Pearson

DIGITAL EXHAUST

WHAT EVERYONE SHOULD KNOW ABOUT BIG DATA, DIGITIZATION, AND DIGITALLY DRIVEN INNOVATION

数字经济2.0

引爆大数据生态红利

[美] 达尔·尼夫（Dale Neef）◎著
大数据文摘翻译组◎译

世界数据权威专家关于**深化大数据应用、助力数据生态圈**打造的力作

个人与企业数字经济深度学习的首选读本

中国人民大学出版社

DIGITAL EXHAUST

WHAT EVERYONE SHOULD KNOW ABOUT BIG DATA, DIGITIZATION, AND DIGITALLY DRIVEN INNOVATION

数字经济2.0

引爆大数据生态红利

[美] 达尔·尼夫（Dale Neef）◎著
大数据文摘翻译组◎译

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

数字经济 2.0：引爆大数据生态红利 /（美）达尔·尼夫（Dale Neef）著；大数据文摘翻译组译 . -- 北京：中国人民大学出版社，2018.4

书名原文：Digital Exhaust: What Everyone Should Know About Big Data,Digitization, and Digitally Driven Innovation

ISBN 978-7-300-25448-7

Ⅰ . ①数… Ⅱ . ①达… ②大… Ⅲ . ①互联网络—应用 Ⅳ . ① TP393.4

中国版本图书馆 CIP 数据核字（2018）第 011648 号

数字经济 2.0：引爆大数据生态红利

[美] 达尔·尼夫（Dale Neef） 著

大数据文摘翻译组 译

Shuzi Jingji2.0: Yinbao Dashuju Shengtai Hongli

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010-62511242（总编室）	010-62511770（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京中印联印务有限公司		
规　　格	170mm × 230mm　16 开本	**版　　次**	2018 年 4 月第 1 版
印　　张	14　插页 1	**印　　次**	2018 年 4 月第 1 次印刷
字　　数	193 000	**定　　价**	59.00 元

阅读成就思想……

Read to Achieve

前言

本书适合每位想了解大数据现象和互联网经济内涵的读者：它是什么；为什么它是不同的；支持它的技术是什么；公司、政府和普罗大众如何从中受益；以及它可能在未来呈现给社会一些什么样的威胁。

这是一个相当高的要求，因为我们在本书中探索的公司和技术是：巨大的互联网科技集团，如谷歌和雅虎；全球零售商沃尔玛；智能手机和平板电脑生产商，如苹果公司；大规模的在线购物集团，如亚马逊和阿里巴巴；社交媒体和短消息公司，如 Facebook 或 Twitter。它们现在是世界上最具创新力、复杂、变化迅速和财力强大的组织。了解这些互联网强大集团最近的表现和可能的未来有助于我们了解数字创新正引领我们去向何方，同时也是了解大数据现象的关键。重要的是，无数的创新框架和数据库技术——NoSQL、Hadoop 或 MapReduce，它们正在极大地改变我们收集、管理和分析数字数据的方式。

鉴于这一主题的复杂性和多学科性，任何关于该主题的书都需要保持在

相当高的水准上。虽然本书不会提供在 Hadoop 中编程或者在大规模并行处理网络中设置节点的方法，但会让读者对什么是大数据、哪些公司在引领我们以及为什么这些技术在未来如此重要有一个更透彻的理解。

写一本关于大数据的书的第二个挑战是，这一领域在迅速变化，几乎每周都有新技术、初创公司和应用程序涌现、合并或崩溃。事实上，在写本书的时候（可想而知还需花费时间来读），有许多新的产品发布，从首次公开募股（IPO）到新的智能手机版本，它们将创新的过程微微推向多个不确定的方向。考虑到这一点，我们会提供本书的定期更新版本。

一般来说，尽管公司、技术、政策和问题会有变化，但大数据和数字经济背后的主要趋势是显而易见的：更强大和灵活的计算机分析、借助移动智能手机和平板电脑的全球扩张、个人信息的大量数字化、专注于基于云的应用和存储以及公司和政府机构对个人客户和用户数据的收集。希望本书有助于所有的读者，无论是技术和非技术人员、公司经理和小企业主，还是学生或者其他感兴趣的人士，以更加明智的方式处理这些问题，因为我们正处于大数据和数字经济时代。

目 录

Digital Exhaust

What Everyone Should Know About Big Data, Digitization, and Digitally Driven Innovation

第 1 章

大数据大爆炸

- 数字经济五个方面的共同推动产生了大数据现象：
  1. 消费者互联网；
  2. 工业互联网；
  3. 物联网；
  4. 不断增长的数字数据收集产业；
  5. 收集、解读非结构化及基于互联网的数据的新技术。
- 大规模的、前所未有的数字数据正在全世界范围内被收集着。
- 这为在以下四个方面利用大数据提供了可能：
  1. 在趋势和相关性方面提供之前无法获取的独特洞察力；
  2. 通过收集大量的用户数据提升销售和服务，或用于定向广告营销；
  3. 将出售用户数据作为单独的盈利手段；
  4. 大量使用机器交互数据创造行业供应链效率。

如今大数据这个词频繁地出现在各种媒体上。大数据过去常常被描述为遗传学、流行病学项目或DNA序列的统计方法，也会用于解释新的搜索和存储技术，这些技术能够让公司扫描各类在线媒体以获取“舆情”数据。很多人听到这个词都会想到谷歌，或者美国国家安全局（NSA），暗示他们收集、售卖与公民权利及隐私有关的个人数据。也有一些人认为大数据主要表现在新技术的应用上，如Hadoop、云计算或即将到来的物联网。有人说大数据是一场革命，也有人说那只不过是在信息技术发展的下一个变更阶段中找到兴奋点，利用过分渲染获取更大利益的借口（感谢这么多人都在热情高涨地谈论大数据，我才得以出版这本书）。

引发讨论的是：大数据包括所有这些事物，但又不能局限于任何一个特征定义，使得它成为一个“大事物”，所有这些不同的解释都反映了技术和经济变化的长期过程，它现在才刚刚开始变得成熟，开始证明自己是大数据智能复合体（这是我起的称呼）：一群有钱且有影响力的公司和政府机构负责大量的技术开发，助力经济发展和创新，并且在人们的交流方式、自我娱乐方式以及在与遍布世界的其他人的交互方式等方面产生变革。

大数据是一场革命或变革吗？按照历史标准，也许是。它是否是进化或变革过程并不重要，甚至比起电力、电话或内燃机，是否称得上是一场技术革命也并不重要。这些争论于我而言，不过是市场营销人员努力标榜或者美化自己的一种手段——用这些来证明自己是技术快速变化中的参与者。不过无论大家如何描述大数据，重要的是要考虑大数据在未来的几十年会将我们带向何方，因为大数据智能复合体的出现已经证明它无法控制且不可阻挡，创新以无法追随的速度出现，而且毫无规律。它不仅会影响公民权利和个人隐私（这两项已经变得非常重要），还会影响到我们的商业、全球经济、法律，甚至国家关系和未来发展。

大数据生态系统

在本书中，我们将能看到大数据在五个方面的融合，这五个方面看起来互有区别，但实际上是实力机构、新技术和消费趋势的聚集。这五个方面分别是什么呢？

首先和最突出的是较为熟悉的消费者技术：互联网、电子商务、远程信息处理、社交媒体以及移动技术共同创造了消费者驱动的大数据行业。这些都跟娱乐、智能手机和即时消息有关。我们在生活中的每一天都被这些围绕着：争论、合并和首次公开募股，金钱围绕着这些工具和玩具，数十亿的资金被投入到最新的应用程序上。消费者驱动的大数据行业很有价值，很有娱乐性，甚至让人分散注意力，正因为如此，我们会趋向于在某些程度上越来越轻视它，认为它没有比常规的生产力经济更重要或者有更强的经济属性，也不是促成就业和经济增长的要素。但是消费者驱动的大数据行业并不是微不足道的，它涉及了推文、照片共享和愤怒的小鸟。这是大生意，也汇集了大笔的钱，汇聚了广告、应用程序和游戏领域最顶尖、最前沿的思想，以前所未有的聪明手段捕获大量的个人数据。

就算不是技术性的革命，实际的技术也是非常引人瞩目的。但更有可能具有变革性的是在用户驱动的大数据经济领域中，巨头公司越来越多：谷歌、亚马逊、Facebook、阿里巴巴、Twitter、苹果公司以及众多遍布全球的支持或依靠这些大公司生存的互联网初创企业在未来十几年都将会主导行业领域的经济，或至少会产生巨大的影响。在可预见的未来，经济的增长不仅会发生在发达国家，也会发生在发展中国家（如果这种分类依然适用），这将取决于这些大数据巨头会把我们带向何方。有些人可能会感到焦虑：全球经济的未来在很大程度上依赖于屈指可数的网关技术公司，这是令人不安甚至恐惧的。

但是我们每天看到的与消费者相关的大数据只是一个方面，在消费者驱动的大数据行业风生水起的同时，另外一个大数据发展的重要领域也悄然萌

芽，那就是大数据在工业方面的应用，大数据也同样适用于所谓的“旧”经济。正是因为机电一体化和互联网技术的结合，才改变了传统收集和分析业务数据的方式。新的自动报警传感器、组件和系统能够把性能数据放到无比复杂的企业级计算系统中，使得销售、财务、库存和后勤等传统业务功能更加高效（使用更少的员工即可完成相关的工作）。这些创新性的、基于机器的数据收集和分析技术背后是爆炸式增长的工业互联网和物联网，这两个方面导致数字数据的产生和收集并行发展（有时也会重叠）。

尽管硬件依然是由美国通用电气公司、西门子或埃里克森这些“旧经济”主导者生产和制造，并且这些公司也很可能会推动物联网的发展，但当物联网真正来临时，能控制它且从中获利的，却可能是那些新兴的少壮派，如谷歌、Facebook 和亚马逊。因为这些公司掌握了大数据挖掘和分析的核心能力。那些曾经担心 IBM 会成为他们老大哥的人应该重新思考这个问题了。当互联网的战役结束，IBM 和美国通用电气公司将只能提供支持性的基础架构，帮助亚马逊、谷歌和 Facebook 控制业务系统、家居、车辆和智能手机中数字数据流的流入和流出。

同样，这些工业互联网大数据技术本身并非是革命性的。我的萨博（Saab）汽车 6 年来一直提供与车辆相关的机械和电子故障预警。预测性诊断很有用，但它们仍然不能修复车辆。而谷歌就能很快地帮我启动车辆，还能告诉我应该去哪里修车以及如何到达那里。如果我带上谷歌眼镜，我还能通过语音命令或点头激活 Eaze 应用，然后通过摄像头扫描二维码激活我的虚拟苹果支付（Apple Pay）或比特币（Bitcoin）钱包，完成转账和付款。当我离开家的时候，谷歌利用 Nest 技术调整家中的恒温器，我还能通过亚马逊预订零部件或安排维修，或预订生活用品和洗衣服务。我通过运行谷歌 Android 操作系统的手机就能监控或发送指令，能够让谷歌监控和捕捉所有活动（包括我发送到服务中心邮件带有的情绪），这些信息都会增加到我的数字资料中，让我收到定制化广告（有可能会是一个来自汽车厂商竞争对手服务中心的优

惠券）。谷歌会记录我对这些优惠券的反馈，以及我是否通过“喜欢”按钮将这个服务中心推荐到社交媒体的朋友圈，然后谷歌就会通过 cookie[①] 找到我朋友圈的那些用户，相应的优惠券就会出现在他们的 Facebook 站点上，他们也会被邀请到服务中心体验，与此同时数字信息收集还会持续并不断增长。

这个杜撰的老旧萨博汽车故事提出了一个很重要的观点。当三个大数据的趋势：当消费者互联网、工业互联网和物联网高度融合的时候，大数据的新纪元就将开启。

数字数据收集作为另一个强有力的行业，也与这些主要的大数据趋势并行发展起来。这里包含了很多大型互联网公司如谷歌、雅虎、Facebook 和 Twitter，还有在线零售商如亚马逊和苹果公司。此外还有大部分主要的线上或线下零售商（之前和现在的传统实体企业），如在美国的沃尔玛、塔吉特公司和沃尔格林公司，他们收集并售卖客户的个人信息和交易数据。还有几百家的在线数据跟踪软件和服务公司，可能大部分人从未听说过这些公司，但它们却监控着我们每天的线上活动、跟踪我们的数字化足迹，并且将数据（既有整合的，也有个人的可识别信息）出售给广告商、人力资源代理机构、收债方以及任何愿意为之付钱的买主。当然，这里还包含了大部分的广告代理机构以及大型数据整合公司，如 Experian、FICO 以及 Acxiom，这些公司最早是做信用报告的，但它们现在维护着巨量的数据库，里面包含了全世界数以百万计的个人信息和隐私数据。

当这些数据的持有者们在一起的时候，他们有时互相竞争，有时又互相支持与合作，形成了一个既强有力又阴暗的经济力量，他们通过解读和售卖与消费者相关的数据让更多的公司以一种以前不可能想到的方式更加了解他们的消费者。这些数据收集者从两个方面体现他们的价值：一是他们拥有这

① cookie，有时也用其复数形式 cookies，指某些网站为了辨别用户身份、进行 session 跟踪而储存在用户本地终端上的数据（通常经过加密）。——译者注

些数据库，这是事实；二是他们有工具来控制那些想要拿到数据的人们，他们来决定数据如何被分发、谁能够看到以及如何使用。他们对于大数据经济的成功是至关重要的，因为他们能将原始数据转化为以用户为目标的广告黄金。

这四个趋势的交汇造成了数据生产和收集活动的疯狂以及数字数据的浪费。事实上，对大多数公司来说大数据意味着大数据过载。这些公司被通过在数字数据市场售卖与客户相关的数据能够创造新的盈利点所诱惑，被告知要收集所有能收集到的客户和交易数据，并把它们都存储起来，以备在日后能够从中提取到对广告或者产品销售有用的信息，或者干脆直接把这些客户数据卖给其他公司。这些都已经变成了如今这些机构的咒语（特别是像美国国家情报局这样的部门），他们会收集所有的数据。这意味着电子邮件、信用卡号、在线购物、用户在网上查看和拒绝的内容、广告点击和页面停留时间、客户投诉电话、Twitter 上提到过某个公司或产品的记录等都会被收集。因此在大数据市场，对个人数据的价值、所有权和神圣性的期望正在发生巨大变化。

现在我们来谈一谈大数据现象的第五个特征：支撑大数据的技术也正在不断涌现。要能从所有的数字数据中提取价值，数据首先需要被存储、组织、检索和分析，而当前的系统不能很好地处理大量非结构化数据。我们这些做 IT 的人士比较了解，大多数公司的系统都已经到达一个极限，都受限于传统数据库技术。

这意味着如果想从大数据中获益，那么这个大数据现象中所涉及的众多方面都需要达到目标才能确保新的大数据技术能够成功，这些方面包括：

- 数据搜索和索引技术允许挖掘大量独立的数据集。它们以 NoSQL、Hadoop 和 MapReduce 类技术的形式出现，这些技术为谷歌、雅虎、Facebook 和亚马逊的大数据搜索功能提供了相同的工具。

- 易于访问的数据存储技术，部分需要促进向基于云的外包迁移，以及由亚马逊和谷歌等互联网强权建立的庞大的全球数据存储能力。
- 改进的分析工具，有助于筛选大量数据，以发现重要的关系和相关性。

这些新技术的发展是重要的，因为它们使得 IT 行业和风险资本市场将注意力从计算能力（利用更强大的计算系统的数字处理能力）的持续增长（如果是平稳的）转到专注于非常不同的大数据。

大数据的特征定义

大数据和大计算有什么不同呢?

首先，大数据不仅仅是数字运算，大数据是要收集和利用前所未有并且几乎是难以想象的数量级的数字数据，现在这些数据已经存在并且可以使用新的分析工具形成数据洞察。之所以称之为大数据，数据量是一个关键因素，前提条件是世界各地每一秒的数据都呈现爆发式的增长，这已经成为事实了。

简单来说，数字化的爆发式增长已经达到 GB 或 TB 的边界，这在以前已经是非常惊人的数据量，但今天我们谈论的已经是 PetaByte（PB），ExaByte（EB），YottaByte（YB），ZettaByte（ZB，1ZB = 1 万亿 GB）这样的形式。我个人比较喜欢 BrontoByte（1000 YottaByte），部分原因是这听起来有些像《摩登原始人》（*the Flintstones*）[①] 里的人订了一个外卖。

对我们大多数人来说，这些数字没有透露很多信息。大部分时间我都在做数据管理工作，当 IBM 估计每天生成额外的 250 万的三次方字节的数据时，我不知道这意味着什么。图 1-1 的这种数字数据爆发式增长趋势图能让大家

① 《摩登原始人》是一部美国动画电视剧。——译者注

好好体会一下。

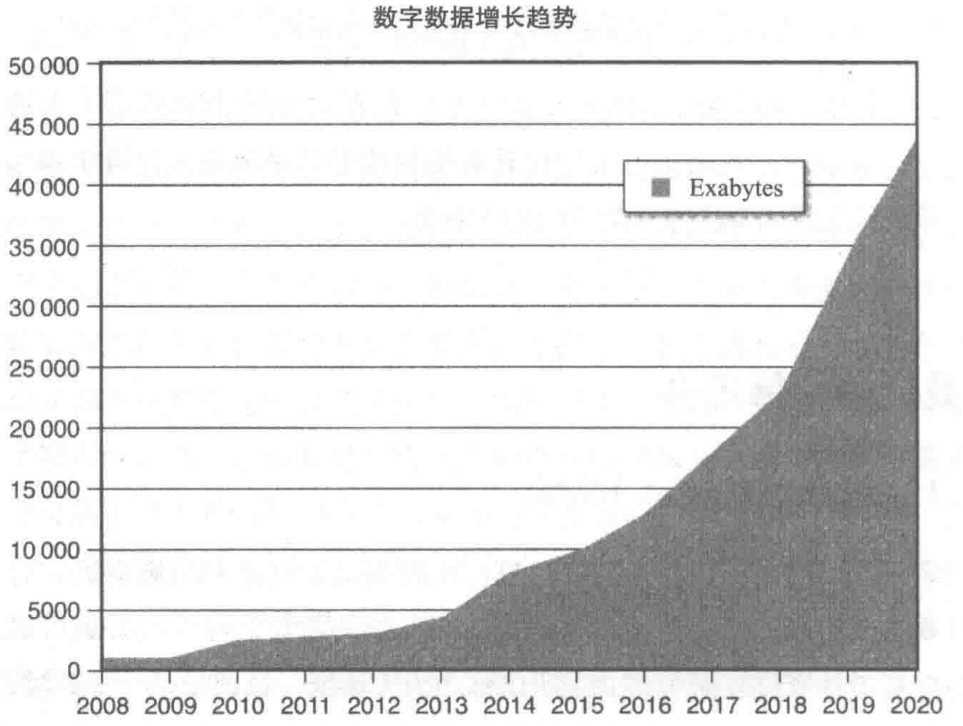

图 1-1　数字数据的爆发式增长

数据来源：IDC

做一个对比会有帮助。例如，据估计，在 2011 年左右的时候，全球产生的数据量大约超过 1.8ZB（1.8 万亿字节），那时电子化存储的字节数像宇宙中的星星一样多。或者想象一下每天有 25PB 的新数据进入互联网，这比美国国会图书馆所有收藏的 70 倍还多。IDC 预测数字化宇宙将增长 10 倍，数据量将从 2013 年的 4.4EB 增长到 2020 年的 44EB。

最好是想想我们比较熟悉的交易形式。例如，每分钟就会有 48 小时时长的新视频上传至 YouTube。同样在这 60 秒内，Facebook 记录了 34 722 个“喜欢”（likes），全球范围内创建了 571 个新站点。在 1 小时里，沃尔玛的 POS 机记录了 100 多万个客户交易。全球范围内每三天就会有 1800 多亿次的邮件