数据策略的独到之处，并能够结合自身的需求，将大数据分析与挖掘应用到实际工作中。

作为译者，本人非常愿意向读者朋友们推荐本书。因为在你没有任何技术背景的情况下，想要了解大数据革命和相应的应用情况，本书绝对不失为一位不折不扣的导师。如果你是一名 IT 工作者，也请将本书推荐给你的老板、投资人，认识的媒体人和创业者，这将可以帮助他们了解你的价值所在，并促使他们更多地在大数据领域进行投资。

在此，我非常感谢清华大学出版社的编辑们为本书付出的心血，你们卓越的见识，才使得这本好书有幸能与中国读者见面。你们的专业与严谨也令我非常佩服，没有你们，本书就不可能顺利出版。同样感谢我的家人对我翻译工作的支持和鼓励。感谢我可爱的儿子，他给予我翻译这本书的动力。

由于本书覆盖面广，尽管一直努力保证翻译的准确性，但由于水平有限，难免会有疏忽之处，恳请各位读者批评指正。对于本书的任何想法和意见都欢迎发送邮件至 celene1122@gmail.com。

于楠

作者简介

帕姆·贝克(Pam Baker)是一位著名的商业分析师、科技自由撰稿人，以及在线出版刊物和电子杂志 *FierceBigData* 的编辑。她在多家受人尊敬的出版刊物供职，包括但不限于 *Institutional Investor* 杂志、科技博客网站 ReadWriteWeb、CIO(印刷版)科技博客 CIO.com、*Network World*、*Computerworld*、*IT World*、*Linux World*、*iSixSigma* 杂志和网站 TechNewsWorld。她以前曾担任总部设在伦敦的 VisionGain Research 公司和总部设在加州圣克鲁斯的调研机构 Evans Data Corp 的签约分析师。她还曾任研究员、作家以及设计公司 Wireless I.Q.和总部设在纽约的市场研究公司 ABI Research 旗下的 Telematics Journal 的执行主编。

有兴趣的读者可以从下面这些网站上查看各种出版物，了解更多有关帕姆·贝克的工作：互联网媒体公司 Mediabistro 外包平台，网址为 <http://www.mediabistro.com/PamBaker>；互联网新闻公会(Internet Press Guild)，网址为 <http://www.netpress.org/ipg-membership-directory/pambaker>。她的 LinkedIn 个人页面介绍中也有很多专业参考，网址为 <http://www.linkedin.com/in/pambaker/>。

她还写了许多电子书，出了几本纸质书。其中 6 本在亚马逊作家中心在销。此外，贝克与他人共同为美国的大学科技经理人协会(Association of University Technology Managers, AUTM)合著了两本有关生物科学的书籍，该协会是一个由技术经理和企业高管组成的全球非营利组织。这两本书主要是由比尔和梅林达·盖茨基金会资助。

在贝克获得的众多奖项中，其中一项是她制作的有关造纸工业的纪录片赢得了国际赞誉。佐治亚州哥伦布市授予决议，就她在《佐治亚州趋势杂志》(*Georgia Trend Magazine*)中对哥伦布市所做的一组新闻报道予以表彰。另一位获此殊荣的是传奇作家卡森·麦卡勒斯(Carson McCullers)。贝克是全国记者俱乐部(National Press Club, NPC)和互联网新闻公会(Internet Press Guild, IPG)的成员。可以关注她的 Twitter 账号@bakercom1 或 Google+ [google.com/+PamBaker](https://plus.google.com/+PamBaker) 与她交流。也可以通过她供职的 FierceBigData 网站联系表单联系她(参见 <http://www.fiercebigdata.com/>)。

鲍勃·格雷(Bob Gourley)为本书特约作者，他主要研究国防情报领域中的大数据，撰写了第 10 章的大部分内容。格雷还撰写了第 20 章。他是 CTOvision.com 网站的主编，也是技术研究和顾问公司 Crucial Point LLC 的创始人和首席技术官。

《信息世界》(*InfoWorld*)将他评为全球 25 位最具影响力的首席技术官之一，《华盛顿人》(*Washingtonian*)杂志将其命名为华盛顿地区“技术巨头”之一。政府 2.0 社区 GovFresh

将他评为“政府 IT 界 25 位最有魅力的沟通者”之一。2012 年,《福布斯》将他评为“Twitter 上研究大数据最有影响力的人物”。

鲍勃拥有三个硕士学位,包括美国海军研究生院科学与技术情报硕士、美国海军陆战队大学军事科学硕士和詹姆斯·麦迪逊大学计算机科学硕士。鲍勃发表了 40 多篇文章,涉及内容广泛,他也是 2009 年 1 月出版的 *Threats in the Age of Obama* 一书的特约作者。WashingtonTech 网站将他的博客 CTOvision 列为美国顶尖科技博客之一。

鲍勃是网络冲突研究协会(Cyber Conflict Studies Association)的创始成员和董事会成员。该协会为非营利组织,着重研究权威学术机构中的网络冲突,同时提高国家理解网络空间中复杂动态冲突的能力。可以关注他的 Twitter 账号@bobgourley 与他聊天。还可以通过 Twitter 账号@AnalystReport 和@CTOvision,或访问 <http://ctovision.com/pro> 联络他。

致 谢

首先，我想感谢这段时间以来圣智学习出版公司(Cengage Learning PTR)团队的辛勤工作和耐心。具体要感谢史黛西·伊凯(Stacy Hiquet)对本书付梓付出的努力。还要感谢希瑟·赫尔利(Heather Hurley)，她工作能力超群，促进本书得以成书。她专业素养一流，一直充满乐趣地工作着。感谢凯姿·娅恩兹利(Kezia Endsley)有条不紊地监督整个编辑与付印过程，还要应付我不经意间制造的种种麻烦和挑战。感谢整个出版团队的各位同事，他们的耐心、宝贵的建议成就了本书出版。她是迄今为止我有幸合作过的最好的编辑，真诚地希望再次与她合作。感谢理查德·桑塔丽莎(Richard Santalesa)详细的技术审查和合理的建议。他也将永远是作家最好的法律资源和技术编辑，而且是最真挚的朋友。感谢他加入圣智的整个出版团队。

最后还要感谢我的家人这段时间对我的耐心和支持。花费冗长、看似无休止的几个小时整理书稿不仅对作者是一种考验，对家人也是如此。特别感谢我的两个兄弟——斯蒂文·达菲(Steven Duffey)和约翰·达菲(John Duffey)——他们协助我创建屏幕截图、制定严格的规格，得以在时间紧迫的情况下，加速完成书稿。

前言

当下有关大数据的讨论、文章和会议演讲中，始终有一个悬而未决的问题：到底可以用大数据做什么？当然，给出的回答通常非常概括，字眼含义往往含糊不清。很少有人阐明大数据的来源，更不用说大数据应用了。相关问题的答案更是少见，比如从如何计算大数据项目的投资回报率(ROI)，将大数据化为实际收益到如何开发一个成功的策略和最终如何运用分析来改变整个组织和行业。本书将回答目前有关大数据的最紧迫的以及更高层次的问题。

本书读者对象

本书的读者对象是那些对大数据实际应用而非技术细节更感兴趣的人。无论你是独自经营公司还是跨国运营企业，都会在本书中找到切实可行的建议：何时以何种方式利用大数据为所在机构赢得最理想的效果。无论是数据科学家、部门主管、律师、小企业主、非营利机构负责人，还是企业高管或董事会成员，本书都可以帮助他们将大数据技术应用于工作中，进行辅助决策。

此外，书中用大量篇幅描述了具体行业的大数据实际应用案例，用以指导不同行业和企业大数据的有效利用和潜在应用。不同的章节将针对 10 个不同行业有更详细的论述。本书将具体讨论大数据在包括政府、国防和情报、安全、医疗保健、小企业和农场、交通运输、能源、零售、银行保险及制造业等部门或领域的案例、策略、潜在因素和新兴趋势。然而，仅仅阅读论述所在行业的章节是不够的。不久的将来，大数据对其他行业的变化也将对你自身所处行业的未来发展产生影响。

如果说到大数据的发展方向的话，那便是催生变革，促进行业发展进程全面转折。事实上，大数据正在促进整个产业的融合，可以说，跨行业融合正在以前所未有的态势高速发展。因此，有必要关注那些与所在行业正在融合的产业，关注哪类客户正在减少或完全取消所提供的服务。在这方面，本书极有可能带给你更多惊喜。

策略就是一切

本书涉及多个不同的行业主题，讲述关于如何开发成功的大数据策略及其成果。书中回答了如下问题：如何计算投资回报率；搭建数据团队；筹措数据货币化；提出有吸引力的商业命题；制定出正确的问题；从分析中找到可行性方案；为企业和行业预测未来发展；有效地处理隐私问题；利用可视化效果找到最优数据表达方式；明确产品和服务创新的场所、时机和渠道；以及如何改变整个机构。

在书的最后应该能够很容易认识到应从何种角度开始，如何做好大数据应用，以及今后大数据的发展方向。

书中仅介绍了很少一部分大数据分析工具。和所有的技术一样，分析工具会随着时间的推移而老化，但大数据策略将胜出。大数据策略将根据需求不断变化，是适用于企业经营的基本策略。策略是需要关注的重点，本书对此也有相应论述。

利用手中的策略，我们会知道投资哪些工具，把它们用在何处，以及如何利用它们。除去有助于选择正确的分析工具，增加利润，大数据策略还会帮助我们度过迫在眉睫的巨变。改变很多，不可避免。唯一的办法是做好准备，积极主动地选择前进的道路。本书将尽力利用大数据展示更多可行方案，以期全部实现。

目 录

第 1 章	何为真正的大数据	1
1.1	技术层面的定义	1
1.2	为什么数据规模无关紧要	4
1.3	大数据对管理层意味着什么	4
1.3.1	“大数据是万能的”	4
1.3.2	“数据只是另一种电子表格”	5
1.4	大数据的执行方式	5
1.5	小结	10
第 2 章	如何制定成功的大数据策略	11
2.1	转不出的死命循环	11
2.2	如何解开“谁是第一次”这个难题	13
2.2.1	改变大数据视角	13
2.2.2	用户认知与数据采集	13
2.2.3	Facebook 预测性分析的现实	14
2.2.4	Facebook 数据收集走得 更远	15
2.2.5	使用 Facebook 坦诚认知 大数据发展潜力	16
2.2.6	专业认知与大数据现实	16
2.2.7	从感知到认知偏差	17
2.2.8	寻找大数据占卜师	17
2.3	下一步：拥抱无知	19
2.4	始于何处	19
2.4.1	在结束时开始	20
2.4.2	当行动变为无为时	21

2.5	确认目标，瞄准目标	22
2.6	如何获得最佳实践方法，让落后 观念远离前进的道路	24
2.6.1	解决人们对大数据的恐慌	24
2.6.2	终结未知的恐惧	24
2.6.3	接受改变，融入改变	25
2.6.4	机器统治并不确定，人类 仍然起作用	26
2.6.5	接触少数固执的人	26
2.7	回答没人提出的问题	26
2.7.1	持续询问可能性	27
2.7.2	寻找最终目标	27
2.8	与解说团队交叉合作	28
2.8.1	为团队增加业务分析师和 关键终端用户	28
2.8.2	为收集和管理数据增加 首席数据官	29
2.9	小处着手、逐步发展并扩张	30
2.10	原型和迭代策略	31
2.11	谈谈向数据策略中添加 预测分析	31
2.12	民主化数据，但预计几乎 无人使用(目前)	31
2.13	策略就是一个活的文档； 充实它、滋养它	32
2.14	小结	32
第 3 章	提出“正确”的问题	33
3.1	协作努力，提出问题	34

- 3.2 魔法 8 号球效应 35
- 3.3 用数学软件来分析现实问题 36
- 3.4 “正确”问题的清单 36
- 3.5 小结 36
- 第 4 章 选择“正确”数据源的方法 37
 - 4.1 需要更多的数据源(数据类型)而非数据本身(数据容量) 37
 - 4.2 为什么无论数据规模多大,生成的数据量都会不足且永远不足 38
 - 4.3 数据囤积与先捉再放 38
 - 4.4 不可思议的大数据案例: 购买尿布的狗主人 39
 - 4.5 升级事务性数据的价值 39
 - 4.6 社交媒体数据分析的局限性 40
 - 4.7 大数据买卖的货币价值 40
 - 4.8 利用黑客技术赚钱碰到麻烦 41
 - 4.9 评估数据源 42
 - 4.10 过时的模型招致竞争对手 42
 - 4.11 购买数据时的考量 43
 - 4.12 确定所需的外围数据 43
 - 4.13 谈谈结构化与非结构化数据 44
 - 4.14 防止人为偏见对数据选择的影响 46
 - 4.15 数据孤岛的危险 46
 - 4.16 使用所需数据源的必要步骤 47
 - 4.17 小结 48
- 第 5 章 解答大数据问题如同玩魔方 49
 - 5.1 可行性数据的概念 49
 - 5.2 描述性、预测性和规范性数据分析类型的差异 51
 - 5.2.1 描述性数据分析 52
 - 5.2.2 预测性数据分析 53
 - 5.2.3 规范性数据分析 53
 - 5.3 已有明确答案的问题 54
 - 5.4 解释会导致更多的问题 55
 - 5.5 需要解读的问题——魔方 55
 - 5.6 小结 57

- 第 6 章 实时分析在动态化策略中的作用 59
 - 6.1 检查实时错觉和时间胶囊 60
 - 6.2 静态策略与动态策略 61
 - 6.3 谈谈转向动态策略的变革管理 62
 - 6.4 选择分析方式 62
 - 6.5 利用专家经验, 加速数据分析 65
 - 6.6 实时分析来得太迟时该怎么办 66
 - 6.7 小结 66
- 第 7 章 大数据的价值主张和货币化 67
 - 7.1 确定未知领域的投资回报率(ROI) 67
 - 7.2 滥发的货币和模糊的投资回报率 69
 - 7.3 成本核算的困惑 70
 - 7.4 成本不是问题 71
 - 7.5 先考虑大数据项目再谈商业案例 71
 - 7.6 计算实际成本 72
 - 7.7 价值所在 73
 - 7.7.1 从技术角度看待商业案例 73
 - 7.7.2 从非技术角度看待商业案例 74
 - 7.8 项目回报率的计算公式 74
 - 7.9 重要问题: 是否应该出售数据 76
 - 7.9.1 销售数据解析 77
 - 7.9.2 物以稀为贵 77
 - 7.10 小结 78
- 第 8 章 协同经济的兴起和盈利方式 79
 - 8.1 数据等于知识和财富 79
 - 8.2 大数据带来的最大冲击: 颠覆原有模式 80
 - 8.2.1 分享经济 82
 - 8.2.2 创客运动 83
 - 8.2.3 合作创新 84

8.3	新模式在新协同经济中兴起.....	85	10.2	信息相关性问题的处理(“了解情况”问题).....	121
8.4	强调流畅性,摒弃灵活度.....	87	10.3	海量数据中信息搜索和发现(“海底捞针”问题).....	124
8.5	应用大数据制定战略新模式.....	89	10.4	企业网络安全数据管理.....	127
8.6	小结.....	90	10.5	后勤信息(包括粗放型/动态性企业资产目录).....	127
第 9 章	隐私难题.....	91	10.6	加强卫生保健.....	127
9.1	真相揭开的那天预示着个人隐私神话的失败.....	92	10.7	开源信息.....	129
9.1.1	危险汇总.....	94	10.8	内存数据的现代化.....	130
9.1.2	可在世界各地接听的手机通话.....	94	10.9	企业数据中心.....	130
9.1.3	公民和退伍军人的数据如何帮助其他国家策划袭击.....	96	10.10	武器装备与战争中的大数据用例.....	130
9.1.4	数据扩散逐步升级.....	97	10.11	小结.....	131
9.1.5	为个人隐私画一条底线.....	98	第 11 章	政府大数据管理用例.....	133
9.1.6	企业的隐私难题.....	101	11.1	大数据趋势对政府数据的影响.....	134
9.2	数据收集中的 4 大转变.....	102	11.2	联合国“全球脉动”计划用例.....	135
9.2.1	数据入侵性改变.....	103	11.3	联邦政府(非国防部或情报界)用例.....	137
9.2.2	数据多样性的改变.....	104	11.4	州政府用例.....	139
9.2.3	数据整合性的改变.....	105	11.5	当地政府用例.....	142
9.2.4	数据作用范围的变化.....	105	11.6	法律实施用例.....	144
9.3	必须质疑的商业问题.....	110	11.7	小结.....	145
9.4	谁是真正的数据拥有者.....	110	第 12 章	安全行业用例.....	147
9.5	当前法律和措施在设定先例中的作用.....	111	12.1	一切都在互联网上.....	147
9.6	授权允许的误区.....	113	12.2	亦敌亦友的数据.....	148
9.7	个人价值与混合数据.....	113	12.3	防病毒/反恶意软件用例.....	149
9.8	匿名数据的误区.....	114	12.4	目标如何击中要害.....	151
9.9	个人隐私与个人利益之间的平衡.....	115	12.5	虚拟和现实世界的碰撞.....	156
9.10	数据收集何时会使你或你的公司承担责任.....	115	12.6	纷乱的机器数据.....	157
9.11	商业价值的透明度.....	117	12.6.1	农民面临的信息安全困境.....	157
9.12	数据从业人员必须铭记的事实.....	118	12.6.2	物联网中农民面临的安全困境周而复始.....	158
9.13	小结.....	118	12.7	当前和未来信息安全分析法.....	159
第 10 章	国防情报部门中的用例.....	119			
10.1	态势感知和可视化.....	120			

12.8	小结	162
第 13 章	医疗保健领域用例	163
13.1	解决抗生素危机	163
13.2	使用大数据治病	165
13.3	从谷歌到疾病预防控制 中心	165
13.3.1	美国疾病预防控制中心 (CDC)的糖尿病交互 图谱	168
13.3.2	项目数据领域	171
13.3.3	赛智生物网络	172
13.4	另一方：生物黑客	173
13.5	电子健康记录(EHR)、电子 病历(EMR)和大数据	175
13.6	公布医疗保健数据	176
13.7	小结	179
第 14 章	小企业和农场用例	181
14.1	大数据适用于小企业	181
14.2	炒作和真实世界局限性 之间的界限	182
14.3	为工作选择合适的工具	182
14.4	可能会使用的外部数据源 示例	187
14.5	给使用共用或共享数据农民的一 句忠告	192
14.5.1	说法一：数据属于农民	193
14.5.2	说法二：数据只用于 “帮助”农民从农场中 更加受益	194
14.5.3	说法三：农民的数据将会 保密	194
14.6	钱、钱、钱：大数据扩大借贷 能力的方式	195
14.6.1	PayPal 信贷	196
14.6.2	亚马逊资本服务	196
14.6.3	数据驱动型贷款公司 Kabbage	197
14.7	小结	197

第 15 章	交通运输中的用例	199
15.1	加速发展大数据赚取利润	199
15.1.1	美中不足的事	200
15.1.2	依靠数据获胜不会长久	201
15.1.3	火车、飞机和船舶中的 数据使用	201
15.2	车联网：很可能不是你 以为的那样	203
15.2.1	数据引导创新和自动化	206
15.2.2	智能城市的崛起	206
15.2.3	正在发生的交通创新 实例	207
15.3	数据和无人驾驶汽车	208
15.4	互联的基础设施	210
15.5	汽车保险品牌数据收集 设备	212
15.6	交通领域无法预料的数据 可靠性	214
15.7	小结	215
第 16 章	能源领域中的用例	217
16.1	关于能源神话和假设的 大数据	217
16.2	美国能源信息署(EIA)能源 数据存储库	219
16.3	EIA 能源数据表格浏览器	220
16.4	失踪的智能电表数据	222
16.5	EIA 的 API 和数据集	222
16.6	国际意义与合作	223
16.7	公私合作下的能源数据 变革	224
16.8	公用事业用例	225
16.9	小结	227
第 17 章	零售业大数据用例	229
17.1	在大数据中重新运用 老战术	229
17.1.1	零售业没搞砸，对象客户 发生了变化	231
17.1.2	品牌叛变和恶魔客户	231

17.1.3	客户体验又成为一个 问题.....	232
17.1.4	大数据与恶魔客户复兴.....	232
17.2	零售业与大数据博弈的 原因.....	234
17.3	大数据帮助零售业的方式.....	234
17.3.1	产品选择和定价.....	235
17.3.2	当前市场分析.....	236
17.3.3	利用大数据开发新的 定价模式.....	236
17.3.4	寻找更好的方法获取更多、 更好和更清洁的客户 数据.....	237
17.3.5	研究和预测客户接受度 和反应.....	237
17.3.6	预测并规划应对更广泛的 市场发展趋势.....	241
17.4	预测零售业未来.....	243
17.5	小结.....	244
第 18 章	银行和金融服务业用例.....	245
18.1	定义问题.....	245
18.2	银行和贷款机构的用例.....	246
18.3	大数据如何在借贷领域点燃 新竞争.....	248
18.4	新型可选择贷款方式.....	248
18.4.1	贝宝(PayPal)贷款项目.....	248
18.4.2	人人贷和贷款俱乐部.....	249
18.5	零售商与银行的较量; 信用卡品牌规避银行.....	250
18.6	征信局所面临的大数据 问题.....	250
18.7	谈谈保险公司.....	252
18.8	小结.....	254
第 19 章	制造业用例.....	255
19.1	经济形式与机会展望.....	256
19.2	制造业的十字路口.....	258
19.3	3D 打印与大数据的相 交点.....	260
19.4	3D 打印是如何影响制造业 并扰乱客户的.....	261
19.4.1	盈创公司一天打印 10 所 住宅.....	261
19.4.2	3D 打印的景观别墅.....	262
19.4.3	3D 打印的傍水小宅.....	263
19.4.4	3D 家庭打印对制造业的 影响.....	263
19.5	增材制造的转变将是巨大的, 并会波及所有部门.....	263
19.6	个性化制造将如何改变一切, 甚至创造更多的大数据.....	265
19.7	制造业内部新数据源涌出.....	266
19.8	此行业的用例.....	267
19.9	小结.....	267
第 20 章	下放权力.....	269
20.1	数据民主化.....	269
20.2	4 步措施.....	270
20.3	其他 4 步.....	272
20.4	小结.....	273
第 21 章	摘要.....	275
21.1	何为真正的大数据.....	275
21.2	如何制定成功的大数据 策略.....	276
21.3	提出“正确”的问题.....	276
21.4	选择“正确”数据源的 方法.....	277
21.5	解答大数据问题如同玩 魔方.....	277
21.6	实时分析在动态化策略中的 作用.....	278
21.7	大数据的价值主张和货 币化.....	279
21.8	协同经济的兴起和盈利 方式.....	279
21.9	隐私难题.....	280
21.10	政府大数据管理用例.....	280
21.11	国防情报部门中的用例.....	281

21.12	安全行业用例	282	21.17	零售业大数据用例	286
21.13	医疗保健领域用例	282	21.18	银行和金融服务业用例	287
21.14	小企业和农场用例	283	21.19	制造业用例	288
21.15	能源领域中的用例	284	21.20	下放权力	289
21.16	交通运输中的用例	285			

何为真正的大数据

似乎每个人都把大数据挂在嘴边，甚至对许多人来说，大数据是需要考虑的头等要务。因此，有人认为，每个人都知道何为大数据。其实不然。尽管大数据有各种各样技术上的定义，但大多数人仍然不确定大数据的大小与常规数据库的大小之间确切的界限。这给介绍和探索大数据的发展，尤其是定义大数据项目参数带来一定困难。

本章将探讨对“大数据”概念的不同解读。

1.1 技术层面的定义

大数据并不只是存储规模从吉字节(gigabyte, GB)到太字节(terabyte, TB)的简单的数量级增长，尽管数据集仍如预期增长迅速，这一点将在下一章详细阐述。更确切地讲，大数据是各类数据集合的汇总，包括一些结构化和非结构化数据，一些由物理数据源转换为在线数据集的数据集，以及事务型和非事务型数据库。来源多种多样，包括一些自产数据和第三方数据。通常数据集的存储模式存在差异，缺乏一致性。一般来说，大数据的处理繁冗而复杂，支出明显过高，即便并非完全行不通，现有的运算技术也很难支持大数据计算。

此外，在技术上，数据集达到何种规模才符合大数据标准尚未达成共识。而技术领域内部更倾向于从描述数据特征、衡量数据规模、计算处理大规模数据量来定义大数据。

2001 年美国 Gartner 公司(原为 Meta 集团)的一份报告中对大数据进行了定义，强调大数据必须具备 3V 特征，即容量大(Volume)、多样化(Variety)和速度快(Velocity)。现在，机构在 3V 之外又定义了第 4 个“V”——真实性(Veracity)，用以强调数据质量的重要性。

本质上，大数据是需要使用新工具来计算的任何规模的数据集。因此，大数据只是一个相对的概念，今天被定义为“大”的数据未来可能只是“一般”甚至“小”规模的数据。

所以，在数据前面加上“大”字作为修饰语并不恰当，也并无大用。大多数业内专家预计不久的将来，这个“大”字将会被完全舍弃，因为它无法准确地描述大数据概念的本质。真正让大数据产生价值的是它将人、地点和事物广泛地连接起来，这在以前无论如何