

DATA REVOLUTION

数据革命

大数据价值实现
方法、技术与案例

范煜◎著

宏观和微观、人文和技术、启迪思想
和关注实用并举

既适合宏观层面的领导启迪思维，提出工作目标，又适合
微观层面的执行人员找到实现的方法和路径

清华大学出版社



数据革命

大数据价值实现
方法、技术与案例

范煜◎著

DATA
REVOLUTION

清华大学出版社

北京

内 容 简 介

在信息技术革命之后,我们将迎来数据革命。在大数据的概念、性质和价值已得到政府和社会的认可之后,大家关注的是数据如何获取,以及有了数据以后如何挖掘数据的价值。仅适合特定行业、满足特定需求的技术不足以应对一场革命,大数据不但是超出计算机软硬件处理的能力,更是超出人类的认知能力。只有实现对数据的认知,利用数据辅助决策,才是适合不同行业数据价值实现的通用手段。本书描述了数据革命的起源、实现的思路、所用的技术和要达到的目标,针对当今社会热点描述了在数据时代的应对之策。

本书宏观和微观、人文和技术、启迪思想和关注实用并举,既适合宏观层面的领导启迪思维,提出工作目标,又适合微观层面的执行人员找到实现的方法和路径。本书介绍的理论和技巧均可在智慧城市、智能制造领域实际使用。

本书适合政府、企业决策者和 CIO,及其他对大数据应用感兴趣的人阅读。

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

数据革命:大数据价值实现方法、技术与案例/范煜著. — 北京:清华大学出版社,2017
ISBN 978-7-302-46693-2

I. ①数… II. ①范… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 038741 号

责任编辑:刘 洋

封面设计:李召霞

版式设计:方加青

责任校对:宋玉莲

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 装 者:三河市金元印装有限公司

经 销:全国新华书店

开 本:170mm×240mm

印 张:15.75 字 数:231 千字

版 次:2017 年 5 月第 1 版

印 次:2017 年 5 月第 1 次印刷

印 数:1~3000

定 价:49.00 元

产品编号:072823-01



前言

自 2008 年经济危机爆发以来，从美联储开始，世界各国央行都在通过货币政策防止更大经济危机的爆发。大家同时也认识到货币与经济政策只能缓解经济危机的严重程度，争取时间，而不能从根本上走出经济危机。只有新的技术革命的发生，世界才能真正地走出经济危机。

那么下一场技术革命是什么呢？现在有很多推测，如机器人、工业 4.0、虚拟现实（VR）、3D 打印、人工智能等。如果研究一下这些技术的影响范围，就会发现它们都难以担当下一次技术革命的重任。

能把世界拉出经济危机的技术革命是怎样的呢？它应该具有以下特征：①应用领域非常广泛，而不局限在某个行业；②影响的人群非常大，会产生大量新兴的职业；③和日常生活息息相关，具有很大的渗透性。

显然，我们正处于信息技术革命的末期，越来越多的迹象表明，信息技术革命已经完成了它的历史使命：①英特尔公司已经走上了巅峰，摩尔定律即将失效，难以开发出更高性能的 CPU，即使推出更高性能的 CPU，市场需求也不大，市场普遍认为现在的 CPU 已经能满足现有需求；②智能手机经过高速发展后，市场进入饱和期，功能也基本满足需求，大家更换手机的动力减弱，高速发展的苹果公司销量停止了增长；③微软公司推出的 Windows 10 操作系统免费升级，但即使使用多种强制手段，大家升级的愿望也比较弱，甚至认为即使很老的 Windows XP 系统也能满足基本需求。当初人们对信息技术革命的期望，能够做到的已经做到，不能做到的

现在暂时也遥遥无期。

信息技术革命给我们留下什么呢？无论是初期的大型计算机，还是后来的个人计算机、笔记本电脑、智能手机，或者现在很时新的物联网，都会在使用过程中产生大量的数据。这种数据分散在不同的地方，很多数据用完即扔。数据只在产生的过程中发生了作用，历史数据价值并没有被发掘。特别是物联网技术产生以后，收集到越来越多的数据，但很多数据还没有找到用处。

在计算机出现的初期，人们就对利用计算机实现决策支持抱有非常大的期待。关于决策支持系统、专家系统的研究，有一段时间也非常红火，但现在看来这个期望还没有实现。以实现决策支持为目的的商业智能领域虽然积累了许多有价值的理论和技术，但达到的效果还不能满足人们的预期目标。

综上所述，作为信息技术革命成果输出的大量数据应该是下一场革命的输入，是新革命的原料，信息技术革命未完成的使命就是下一次革命的目标。因此，下一场革命无疑应该是数据革命。

数据革命是信息技术革命的延伸，它会对社会产生巨大的影响，它完全是从另一个角度去完成信息技术革命不能完成的目标，还会对信息技术革命的产物（比如计算机网络、云计算、物联网）产生更大的需求，导致现在看来过剩的计算能力又变得不足。

数据革命的影响巨大，会深入到社会经济的方方面面。现在凡是用到计算机的地方，都涉及数据的问题。数据的处理对很多人来说都是一个崭新的领域，有很多新知识需要学习，并因此产生许多新兴职业。

目前已经出现一些与数据革命相关的事情，比如政府开放数据。现在欧美很多国家政府制定了政策，要求政府数据、有政府基金资助的科研数据对外开发，把原来保密的数据变成共享。政府建立起开放数据网站，提供可机读的数据。但是他们的工作局限在数据的供应上，对数据如何利用、如何产生价值还依赖社会的创新，并没有找到通用的方法，也没有找到通用的价值创造机制。

数据革命中的数据不一定是大数据，虽然对海量数据的认知离不开大

数据技术，但大数据只是数据的一个特例。

本书的目的不在于研究数据的获取，因为社会上已有足够的数据，无数运行的软件日夜不停地产生着新数据，无数程序员在编写着程序准备产生更多的数据。本书更多地是放眼于数据时代对数据的存储和应用，以及数据应用会产生哪些改变，这些改变包括政治的、经济的、社会的等；并且探讨了一个通用的数据产生价值的途径——决策支持，其中涉及一个被称为“鹰眼”的核心技术，这个技术的推广应用将对数据使用发挥很大作用。

利用数据进行革命的最终结果应该是：人们通过对数据的分析，解决了在经济生活中遇到的一些难题，反过来推进更多的数据的产生和存储，再进一步推动更多的信息技术产品的生产和销售，吸引更多的人从事与数据相关的工作。



目 录

1 开编故事

5 第 1 章 迎接数据革命

- 1.1 信息技术革命 / 6
 - 1.1.1 未完成的第三次工业革命 / 6
 - 1.1.2 从智力替代到辅助决策、自主决策 / 7
 - 1.1.3 三次工业革命的比较 / 8
 - 1.1.4 数据是信息革命的主要遗产 / 10
- 1.2 为什么是数据革命 / 11
- 1.3 社会需要数据革命 / 13
 - 1.3.1 发展需要资源配置均衡 / 13
 - 1.3.2 数据促进社会平等 / 14
 - 1.3.3 不均衡导致中国古代王朝更迭 / 15
 - 1.3.4 熵增原理 / 16
 - 1.3.5 中国国内市场的完善 / 17
 - 1.3.6 新的就业机会 / 18

- 1.3.7 建立社会经济运行的反馈机制 / 19
- 1.3.8 权威的信息交换平台 / 20
- 1.3.9 分享经济模式的扩张 / 21
- 1.4 从海关数据看数据价值 / 23
- 1.5 美国的启示 / 27
- 1.6 数据的价值与变现 / 30
 - 1.6.1 数据的变现 / 30
 - 1.6.2 决策产生价值 / 31
 - 1.6.3 数据的价值特点 / 32
 - 1.6.4 数据服务的商业模式 / 33
- 1.7 信息时代遗留的问题 / 34
 - 1.7.1 缺乏原始数据 / 34
 - 1.7.2 难搞的需求 / 35
 - 1.7.3 自助分析的陷阱 / 37
 - 1.7.4 难以满足的客户 / 38
 - 1.7.5 完全不一样的需求 / 40
 - 1.7.6 心有余而力不足的数据挖掘 / 41
 - 1.7.7 跳出事务处理的红海 / 43

45

第2章 认识数据革命

- 2.1 认识数据 / 46
 - 2.1.1 数据分类 / 46
 - 2.1.2 数据来源和存储 / 47
 - 2.1.3 非结构化数据 / 49
 - 2.1.4 数据处理的三个层次：产生、获取和分析 / 49
 - 2.1.5 数据比图像、视频更有价值 / 50
 - 2.1.6 数据与程序要分离 / 51

2.1.7	SQL是访问数据的通用语言	/ 52
2.1.8	需要标准并开源的数据库设计	/ 55
2.2	关于数据	/ 56
2.2.1	数据和信息的区别	/ 56
2.2.2	数据含金量	/ 57
2.2.3	用于理解大数据的小数据	/ 58
2.2.4	广义和狭义大数据技术	/ 58
2.2.5	看懂数据的认知计算	/ 60
2.2.6	数据的冷态、温态和热态	/ 60
2.3	走出大数据应用误区	/ 61
2.3.1	从个性化需求到普遍服务	/ 61
2.3.2	走出结果导向	/ 62
2.3.3	从有方向到无方向	/ 64
2.3.4	自助分析工具与自助分析系统的区别	/ 65
2.4	信息系统总体规划	/ 67
2.4.1	基于数据的规划	/ 67
2.4.2	用规划展示数据不足	/ 69
2.4.3	以市长为核心的智慧城市总体规划	/ 69

第3章 推动数据革命

3.1	数据的立法	/ 74
3.2	数据的公开	/ 75
3.2.1	对信息公开的认识	/ 75
3.2.2	政府开放数据	/ 76
3.2.3	对开放数据的要求	/ 77
3.2.4	政府主导的公共数据库	/ 78
3.2.5	科研数据的公开	/ 79

- 3.3 有时数据隐私只是借口 / 80
- 3.4 数据基础设施 / 82
 - 3.4.1 数据作为基础设施 / 83
 - 3.4.2 数据垄断的“滑铁卢” / 84
 - 3.4.3 公共数据服务与中介 / 85
 - 3.4.4 农产品交易数据的案例 / 86
- 3.5 建立数据图书馆 / 88

第4章 进行数据革命

- 4.1 数据用于决策支持 / 94
 - 4.1.1 数据分析需要统计而不是检索 / 94
 - 4.1.2 数据通过辅助决策产生价值 / 95
 - 4.1.3 两类完全不同的程序 / 96
 - 4.1.4 传统商业智能模式的沦落 / 97
 - 4.1.5 像鹰一样看数据 / 99
 - 4.1.6 数据一致性不是分析的先决条件 / 100
 - 4.1.7 从数据比较中发现价值 / 101
 - 4.1.8 保障决策者的决策思维流 / 102
 - 4.1.9 建立基于可视化数据的指挥室 / 104
 - 4.1.10 组织的决策支持流程 / 105
 - 4.1.11 宏观和微观的融合 / 107
 - 4.1.12 用过度设计满足任意需求 / 108
- 4.2 建立数据模型 / 110
 - 4.2.1 存储数据的数据仓库 / 110
 - 4.2.2 可以推导需求的维度模型 / 112
 - 4.2.3 维度模型原理 / 114
 - 4.2.4 分主题进行数据分析 / 120

4.2.5	离不开的时间维度 /	121
4.2.6	通过时间分析数据 /	122
4.2.7	空间维度直观地显示数据 /	124
4.2.8	数据的可视化钻取 /	125
4.2.9	用OLAP提升统计速度 /	127
4.2.10	数据可视化加快对数据的认知 /	129
4.2.11	用内存数据库实现实时数据分析 /	131
4.3	改变思路 /	132
4.3.1	建立基于真实数据的KPI /	132
4.3.2	为实现工业4.0建立数据基础设施 /	133
4.3.3	主动抽取数据实现数据集中 /	136
4.3.4	统计数据从报送到抽取 /	137
4.3.5	改进数据分析工作流程 /	137
4.4	适应数据分析的硬件 /	140

143

第5章 实现数据革命

5.1	数据革命的作用 /	144
5.1.1	对国家治理的作用 /	144
5.1.2	对国有企业改革的作用 /	145
5.1.3	对政府“三公”经费管理的作用 /	148
5.1.4	对“一带一路”战略的作用 /	149
5.1.5	对医疗改革的作用 /	150
5.1.6	对银行信贷风控的作用 /	153
5.1.7	对降低社会成本的作用 /	156
5.1.8	对防止欺诈上市的作用 /	158
5.2	数据革命的后果 /	159
5.2.1	竞争机制的替代 /	159

5.2.2 计划经济和市场经济的融合 / 161

5.2.3 经济危机的消除 / 162

5.3 数据革命后的技术 / 163

5.3.1 以数据检索为主的搜索引擎 / 163

5.3.2 基于数据的云服务 / 164

5.3.3 可以检索数据的浏览器 / 165

167

第6章 工业数据革命

6.1 智能制造首先要解决数据问题 / 172

6.2 工业企业数据总体架构 / 175

6.3 财务数据分析 / 177

6.3.1 四个层次 / 177

6.3.2 阿特曼Z-score模型 / 178

6.3.3 财务比率 / 179

6.4 经营数据分析 / 180

6.4.1 名词解释 / 181

6.4.2 经营数据中心 / 182

6.4.3 销售数据分析 / 186

6.4.4 毛利数据分析 / 189

6.4.5 应收款数据分析 / 190

6.4.6 采购数据分析 / 192

6.4.7 应付款数据分析 / 193

6.4.8 库存数据分析 / 195

6.5 与上市公司外部数据比较 / 197

6.6 控制数据分析 / 199

6.6.1 从工业大数据中找到故障 / 199

6.6.2 从检测大数据中发现质量问题 / 201

- 7.1 政府房产数据分析 / 206
 - 7.1.1 监控中心 / 206
 - 7.1.2 预售数据分析 / 208
 - 7.1.3 成交数据分析 / 209
- 7.2 医院管理决策支持系统 / 211
 - 7.2.1 监控中心 / 212
 - 7.2.2 医药收费数据分析 / 213
 - 7.2.3 门诊数据分析 / 216
 - 7.2.4 住院数据分析 / 220
 - 7.2.5 手术数据分析 / 221
 - 7.2.6 用药数据分析 / 223
 - 7.2.7 医疗项目收入数据分析 / 224
 - 7.2.8 大型诊断检查数据分析 / 224
 - 7.2.9 体检数据分析 / 224
 - 7.2.10 物资出入库数据分析 / 225
- 7.3 政府财政数据分析 / 227
 - 7.3.1 监控中心 / 227
 - 7.3.2 收入数据分析 / 228
 - 7.3.3 支出数据分析 / 229
 - 7.3.4 收支执行数据分析 / 230
 - 7.3.5 预算执行用款数据分析 / 231
 - 7.3.6 政府采购数据分析 / 231



开编故事

2030年，中国经济经过深度调整，借助数据革命的春风，走上了一条健康发展的道路。虽然GDP的增长率已经大大下降，但由于整个社会发展比较均衡，增长后劲十足。

中国不但拥有完善的基础设施，而且基础设施的收费大大降低，使所有生活、生产必需品在整个生活、生产成本中的占比大大降低，原来家里有空调不敢开的现象已基本不存在了，因为电力成本在收入中已经低到可以忽略不计，所以有很多人即使不在家也整天开着空调。很多人出于环保意识经常在网上呼吁节约能源，但节约能源完全是出于公益心而不是基于成本的考虑。高速公路已基本上实行免费，即使收费也非常廉价，更多的收费是用于不同的时段调节高速公路上的车流，避免高峰时期车流太大导致车速降低、通行效率下降。现在，人力资本已成为所有企业首要考虑的成本。

由于传统的高速公路、高速铁路、电力线路等基础设施建设已非常完善，所以在2030年人们的概念中，典型的基础设施已经不是这些东西，而是被称为数据基础设施的数据仓库。通过国家建立的众多的数据基础设施，无论人们在工作还是在生活中都能获取大量既准确又及时、详细的数据，从而大大提高人们的生活和工作效率，一如过去的高速公路提升人们的出行效率。

张开是西部地区一所著名高校的毕业生，读研究生的时候，在导师指

导下开发出一种先进的医疗器械，技术含量很高，市场前景广阔。他在家乡成立了一家公司，并且具备了一定的生产能力，准备到上海去开拓东部市场。

由于准备在上海住一段时间，张开需要租一套房子。张开从上大学开始一直在西部地区，对上海不熟悉，所以他只有在网上寻找相关的信息。他首先打开上海官方的租房数据网站，先研究出租房源和价格的分布。由于他决定租一套两居室的房子，所以从分析不同区域房源数量和房屋租金的分布着手，看到不同租金段的房子在地图上用不同颜色标识出来，颜色密集程度表示房源数量。根据自己的心理价位，张开觉得浦东金桥的房源比较多，房租价位也比较合适。他再分析一下租金历年随时间的变动情况，发现现在是一年中租金的低谷，而下个月由于很多毕业生要求职，每年这个时候租金都会上升，所以他认为这个月应该赶快租好房子。他查询了几套在租房网上的详细信息，比较满意，决定看一下这几套房子，并从中选择一套。

张开之所以对这几套房子的信息真实性没有怀疑，是因为这个平台是政府建立的，统计数据完全是根据房东签约后缴纳税收的数据进行统计，所以数据的准确性毋庸置疑。出租房源的信息虽然是由房东提供通过中介发布的，但中介对信息的真实性负有完全的责任，因为根据国家数据安全法规定，如果发布的信息不真实或者失去时效，一旦租房者为错误或失效数据付出成本，则信息发布者需按租房者为此实际付出成本的十倍进行赔偿。也就是说，如果张开为了租房从西部地区飞到上海而没有租到房，那么信息发布者要按来回机票的十倍价格进行赔偿，或者当天由于信息不真实没有租到房子而住在宾馆，这个宾馆房费的十倍要由信息发布者来承担。如果张开选择入住一家五星级宾馆，即使耽误一天赔偿额也是挺高的。所以信息发布者是非常谨慎的，张开完全不需要担心信息的真实性。

张开在预订飞机票的时候看了一下飞机的航班时间，从他所在城市到上海每天有十个航班，在分析了不同航班的机票价格和误点率等信息之后，他选择了一个很早的航班，因为他想早点将房子定下来就可以不用住宾馆，另外也由于早晨的航班是一天中误点率最低、价格最便宜的。

张开搭乘这趟早班航班到达上海，他在上海很顺利地找到中介，在几套房子中选择了一套满意的租了下来。

在上海的第二天，张开就开始进行市场的拓展工作。因为他的目标客户是医院，所以他要找到上海所有医院的资料。在2030年，由于数据基础设施的发达，以前的人脉已经失去作用，如果像以前要依靠人脉的话，像张开这种在上海举目无亲的人是无法在上海开拓市场的。而现在的政府数据平台上，公开了上海所有医院的信息和数据。张开先分析了一下所有相同功能医药器材的采购、使用统计和增长情况，发现使用量庞大而且增长迅速，所以他对做好这个市场很有信心。然后他又在这些数据中，找到使用量排名前三的三家医院，准备从这三家医院着手。由于他的生产能力有限，他准备先打开市场，前期先占据10%的市场份额，回去再慢慢扩大生产规模。在数据平台上，张开除了看到使用量以外，还看到每家医院采购的平均价格。他认为他的产品除功能有创新外，价格还是有竞争力的。他的价格比现在的采购价格大约便宜10%。

张开从网上顺利找到这三家医院采购部门的联系人和联系电话。他打电话过去，和第一家医院顺利地约好见面时间。第二天按照约定的时间，他拿着自己的样品到达那家医院，拜访了这家医院的采购负责人。这家医院的采购负责人是一名非常内行的专家，在看到样品之后，对产品的功能和质量非常认可，当即决定先采购小批量产品在医院进行试用，满意后再扩大采购规模。

首战告捷，张开非常高兴。回去之后他和第二家医院也约好时间，前去拜访。这家医院的采购负责人并非专业人员，对产品的质量难以把关，当他看了样品之后对产品质量产生怀疑。但他为人很好，建议张开到政府的检测中心检测一下产品质量，并明确表态，若是质量没有问题他就会采购小批量试用。于是张开根据他的推荐，找到一家检测机构，将自己的产品样品送过去检测。检测机构在三天之内给他产品的检测结果。张开将检测结果拿到医院之后，采购负责人也同意试用。张开承诺交付产品的质量和检测结果是一样的，如果他的产品和检测样品有差异的话，根据质量法，他将会承担巨额的罚款，他的公司也可能由于这项罚款而倒闭。所以第二

家医院看到他能够生产出这个样品，并不担心批量产品质量的稳定性。

第三家医院的采购负责人开始拒绝和张开见面，也不说理由。张开把自己产品的功能优势、质量检测结果，根据该医院使用量统计数据计算的成本节约数据告诉这名负责人，并表示将向政府公平竞争管理部门投诉，该负责人答应过几天见面。张开后来才知道该负责人家里遇到事，心情不好。

张开正是由于在开拓市场方面所费功夫甚少，所以他把精力主要放在产品功能的完善、生产质量的保证和产量提升上，根本不需要在销售方面下太多功夫，而且他也知道随着其产品的供货扩大，其他供应商也会在产品的性能和价格上和他竞争，也就是说，他现在的优势只能维持一年左右，所以他还必须花大量的功夫在新产品的研发上。

张开的东部市场开拓异常顺利，他给一直向他泼冷水的舅舅打电话报喜。舅舅对他这么快就能打开市场觉得不可思议：“数据基础设施建设效果真的这么神奇？”

张开对未来充满信心，计划在三年内让公司上市，并走向国际市场。