

大数据主义

DATA-ISM

THE REVOLUTION TRANSFORMING
DECISION MAKING, CONSUMER BEHAVIOR,
AND ALMOST EVERYTHING ELSE

大数据主义

一场发生在决策、消费者行为
以及几乎所有领域的颠覆性革命！

[美] 史蒂夫·洛尔 (Steve Lohr) ◎著

胡小锐 朱胜超◎译



图书在版编目 (CIP) 数据

大数据主义 / (美) 洛尔著; 胡小锐, 朱胜超译. —北京: 中信出版社, 2015.9

书名原文: DATA-ISM

ISBN 978-7-5086-5371-6

I. ①大… II. ①洛… ②胡… ③朱… III. ①经济管理—数据管理—通俗读物 IV. ①F2-39

中国版本图书馆CIP数据核字 (2015) 第 173636 号

DATA-ISM

Copyright © 2015 by Steve Lohr

Simplified Chinese translation copyright © 2015 by CITIC Press Corporation

ALL RIGHTS RESERVED

本书仅限中国大陆地区发行销售

大数据主义

著 者: [美] 史蒂夫·洛尔

译 者: 胡小锐 朱胜超

策划推广: 中信出版社 (China CITIC Press)

出版发行: 中信出版集团股份有限公司

(北京市朝阳区惠新东街甲 4 号富盛大厦 2 座 邮编 100029)

(CITIC Publishing Group)

承 印 者: 北京诚信伟业印刷有限公司

开 本: 787mm×1092mm 1/16

版 次: 2015 年 9 月第 1 版

京权图字: 01-2015-5630

书 号: ISBN 978-7-5086-5371-6/F·3445

定 价: 49.00 元

印 张: 20.25 字 数: 180 千字

印 次: 2015 年 9 月第 1 次印刷

广告经营许可证: 京朝工商广字第 8087 号

版权所有·侵权必究

凡购本社图书, 如有缺页、倒页、脱页, 由发行公司负责退换。

服务热线: 010-84849555 服务传真: 010-84849000

投稿邮箱: author@citicpub.com

大数据何以成为“主义”？

数据者，有广义与狭义之分。狭义的数据，就是数字或数值，如 1、2、3、4、5……；广义的数据，则可概括为人类观察、实验、计算等的记录。作为这些记录的符号，或数字，或文字，或图像，或音视频，从上古时代的结绳记事、楔形文字、甲骨文，到古代乃至现代以竹简、布帛、羊皮、纸张等为载体的图文，直至现在以比特为单位的电子信息，可谓无所不包。

也许，正是由于互联网技术工程师们习惯于把以电子信息方式存在的内容统称为“数据”，于是，“数据”一词便由狭义的“数字”或“数值”演变为主要指向通用的广义“数据”。

随着计算机、互联网、现代通信以及相关软硬件技术的飞速发展，大数据和云计算，如同一枚硬币不可分离的两面，成为我们这个时代的高频词。

大数据之大，不仅大在巨量或海量——由人们熟知的千字节（KB）、兆（MB）、千兆（GB）和太字节（TB），跃升为专业人士才了解的拍字节（PB）、艾字节（EB）、泽字节（ZB），乃至尧字节（YB）。（1KB=1 024B，1MB=1 024KB，1GB=1 024MB，1TB=1 024GB，

1PB=1 024TB, 1EB=1 024PB, 1ZB=1 024EB, 1YB=1 024ZB)。其中,不变的是基本单位B,即比特或字节,而从K到Y,则是成千倍的递增,以致有人据此宣称,目前的数据计算已经进入“PB时代”。

大数据之大,还大在数据结构的有容乃大——它不再需要传统的数据库表格来整齐排列,几乎可以无所不包地记录、存储和计算各种规则的结构化数据和不规则的非结构化数据,于是便有了逐步演变为一个数字化世界的可能。

如此庞大和复杂的数据,远远超出传统计算机的处理能力,于是,建立在互联网基础上的云计算技术应运而生,承担起存储、传输、计算和应用大数据的重任。而正是大数据与云计算的有效互动,打开了世界观、方法论乃至价值观的新视野。

在本书中,作者引用专业研究机构的统计,揭示了大数据的规模与速度:一方面,到2014年,全世界电子化数据已增至4.4ZB,即4.4亿万亿字节,如果将如此之巨的信息量存入只有7.5毫米厚的苹果平板电脑,后者叠加起来的厚度可达地球与月球间距离的2/3;另一方面,有史以来90%的数据量,都是在过去两年的时间里产生的。

由此不难预期,一个电子化的、独立于物质世界的“数字世界”,正在大数据和云计算的互动中迅速构建,它虽然不可能穷尽物质世界全部存在,越来越逼近物质世界本体却是不争的事实。

尤为值得注意的是,许许多多以往被闲置的数据,由于一些精明商家的开发和利用,开始“变废为宝”。一个耳熟能详的案例,就是那个“尿片+啤酒”的商业发现与行动。世界最大零售商沃尔玛通过大数据统计和分析发现,男性顾客在购买婴儿尿片时,常常会顺便买上几瓶啤酒,于是推出将啤酒和尿片捆绑销售的促销方式,从而有效地提高了啤酒销量。

凡此种种表明，如同宇宙大爆炸般飞速扩张的“数字世界”，不仅日益成为外在的客观物质世界的“镜像”，而且正在越来越多地包含对人类自身行为的追踪和记录，成为人类观察和认识自我的“镜子”。

二者的叠加，形成一个有趣的悖论：由大数据构筑而成的数字世界，在日趋脱离客观物质世界的同时，又越来越接近世界的本原。因此，人们在解码这样一个虚拟世界的同时，也在一定程度上改变着对世界的看法。

然而，就在哲学家们对数字世界的属性还没有来得及给出明确界定之际，为利益所驱动的商业家们却迫不及待地启动了对这一新矿藏的发掘。

它们是如此急切：还没来得及弄清两个相关现象之间的互动机理或因果关系，便急匆匆地将其中的商机转化为提升经济效益的手段；云计算技术刚刚出现，便迅速地将统计分析对象由随机采样拓展为可获取的全部数据；为寻求“大数据的高效率”，不惜置“小数据的精确度”于不顾……

回过头来看，正是这近乎“饥不择食”的匆忙，竟在无意中成就了认识数字世界的锁钥：不再执着于因果关联，不再满足于抽样分析，不再一味地追求精确度的提高，转而直面模糊与混杂，关注看似不相关的相关现象。这一系列有别于以往的方法，为人类认识世界、解决问题提供了传统工具箱中没有的新工具。

诚如史蒂夫·洛尔在本书中的比喻，这些大数据时代的新工具，犹如“望远镜”或“显微镜”。“望远镜”让人们看得更远，发现新的星系；“显微镜”则将比细胞更加微小的世界展示在人们面前，人们据此看到并计量之前一无所知的事物。

抛开学术和技术层面的研讨，大数据及其应用几乎与生俱来就

伴随了喋喋不休的争论。

其中有两个关键词，一是“开放”，一是“保护”。如果说开放就是要打破垄断分割，推动信息与数据互联互通；变革体制机制，实现数据资源共有共享；鼓励技术创新，促进大数据资源开发利用……最大限度地拓展数字世界“公共空间”，让大数据和云计算普惠大众，造福人类；那么保护则意味着要在数字世界为个人留下一方“私密领地”，或者为公权力画上一道不能逾越的“红线”——“风可进，雨可进，国王不能进”。

开放与保护，“公共空间”与“私密领地”，在这里构成既对立又统一的关系。对立在开放与保护“井水不犯河水”，统一在“公共空间”与“私密领地”共存于同一个数字世界，且双方都以对方的存在为自身存在的证据，正所谓没有“公”即没有“私”，没有“私”亦没有“公”。

一言以蔽之，数字世界与现实世界理应奉行同样的价值理念：该开放的一定最大限度开放，该保护的必须严格加以保护。

本书向读者展示了这样一幅图景：不管你自觉还是不自觉，乐意还是不乐意，大数据正以空前的速度和规模渗透到人类社会生活的方方面面，它在一定程度上已经和正在改变人们观察、认识、思考乃至生存与发展的方式。特别是这后一方面的变化，或许就是“大数据”之所以成为“主义”的原因。

孤陋寡闻，借题发挥，难免贻笑于大方。

忝为推荐。

《经济参考报》总编辑
杜跃进

推荐序 / IX

第1章 大数据到底有多“大”？

我们已经进入大数据时代 / 006

技术进步与摩尔定律 / 009

大数据主义的兴起 / 013

第2章 人的潜力与数据的潜力

一位年轻的明星数据科学家 / 025

一个阅读兴趣广泛的“T型人” / 028

聪明人也会犯低级错误 / 039

从华尔街到脸谱网 / 046

第3章 蓝色巨人与大数据的博弈

- 大数据技术在医疗领域的应用 / 054
- IBM公司的大数据战略 / 060
- 从“蓝色巨人”到“智慧的地球” / 067
- 大数据时代的企业生存法则 / 076

第4章 企业决策与数据挖掘

- 数据、直觉与好的决策 / 093
- 大数据将引发“管理革命” / 103
- 数据驱动型企业的赢利之道 / 109

第5章 数据科学的崛起与发展

- 社交网络大数据的价值何在？ / 123
- 用户数据分析技术是脸谱网的引擎 / 127
- 数据科学家们的颠覆式创新 / 134

第 6 章 人工智能与机器学习

厘清大数据中的相关关系和因果关系 / 153

人类与人工智能的未来 / 162

人工智能系统要学会讲故事 / 169

第 7 章 用大数据推动企业与行业发展

当传统的葡萄酒业遇上大数据 / 180

物联网与大型工业机器的数字化 / 188

大数据可以产生巨大的经济效益 / 194

第 8 章 物理世界与数字世界的融合

人与智能机器的合作博弈 / 205

大数据营销的艺术与科学 / 215

用大数据建立人类行为模型 / 225

第 9 章 医疗行业的大数据革命

- 最棘手也是最有价值的问题 / 233
- 建立数据驱动的医疗保健体系 / 239
- 用数据杠杆撬动医疗行业的进步 / 247

第 10 章 大数据时代的隐私“黑洞”

- 个人隐私权受到前所未有的挑战 / 262
- 大数据会泄露多少我们的隐私？ / 268
- 是个性化服务还是经济歧视行为？ / 272
- 能精准量化人类性格特征的数据技术 / 278
- 如何将隐私风险降至最低？ / 286

第 11 章 大数据主义与美丽新世界

- 大数据市场的野蛮生长 / 297
- 大数据技术引发的关键思考 / 303

致 谢 / 307

DATA

第1章 大数据到底有多“大”？

ISM

在孟菲斯的郊区，人与各种机器正在来回穿梭，搬运货物。机械臂从传送带上抓取一捆捆塑料薄膜收缩包装的货物，然后由叉车搬到卡车上运往远方。工人们的血肉之躯在力量上比不上各类机械，因此分配给他们的任务是监控货物的流向，驾驶那些叉车与卡车，利用条形码扫描器与发射无线电波的芯片追踪货物上的识别标志。在井然有序的忙碌中，人与机器演奏出一首和谐的工业交响曲。

把这幅略显忙乱但是效率极高的画面呈现在我们眼前的是麦克森（McKesson），一家完成了全美约 1/3 的药品分销工作的大型公司。公司的各类建筑占地超过 8 个橄榄球场的面积，形成了麦克森全美分销网络的中枢，各类商品从这里被配送到包括社区

药店与沃尔玛超市在内的 26 000 个客户那里。货物主要由药品构成，每天需要配送大约 2.4 亿片。药品分销的工作量极大，但是利润空间却非常小。因此，效率是麦克森几十年来孜孜不倦追求的目标。

但是最近几年，麦克森在追求效率的道路上取得了一个令人瞩目的突破，将任意时间的分销网络库存减少了 10 亿美元。之所以能取得这一成绩，得益于产品、位置及运输数据给麦克森的启示。麦克森利用各种扫描器、传感器收集这些数据，然后运用智能软件从数据中挖掘各种省时减耗的潜在机会。这种通过技术加强经营的理念是一个重大突破，麦克森公司的高级主管唐纳德·沃克尔形容它可以“让一切潜在机会无所遁形”。

镜头切换到亚特兰大。我来到埃默里大学医院的 5 楼，站在一间玻璃围封的重症监护室外。病房中摆满了密密麻麻、林林总总的医学计算电子设备，包括呼吸机、透析机、输送抗生素与镇痛剂的输液器，以及监控心率、呼吸、血压、血氧饱和度等重要指标的器具。几乎每台机器都配备了电脑平面显示器，电子元件不停地发出“哗哗”的声音和各种警报声。我数了数，大的电脑平面显示器加上跟智能手机差不多的小尺寸的平面显示器，一共有 12 台。

通常，一个有 20 张床的重症监护室每秒钟估计可以生成 16 万个数据点。埃默里大学医院的研究表明，医生与护士每天都要根据这些数据，针对每位病人迅速做出约 100 个决定。换言之，一个重症监护室全年要做出 930 多万个跟护理有关的决定，因此，做出错误决定的概率非常高，必须采取某些措施帮助他们摆脱这种困境。目前，有多家医学研究中心正在致力于研究如何利用数据改进成人及新生儿重症监护室的关键性护理工作，埃默里大学医院就是其中的一家。这家医院利用监控病人病情变化的医学设备得到相关数据流，通过软件加以分析，以捕捉代表病情恶化的预警信号。

计算机与软件善于消化大量数据，并从中发现极不明显的规律，在这方面它们的效率高于人类。在埃默里大学医院，负责此项工作的是蒂莫西·布克曼博士。布克曼是一名外科医生、科学家，还是一名经验丰富的飞行员，因此，他用飞行来打比方，向我们解释他的研究目标。如果飞机偏离航向（用布克曼的话说是“偏离轨道”），飞机上的全球定位系统（GPS）位置数据就会转变成图像，显示在空中交通管制员的屏幕上，这项工作在飞机坠毁之前就早已完成。布克曼希望医学上也能建立同样的预警系统，在病人“坠毁”之前，及时捕捉重要指征的某种规律偏离医学轨道的信号。布克曼说：“这就是大数据的意义所在。”

我们已经进入大数据时代

早在谷歌、脸谱网等硅谷互联网孵化器问世之前，大数据时代就已经拉开了帷幕。起初，大数据是由字节构成的纯数字形式，然后它迅速迈进了由原子构成的物理世界，最后融入了社会主流。麦克森分销中心与埃默里大学医院重症监护室的例子告诉我们，大数据不仅可以节省时间，还可以挽救生命。从长远来看，大数据技术必将发展成为数据驱动的人工智能，驻留于数码世界与物理世界的顶层。现在，我们扬帆起航，朝着这个愿景进发。大数据技术正在引导一场计量方面的革命，有望为下一个席卷整个经济领域的效率与创新浪潮奠定基础。但是，技术并不是孤军作战。将来，我们在考虑如何决策，或者说应该如何决策这个问题时，同样需要大数据为我们出谋划策。我在《纽约时报》的同事戴维·布鲁克斯把这种正在兴起的观点称作“大数据主义”，这个词可以形象地反映大数据的深远影响。经验一再告诫我们，创新工具非常重要，它们不仅有利于经济增长，还可以帮助我们重塑世界观，以及做出相关决策。

大数据催生了大量技术，首先是各种新老数据源，包括 Web 网页、浏览习惯、传感器信号、社交媒体、智能手机的 GPS 位