大数据与智慧社会：数据驱动变革、构建未来世界

作者：张克平;陈曙东

版权信息

书名：大数据与智慧社会：数据驱动变革、构建未来世界
作者：张克平；陈曙东
出版社：人民邮电出版社
出版时间：2017年6月
ISBN：9787115456243
本书由人民邮电出版社授权得到APP电子版制作与发行

内容提要

大数据正在改变人们的生活、社会的运行方式以及各行业的竞争生态，是提升政府治理水平和企业竞争力的核心要素。然而，政府和企业如何才能抓住大数据带来的宝贵机遇，改善公共服务、激发商业创新？推进大数据应用的进程对现有技术框架、管理机制、评价体系又有哪些新的要求？

针对这一系列问题，《大数据与智慧社会》一书做出了系统的回答。本书从全局出发，对大数据的基本内涵进行了系统描述，概括了大数据的前世今生，揭示了其哲学本质；以技术为主线，深刻剖析了大数据的技术框架，预测了大数据的技术发展趋势；理论与实践相结合，形成大数据系统评价标准；选取大数据在生活、政务、交通、医疗、金融领域落地应用的实战案例，进行深入分析和解读，以期为我国的政府治理、经济发展、企业创新提供有效的指导和帮助。

本书适合政府决策者、企业管理者、IT实施者（CTO、CDO、技术人员等）以及高等院校相关专业的师生阅读。

本书编委会

主编　张克平　　陈曙东

编委（按姓氏笔画排序）

孔聪聪　　张　亮　　杜　蓉　　李伟炜　　倪　民

推荐序一

“大数据”的概念从问世到现在仅几年时间，却在全球引起了一次又一次热潮。这其中有两个重要因素在起推动作用。第一个是人类社会在发展过程中对信息的渴求。但是为什么直到5年前才“突然”出现大数据的概念？这就引出了第二个因素——以传感技术、互联网、移动智能终端为代表的一系列新信息技术，使信息的获取、利用和集聚在数量、作用和影响力等方面发生了突飞猛进的变化，成为推动历史进入新阶段的根本原因之一。这一切，深深植根于大数据的内在含义中。

认识大数据的本质，就是认识信息资源的本质。信息资源在人类发展的全部历程中扮演着极其重要的角色。语言这种特定的信息形式使人类摆脱了相互交流的障碍，个体的发现和能力可以在一个群体扩散利用，加快了人类进化的步伐。从结绳记事、岩画到文字的诞生，这些方式的出现使人类信息的交流摆脱了口口相传的时空约束，使经验和知识有了客观载体，可以跨越时空，显著加速了人类文明的进化。活字印刷、机器印刷的发明提升了信息生产和传播的效率及质量，推动了农业文明的快速发展。各类书籍、报刊、杂志的出版发行，各种藏书楼、图书馆的产生，为人类知识的汇集和利用提供了新的平台，加速了科技和文明的发展，推动了农业社会向工业社会迈进。

1970年，哈佛大学的奥汀格教授和他的研究队伍提出了信息、材料、能源是推动人类社会进步的三种基本资源的论断。40年后，全球经济发展的实践证明了这一论断的正确性。把握大数据本质，就是要深刻理解“信息资源是推动人类社会进步的一个基础资源”这一观点。

传感技术、互联网、虚拟现实、大数据等一系列新信息技术的诞生和发展，使人类对信息的处理、传输、利用能力得到全面的提升，信息资源在社会发展中的作用日趋重要，推动着工业社会向信息社会迈进。2008年金融危机后，一些欧洲国家又相继发生主权债务危机，与贸易保护主义、恐怖主义、南北不平衡等重大全球性问题纠缠在一起，全世界的理论家、战略家、政治家都在思考同一个问题，如何使人类社会摆脱危机，走向新的发展阶段。从2011年开始，新工业革命、第三次工业革命、互联网能源、工业革命4.0、CPS、两化融合和两化深度融合、第二个机器时代等概念不断产生和发展。麻省理工学院（MIT）的埃里克·布林约尔松（Erik Brynjolfsson）和安德鲁·麦卡菲（Andrew Mc Afee）合著的《第二个机器时代》（The Second Machine Age）提出，我们将经历人类历史上两个最神奇的事件：创造真正的机器智能，以及全体人类通过一个共同的数字网络互联互通、从根本上改变地球经济的格局。第二个机器时代与第一个机器时代的不同之处在于智能化。第一个机器时代的机器取代并倍增了人类和动物的体力劳动，第二个机器时代的机器将取代并倍增我们的智慧。

这些现象和趋势的共同指向就是经济社会正在发生重大变革，这个变革的核心是信息技术体系和工业技术体系的融合，信息资源与能源、材料的协同，人类社会的经济和社会活动将以赛博–物理空间为依托。大卫·兰德斯指出：“工业革命是指生产方式上的深刻变革。即通过用机器代替人工、用非生物力代替人力和畜力，实现从手工工业向机器大生产的转变。”由于工业技术体系本身已经不足以从根本上继续提供推动历史转型的技术能力，人类需要构建赛博–物理空间，将信息、材料、能源三种资源利用综合起来，提升到一个新的水平。信息、材料、能源是推动经济社会发展的三驾马车，工业革命形成的生产力和信息革命形成的生产力推动人类社会进入一个新的历史阶段，已成为历史发展的必然要求。

面向未来，抓住大数据技术带来的机遇，主动推进社会发展变革，要特别重视技术、产业和应用。从技术的角度来看，主要有两大问题：一是大数据每隔几年就会提升一个数量级，从这个角度看，如今的计算机处理体系不符合大数据处理的需求，所以要从芯片开始重构适合大数据发展的处理系统，要有新的芯片和新的处理结构，这是技术问题的一个制高点；第二个制高点是大数据的语义处理能力，也是智能技术的核心部分，这一技术将成为今后一个阶段信息技术创新的核心内容。从产业角度看，大数据产业大概可以分为两类：一类是“技术变成产业”，就像当年数据库管理系统变成了数据库公司，当真正的大数据处理芯片和计算架构形成时，还将会形成新的产业；另一类是各个企业、机构甚至个人（以后我们很多人）都可以变成大数据的拥有者和大数据产业的从业者。从应用的角度看，大数据最重要的意义在于，所有企业、机构和个人如何将大数据变成自身提升能力、提升竞争力、提升生活质量的来源，“以信息化培育新动力、以新动力推动新发展”，用“信息流引领技术流、资金流、人才流、物质流”，使其

成为资源配置优化、全要素生产率提升、经济社会发展转型、经济结构调整的新动能。

《大数据与智慧社会》一书系统地介绍了什么是大数据和大数据技术框架，详细分析了以Hadoop和Spark为代表的典型技术，介绍了大数据在生活、政务、交通、医疗和金融领域的应用，为我们认识大数据、利用大数据提供了又一份精神食粮。更要指出的是，本书出自几位基层信息岗位的主管，着实难能可贵。

是以为序。

杨学山

工业和信息化部原副部长

北京大学教授

推荐序二

随着信息技术的飞速发展，“大数据”已被认为是继互联网、云计算、物联网之后又一大颠覆性的技术革命。通过对海量、动态、高增长、多元化数据的高速处理，大数据正在引发全球范围内的经济和商业变革，其应用涉及金融、交通、教育、医疗、制造、环保、零售、文化、娱乐等各行各业。大数据更带来了一场政府治理方式的变革，在提高公共决策能力的同时，改变着国家治理的架构和模式。可以毫不夸张地说：“大数据时代没有旁观者。”

研究发现，即使在缺乏精准的数据分析模型和算法的情况下，只要拥有足够多的数据，也能揭示事物的内在联系，引出重要的结论。这给计算衍生学科带来了里程碑式的启示：大数据本身可以保证数据分析结果的有效性。因此，大数据被誉为新的生产力。在大数据时代，大数据生产力将会推动生产关系和社会的发展，创造无穷无尽的价值，给人类思维的发展带来变革。

大数据使我们至少拥有了四个方面的核心能力。

首先是海纳百川的数据融合能力。通过将各种数量庞大、分布广泛、形式多样、变化迅速的异构数据资源汇聚、融合在一起，资源数量和质量的巨大提升引发资源价值的巨大提升，使大数据成为现代社会的巨大财富。

其次是基于大数据的科学研究能力。基于这种能力的科研范式有别于传统的实验归纳、模型推演、仿真模拟等范式，被称为数据密集型科学发现，即第四范式。应当指出，运用这种能力，当数据达到一定量时，传统计算架构已经不再适用，云计算应运而生。实际上，大数据与云计算相辅相成，两者之间互相推动与促进。没有云计算能力，大数据的价值就无法被挖掘出来；没有大数据，云计算也就没有用武之地。

再次是明察秋毫的洞见力。透过数据发现隐藏在事物表面下的本质规律，发现事物之间的关联，揭示事物发展的规律，便于人类发现新的原理或者产生新的科学创造。这种运用第四范式获得的科学发现，既不像理论和模拟那样在一定程度上告诉我们“为什么”，也不像实验那样明确地告诉我们“是什么”，只能告诉我们“与什么相关”。第四范式强调了以大数据为基础的数据密集型研讨方法，这种方法将会在越来越多领域的研讨中发挥至关重要的甚至是决定性的作用。

最后是高瞻远瞩的预测和决策能力。在过去的商业决策中，管理者凭借自身的经验和对行业的敏感来决定企业的发展方向和方式，这种决策有时候仅仅参考一些模糊的数据和建议。而大数据和大数据分析工具的出现，让人们找到了一条新的科学决策之路。以数据为依据，立足事实，既观全局，又见未来。

大数据正是因为能赋予我们这四种核心能力，才受到越来越多的关注。我国已经将大数据提升到国家战略高度，在“十三五”规划纲要中指出：实施国家大数据战略，把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新；深化大数据在各行业的创新应用，探索与传统产业协同发展新业态、新模式，加快完善大数据产业链；加快海量数据采集、存储、清洗、分析发掘、可视化、安全与隐私保护等领域关键技术攻关。习近平总书记强调，机会稍纵即逝，抓住了就是机遇，抓不住就是挑战。我国发展大数据有非常好的机遇，同时我们也应该清醒地认识到，我国大数据产业刚刚起步，从技术上、观念上、法律上等多个层面都需要变革，才能满足大数据的发展需要。

由张克平局长、陈曙东研究员主编的《大数据与智慧社会》一书，顺应时势，系统地从大数据起源、大数据哲学本质、大数据技术框架、大数据应用案例等不同的角度为读者展示了一幅大数据技术图谱。该书首先概述了大数据的哲学本质、技术现状和发展趋势，然后详述了大数据的技术框架、大数据存储和大数据处理技术。“科学家要多做实践中的研究”，当前，大数据应用处于起步期，产业生态处于酝酿期，必须在实际应用中发挥大数据技术的作用，推动大数据产业的发展。因此，作者们又详述了大数据在生活、政务、交通、医疗和金融等相关领域的应用实战，为读者使用大数据指出了一条探索之路。

我相信本书将受到关注大数据的“政产学研用”各界的欢迎，为大数据在中国的发展助一臂之力。

倪光南

中国工程院院士

第1章
大数据概述

1.1 什么是大数据

1.1.1 大数据的定义和特征

什么是大数据？目前业界有多种定义和理解方式。

大数据的基本定义

最早进入人们视线的大数据的3V定义，由高德纳（Gartner）分析师道格·莱尼（Doug Laney）在2001年提出。3V分别代表Volume（数据规模大）、Velocity（快速的数据流转和动态的数据体系）、Variety（多样的数据类型）。2012年，高德纳修改此定义为“大数据是大量、高速、和/或多变的信息资产，它需要新型的处理方式来促成更强的决策能力、洞察力与最优化处理”。

美国咨询公司麦肯锡在其报告《大数据：下一个创新、竞争和生产力的前沿》（Big data：The nextfrontier for innovation,competition,and productivity）中对大数据给出的定义是：大数据是指大小超出常规的数据库工具获取、存储、管理和分析能力的数据集合。但它同时强调，并不是说一定要超过TB值的数据集才能算是大数据。

大数据企业Hortonworks 公司战略副总裁肖恩·康诺里（ Shaun Connolly）认为，过去的资料大部分是人工手记下来的交易资料（Transactions），现在则是机器替我们记录下来的交易资料；除此之外，还有人们跟事物、企业间的互动资料（Interactions），如人们在网络上点击网页和链接的记录；最后则是由机器自动生成、累积下来的观察资料（Observations），如智慧家居产品记录下来的室温变化数据等。因此，肖恩·康诺里定义大数据是由交易资料、互动资料、观察资料所组成的资料类型。

著名的摄影师和出版人里克·斯莫兰（Rick Smolan）在其著作《大数据的人性面孔》（The Human Face of Big Data）一书中从哲学角度对大数据进行了定义：大数据是帮助地球建构神经系统的一个过程，在这个系统中，我们（人类）不过是其中一种感测器。

大数据的5V特征

要想充分了解大数据，除了各种定义以外，我们还必须了解大数据的特征。在高德纳给出大数据的3V特征之后，业界人士在此基础之上陆续提出了更多“V”，如Veracity（真实性）、Validity（可验证性）、Value（数据有价值）、Visibility（可视化）等，其中以 Value最被普遍认同，这就是IDC（国际数据公司）给出的大数据4V特征。还有其他一些机构将Veracity（数据真实性）也纳入大数据的特征描述中来，形成了5V特征。

本书将分别从数量（Volume）、多样性（Variety）、速度（Velocity）、价值（Value）以及真实性（Veracity）这五个方面来剖析大数据的特征。

数量（Volume）：这个特征描述的是汇聚在一起进行分析的数据规模非常庞大。那么，大数据时代的数据规模究竟有多大呢？我们来看一组测算数据。全球的数据以每年40%的速度增长，截至2011年4月，美国储存数据量最大的国会图书馆拥有235TB（1TB=1024GB）的数据；2010年，全球企业硬盘上存储的数据超过7EB（1EB=10亿GB），相当于美国国会图书馆中存储数据的3万倍；全球消费者在个人电脑上存储了超过6EB的数据；美国的17个行业大类当中，有15个行业的数据储藏量超过了美国国会图书馆。

多样性（Variety）：数据形态多样，可以被归类为结构化、半结构化和非结构化数据。相对于以往能够用数据或统一的结构加以表示的结构化数据，半结构化和非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这些多类型的数据对数据的处理能力提出了更高要求。

速度（Velocity）：这个特征描述了大数据的高速流转和其动态的数据体系，表现为数据量的增长速度快。如今的数据量大约每20个月就能增长一倍。同时，系统要能够快速地处理这些数据，时效性要求高。例如，搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法要尽可能地实时完成推

荐。这是大数据分析区别于传统数据挖掘的显著特征之一。

价值（Value）：这个特征包含两个方面的含义。一方面指大数据价值密度低。尽管我们拥有大量数据，但能够真正发挥价值的仅是其中非常小的部分。以监控视频为例，一段时长为1小时的不间断视频，其中关键视频数据可能时长仅为1~2秒。另一方面指大数据背后潜藏的价值巨大。例如，美国社交网站Facebook有10亿用户，网站对这些用户信息进行分析后，广告商可根据结果精准投放广告。对于广告商而言，10亿用户的数据价值上千亿美元。

真实性（Veracity）：对于虚拟网络环境下如此大量的数据采取措施确保其真实性、客观性，是大数据技术与业务发展的迫切需求。同时，通过大数据分析真实地还原事物的本来面目、预测事物的发展规律，也是大数据应用未来发展的趋势之一。

大数据的十字特征

大数据是一个宽泛的概念，以上任何一个定义和特征均无法全面地概括大数据的特点。因此，本书用“大杂全多快，久活简稀联”概括描述大数据从产生、存储、处理到应用这一全生命周期内区别于传统数据的特征，称为大数据的“十字特征”。

一“大”：数据体量巨大，对应于4V描述中的“Volume”。截至目前，人类生产的所有印刷材料的数据量是200PB（1PB=1024TB），而历史上全人类说过的所有话的数据量大约是5EB（1EB=1024PB）。

二“杂”：数据类型多种多样，对应于4V描述中的“Variety”。

三“全”：大数据应用为决策者提供一个业务的全局视图，这里主要是强调数据的业务完备性。

四“多”：大数据的数据来源多、维度多，不仅包含企业内部业务数据，而且包含许多相关的外部数据，如政策数据、经济数据、气象数据、环境数据等。这里主要强调引入外部数据源构建数据的多维性。

五“快”：大数据强调的是在线数据的实时分析处理，这是大数据区分于传统数据分析的最显著特征。从这个角度来说，对应于4V描述中的“Velocity”。在如此海量的数据面前，处理数据的效率就是企业的生命。由于业务变化速度加快，数据的贬值速度也被加快。数据如果不能被及时地分析利用，其价值会快速贬值。只有及时地挖掘数据背后的价值，数据之间的联系才会随之变得更加紧密，就像滚雪球一样更加有利于发现数据背后的价值。

六“久”：大数据应用十分重视数据的长期积累，数据积累时间越长，越有利于发现数据间内在的相关性。

七“活”：数据是在线的，可以随时调用和计算，这是大数据区别于传统数据的最大特征。放在磁盘或磁带中的离线“死”数据，其商业价值远远不如在线的“活”数据。

八“简”：在使用的分析算法上，大数据算法突出了简单、易行的特点。这也是不同于传统数据挖掘采用小数据复杂算法的显著特征之一。

九“稀”：对应于4V描述中的“Value”，即大数据应用中真正有价值的数据占比极少。有句谚语可以形象地比喻大数据“稀”的特点：“为了那一点点的金子，我们不得不保存所有沙子。”因此，如何通过有效的计算更迅速地完成数据的价值“提纯”已成为大数据亟待解决的问题。

十“联”：大数据更加关注数据间的关联性。近现代科学最重要的特征是寻求事物的因果性，无论是唯理论，还是经验论，区别只在寻求因果关系的方式不同。大数据最重要的特征是重视现象间的相关关系，并试图通过变量之间的依随变化找寻它们的相关性，从而不再一开始就把关注点放在内在的因果性上，这是对因果性的真正超越。

不管如何定义大数据、如何描述其特征，各大研究机构、企业都对其影响力进行了评估。我们从中可以看到其背后都隐含的一个共识：大数据已经成为科学和创新领域的前沿话题，需要全新的基础设施、分