

BIG DATA

大数据与 商业分析

【美】杰伊·利博维茨 (Jay Liebowitz) ●主编
刘斌 曲文波 林建忠 等●译



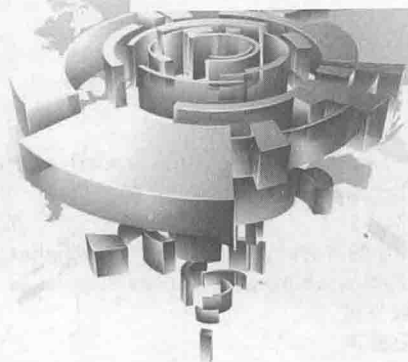
& BUSINESS ANALYSIS

清华大学出版社

大数据与商业分析

BIG DATA & BUSINESS ANALYSIS

【美】杰伊·利博维茨 (Jay Liebowitz) ● 主编
刘斌 曲文波 林建忠 等 ● 译



清华大学出版社
北京

北京市版权局著作权合同登记号 图字:01-2013-8011

Authorized translation from the English language edition, entitled BIG DATA AND BUSINESS ANALYTICS, 9781466565784 by JAY LIEBOWITZ, published by CRC Press Taylor & Francis Group, copyright ©2013

All Rights Reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from CRC Press Taylor & Francis Group, Inc. CHINESE SIMPLIFIED language edition published by TSINGHUA UNIVERSITY PRESS Copyright ©2015.

本书中文简体翻译版由清华大学出版社出版发行。未经许可,不得以任何方式复制或抄袭本书的任何部分。

本书封面贴有清华大学出版社防伪标签,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

大数据与商业分析/(美)利博维茨(Liebowitz,J.)主编;刘斌等译.--北京:清华大学出版社,2015

书名原文:Big data and business analytics

ISBN 978-7-302-39079-4

I. ①大… II. ①利… ②刘… III. ①商业管理—数据管理 IV. ①F712

中国版本图书馆 CIP 数据核字(2015)第 017240 号

责任编辑:刘志彬

封面设计:汉风唐韵

责任校对:王荣静

责任印制:王静怡

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 刷 者:三河市君旺印务有限公司

装 订 者:三河市新茂装订有限公司

经 销:全国新华书店

开 本:170mm×240mm

印 张:17.5

字 数:235千字

版 次:2015年4月第1版

印 次:2015年4月第1次印刷

印 数:1~4000

定 价:49.00元



作者序

“能够比竞争者学习得更快是你仅有的、长期有效的竞争优势，……”这是 AriedeGues《长寿公司：商业“竞争风暴”中的生存方式（*The Living Company: Habits for Survival in a Turbulent Business Environment*）》一书中经常被引用的一句名言。

最近几年，由于政府的重视、媒体的宣传和大企业的应用，“大数据”突然成了一个家喻户晓的名词和炙手可热的话题。确实，随着信息、通信和媒体技术的发展，数据规模之大、形式之杂、增长之快，已远远超出许多企业在数据采集、存储、分析、管理和应用方面的能力。但社会所面临的更大挑战是：许多企业对数据和数据能带来的价值还缺乏最基本的理解和应用体验。

在美国，用数据来推动和优化政府或公司决策已经历了几十年的历程。对数据价值的认识和应用的成功，在很大程度上推动了美国数据产业的“基本设施建设”、数据产业链和供应链的形成以及相关政府政策和法律的制定。大量资本和资源投入数据领域，以帮助技术和商业模式的创新。当大数据的机会到来时，政府和企业都已经做好了充分的准备，大数据市场的建立和相关产品服务的创新因而随即得到迅猛发展。

然而，尽管数据的潜在价值很高，很多的企业也已拥有大量数据，但是这些数据价值的实现常常很困难。具体原因包括：企业的数据库管理，分析能力和投入不足，以及数据处理不够标准化和自动化，最关键的是缺乏管理层对用数据指导决策的认识与热情。在这些方面美国企业已在最近十年中有了可观的改进，使得它们能从容地驾驭

大数据运营的机会。

在高科技革命加速国际竞争各国面临洗牌的重要时机,政府需要对大数据时代具有深刻的认识;企业需要具有大数据实践经验以避免和少走弯路;大众需要相关知识帮助它们在大数据时代做出最佳个人决策。

目前中国图书市场如雨后春笋般涌现出大量关于大数据的著作或译作,但对于如今的企业作为一个整体在大数据方面面临的前所未有的挑战和机遇还没有清晰的认识,从跨行业层面来看,企业管理和商业智能分析人员还存在盲点、偏见,甚至是一些误解。作为《大数据与商业分析》原作者之一,我感到这本书会对中国企业和政府在大数据学习和应用起到很重要的启发和指导作用。

这本《大数据与商业分析》是从事大数据研究和应用的多名一线专家共同努力下的力作。它涵盖了不同的视角,既有学术性讨论,又有企业应用总结;既有单一行业内的实践洞察,又有跨行业的“宇宙观”;既有涉及IT和平台技术的前沿知识,又有专门聚焦于数据和分析的评论。此外,书中的很多案例又正是中国企业转型期需要的各类问题的综合解答。因此,我向中国的读者特别推荐这部群体智慧结晶的著作,并衷心希望它能为大数据在中国企业的广泛和成功的应用作出贡献。

最后,感谢上海商业发展研究院对本书翻译、校对和出版的大力支持。在此还要感谢上海交通大学的朱文康、张轶、雪洋、包芊颖、杨宇弘、陶文武、张译文、刘奇和罗群为此书的翻译投入的大量的时间和精力。

程杰(Jie Cheng)

2014年8月于美国旧金山



序

乔·拉古纳(Joe LaCugna)博士
星巴克咖啡公司企业分析和商业智能部

大数据和智能分析的前途和潜力已经通过其能够带来的更好的决策和更强的经营绩效而被人们所熟知。然而,好的想法却很少能够转变为现实,沉重的历史意味着失败的实践和过时的处理方式仍在继续。有一些组织机构,它们因自己具备一个充满活力的思想交流平台而骄傲。即便如此,能够将数据和见解转变为较好的经营绩效对它们中的高级主管而言也是一个急迫的、富有战略意义的挑战。

一个组织机构如何从具有丰富的数据转变为具有丰富的见解,并且将那些最好的见解付诸实际?大数据和精确的分析都不能保证组织机构能基于那些分析学教授提出的见解来做出较好的决策。部分分析学家的工作直接影响了经营绩效,另一部分分析学家的贡献却很少。而深刻的洞悉和庞大的数据集使得这一影响很少具有差异。开发灵活的、可变通的方法来证实那些最好的见解,并将它们融会贯通,与此同时避免慢吞吞的挖掘数据或是“分析瘫痪”时出现的时间陷阱,这是一种全新的、关键的执行能力。

一、信息超载和一个转换工作

怎样才能使数据、决策以及效果更加紧密地联系在一起?1971年,诺贝尔奖获得者赫伯特·西蒙(Herbert Simon)首次发现了一件重要的且具有讽刺意义的事情:当数据较为丰富时,组织内的高级决策

制定者的时间和精力却是最稀缺、最珍贵的资源。我们可能从未有过足够的时间,但是我们肯定会有足够的数据库。在高管普遍的模糊与分析学家预测和技术明显的精确性之间,同样有一个困难的转换工作。分析学家的见解和指示经常无法与重要战略决策中本身的不精确性、非结构化以及具有时限这些性质挂钩,而高管有时无法完全领会到那些可以用抽象算法表达出来的机遇和风险,并且分析学家们经常不能成为这些高管们值得信赖的顾问。大多数高管认为,模型和分析只是高度复杂模式的还原和简化,而有时这些模型提供的是过度简化的夸张描述,并非有用的精度。总之,尽管高级分析技术对于制定决策正变得越来越重要,精明的高管却强调,数学和模型只有在亲身经历、对一个领域有深刻的了解、具备均衡的判断之后,才是最有价值的。

二、数据驱动分析的限制

数据越多,反而会使决策制定者更头疼,而非更轻松,因为有时大量的数据会推翻某些梦寐以求的观点,并且对一些完善的做法给出调整的建议。智能分析同样会忽略某些影响,或是在没有影响的地方建立问责制。但有时,正如安德鲁·朗(Andrew Lang)所言,统计学的应用就像醉汉利用街道旁的灯柱一样——是用来倚靠而非照明的。有时,正如在房地产、抵押贷款银行和国际金融方面的危机所证实的一样,分析学家对于他们的模型和算法太自信了,而忽视了“黑天鹅”事件以及所谓的结果“非正态”分布发生的可能性。我们常常忘记未来和不久之前发生的事是大不相同的,而对未来如何变化知之甚少。马克·吐温忠告我们:“历史不会重演,但总会惊人地相似。”统计学和分析学家几乎不可能看清未来将何时重现或是以散文的形式撰写。

一些最为重要的组织决策并不适合传统的分析技术,也不能以现有的数据库来有效地描绘。比如说,创新投资或是与其他组织合作伙伴间的决定事先很难评估,有限的数据库和不可度量的风险会推翻这样的战略决策。当然,没有数据库来支撑这样的非结构化的战略决策并不意味着这些决策不好,只不过判断和观察是更好的决策指南。

许多组织会发现,较为明确地分清各种类型的决策、授予谁来制定它们以及如何制定将会是有好处的。不少诸如人员配置、库存计划,或是后台操作这样的日常工作和战略决策,可以通过对数据的信赖以及自动处理决策制定过程中的关键部分来加以改善。比如说,可以利用最优化技术。这些规则和决策经常由现场经理或者总部员工来实施,并不需要由高管来负责。当含义不明确、缺少先例或是交易不能很明确地被量化时,更多重要的决策确实需要高管们来参与制定。在这些混乱的、可能造成严重后果的情况下,当未来的状况与近期可能截然不同,预测模型和最优化技术的价值就显得十分有限了。其他一些定性的分析技术,诸如现场调研或是聚焦群体,以及诸如情绪分析和社交网络图等崭新的分析技术,能够提供可行的、近实时的见解,这些不太可能通过模拟或是大规模的数据挖掘来实现,但却是强大的诊断途径。

即便在高不确定性、高风险的情况下,当判断和经验是最有用的指南时,高管们往往会通过在他们空气稀薄的处于角落的办公室外征求意见而受益。大量的学术和应用研究显示,通过不同阶层、不同工资等级以及不同纪律条例制定出的决策通常比那些尚未经过值得信任的专家们审核而制定出的决策要好。当新投资的商业案件被置于不切实际的乐观假设中,或者当一个经理关注她的商业个体中的积极影响而非整个机构时,那些觉得自己身陷不全面、有偏见的信息“泡沫”中的高管们很有可能会误入歧途。为了减少这种博弈和次优化风险,通过非传统的途径发现异议就显得具有重大价值,同样,这也是一种重要的方法。在具有战略重要性的,同时又模棱两可的情况下,定性的“群体的智慧”往往是一个比较好的制定明智决策的指引,而非盲目地依赖于大量数据分析,或是依赖那些受高管们青睐的目光短浅的、有局限性的一部分观点。优秀的分析学家同样能扮演重要的角色,因为他们对于前前后后所拥有的数据始终坚持科学方法中的严密性、条理性。机遇的目的就是避免太过寻常的评价:我们这么做是因为 CEO 让我们这么做的。

许多高管可能会遇到信息失真的问题,往往这会以贮存或者不愿意自由地、广泛地在组织间分享信息的形式出现。其无益的“双胞胎”“向上管理”也会表现为:分享那些有选择性的、已经过滤的、正面的信息,以讨好更高级的决策者。这些做法会有损决策、滋养祸患、终止学习的进行、突出不和谐因素,并且延迟学习群体的出现。过去,贮存和向上管理是合理的并且有时会被批准;而现在,领导力意味着,坚持不同级别间清晰地、开诚布公地共享信息是一种新的普遍规律。这在见解与现有观点和行为相一致以及与数据和分析相冲突时都是正确的。有冲突的意见和相互竞争的利益可以通过公开它们、演说争辩以及认识到它们能改善决策而得到最好的解决。

三、逐步形成一个数据驱动的学习文化

有些机构依赖于来之不易的经验、难忘的事件以及其他令人满意的探索结果,而数据驱动决策制定的法则对于它们来说是一种全新的改善企业绩效的方法。正如本书中部分章节所指出的那样,在一个公司文化中强行灌输一种分析方法有时是不太可能的。学会通过分析学来改进企业绩效通常是零散的过程,是一个主题接着一个主题、一个过程接着一个过程地分组完成的,往往时断时续。但如果没有强大的执行力、群众的拥护以及关注,它将无法实现——而建立一个数据驱动的满意决策的意愿,甚至会延迟回答重要商业问题的进程。

那些专注于加强分析学的影响,希望它引起人们关注的高管们应该注意机构调整的规模和范围,因为这种调整关系到能否把握数据驱动决策制定的价值。这可能需要涉及文化上的改变,比如提高知名度、资历以及分析团队在全公司所享有的关注度,也许这将意味着以其他项目和团队为代价,对分析学领域的额外稀缺资源进行投资,就像宝洁公司近几年所做的那样,这种做法至今为人称道。同样,决定最佳的组织分析才能也需要反复地尝试探索:比如它们是否属于信息技术(IT)的一部分、是否应归入业务单位、是否应成为总部“卓越中心”(center of excellence)的核心,或者是否应分散到全球?构建这些

功能需要花费一定的时间,也需要一个变通的方案,因为到目前尚未出现加快这方面成熟发展的一贯有效的做法。同样,分析的优先级和投资项目也会因公司而异,因此对于高管们而言,明显需要做的事情是决定第一优先级的分析目标、如何对数据和分析学家们进行分配和组织,以及在其公司内部如何形成决策。

四、掌控组织的复杂性并不简单

本书的部分章节提供了一些有用的案例研究、技术路线的经验教训,以及一些“这样做,避免那样”的指导。然而,制定好的决策有很多方法,并且决策制定是高度特质的,需要依赖于具体的环境。比如:一个组织适用的体制、方法在另一个组织中就可能不适用,即便是在市场中同一家企业里工作的同一类人都不一定有相同的特质。这具有相当大的讽刺意义:我们知道,强大的分析能力能够改善经营绩效,但我们对于组织机构构建这些能力的最佳方法,尚未有一个严谨的理解。关于如何能最有效率地、以最大的影响力构建那些能力,目前鲜有案例可循。

通常,明智的决策需要的不仅仅是熟练的分析,组织学习的技术可能比庞大的数据更重要。高效能的团队会认识到他们对问题的偏见,从而有建设性地做出否决、综合各种反对的观点,并且比其他团队学得更好更快。相对来说,学习速率十分重要,因为比竞争者拥有更快速的学习能力有时被认为是持续竞争优势的唯一来源。还有一个相应的、至今未得到好评的组织技术:一个公司忘记过去的的能力。忘记过去确实重要,因为对现状的过度负责会限制所考虑的选择范围、影响创新,并且会死死地记住那些已经认为是理所应当的惯例。这些“核心刚性”对于一个组织的“核心竞争力”而言是不受欢迎的负面因素,并且很难根除,尤其是那些已取得成功的企业。一些非常成功的企业一次又一次地、在一个又一个市场上被由新兴挑战者推出的产品和技术所取代。由于被过去的成功和之前的投资冲昏了头脑,这些现有的公司具有盲目的自信:它们认为过去适用的东西放到未来依然适

用。简言之,尽管大数据和复杂分析对于更好地进行决策正变得越来越重要,但以数据为基础的远见将带来更好的业务绩效,如果具备了有效的团队学习技巧、比他人学得快的能力,以及强烈的挑战现状的意愿,这种可能性将大大增加。

当高管们考虑数据驱动的决策制定的方法时,他们面对的是不止一种客观存在的约束:普遍缺乏深入分析的能力,并且我们不能简单地靠培养足够的能力来弥补这个缺陷。对于这种能力缺陷的评估因人而异,但是,考虑到在正规教育中所花的时间,以及在商业领域亲身体验的重要性,认为这个缺陷在不久的将来可以弥补的说法是毫无根据的。而在商业领域的亲身体验可以让分析学家们成为值得信赖的顾问。讽刺的是,谷歌公司的哈尔·范里安(Hal Varian)相信,统计学家将享受“未来十年最性感的工作”。那些将强大的技能与扎实掌握的业务问题相结合的分析学家将会有最佳的选择,并能通过解决最有趣的问题发现最优秀的企业。

另外,还有一个新出现的共识,那就是:许多认为自己已经属于“数据驱动型”的经理和高管们需要变得更加“数据驱动”,并且需要更深入的分析技能,来更加细致入微地对他们的顾客、竞争者以及新出现的风险和机遇进行揣摩。就像MBA已经成为进入管理层的敲门砖一样,人们也越来越期望高管们对于研究方法和分析技术具备更加深入的了解。这种全新的必备技能并不是去开发高端的预测模型,或是很有底气地谈论什么置信区间,而是能够批判地评估别人所提出的见解:核心假设是什么?什么情况能使它们不成立?边界条件是什么?是A导致了B发生还是反之?一系列的结论在统计意义上都成立吗?调查结果在规模上是否可操作、可重复?Cronbach置信系数 α 为5%是好还是不好?

抓住大数据和智能分析潜在的价值并不是件轻松的事。在许多行业、市场和技术间,少数公司已经能够通过培养组织的能力,发掘有价值的见解并依据其中最好的见解采取行动,从而为它们自己创造竞争优势。它们中的大多数都是大家耳熟能详的,如,星巴克、沃尔玛、

联邦快递、哈瑞氏(Harrah's)集团、艾派迪公司等,而且强有力的证据表明,这些投资从金融上而言是精明的,是相当有策略的,并且有竞争力价值。如果没有强大的、持续的执行赞助,这很难实现。这些杰出的企业在培养可变通的分析能力方面进行投资,同时还在分析学家和经理群体方面投资,这些人可以整理数据、制定决策,并影响高管。这些企业并不满足于早期的成功,它们积极进取,开创全新的分析技术,并且将更加严格的方法应用于其更多的业务操作中去。采用并延伸这种数据驱动方法被称为“未来的一切”。如今,这个机遇正以同样的方式被其他公司的高管们所抓住:通过严谨的分析和较优的决策来挖掘他们信息资产的价值。除了更有效率的业务操作,这同样是一条前途光明的道路,可以识别新的市场机遇、提出竞争的弱点、赢得更多忠实的顾客,并且改善接近盈亏预算线的市场结果。

大数据是一场大交易,而高管们的判断以及明智的组织机构内的学习习惯将使得大数据变得更为重要。



前 言

那么,为何要写这本《大数据与商业分析》呢?是因为2012年3月29日白宫科学与技术政策办公室举办的一次会议,声称:对于大数据和相关分析进行研究和发展的可以奖励2亿美元吗?是因为像KM-World公布的那样,研究大数据所得的收益将由2011年的50亿美元增长到2017年的500亿美元吗?抑或仅仅是因为“3V”在我们脑海中已经根深蒂固:数据的容量(volume)、数据的多样性(variety)以及数据的传播速度(velocity)?

随着诸如网络安全、应急管理、医疗保健、经济金融、交通运输等各个领域大量数据的使用,对于组织机构而言,及时有效地弄清数据和信息的意义来改善决策制定的过程将变得尤为重要,这就是分析学的用武之地。研究表明,到2018年,单单美国就将有14万~19万商业数据分析学家的短缺。而为了弄清诸如结构化、非结构化、文本、数字、图像及其他各式各样数据的含义,这些分析学家应该对机器学习、高级统计技术,以及其他预测分析学等有所掌握。

就更好地理解组织机构的案例研究、发展趋势、主要问题、挑战以及与大数据和商业分析相关联的技术而言,本书专为填补这一领域而量身定制。我们十分高兴地能邀请到一些杰出的专家学者和机构为本书赐稿。而来自行业、政府、非营利机构以及学会所写的章节提供了有关大数据和商业分析这块新兴领域的许多有趣的观点。在星巴克咖啡公司负责监管公司分析和商业智能的乔·拉古纳博士,根据他在这块领域——无论是在行业还是在学会中多年工作的经验,为本书作序,对此我们同样非常高兴。

没有约翰·维查来克(John Wyzalek)以及他的两个同事泰勒(Taylor)和弗朗西斯(Francis)的深思熟虑,编著本书的成就也不可能实现。我尤其还要感谢我的家人、学生和在美国马里兰大学学院的同事们与我进行专业的交流,你们使我对该领域的理解更加深刻。

祝好!

杰伊·利博维茨, DSc
欧坎德管理与技术讲席教授
马里兰大学学院研究生院
Adelphi, Maryland
Jay.liebowitz@umuc.edu



关于编者

杰伊·利博维茨博士是马里兰大学学院研究生院 Orkand 管理与技术讲席教授,之前曾任约翰霍普金斯大学凯瑞商学院的教授。据2010年1月《知识管理期刊》所载,他在全球11 000名知识管理方面的研究员和从业者中排名前十,并在全球知识管理策略领域排名第二。在约翰霍普金斯大学,他是竞争情报毕业证书的创始项目总监,也是商业项目信息和电信系统的最高负责人。在该项目中,他吸纳了30多个行业机构、政府机关以及非营利机构加入到他的团队。

在进入约翰霍普金斯大学之前,利博维茨博士是美国国家航空航天局戈达德太空飞行中心的第一位知识管理高管。在这之前,利博维茨博士担任过马里兰大学巴尔的摩分校信息系统的 Robert W. Deutsch 特聘教授、乔治华盛顿大学管理科学教授,以及美国陆军战争学院人工智能讲席教授。

利博维茨博士是《专家系统与应用国际性学报》(由 Elsevier 出版)的创始人及主编,根据最近的 Thomson 影响因素来看,它在全世界有关智能系统、人工智能方面的杂志中排名第三。2011年,该杂志刊登的文章在全球范围内有180万的下载量。利博维茨博士是富布赖特学者(Fulbright Scholar),享有电气和电子工程师学会(IEEE)——美国联邦通信委员会执行伙伴的美誉。另外,他还享有年度计算机教育专家(国际计算机信息系统协会,IACIS)的称号。他已出版了超过40本著作,在知识管理、智能系统以及IT管理方面发表了许多篇期刊论文。他最近出版的著作有:《知识储存:策略和方法》(Taylor & Francis, 2009)、《公共卫生知识管理》(Taylor & Francis, 2010)、《知识管理和

电子学习》(Taylor & Francis, 2011)、《知识管理之外:领导必须了解的东西》(Taylor & Francis, 2012),以及《知识管理手册:合作与社交网络(第二版)》(Taylor & Francis, 2012)。2011年10月,因利博维茨博士指导的学生研究论文在IACIS年度大会上被评为最佳,国际计算机信息系统协会授予其“杰出学生研究奖”。如今,他在全球范围内从事讲座和咨询的工作,可以通过 jay.liebowitz@umuc.edu 联系他。

目录

第 1 章 通过大数据管理企业	1
1.1 引言	1
1.2 挑战	2
1.3 正在发生的现象	2
1.4 社交网络	3
1.5 个性化服务和社群	3
1.6 科技驱动和商业分析	4
1.7 从数字到大数据	4
1.7.1 我们是如何走到这一步的?	4
1.7.2 为什么它如此重要?	9
1.7.3 科技的升级换代如何满足需求?	10
1.8 重新定义组织结构	11
1.8.1 关于重新定义	11
1.8.2 一些挑战	12
1.8.3 一些机遇	12
1.8.4 重塑的机会	13
1.9 为大数据时代做好准备	16
1.9.1 科学、技术、工程和数学	16
1.10 一些建议	17