Big Data Revolution:
How Data Science Changes the World

缺少数据资源　无以谈产业
缺少数据思维　无以言未来

大数据时代的历史机遇

产业变革与数据科学

赵国栋　易欢欢　糜万军　鄂维南　著

清华大学出版社

大数据时代的历史机遇

——产业变革与数据科学

赵国栋　易欢欢　糜万军　鄂维南　著

清华大学出版社
北京

内 容 简 介

大数据正以前所未有的速度，颠覆人们探索世界的方法、驱动产业间的融合与分立。本书力图系统、全面的阐述大数据在社会、经济、科学研究等方方面面的影响，或许可以帮助大家澄清一些认知误区，有助于大数据在各行各业落地生根。全书分为三大部分，第一部分重点讲述大数据时代产业发展的三大趋势以及驱动产业融合、升级、转型的根本因素，并给出践行大数据的最佳范式。第二部分首次完整阐述“数据科学”的基础性价值，论述数据科学对科学研究、社会研究、产业发展的影响，并提出数据科学的教育体系。第三部分全景式的介绍重点国家、经济体、新兴企业在大数据领域取得的进展，展示一幅真实的大数据图景，把判断留给读者，看谁拥有未来！

本书面向资本市场、产业界和学术界，成为链接三方的纽带。有助于投资人了解产业趋势、评估公司价值；有助于产业界确立公司战略方向；有助于学术界了解产业需求，促进产学的协作。

图书在版编目（CIP）数据

大数据时代的历史机遇：产业变革与数据科学 / 赵国栋等著. —北京：清华大学出版社，2013

ISBN 978-7-302-32535-2

Ⅰ. ①大… Ⅱ. ①赵… Ⅲ. ①数据管理-研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2013）第 104803 号

责任编辑：夏兆彦
封面设计：胡文航
责任校对：胡伟民
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：清华大学印刷厂
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×230mm　**印张**：26.5　**插页**：2　**字　　数**：500 千字
版　　次：2013 年 6 月第 1 版　　**印　　次**：2013 年 6 月第1 次印刷
印　　数：1～15000
定　　价：49.00 元

产品编号：049306-01

目录
Contents

第一章　大数据概述

大数据是“在多样的或者大量的数据中快速获取信息的能力”，其关乎国计民生、产业兴衰、公司存亡，不可不察。

第一部分　产业大势

第二章　大数据时代已经到来

资本市场、产业界、学术界、政府都在紧锣密鼓地行动，四方联手推动 2012 年成为大数据时代的元年。

第三章　数据成为资产

大数据时代公司的价值与其数据资产的规模、活性成正比；与其解释、运用数据的能力成正比。

第四章　大数据颠覆媒体行业

传统平面媒体业正在经历历史上最严重的倒闭浪潮，取而代之的是新兴的互联网媒体公司。以谷歌为代表，他们以数据资产为中心，创造了迄今为止最完美的商业模式之一。

第五章　大数据冲击金融行业

比尔·盖茨曾说："传统银行若不能对电子化作出改变，将成为21世纪行将灭绝的恐龙"，从小微信贷、众筹、互联网金融等新兴的金融服务模式来看，金融业不得不经历痛苦的嬗变过程。

第六章　大数据加剧产业的垂直整合趋势

大数据时代，消费者真正登上了舞台中央。哪些越靠近最终消费者或者用户的公司，在产业链上就拥有越来越大的发言权。产业生态将围绕消费者重构。

第七章　泛互联网化是发挥大数据价值的最佳范式

那些仅仅拥有产品，无法形成终端、平台、应用、数据一体化的公司，将难逃被颠覆的命运。泛互联范式成为累积数据资产、发挥数据资产价值的最佳范式，也是构成大数据思维的重要组成部分。

第八章　大数据掀起的企业组织变革

大数据首先是一种思维方式，必须融入到企业的每一个毛细血管中。运用大数据思维必将审视企业与客户的关系，企业的战略、组织、文化都将因大数据而彻底改变。

第二部分　数据科学

第九章　数据科学

大数据在科学领域的表现是数据科学的兴起，数据科学将逐渐达到与其他自然科学分庭抗礼的地位。用数据研究科学，科学的研究数据。

第十章　数据技术：当前进展及关键问题

欲工其事必先利其器。促进大数据在各行各业落地的重要因素，除了建立大数据思维以外，必须掌握新兴的处理技术。需要重新审视企业的软件开源策略、数据处理技术、人才培育计划。

第三部分　全景扫描

第十一章　国家选择

开放、共享是大数据时代的核心精神。但是于政府而言，大数据是把双刃剑，它既能促进政府开放、透明，又能帮助加强集中管控。选择考验智慧！

第十二章　巨头碰撞

新兴的产业巨头凭借独一无二的数据资产，正在重新定义产业生态和竞争格局，老牌科技公司沦为看客，围观的传统产业逐一被颠覆。

第十三章　创新凶猛

新兴的大数据公司如雨后春笋，观察他们的成长，我们才深深体会到产业的脉动、变化的节奏和演变的方向。毫无疑问，他们正在重新定义未来。

附录　大数据发展大事记

后记

参考文献

引子

大数据总统奥巴马

2012 年 8 月份，美国总统大选正如火如荼。出人意料的是，奥巴马总统的数据团队要求他去一家叫 Reddit 的新闻网站去回答问题。对许多人来讲，Reddit 是一个陌生的名字，总统的高级助手们对它也不甚了解。但是来自数据团队的回答却非常简单："因为我们需要动员的一些人，经常在 Reddit 上。"

这仅仅是选战过程中一件毫不起眼的数据决策案例。事实上，奥巴马的数据团队非常神秘、低调，但其触角又无处不在，几乎左右了整个大选，他们被内部人士戏称为"核编码"。他们创建了单一的巨大系统，可以将从民调专家、筹款人、选战一线员工、消费者数据库、以及"摇摆州"民主党主要选民档案的社会化媒体联系人与手机联系人那里得到的所有数据都聚合到一块。这个组合起来的巨大数据库令奥巴马的数据团队工作极富成效，令人惊叹[①]。在这个组合的数据库中，每个选民甚至被精确地划分为 1000 多个特点，通过建模和算法分析，系统能为每个选民找出

① 英文原文参见 CNN 网站 http://edition.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team。

一个最能说服他的理由；每晚进行 6.6 万次模拟选举，在个体水平上，计算出奥巴马在任何一个摇摆州的胜率。事实上不仅如此：

他们建立的模型能够预测谁会在线捐款。

他们用来网上筹款的邮件，也充分利用了数据收集和分析。

他们借助模型帮助奥巴马筹集到创纪录的 10 亿美元。

他们帮助优化电视精准投放广告的模式。

他们创造出了摇摆州选民的精细模型。

他们计算出第一夫人发的拉票邮件在春天最受欢迎。

他们利用数据来详细分析关键州的选民。深入分析各个族群的选民在任何时刻的趋势。在总统候选人的第一次辩论之后，他们分析出哪些选民倒戈，哪些没有。

他们利用熟人效应，开发 Facebook App 拉票。

他们为竞选团队购买广告提供决策参考。

他们通过一些复杂的模型来精准定位不同选民，他们购买了一些冷门节目的广告时段，而没有采用在本地新闻时段购买广告的传统做法。广告效率相比 2008 年提高了 14%。

他们导致经验主义的竞选专家的作用急剧下降，能够分析大数据的量化分析专家和程序员的地位却大幅提升。

他们让政客们，尤其是对手知道政治领域的大数据时代已经到来。

一瓶茅台酒的旅程

消费者最头疼的恐怕还不是茅台酒的价格，而是能否买到货真价实的茅台。“道高一尺魔高一丈”，茅台历来的防假手段，除了推高茅台酒瓶的回收价格以外，似乎并没有真正让消费者放心。

为每一瓶茅台建立“档案”，消费者可以轻松方便地查询到任何一瓶茅台酒的档案材料，是防假的终极解决之道。每一瓶酒都有一个独立的“身份证号”，铭刻到酒

瓶上，在信息系统中记录下从灌装到出厂、运输、批发、零售所有环节的信息。人们只要把“身份证号”传输到网站一查，真伪立辨。这个办法看起来容易，但是真正实施，我们立刻会被淹没在大量的数据之中。

不仅仅是茅台，中国目前所有食品面临“安全、卫生”的大难题。如果能把茅台酒的做法推而广之，无疑是全民之福。但是这些海量的数据记录，对传统的信息处理技术提出了巨大的挑战。

茅台的故事，其实可以引发管理理念的变化。这是管理日益精细化的具体体现。原来“茅台们”的管理都是按照生产批次，通常认为同一个生产批次的产品，是没有差别的。现在的管理理念则不同，要求对每一件单品实行差别化管理。

城市治理中，也在发生同样的事情。小到每一个下水道井盖都被仔细编号、追踪。这当然另我们的生活更加便利，但产业界首先需要应对的则是大数据的挑战。

导读：

1. 大数据正以前所未有的速度颠覆人们探索世界的方法，引起社会、经济、学术、科研、国防、军事等领域的深刻变革。
2. 数据成为资产、产业垂直整合、泛互联网化是大数据时代的三大发展趋势。数据资产成为和土地、资本、人力并驾齐驱的关键生产要素。围绕数据资产可以演绎跌宕起伏的产业大戏。
3. 数据科学应运而生并将成为科研体系中的重要组成部分，逐渐达到与自然科学分庭抗礼的地位。数据科学既可以推动数学、计算机科学、统计学、天体信息学、生物信息学、计算社会学等学科的发展，又能够助力产业界升级转型。
4. 需要在宏观尺度拓宽大数据视野、建立完整的大数据思维；正视普遍存在的三大数据治理问题（数据割据、数据孤岛和数据质量）及人才短缺的现状。

第一章

大数据概述

大数据是“在多样的或者大量的数据中快速获取信息的能力”。

——笔者

大数据，事关国计民生、产业兴衰、公司存亡，不可不察。信息科技经过 60 余年的发展，数据（信息）已经渗透到国家治理、国民经济运行的方方面面。经济活动中很大一部分都与数据的创造、传输和使用有关。2012 年 3 月，奥巴马公布了美国《大数据研究和发展计划》[①]，标志着大数据已经成为国家战略，上升为国家意志。

国家竞争力将部分体现为一国拥有数据的规模、活性，以及解释、运用数据的能力。国家数字主权[②]体现为对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间[③]。没有数据安全，也就没有国家安全。华为、中兴开拓美国市场受挫，就是非常明显和清晰的信号。美国政府对自家数据安全的重视程度，已经到了不能让任何外国信息基础设施产品供应商染指的地步。华为此前一直希望通过竞标和并购等方式进入北美市场，多年来未能如愿。2008 年，华为与贝恩资本联合竞购 3COM 公司，却因美国政府阻挠未能成行；2011 年，华为被迫接受美国外国投资委员会的建议，撤消收购 3Leaf 公司特殊资产的申请；同样是在 2011 年，美国商务部阻止华为参与国家应急网络项目招标。

再看美国国防部立项的几个大数据项目[④]：多尺度异常检测（ADAMS）项目，解决大规模数据集的异常检测和特征识别的问题；网络内部威胁（CINDER）计划，旨在开发新的方法来检测军事计算机网络与网络间谍活动，提高对网络威胁检测的准确性和速度；Insight 计划，主要解决目前情报、监视和侦察系统的不足，进行网络威胁的自动识别和非常规的战争行为……参见附录四。其他部门包括国土安全部、

① 《大数据研究和发展计划》原文网址：http://www.whitehouse.gov/blog/2012/03/29/ big-data-big-deal，中文译稿参见本书附录四。

② 通过搜索引擎，并未发现其他文献强调“数字主权”。之所以采用“数字主权”，而非“数据主权”，主要因为构成信息科技的基础是“0”、“1”两个二进制的数字。所有的数据在本质上都是“0”、“1”的排列组合。

③ 参见国金证券大数据系列报告第三篇《以数据资产为核心的商业模式》，第 1 页。

④ 原文参见 http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data _fact_sheet_final_1.pdf。

能源部、卫生和人类服务部、国家航天总局、美国国家科学基金会、美国国家安全局、美国地质调查局纷纷推出大数据项目。奥巴马指出："通过提高我们从大型复杂的数据集中提取知识和观点的能力，加快科学与工程前进步伐，改变教学研究，加强国家安全。"

产业层面，大数据技术虽然发源于信息科技，但其影响已经远远超出信息行业。数据已经存在于全球经济中的每一个部门，就如固定资产和人力资本等生产要素一样，如果没有它许多现代经济活动根本就不会发生。笔者观察到一些新兴的互联网公司，利用新技术大规模地收集数据，预判客户行为，然后在不同的行业纵横捭阖。它们剑锋所指，现代服务业无不受其锋芒所迫，或随波逐流，或奋起反击。但缺少数据资产、缺少强大的数据分析能力，这类公司无疑处在被颠覆的边缘。笔者也看到传统行业的公司，数十年如一日坚持积累当时被视作"废料"的数据，现在回头审视这些数字化的资产，居然一跃成为人类的宝库。凭借独一无二的"数据资产[①]"，公司进入相关行业，易如反掌。

当笔者回头审视产业的起起伏伏时，就会发现决定产业兴衰的根本性因素已经不是一城一地的争夺了。土地、人力、技术、资本这些传统的生产要素，甚至需要追随"数据资产"重新进行优化配置。封建时代，往往是裂土封王，权贵都是大地主；工业革命后，制造业巨子成为偶像；资本市场化后，受到追捧的是拥有大量钱财的投资家。但是在大数据时代，"数据资产"成为最重要的生产要素，拥有大量数据资产的人，已经成为美国总统的座上宾[②]。

产业的分分合合，一直是资本市场非常喜欢的故事。不管是分拆也好，整合也罢，资本市场都有钱赚。以往产业的整合基本围绕产业链展开，要么向上游扩展，要么向下游兼并。但是在大数据时代，人们看到的商业图景是围绕"数据资产"拉

① 数据成为资产，参见国金证券大数据系列研究报告《大数据时代的三大发展趋势及投资方向》。

② 美国总统奥巴马于 2011 年 2 月 17 日与多名科技界领袖共进晚餐。总统左侧是苹果公司创始人史蒂夫·乔布斯，右侧是 Facebook 的创始人马克·卡克伯格。

开产业并购的大幕。谷歌所有的收购或者推出的新产品，都是为了增加数据资产的“维度”和“活性”[①]。所有观察公司发展、产业未来的机构或者个人，如果忽略“数据资产”，或者对“数据资产”认知肤浅，必将导致错误的判断。大数据将是决定产业未来的战略性资产。未来产业间的整合并购，将会在很大程度上围绕“数据资产”展开争夺。

企业家、投资人、咨询顾问、分析师，必须要从战略层面思考大数据对产业、公司的影响。2012 年初，笔者曾经和恒安国际的董事会一道交流大数据对制造业的影响。会上许连捷[②]总裁说：“在大数据时代我们收集数据，研究消费者行为，推出新的产品，改善供应链，降低库存。一句话就是把大数据融入到经营中去。也许有可能把库存降到近乎‘0’的水平。”所以，我们谈大数据，首先是思维方式的问题，要建立全面、系统的大数据意识，其次才是落实到公司战略。大数据对公司的影响是多方面的，涉及组织、文化、流程、技术等。本书第八章将专门详细论述大数据对公司组织结构的影响，在此不赘言。

具体到中国信息产业，发展速度一直落后于国外的巨头，长期处在产业链的末端，赚取刀片一样的利润，积累到最后发觉只形成了简单可替代的“中国制造”而非具备革命性创新性的“中国智造”。国家拿出大笔资金扶持上游环节的拓荒者，如 CPU、操作系统、办公软件，但是相关领域国内外的差距过于遥远，也缺少大规模的商用市场，花了国家的钱，却鲜有在商业上大获成功的先例。但是在新兴的大数据处理领域，中外公司几乎站在同一起跑线上。中国作为数据的巨大产生国，有着更广阔的应用空间。比如，中国移动、工商银行、淘宝，已经具备世界级的产业应用环境。有业内人士表示，单纯考虑狭义的大数据处理技术（如 Hadoop、MapReduce、模式识别、机器学习等），中外差距仅有 5 年左右。如果考虑数字资产规模以及利用的技术，中外差距更多体现为意识上的差距。美国在数据开放、跨部门共享方面做出了表率，而我国对大数据的价值和应用，政府、学术界、产业界

① 维度、活性等概念将在数据资产章节详细说明，是数据资产评估模型的一部分。
② 许连捷现任中国民间商会副会长，泉州市工商联主席，第十届全国工商联副主席。