BIG DATA

大数据

引爆新的价值点

孙 静 ◎ 主编　栾大龙 ◎ 主审

清华大学出版社

BIG DATA

大数据

引爆新的价值点

孙静◎主编

清华大学出版社

北京

内 容 简 介

大数据是“互联网＋”浪潮下的重要产物，也是推进“互联网＋”战略的关键技术。本书分为4篇，共10章，搜集了来自互联网企业、运营商、旅游、交通、电力、税务多个领域的真实案例，通过理论概念和应用案例相结合的方式逐步展开。从认识大数据的概念及其特征出发；定义大数据的经济资源属性，所涉及的隐私博弈和开放共享成为大数据发展的关键瓶颈；从实际应用中发现大数据的价值，不仅存在于产业经济中，还存在于社会政务中。

图书在版编目(CIP)数据

大数据：引爆新的价值点/孙静主编. —北京：清华大学出版社，2018
ISBN 978-7-302-48458-5

Ⅰ. ①大…　Ⅱ. ①孙…　Ⅲ. ①数据管理　Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第227150号

责任编辑：贾　斌　薛　阳
封面设计：刘　键
责任校对：梁　毅
责任印制：杨　艳

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
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：11　　**字　　数**：262千字
版　　次：2018年10月第1版　　**印　　次**：2018年10月第1次印刷
印　　数：1～2000
定　　价：45.00元

产品编号：072282-01

编辑委员会名单

前言
PROFACE

1980年，未来学家阿尔文·托夫勒将大数据称作“第三次浪潮的华彩乐章”，与大数据相关的概念、技术、应用开始进入人们的视野，人们开始重新认识在互联网时代各类信息与行为数据所具有的更深层的意义和价值。2016年，在中国杭州召开的二十国集团(G20)领导人第十一次峰会，大数据流动的便利化，已被认为是未来国家经济发展最重要的动力。

在微信、QQ、陌陌、云集、淘宝、京东、手机银行、高德地图、滴滴、携程、网盘、云端、手游等各类应用占满我们的手机屏幕的时候，每个人都身处大数据的旋涡之中。谁在产生数据？谁在获取数据？谁在交易数据？谁在分析数据？谁又在从数据中获取价值呢？

大数据是新兴的概念，却不是新的事物，因为数据是对客观事件进行观察或记录的结果，是我们生活或生产过程中每一个选择、决策、交流、发布、分享等行为的产物，可以说数据在人类之初就有，直到现代数字信息技术成熟，当TB甚至EB量级的数据可以被采集、存储和处理时，才有了大数据。

作为未来经济发展的“新型石油”，数据的产生是广泛的、随机的、多源的、离散的。从事大数据工作就是要发现、采集、分析、研究TB甚至EB量级的数据，从无序的数据中找到微妙的关联和规律，例如“啤酒与尿不湿”，从而对当前决策进行优化，对未来趋势进行预测。

本书编写的初衷是在与很多大数据从业者交流中，发现大数据的应用已比人们认识的更广泛，发展得更迅速，尽管业内生态仍需要完善，数据孤岛依然存在，相关政策和标准还有待制定，但是总能听到两个声音：“我们有好的实践案例想与大家分享，听取更多的建议。”“我们想看到更多、更新、更详细的案例，想将我们的数据变成真的黄金。”因此，得到北京邮电大学、阿里云研究中心、电信科学技术研究院、国家信息中心、京东物流、美林数据、明略数据等单位的大数据研究人员的支持和无私的分享，将各个领域的应用案例推荐给广大的读者。期望通过本书，能够更好地帮助大家认识大数据，理解大数据，运用大数据，在数据的海洋中挖掘到更多的宝藏。

“互联网+”时代，信息的数字化、移动应用和支付的普及化、位置和医疗等个人信息的实时获取、物联网万事万物的互联互通……新科技、新概念如同海浪般，一浪推着一浪翻滚前进，大数据就如同海中的浪花一般随波前行，且随着一浪一浪的相互推动而越发丰富绚烂。

编　者

2018年6月

目录
CONTENTS

第一篇　大数据，新时代的代名词

第二篇　大数据，一种经济资源

第三篇 大数据，价值创新的土壤

第一篇　大数据，新时代的代名词

人类历史上从未有哪个时代和今天一样产生如此海量的数据，数据的产生已经完全不受时间、地点的限制，尤其是社交网络、电子商务、移动互联网等领域的飞速发展，把人类社会带入了一个以 PB(1PB＝1024TB，1TB＝1024GB)为单位的结构和非结构数据构成的网络化、数字化时代。一个大规模生产、分享和应用数据的时代正在开启。

2016 年 10 月，在杭州举行的“互联网大数据高峰论坛”上，阿里巴巴原副总裁、《数据之巅》的作者涂子沛指出：“弄潮儿向涛头立，手把红旗旗不湿。我们今天的涛头就是互联网、大数据。”

中国科学院大学经管学院教授吕本富，在《G20 国家互联网发展研究报告》中指出“我们认为，未来数据的流动将是世界经济增长最重要的动力来源。如果过去是服务的便利化、贸易的便利化，以后一定是大数据流动的便利化，成为这个国家经济发展最重要的动力。”

大数据已掀起了时代的浪潮，任何人和事都需要用数据说话！

第1章 数据时代

“大数据”一词近五年在百度搜索指数中的整体趋势从2012年开始呈快速增长的态势，并在2016年5月周平均值最高达7287，如图1-1所示。

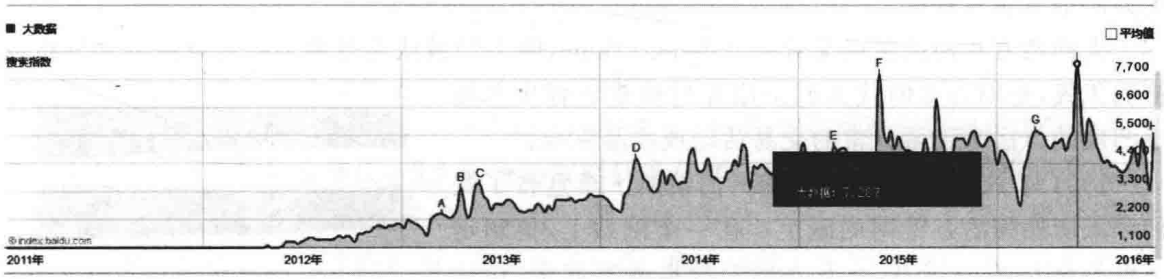

图1-1 “大数据”的百度搜索指数

（数据来源：百度指数，http://index.baidu.com/）

2014年8月，在中央电视台财经频道、综合频道、纪录频道、科教频道播出了一套10集的纪录片《互联网时代》（别名《大数据时代》），这是中国第一部，甚至也是全球电视机构第一次全面、系统、深入、客观地解析互联网的大型纪录片。这部在开播前没有密集的节目宣传，没有明星、噱头与谈资的纪录片，仅是在社交网络上的口口相传，百度搜索的指数就从0开始直线攀升至15 127。2015年8月31日，国务院印发《促进大数据发展行动纲要》（国发〔2015〕50号）。指出主要任务是：加快政府数据开放共享，推动资源整合，提升治理能力；推动产业创新发展，培育新兴业态，助力经济转型；强化安全保障，提高管理水平，促进健康发展。

这些数据和现象都在说明一个客观事实、一个社会热点、一个发展趋势——大数据的发展已成为国家发展战略的重要组成部分，大数据正在或已经成为时代前进的代名词。如何认识大数据，如何应用大数据，是从IT时代走向DT时代的必要课题。

1.1 大数据溯源

早在1980年，著名未来学家阿尔文·托夫勒在其所著的《第三次浪潮》中就提出“数据就是财富”，并热情地将“大数据（Big Data）”称颂为“第三次浪潮的华彩乐章”。但是到2008年，学术界、工业界甚至于政府机构才开始密切关注大数据问题。*Nature*杂志在2008年9月推出了名为“大数据”的封面专栏，*Science*则在2011年推出了专刊*Dealing with Data*，

主要围绕着科学研究中大数据的问题展开讨论，说明大数据对于科学研究的重要性[①]。

大数据的概念和技术不是凭空出现的，人们对于大数据的认知或许最早来自托夫勒在其所著的《第三次浪潮》，但是人类对于数据的搜集、存储可以追溯到远古时代，对于事物的数据化发展于计算机的出现。“大数据”并不是作为一个全新的事物出现的，它是基于人类发展过程中，对于数据搜集、存储、分析能力的提升而出现的一种新的思维方式，一种新的服务模型，一股推动经济社会发展新的助力。

1.1.1 数据起源

数据(data)是对客观事件进行观察或记录的结果，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合，是对客观事物的逻辑归纳，用于表示客观事物的未经加工的原始素材。它可以是数字，也可以是具有一定意义的文字、字母、数字符号的组合、图形、图像、视频、音频等，是可识别的对客观事物的属性、数量、位置及其相互关系的抽象表示符号。

大约两万年前的伊尚戈骨头(Ishango Bone，图 1-2)被认为是最早的记录数据和分析数据的工具，是旧石器时代人们采用在树枝或者骨头上刻下凹痕的方法来记录日常的交易活动或物品供应。

图 1-2 伊尚戈骨头

1991 年，计算机科学家蒂姆·伯纳斯·李宣告了我们今天所熟知的万维网的诞生。在一个网站上，他制定了世界网络的协议书，使互联网的数据联通起来，让任何人可以在任何地方进行通信。互联网时代的开启，带动了各行各业的网络化发展。人、物、机器等都可以通过一个终端接入这个不受时间、空间限制的虚拟网络中。在商业、生活、生产、农业、医疗、金融等领域网络化的过程中，带来了以几何倍数增长的数据量。

2004 年，Facebook(脸书)、Twitter、Instagram 等社交网络的相继问世迎来了开放共享的 Web 2.0 时代。网络平台不再是自上而下地由少数资源所有者控制，而是自下而上地由广大用户的智慧和力量主导。在 Web 2.0 模式下，网络用户出于对某个或某些问题的共同兴趣而聚集，这促使他们主动积极地参与问题讨论和信息分享。全球数据量预测如图 1-3 所示。

根据知名市场研究机构 IDC(International Data Corporation，国际数据公司)的研究报告表明，2011 年全球数据总量已经达到 1.8ZB，未来全球数据总量年增长率将维持在 50% 左右，到 2020 年，全球数据总量将达到 40ZB，如图 1-3 所示。

1.1.2 数据存储

人们在生产生活过程中所创造的各种数字、图像、文字、记录等需要被采集并保存下来，才能够形成数据。一个坚持 30 年，每天走一万步的人，他的个人运动数据和位置数据，在微信运动或计步 App 等出现后，同样的行为才被采集并存储成为数据。

亚历山大图书馆(公元前 300 年—公元 48 年)可能是古代最大的数据储存地了，这里

① 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，50(1)：146-169.

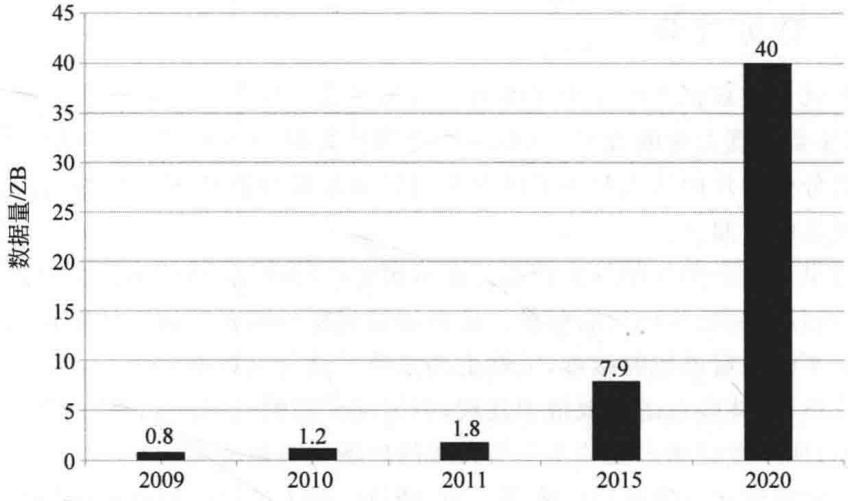

图 1-3　全球数据量预测

（数据来源：IDC）

50 万卷的藏书几乎涵盖了当时人们学习的各个领域。

1928 年，工程师波弗劳姆(Fritz Pfleumer)发明了一种用磁带来存储信息的方法。他发明的这个原理今天依然在使用，绝大部分的数据就是存储在有磁性介质的计算机硬盘上。

1965 年，英特尔(Intel)创始人之一戈登·摩尔(Gordon Moore)提出了摩尔定律，揭示了信息技术进步的速度。其内容为：当价格不变时，集成电路上可容纳的元器件的数目，约每隔 18～24 个月便会增加一倍，性能也将提升一倍。在摩尔定律的推动下，计算存储和传输数据的能力在以指数速度增长，每 GB 存储器的价格每年下降约 40%。

1965 年，美国政府计划在世界首个数据中心的磁盘上存储 7.42 亿的纳税申报单和 1.75 亿的指纹信息。1967 年，IBM 公司推出世界上第一张“软盘”，是最早的可移动数据存储介质。

2010 年印刷版《大英百科全书》，共 32 册，重达 58.5kg，然而它的全部内容，还装不满一个 4GB 的 U 盘。

历史的进程进一步证实了摩尔定律，数据存储能力的指数提升如图 1-4 所示。

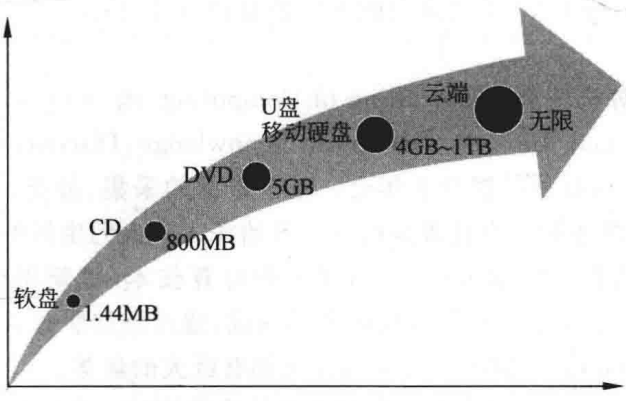

图 1-4　数据存储能力的提升

1.1.3 数据计算

数据分析就是对数据进行分析并得出有用的结论。首先不一定使用统计分析的方法；其次，不一定非要处理大量的数据，也不一定要用计算机；再次，数据分析自古就有。百度百科对于数据分析的片面认识反映了国内人们对于数据分析认识的模糊，也反映了商业利益对于正常观念的扭曲。

数据分析早在两千多年前就在使用。在战国时期的孙庞斗智中，孙膑设计蒙骗庞涓，孙膑命令部队，每日大幅减少炉灶的数量。庞涓通过观察孙膑军队的炉灶数量逐日大量减少，分析得出孙膑军队大量逃散的结论，最终上当战败。这就是数据分析。

在辽沈战役中，林彪在诸多战报中发现，在胡家窝棚附近缴获的短枪与长枪的比例比其他战斗中的高，那里缴获和击毁的小车与大车的比例比其他战斗中的高，在那里俘虏和击毙的军官与士兵的比例比其他战斗中的高。他就断定，敌人的指挥所就在这里。果不其然，敌军司令廖耀湘在胡家窝棚附近被逮个正着。这也是数据分析。

数据分析发展自古到今，已经涵盖了最朴素的数据分析，也涵盖了数据统计、数据挖掘和大数据处理的所有内容。这两个案例都说明了数据分析古已有之，且数据分析不一定要有海量数据，也不一定要用复杂度统计分析方法，只要统计数据分类（统计口径）正确；同时还说明了数据分析极其重要，更说明了数据意识和素质的重要。

当各类数据能够被采集并得以保存时，提升计算和分析数据的能力，成为实现数据价值的必要手段。

安提凯希拉（Antikythera）机器，是最早被发现的机械计算机[①]，也代表了数据分析能力从人工计算向机械计算的提升。

1663 年，约翰·葛兰特（John Graunt）在伦敦用记录下的当时肆虐欧洲的黑死病死亡人数信息，建立起了早期预警系统的理论，是第一次有记录的统计数据分析实验。1865 年，银行家亨利·福尼斯（Henry Furnese）用结构化的方式收集和分析有关竞争对手的商业活动来取得竞争优势，这被认为是第一次将数据分析用于商业目的。

1881 年，美国人口普查局聘用了一位年轻的工程师赫尔曼·何乐礼（Herman Hollerith），他发明了著名的打孔卡片制表机，被认为是现代计算机的雏形，将原本预计需要花费 10 年时间去分析的 1880 年收集到的人口普查数据工作缩短为三个月，数据处理速度提升了近 40 倍。

1989 年，美国计算机协会（Association of Computing Machinery，ACM）数据挖掘机知识发现委员会（Special Interest Group on Knowledge Discovery and Data Mining，SIGKDD）主办了第一届数据挖掘学术年会。基于数据的采集、分类、估值、语言、相关性分组或关联规则、聚集、描述和可视化等分析方法开始深入到人们生活的方方面面。

2004 年，谷歌公开的 MapReduce 分布式并行计算技术，是新型分布式计算技术的代表。一个 MapReduce 系统由廉价的通用服务器构成，通过添加服务器节点可线性扩展系统的总处理能力（Scale Out），在成本和可扩展性上都有巨大的优势。

2005 年，Hadoop 诞生，它是专门为存储及分析大数据的开源框架。它能够灵活管理人

① Antikythera mechanism[OL]. Wikipedia，https://en.wikipedia.org/wiki/Antikythera_mechanism.

们不断产生和采集的非结构化数据，例如语音、视频、文档等。以 Hadoop 为代表的分布式存储和计算技术迅猛发展，极大地提升了互联网企业数据管理能力，互联网企业对“数据废气”的挖掘利用大获成功。

2007 年，《连线》(*Wired*)杂志在文章《理论的终结：数据洪流让科学方法过时》中将“大数据”的概念引进了大众的视野[①]。

回顾数据的起源和发展，可以清晰地看到今天的大数据是从最朴素的数据分析、数据统计和数据挖掘一步步走过来的，数据分析为社会带来的经济价值越来越高。今天的大数据也好，数据挖掘也罢，都是在做数据分析这件事，只不过是数据的体量在提高，数据的复杂性在提高，数据处理的能力在提高以及数据处理的结果更具有创造性。

从最朴素的数据分析到大数据处理，运用数据的思路与逻辑是一致的。所有的数据分析无非是在寻找：什么是我要找的数据，我要找的数据在哪里能找到，最大(小)的数是多少，最大(小)的数据在哪里，最大(小)的可能是多少，最大(小)的可能在哪里，哪些因素最相关，相关性多大，从大到小的排序，按照时间或位置排列的升降状态等。数据分析的思路就是搜索、对比、概率计算、相关性分析、分类、排序、预测等，最后做出的结果就是预测、聚类与排序。

1.2 初识大数据

在人类社会发展的历史长河中，经济发展往往伴随着技术革命。2013 年称为“大数据元年”。目前，几乎所有世界级的互联网企业，都将业务触角延伸至大数据产业；无论社交平台逐鹿、电商价格大战还是门户网站竞争，都有它的影子。

大数据无处不在，大数据应用影响到了人们的工作、生活和学习，并将继续施加更大的影响。

1.2.1 大数据的定义

在计算机科学中，数据是指所有能输入到计算机并被计算机程序处理的符号的介质的总称，是用于输入电子计算机进行处理，具有一定意义的数字、字母、符号和模拟量等的通称[②]。

数据的基本计量单位是 Byte，按照 1024(2^{10})进率，依次递增为 B、KB、MB、GB、TB、PB、EB、ZB、YB、DB、NB。

1B=8b

1KB=1024B

1MB=1024KB

1GB=1024MB

1TB=1024GB

1PB=1024TB

---

① The End of Theory: The Data Deluge Makes the Scientific Method Obsolete[OL]. 2013, http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/.

② 王珊，萨师煊. 数据库系统概率(第5版)[D]. 北京：高等教育出版社，2014.

1EB＝1024PB

1ZB＝1024EB

1YB＝1024ZB

“大数据”一词本身就是一个比较抽象的概念，单从字面来看，“大”体现了研究或应用的量级规模是庞大的，“数据”则说明了研究或应用对象的实质。但是什么样的数据量级才可以称之为“大”呢？

传统数据库有效工作的数据规模一般为10～100TB，因此麦肯锡和IDC公司对此都有过相近的说法，10～100TB通常成为大数据的门槛。所谓大数据从数据规模上看，大概是指100TB以上的数据体量，100TB相当于现在100部最新笔记本（1TB硬盘）的最大存储总量。但是，数据计算的难度与速度还涉及数据的类型、结构与存储的复杂性，因此以100TB为基准来定义大数据的说法未必科学。

大数据和互联网都是一种通用目的技术（General Purpose Technology），随着技术和应用的发展，其概念也在不断地演进。尽管有很多研究机构和学者给出的定义被广泛认可，但是却没有公认的、唯一的准确定义。

维克托·迈尔·舍恩伯格与肯尼斯·库克耶在他们合著的《大数据时代》一书中指出：大数据是指不用随机分析法这样的捷径，而采用所有数据的方法[①]。

大数据：样本＝全体。

因此，所谓的“大”其实也包含着“全”的含义，不是相对的量级，而是绝对的范围。

对于大数据这一概念比较被认可的定义还有以下几种。

（1）大数据，或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理并整理成为人类所能解读的信息。（维基百科[②]）

（2）一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模（Volume）、快速的数据流转（Velocity）、多样的数据类型（Variety）和价值密度低（Value）4大特征。（麦肯锡全球研究所）

（3）大数据是数据集或信息，它的规模、发布、位置在不同的信息孤岛上，或它的时间线要求客户部署新的架构来捕捉、存储、整合、管理和分析这些信息以便实现企业价值。（EMC公司）

（4）大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产，这些信息资产需要新型的处理方式来强化决策制定、洞察发现和处理优化。（研究机构Gartner，2012）

（5）大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。（中华人民共和国国务院，《促进大数据发展行动纲要》，2015）

---

① ［英］维克托·迈尔·舍恩伯格，肯尼斯·库克耶. 大数据时代［M］. 盛阳燕，周涛，译. 浙江：浙江人民出版社，2013.

② Big data［OL］. Wikipedia，https://en.wikipedia.org/wiki/Big_data.