BIG DATA
BIG TALK

中国大数据发展方向

大数据大家谈

互联网时代谁主沉浮　大数据带你找寻机遇

张华平　商建云　段永朝　白硕　等◎著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据千人会

大数据大家谈

张华平　商建云　段永朝　白　硕　吴甘沙
刘　驰　高　凯　沈　浩　曹　娟　张勇东
郝雅婕　郎清平　吕晓辉　张洪忠　梅其文
邓　宁　刘春阳
著

電子工業出版社
Publishing House of Electronics Industry
北京・BEIJING

内 容 简 介

本书邀请了 14 位国内外大数据产学研有影响力的一线专家学者，总结各自的研究与工作专长，以专题的形式发表了各自的研究成果。本书主要包括了大数据综述、大数据思维、大数据技术与大数据应用四个部分。其中，大数据综述主要介绍大数据的概念、背景、技术与国内外政策等，让读者对大数据有个全景式的了解；大数据思维包括大数据的开放式创新与流动的大数据两个方面；大数据技术分别介绍了大数据平台架构、大数据语义分析、情感分析、大数据可视化、多媒体搜索分析等当前的技术热点；大数据应用主要介绍了新媒体、企业大数据基础设施、金融风控等方向的应用实践。本书适合大数据行业研究者、技术开发工程师与研究人员使用。

图书在版编目（CIP）数据

大数据大家谈 / 张华平等著. — 北京：电子工业出版社，2017.1

ISBN 978-7-121-30181-0

Ⅰ. ①大… Ⅱ. ①张… Ⅲ. ①数据处理－研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 255734 号

策划编辑：李 敏
责任编辑：李 敏 特约编辑：刘广钦
印 刷：三河市双峰印刷装订有限公司
装 订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：720×1000 1/16 印张：17.25 字数：367 千字
版 次：2017 年 1 月第 1 版
印 次：2017 年 1 月第 1 次印刷
定 价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-88254753 或 limin@phei.com.cn。

序

近年来，大数据引起了政府部门、产业界、科技界与学术界的高度关注。2008 年 9 月，*Nature* 杂志发表了文章 *Big Data*: *Science in the Petabyte Era*，“大数据”这个词开始广泛传播。2012 年 3 月 22 日，奥巴马宣布美国政府投资 2 亿美元启动“大数据研究和发展计划”。在此基础上，美国又于 2016 年 5 月发布了《联邦大数据研究与开发战略计划》，其目标是对联邦机构的大数据相关项目和投资进行指导。中国政府于 2015 年 9 月发布了《促进大数据发展行动纲要》，明确指出：坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

目前，我国互联网、移动互联网用户规模居全球第一，拥有丰富的数据资源和应用市场优势，大数据部分关键技术研发取得突破，涌现出一批互联网创新企业和创新应用，一些地方政府已启动大数据相关工作。与此同时，大数据产业也随之蓬勃发展，市场研究公司 Marketsand Markets 公布的报告显示，2013—2018 年，全球大数据市场的年复合增长率预计为 26%，将从 2013 年的 148.7 亿美元增长至 463.4 亿美元。中国大数据产业虽然起步晚，但近年来发展速度快。2014 年，中国大数据市场规模达到 767 亿元，同比增长了 27.8%。预计到 2020 年，中国大数据产业规模将达到 8228.81 亿元。

同时，我们还必须清晰地认识到，当前大数据还处在快速成长期。科学研究、技术开发与产业应用都处在探索阶段，缺乏科学的标准，企业也缺乏明确的评价指标，与成熟产业健康有序发展还有距离。目前大数据产学研均存在一定的炒作和泡沫，遍地开花的大数据产业园、大数据项目和投资，几乎无人不谈大数据，不同专业的学者均会做有利于自己的大数据解释，各类企业纷纷高举大数据的旗帜吸引投资，经过技术炒作周期，大数据已经成为了民众的科学常识。盲目的炒作与投资实际上违背了大数据的科学发展规律，对大数据产学研的健康发展是极其不利的。

大数据涉及方法论层面的哲学思考，也包括大数据的架构、平台、存储与硬件等基础性平台，同时还包括了大数据处理、挖掘、分析与可视化等大数据技术；从数据形态上，大数据又分为结构化大数据与非结构化大数据，从媒体形态上，还包括了大数据文本、语音、视频等；大数据的应用则更加宽泛。国内的大数据论著侧重于大数据处理的分布式架构方面，如 Hadoop、Spark 等平台；而且大部分书籍重在阐述大数据思维，如英国牛津大学的维克托·迈尔-舍恩伯格教授的《大数据时代》、涂子沛的《大数据》。但是，还缺乏大数据相对综合而又理性权威的论述著作。

2015 年年初，笔者组织创立了中国大数据千人会，吸引了国内外大数据政产学研相关的专业人士数千人，并邀请了国内外一线的大数据专家在线演讲，先后做了 30 多期（后因工作繁忙，很遗憾未能持续进行）。为凝炼整理多期的访谈成果，笔者从所有演讲中优中选优，邀请了 14 位大数据产学研有影响力的专家学者，将演讲稿进一步凝练，各负其责，每位专家一章，只写自己专注研究的部分，要求去除水分只留干货，综合写作了《大数据大家谈》。该书名隐含两层意思，一方面指的是本书由大家一起写作，非一家之言；另一方面，每章的写作者基本上都是大数据特定方向上有影响力的“大家”。

本书主要包括四个部分共 14 章，分别是大数据综述、大数据思维、大数据技术与大数据应用。第一部分（第 1 章）大数据综述由北京理工大学商建云执笔，对大数据的概念、背景、技术与国内外政策等进行介绍，让我们对大数据有个全景式的了解。第二部分（第 2～3 章）大数据思维，分别由两位杰出的大数据实践者与思想家完成。驭势科技 CEO 吴甘沙先生写作的《大数据的开放式创新》，提出了开放的数据、基于数据安全流通和定价的数据市场、开放的基础设施、开放的社会化分析服务、跨越领域界限的开放数据思维五点大数据创新过程；吴甘沙先生是英特尔中国研究院前院长，笔者有幸聆听过他关于大数据的开放式创新的演讲，确实脑洞大开。财讯传媒集团首席战略官段永朝对互联网与大数据有过很多冷静的哲学思考，也是网络智酷的发起人，定期的沙龙吸引了大量的专家学者，他所写作的《流动的大数据》一文，延续了段永朝的深入思考。第三部分（第 4～8 章）大数据技术，分别由北理工刘驰教授、北理工张华平副教授、河北科技大学高凯教授、中国传媒大学沈浩教授、中国科学院计算技术研究所的曹娟博士分别介绍了大数据平台架构、大数据语义分析、情感分析、大数据可视化、多媒体搜索分析等当前的技术热点。第四部分（第 9～14 章）大数据应用，主要是介绍大数据的落地实践，我们分别邀请了清博大数据的郝雅婕、上海证券交易所的白硕研究员、美国律商联讯风

险信息公司吕晓辉博士、北京师范大学张洪忠教授、大象金服研究员梅其文、北京第二外国语学院邓宁博士六位做学术与产业的专家分别就新媒体、企业大数据基础设施、金融行业应用、大数据传播第四范式、金融大数据等话题介绍了各种的实践总结分析。

在本书的策划写作过程中，得到了不少专家学者的指点与参与，同时也通过大数据千人会公众号收集了几百万感兴趣的读者反馈。在这里，特别感谢互联网实验室的方兴东博士的前期倡议，感谢北京理工大学黄河燕教授、赵燕平教授，以及大数据搜索与挖掘实验室潘红岩、徐程程、吴松泽、张亚男等多位同学的前期工作。同时，我们还要感谢电子工业出版社的李敏博士的精心编辑与整理。最后，还要感谢我的太太曾飞和孩子的支持。

本书作为大数据的跨界融合之作，希望提供更多视角，以更严谨务实的方式为各位朋友提供冷静的思考。水平有限，敬请批评指正。

张华平

2016 年 9 月

目 录

第 1 章

大数据技术及其相关政策

商建云　张华平　北京理工大学
刘春阳　国家计算机网络应急技术处理协调中心

1.1 大数据产生的背景

随着信息化技术的普及和发展，人类社会积累了大量的数据，早期利用数据库进行数据的存储和分析。随着软硬件技术和各种设备的不断更新，物联网、移动互联网、车联网、手机、平板电脑、PC 及遍布全球的各种各样的传感器，都成为数据来源或承载的方式。数据的格式也因此具有了多样化的特征。

以互联网中的社交网络为例，以微信、新浪微博、腾讯微博、Twitter 与 Facebook 等为代表的新型社交网络的迅猛发展，对经济与社会逐渐产生了重大影响。目前，全球约有 46 亿移动电话用户，有 20 亿人访问互联网。看看我们周边的人们，大家在公共交通工具上，甚至在饭桌上，都在用手机连接互联网。互联网环境也成为必备的基础设施，人们比以往任何时候都更加频繁地与数据或信息进行交互，信息流成为人们生活的重要组成部分。根据国际数据资讯（IDC）公司监测，全球数据量大约每两年翻一番，预计到 2020 年，全球将拥有 35ZB 的数据量，并且 85%以上的数据以非结构化或半结构化的形式存在。

纵观大数据的产生历程，其产生背景可以从人、社会和自然三个层面进行剖析，如图 1-1 所示。下面以每个层面的个别例子进行分析，便可窥见一斑。从人的层面上看，随着技术的发展，人体腕表等可穿戴设备的出现为人们关注个体的生命质量过程创造了条件；从社会的层面上看，大量的个体的人形成的人群的活动和过程记录产生的大量数据，可以用于引导人际活动，如交通状况的实时播报、拥挤地段的疏散等；从自然的层面上看，对环境的演化可以通过各种传感器、物联网等手段获取大量环境演化过程记录数据，如大家关心的 PM2.5，可以进行环境分析，指导排碳期货交易等。可以看出，数据量的快速增长加上数据采集的便利性和成本低廉性，细节数据展现在人们的面前，拓展了人们的认知深度和广度。正如史蒂夫·洛尔在《大数据主义》一书中所说，大数据技术，就是数字时代的“望远镜”或者“显微镜”，使我们可以看到并计量之前我们一无所知的新事物。“望远镜”让我们看得更远，发

现新的星系；而“显微镜”则将比细胞更微小的神秘世界展现在我们眼前。

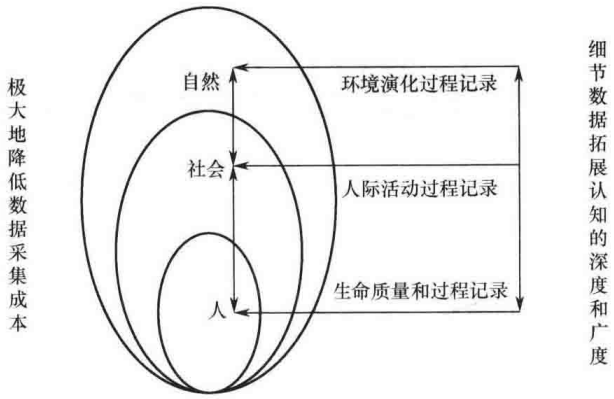

图 1-1　大数据的来源

“大数据”术语的广泛传播始于 2008 年 9 月 *Nature* 杂志发表的文章 *Big Data: Science in the Petabyte Era*。2012 年 3 月 22 日，奥巴马宣布美国政府投资 2 亿美元启动“大数据研究和发展计划”。大数据是与自然资源、人力资源一样重要的战略资源，是一个国家数字主权的体现。2013 年，“大数据”的发展呈现了燎原之势；2015 年是大数据的落地之年。大数据成为新一轮的科技革命，已经引起了政府部门、产业界、科技界与学术界的高度关注，是信息技术发展的新趋势。从政府采购网上公布的项目也可以看出，我国在大数据的应用上发展得如火如荼。正如 19 世纪工业革命的技术变革一样，近期和未来几年大数据正在也将成为新的技术变革，是社会和经济发展的动力。如图 1-2 所示，按 Gartner 2015 年 8 月给出的预测来看，大数据分析中的机器学习技术将在 2017—2020 年达到使用高点。

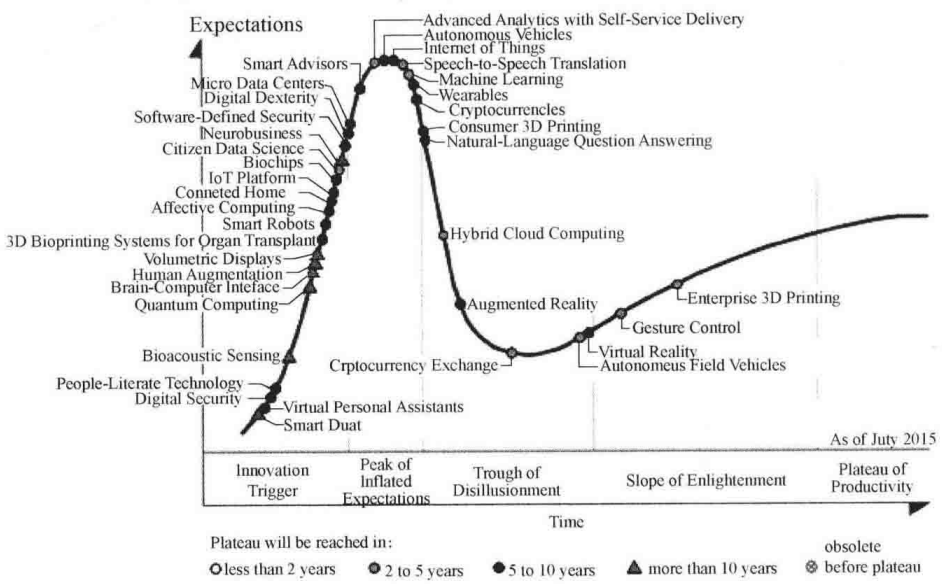

图 1-2　Gartner 新技术预测

大数据的出现带来了人们生活方式的改变、价值观的更新。如何利用好大数据为人类服务，成为自工业革命后信息革命的关键。

1.2　大数据的概念和特征

1.2.1　大数据的概念

关于大数据如何定义，研究机构 Gartner 的定义是：大数据是指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。麦肯锡的定义为：大数据是指无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合。舍恩伯格·维克托的《大数据时代》中的定义为：大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法。

北京理工大学张华平副教授给出的定义是：大数据是指从客观存在的全量超大规模、多源异构、实时变化的微观数据中，利用自然语言处理、信息检索、机器学习等技术抽取知识，转化为智慧的方法学。

无论哪种定义，我们可以看出，大数据并不是一种新的产品，也不是一种新的技术，就如同 21 世纪初提出的“海量数据”的概念一样，大数据只是数字化时代出现的一种现象。那么海量数据与大数据的差别何在？从翻译的角度来看，“大数据”和“海量数据”均来自英文，“Big Data”翻译为“大数据”，“Large-scale Data”翻译为“大规模数据”，“Very Large Data”翻译为“超大规模数据”，“Massive Data”则翻译为“海量数据”。从组成的角度来看，海量数据包括结构化和半结构化的交易数据，而大数据除此以外还包括非结构化数据和交互数据。Informatica 大中国区首席产品顾问但斌进一步指出，大数据意味着包括交易和交互数据集在内的所有数据集，其规模或复杂程度超出了常用技术，按照合理的成本和时限捕捉、管理及处理这些数据集的能力。可见，大数据由海量交易数据、海量交互数据和海量数据处理三大主要技术趋势汇聚而成。

1.2.2　大数据的特征

大数据的特征包含四个层面。第一，数据体量巨大。从 TB 级别，跃升到 PB 级别。第二，数据类型繁多。例如，网络日志、视频、图片、地理位置信息等。第三，价值密度低。以视频为例，在连续不间断地监控过程中，可能有用的数据仅仅有一两秒。第四，处理速度快。1 秒定律，最后这一点和传统的数据挖掘技术有着本质的不同。业界将大数据的特征归纳为 4 个“V”，即 Volume、Variety、Value、Velocity。

1．数据体量巨大（Volume）

大数据通常指 10TB（1TB = 1024GB）规模以上的数据量。之所以产生如此巨

大的数据量，一是由于各种仪器的使用，使我们能够感知到更多的事物，这些事物的部分甚至全部数据就可以被存储；二是由于通信工具的使用，使人们能够全时段联系，机器—机器（M2M）方式的出现，使得交流的数据量成倍增长；三是由于集成电路价格降低，使很多东西都有了智能的成分。

2．数据种类繁多（Variety）

随着传感器种类的增多及智能设备、社交网络等的流行，数据类型也变得更加复杂，不仅包括传统的关系数据类型，也包括以网页、视频、音频、E-mail、文档等形式存在的未加工的、半结构化的和非结构化的数据。

3．价值密度低（Value）

数据量呈指数增长的同时，隐藏在海量数据中的有用信息却没有以相应比例增长，反而使我们获取有用信息的难度加大。

4．流动速度快（Velocity）

我们通常理解的数据流动速度是指数据获取、存储及挖掘有效信息的速度，由于我们现在处理的数据是 PB 级代替了 TB 级，“超大规模数据”和“海量数据”也有规模大的特点，数据是快速动态变化的，因此形成流式数据是大数据的重要特征，数据流动的速度快到难以用传统的系统去处理。

大数据的“4V”特征表明其不仅仅是数据海量，对于大数据的分析将更加复杂、更追求速度、更注重实效。

大数据独立的发展形成特有的市场化与规模化，也充分带动了其他行业与大数据的广泛、充分融合，从而推进了大数据的全面落地。大数据从产业到行业的成熟将推动更多传统企业向科技智能化转型，将推进政府政务大数据发展，也将鞭策大数据行业在中国平稳落地，与其他各行业共襄中华民族伟大复兴之盛举。

大数据之大，还在于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。

1.3 大数据技术发展趋势

1.3.1 大数据带来的决策方式的革命

近半个世纪以来，我们经历了计算机时代计算方式的革命、互联网时代信息传播方式的革命、大数据时代决策方式的革命，如表 1-1 所示。