Big Data and Computational Intelligence

大数据与计算智能

柴园园　贾利民　陈　钧　著

科学出版社

大数据与计算智能

柴园园　贾利民　陈　钧　著

科学出版社
北京

内 容 简 介

本书通过深入探讨计算智能的理论起源和计算本质，归纳大数据处理流程中有待解决的核心问题，总结出基于计算智能的处理范式及算法流程，并对部分模型进行实验分析。全书共六章，主要内容包括：大数据及相关概念、大数据理论研究、大数据面临的主要问题、计算智能基础、计算智能与大数据处理以及计算智能在大数据领域的应用前景展望。

本书可供进行计算智能及其分支算法理论学习及研究的本科生、研究生及科研人员使用，也可供从事大数据相关工作的技术人员参考。

图书在版编目(CIP)数据

大数据与计算智能 / 柴园园，贾利民，陈钧著. —北京：科学出版社，2017.1

ISBN 978-7-03-050616-0

Ⅰ.①大… Ⅱ.①柴… ②贾… ③陈… Ⅲ.①人工神经网络-计算 Ⅳ.①TP183

中国版本图书馆 CIP 数据核字(2016)第 271853 号

责任编辑：耿建业 武 洲 / 责任校对：郭瑞芝

责任印制：张 倩 / 封面设计：铭轩堂

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

中国科学院印刷厂印刷

科学出版社发行 各地新华书店经销

*

2017 年 1 月第 一 版 开本：787×1092 1/16

2017 年 1 月第一次印刷 印张：15 3/4

字数：375 000

定价：78.00 元

（如有印装质量问题，我社负责调换）

前　言

世界的本源是运动的、变化的、由问题组成的。对于科学问题的求解，是人类从事科学研究活动的初衷。问题的提出是人类智慧和想象力的结晶，代表寻找到了一个新的探索方向，从这个视角出发，发现科学问题甚至比解决科学问题更加重要。

伴随 Web 2.0 模式下的互联网信息爆炸和人类行为创造的数据海洋，“大数据”的概念被广泛提出。从各大搜索引擎的搜索量排名，到各大行业的科研调查报告，随处可见“大数据”的影子。大数据的浪潮席卷之际，我们不由自主地对这个未知的概念心存顾虑。如何对大数据复杂的内在结构及其涌现机理进行研究，如何对大数据引发的共性科学问题进行分层次抽象，如何探索区别于传统思维方式的大数据前瞻性理论和方法。这一系列科学问题的解决，势必会给人类的哲学认知和科技水平带来颠覆性的变革。

不同于传统符号主义人工智能的结构模拟，基于连接主义思想的计算智能本着功能实现的研究宗旨。计算智能从人脑和神经系统的生理背景和智能现象出发，模拟它们的工作原理和学习方式，研究非程序的、适应性的信息处理的本质和能力，涉及神经网络、模糊逻辑、进化计算等多学科的交叉和融合，为解决那些不具有纯粹的解析性，无法进行精确描述，难以建立有效的形式化计算（推理）模型进行求解的问题提供了先进的计算理论和框架。

可以看出，基于数值计算和结构演化的计算智能无疑是解决大数据不同层面问题的一个有效工具。针对大数据处理流程中所面临的几类主要问题及其相互关系，本书在讨论计算智能的非线性映射及自适应的算法特性和计算能力的基础上，提出基于计算智能的解决方法和一般过程，形成了大数据背景下计算智能研究的理论体系架构。

人类思想的解放带动了人类科学研究的进步，科学研究的发展是人类了解自然界，改造自然界，并征服自然界的原动力，是人类生产生活，乃至繁衍繁荣的前提。在创新的道路上，永远没有可以嘲笑的提问者。本书的出版希望可以为致力于大数据理论和计算智能研究的专家学者和思想先锋提供一些新的思考和方向。

本书的撰写要感谢我的导师贾利民教授多年来对我研究工作的辛勤指导和大力帮助；感谢中国国防科技信息中心刘林山主任对我工作的鼓励和肯定；感谢北京交通大学轨道交通控制与安全国家重点实验室秦勇教授对我学术研究的支持。其中，毛彬参与了本书第 1 章 1.2 节和 1.5 节的撰写与校对工作，田昌海参与了第 6 章 6.1 节的整理和撰写工作，叶宇铭参与了第 6 章 6.2 节的整理和撰写工作。大数据实验室的罗威、武帅、罗准辰和田丰参与了专著前期的总体架构设计，薛万鹏、谭玉珊、孙鑫和于洋参与了第 1 章和第 2 章的排版和绘图工作，张吉才、高辉、牛海波和孙登峰参与了第 3 章的材料整理和校对工作。此外，国内外同行对本书的部分研究工作也给予了建议，在此一并表示感谢。

感谢鲁汶大学的 Michel Verleysen 教授对我学术水平的认可及理解，这是本书创作的初衷之一。特别感谢我的爱人、儿子和父母给予我的爱和支持，谨以此书献给他们！

由于作者知识水平及能力有限，书中难免存在不足之处，敬请读者批评指正。

柴园园

2016 年 8 月于北京

目　录

第 1 章　大数据及相关概念

1.1　大数据的产生背景

1.1.1　物理空间、信息空间与赛博空间

任何事物都处于一定的时空之中。近代物理学认为，时间和空间不是独立的、绝对的，而是相互关联的、可变的，任何一方的变化都包含着对方的变化。因而，把时间和空间统称为时空，在概念上更加科学和完整。

其实，“空间”一词不够确切，时空（四维）与空间（三维）有着一个维度的区别。如果把宇宙看作四维“时空”，有一个很重要的原因在于它恰好可以全面地描述发生在人类能够认知的三维空间中的一切事件。在本书中，不考虑“时间”维度，我们简要地介绍“物理空间”、“信息空间”和“赛博空间”（cyberspace）三者的含义及联系。

长期以来，人类一直赖以生存和竞争的空间称为“物理空间”。区别于其他生物，人类在物理空间里不断地发明和制造新的工具，扩大自己生存和感知时空的能力。从依靠自身有限的器官去看、去听、去嗅、去品尝、去抚摸……的直接感知，到应用各种工具如听诊器、望远镜、显微镜、超声探测仪、X 射线、CT 断层扫描……的间接感知。工具的发明和使用大大延伸了人类所能感受的时空领域，强化了我们探索自然界、社会以及人类自身生理和心理的能力。

望远镜和显微镜大大强化了人类探索物理空间的能力，带来了观测领域的一场革命，促进了科学的极大发展[1]。

1608 年荷兰人汉斯·利伯希发明了第一部望远镜。1609 年意大利佛罗伦萨人伽利略·伽利雷发明了 40 倍双镜望远镜，这是第一部投入科学应用的实用望远镜。在现代天文学中，望远镜包括了射电望远镜、红外望远镜、X 射线和伽马射线望远镜。近年来天文望远镜的概念又进一步延伸到了引力波、宇宙射线和暗物质的领域。

望远镜的发明和不断改进大大扩展了人类的视野，1990 年美国发射的哈勃空间望远镜（Hubble space telescope，HST）将人类观察宇宙的能力扩大到银河系以外，测量了宇宙中所见过的最远的星系，打破了宇宙距离记录。美国航天局在 2014 年发射了功能更强大的詹姆斯·韦伯太空望远镜（James Webb space telescope，JWST）替代哈勃空间望远镜。它可以按照天文学家的指令去观测宇宙中的任意星体。使我们的观测距离扩展到 130 亿光年，同时使我们可以追溯到 137 亿

年前，宇宙大爆炸以来宇宙的形成和演变的历史。下一步是建造巨型太空望远镜的 Atlas T 计划，预计要到 2025 年才可能运行。

显微镜是人类 20 世纪最伟大的发明之一。在它发明之前，人类关于周围世界的观念局限在用肉眼或者靠手持透镜帮助肉眼所看到的东西。最早光学显微镜是在 1590 年由荷兰的眼镜制造匠人詹森发明的。这个显微镜是用一个凹镜和一个凸镜做成的，制作水平还很低。詹森虽然是发明显微镜的第一人，却并没有发现显微镜的真正价值。也许正是基于这个原因，詹森的发明并没有引起世人的重视。事隔 90 多年后，显微镜又被荷兰人安东尼·范·列文虎克(Antony van Leeuwenhoek)研究成功了，并且开始真正地用于科学研究试验。关于列文虎克发明显微镜的过程，也是充满偶然性的。后来经意大利伽利略的改良，显微镜具有了更佳的效果。现在的光学显微镜可把物体放大 1600 倍，分辨的最小极限达 0.11 微米。

1932 年，德国科学家诺尔和鲁斯卡(Ruska)制成了世界上第一台电子显微镜，将放大倍数提高到 1 万倍。到 20 世纪 90 年代，世界上已经研制出放大率 200 万倍的电子显微镜，人们利用它看到了物质内部的精细结构。看见所有物质都是由一些肉眼看不见的极小的微粒组成的，于是发现了原子世界。1981 年，由格尔德·宾宁(Gerd Binnig)及海因里希·罗雷尔(Heinrich Rohrer)发明了扫描隧道显微镜，也称为扫描穿隧式显微镜。这种显微镜比电子显微镜更先进。自从扫描隧道显微镜发明后，世界上便诞生了一门以 0.1～100nm 这样的尺度为研究对象的新学科，这就是纳米科技。

随着因特网和电子商务的迅速发展，人类正在被带入到一个新的世界环境之中。也就是说，除了生存的物理空间外，一切生物(动植物，包括人)还有另一个生存空间——信息空间(information space)，只不过人类很晚才真正意识到这个空间的存在和重要性。信息空间是人们进行交流、活动的一个新的场所，它是全球所有通信网络、数据库和信息的融合，形成一个巨大的、包罗万象的、相互关联的和相互交流的“景观”。在信息空间中，人们进行数据的获取和处理，传送电子邮件，传播信息和知识，甚至进行情感交流。在不久的未来，全球网络的融合将改变单个网络的特性，网络将不再只是简单地作为一种人们进行交流的中介，而是创造出一个“全球网络生态”，人们能够在“全球网络生态”环境下从事各种活动。

人类生存空间的演进总是与科学知识的积累和科学技术的进步相联系，在每一个历史年代，人类依靠知识和智慧的积累创造各种新的科学技术，使自身不断进化。20 世纪 50 年代以来计算机技术飞速发展，人类发明了各种新型的通信、存储、传感、处理和计算工具，经历了数字化革命，特别是 20 世纪 90 年代后，与现代通信技术结合而形成的互联网络迅猛发展，为人类创造了一个全新的数字化、虚拟化网络空间——赛博空间。人们已经感受到这个空间对人类社会的巨大作用，认识到我们不仅要在物理空间中生存，还要在这个虚拟的赛博空间中竞争。

“赛博空间”一词是控制论(cybernetics)和空间(space)两个词的组合,这个词的本义是指以计算机技术、现代通信网络技术,甚至包括虚拟现实技术等信息技术的综合运用为基础,以知识和信息为内容的新型空间,这是人类应用知识创造的人工世界,一种用于知识交流的虚拟空间。赛博空间由居住在加拿大的科幻小说作家威廉·吉布森在 1982 年发表于 *omni* 杂志的短篇小说《融化的铬合金(burning chrome)》中首次创造出来,它是指在计算机以及计算机网络里的虚拟现实。如今赛博空间已经不再是计算机领域中的一个抽象概念,随着互联网的普及,生活中到处都可以看到它的影子。[2]

赛博空间中被利用的是数据(知识),因此,从某种意义上说,赛博空间的诞生不仅影响着人与人之间的文化交流,而且影响着人和自然的关系。“赛博空间由交易、关系和思想本身构成,它们像一道永恒的波浪,在我们的交流之网上部署着。我们的世界无处不在,又无处可寻,我们的世界不是肉体存在的世界。”

信息空间与赛博空间关系图见图 1.1。需要指出的是,信息空间是由无形的信息所构成的虚拟空间,它相对于有形的物理空间;物理空间中的很多事物会映射到信息空间中。人类构建的赛博空间只是宇宙中信息空间的一个子集,互联网是构成赛博空间的重要组成部分,但绝不是全部。赛博空间中还包含无线通信网、电力网、专用网、工业控制网和特种业务网,它们并不一定应用 TCP/IP 与互联网相连通,内域网、外域网和物联网中也有相当一部分不与互联网相连。

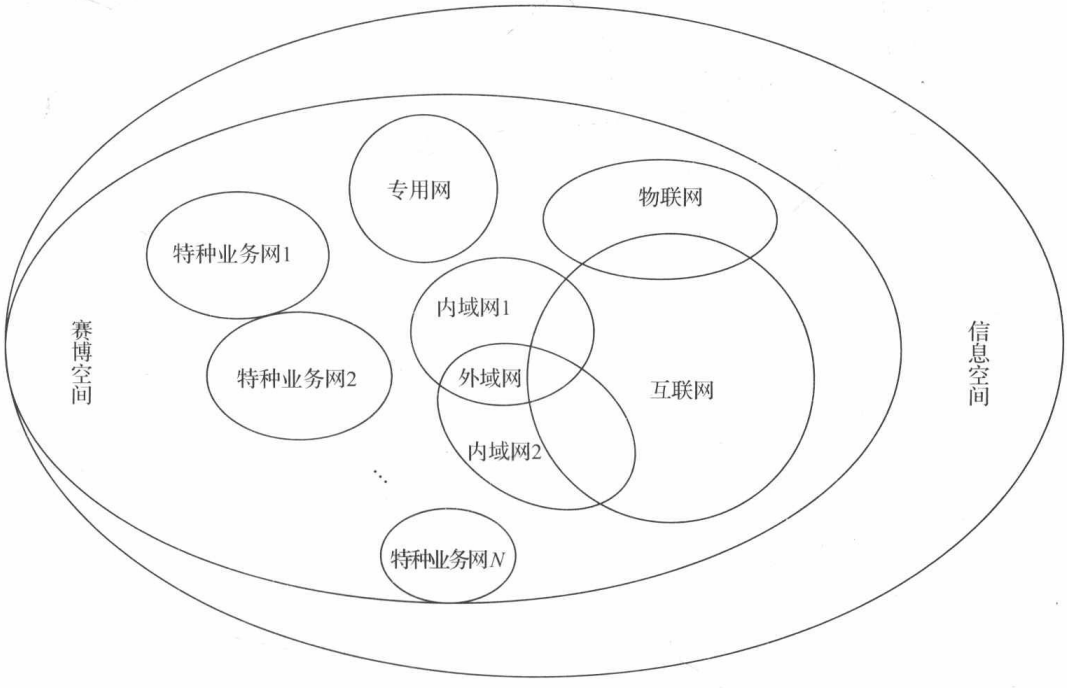

图 1.1　信息空间与赛博空间

尽管人类所创造的赛博空间只是信息空间中很小的一部分，但它对人类社会的发展起着极其重要的作用，是推进人类文明进步的巨大动力。

1.1.2 赛博空间中的数据爆炸

在赛博空间中，人们靠什么来观察和发现各种现象和问题，并进行分析和推断？麻省理工学院斯隆管理学院的经济学教授埃里克·布吕诺尔夫松(Erik Brynjolfsson)称：如果想要理解“大数据”的潜在影响力，首先可以看看显微镜的例子。

显微镜是在四百多年前发明的，可以进行原子层面上的测量，突破人类以往用肉眼观察周围事物的局限，把一个全新的世界展现在人类的视野里。布吕诺尔夫松解释说，数据测量就相当于现代版的显微镜。举个例子，谷歌搜索引擎、基于Facebook帖子和Twitter消息的分析，使得对人们行为和情绪的量化测量成为可能。也就是说，这种虚拟的“显微镜”已经成为人类观测数据、认识赛博空间的有力工具。布吕诺尔夫松进一步指出，在商业、经济及其他领域中，将逐步基于数据和分析作出决策，而不再凭借经验和直觉。“我们将开始变得越来越科学化。”他这样说道。

而作为人类观察和关注的对象，数据正以前所未有的速率和量级迅速增长。若将互联网上的内容抄录到书页大小的纸上，堆起来的高度将为地球到冥王星距离的10倍。全球最大的美国国会图书馆供查阅的书达1.4亿册，但这仅为互联网上数据总量的千万分之一。人类社会每年新增数据量约1～2 EB，其中包括了所有信息存储媒质：书报、杂志、文件、PC、相片、X射线照片、TV、声频、CD、DVD等。人均约为250 MB。每年打印出来的数据约为240 TB，其中约有75亿份办公文件100万种新书、40000种报纸、80000种期刊，而这些的总量不足总数据量的1%。如果在网上搜索“information”一词，用Google只需0.33秒就给出3710000000条搜索信息，用Yahoo只需0.23秒就给出15300000000条，用AOL可给出1220000000条。[3]

根据IDC(国际数据公司)的跟踪分析，全球产生的数据总量2010年首次突破1ZB(10^{21}字节)，2012年达到约2.8ZB，2020年有望达到40ZB。仅就数量而言，40ZB的数据相当于：如果地球上所有海滩上的沙粒有7005000000000000000000颗，40ZB相当于地球上所有海滩上的沙粒数量的57倍；如果把40 ZB的数据全部存入现有的蓝光光盘，这些光盘的重量(不带盒子或包装)相当于424艘尼米兹号航母；2020年，40ZB相当于地球上人均5247GB的数据。

IDC的数字宇宙研究报告预测，到2020年，全球数据总量中有22%将来自中国。持续的互联网、智能手机及社交网络消费者增长；聚焦“物联网”，设备成本的下降；三网融合项目的实施等，促使2014年中国数据总量达909EB(10^{18}字节)，占世界13%的份额，预计2020年将达到8060EB。若用一堆0.29英寸厚、128GB内

存的平板电脑存储中国数据，2013年其总高度为地球至月球距离的9%；到2020年，其总高度将达到地球至月球距离的1.2倍。[4]

从应用角度出发，信息技术指的是一个信息流，从获取、传输、存储、计算到最后的使用。在过去的发展过程中，摩尔定律催生了微电子的快速发展，实际上是通过预测来进一步推动技术的变革。摩尔定律是由英特尔(Intel)创始人之一戈登·摩尔(Gordon Moore)提出的。其内容为：当价格不变时，集成电路上可容纳的晶体管数目约每隔18～24个月便会增加一倍，性能也将提升一倍。换言之，每一美元所能买到的计算机性能将每隔18～24个月翻一倍以上，这一定律揭示了信息技术进步的速度。

在过去二十年里，由于微电子的发展，CPU的计算性能提高了3500倍，但内存和硬盘的价格下降了45000倍和360万倍。当通信的带宽变得越来越廉价且增长速度远远超过摩尔定律的时候，单机就进入了网络计算，离线就进入了在线时代，新的信息技术变革迅速开启。Web 2.0的应用使得过去技术单向交流的方式开始进入了双向交流的时代，我们甚至不用知道服务方在哪里，只需要关注我们需要获取的服务和相应的资源。这也就意味着，通过网络获取信息资源变得越来越快速和低成本，互联网的应用进入了第二次价值挖掘，从而也引发了赛博空间中数据规模的剧增。

在数据发展历程上，“超大”规模一般表示对应GB(1GB=1024MB)级别的数据，“海量”一般表示的是TB(1TB=1024GB)级的数据，而现在的“大数据”则是PB(1PB=1024TB)或EB(1EB=1024PB)，甚至ZB(1ZB=1024EB)级别以上的数据。

可以看出，在大数据时代，数据规模和复杂度的增长已经超出了计算机软硬件能力增长的摩尔定律，这对现有的IT架构以及计算能力来说是极大的挑战，也为人们深度挖掘和充分利用大数据的“大价值”带来了巨大机遇。

数据爆炸是现实，信息爆炸尚可言，但绝不存在知识爆炸，更不可能存在智慧爆炸[4]。作为人类观测数据的“显微镜”，大数据技术仍是这个时代解决各类问题的关键。大数据技术和服务市场代表着全球快速增长的数十亿美元的机会。事实上，2015年年底IDC的预测表明，大数据技术和服务市场将以26.4%的复合年增长率增长，到2018年为415亿美元，大约是总体信息技术市场增长率的六倍。此外，IDC认为到2020年，行业买家可以使分析数据超出其绩效管理的历史值，这关系到非结构化的实时情报和发现探索的两位数增长率。

1.1.3　数据量快速增长的原因

大数据时代已经来临，全球数据量正呈指数级的增长趋势，其主要原因归纳如下。

(1) 各种传感器的剧增及互联网产生的各类数据、高清晰度的图像和视频，导致了数据量的增长。[5]

许多基础学科研究的障碍就在于对象信息获取困难，而一些新机理和高灵敏度的检测传感器的出现往往会成为该学科进展的突破。随着传感器技术的广泛应用及不断发展，传感器获取的数据数量也不断增加，为许多基础研究以及生产实践提供可能。

Google 现在能够处理的网页数量在千亿以上，每天将近 2300 万美元的收入；新浪微博每天有数十亿外部网页和 API 访问需求，夜晚高峰期，新浪微博的服务器群组每秒要接受 100 万个以上的相应请求；中国联通用户上网记录 83 万条/秒，即 1 万亿条/月，对应数据量为 3.6PB/年(10^{15}字节/年)。

同时，近年来互联网服务导致人们的日常生活数据量飙升。据 IDC 公司统计，2011 年全球被创建和被复制的数据总量为 1.8ZB，其中 75%来自于个人(主要是图片、视频和音乐)，远远超过人类有史以来所有印刷材料的数据总量(200PB)。

(2) 自然科学研究产生的数据量剧增。数学、天体物理学、生物学、基因组学和脑科学等都是以数据为中心的学科。这些领域的基础研究产生的数据越来越多。例如，用电子显微镜重建大脑中的突触网络，1 立方毫米大脑的图像数据就超过 1PB。[6]

此外，现在的科学研究比以往任何时候都更依赖将大量数据进行高速可靠的远距离传输及相关实验论证。在过去的 10 年里，连接超过 40 个国家实验室、超级计算中心和科学仪器的能源科学网(ESNET)上的流量，每年以 72%的速度增长。2012 年夏天，疑似上帝粒子——“希格斯玻色子”(Higgs boson)的发现就需要每年 36 个国家的 150 个多个计算中心之间进行约 26PB 的数据交流。

科学研究催生了大数据。如何对其进行收集、管理和分析日渐成为网络信息技术研究的重中之重。以机器学习、数据挖掘和人工智能为基础的大数据技术，将促进数据到知识的转换，形成从知识到行为的跨越。

(3) 企业及商业活动产生的数据量剧增。早在 2007 年，沃尔玛拥有当时世界上最大的数据仓库系统，其存储量高达 4PB 以上。麦肯锡全球研究院估计，2010 年，全球企业在硬盘上存储了超过 7EB 的新数据，消费者在 PC 和笔记本电脑等设备上存储了超过 6EB 的新数据，这些数据总量相当于美国国会图书馆存储量的 5.2 万倍。一项对 531 名独立 Oracle 用户进行的调查发现，90%的企业的数据量在迅速上涨，其中 16%的企业的数据量年增长率达到 50%或更高，不少企业已经感受到失控的数据增长对绩效造成的冲击。

持续增长的数据正在成为一种资源，一种生产要素，渗透至各个领域。而重中之重是，拥有数据的处理能力，即善于聚合并有效利用数据，将会带来层出不穷的创新，从某种意义上说它代表着一种生产力，麦肯锡认为“人们对于海量数据的运

用将预示着新一波生产率增长和消费者盈余浪潮的到来”。

1.2　大数据和大数据时代

数据的通信、网络、传感、存储、搜索、分析和处理等技术及工具的发展促进和催生了大数据时代。“大数据”正在对每个领域都造成影响：在商业、经济及其他领域中的决策行为将日益基于对数据的分析，进而利用这种数据分析来进行指导决策、削减成本和提高销售额；有学者将数学与政治科学联系起来，通过对博客文章、国会演讲和新闻稿件的分析，洞察政治观点的传播方式；在科学、体育、广告和公共卫生等领域，也朝着数据驱动型发现和决策的方向转变。

哈佛大学量化社会科学学院(Institute for Quantitative Social Science)院长加里·金(Gary King)称：“这是一种革命，我们确实正在进行这场革命，庞大的新数据来源所带来的量化转变将在学术界、企业界和政界中迅速蔓延开来，没有哪个领域不会受到影响。”数据已经成为一种新的经济资产类别，就像货币或黄金一样，大数据是一种能帮助人类与贫穷、犯罪和污染等现象展开斗争的智能工具。

1.2.1　大数据定义及属性

1. 定义

自从“大数据”这个术语出现在人们的视野后，与绝大多数枯燥的计算机科学研究不同，大数据在最短的时间内得到了公众和媒体最热烈的关注。标题如“大数据：巨大的收益还是侵犯隐私？(Big data: The greater good or invasion of privacy?)”[7]和“大数据：一个敞开的大门，但是否泄露了太多(Big data is opening doors, but maybe too many)”[8]等各类观点层出不穷。从一开始大数据就与大量的技术和社会问题交织在一起，迄今为止也没有一个确切的定义。从历史上看，最早记录使用这个术语的相关文献来自于许多不同的领域，这也导致了大量，模糊的，甚至是相互矛盾的定义。为了方便进一步的科学研究，给出“大数据”的一个具体定义是一项非常重要的工作。

加利福尼亚大学伯克利分校的研究人员估计，1999年世界已经生产了约115亿GB的信息；2003年的研究发现，信息的数量在3年内翻一番。人类面临的数据量已经越来越大。“大数据”貌似是一个时髦词，但它涉及的很多概念并不是新的，如数据存储和数据分析。因此，这里出现了一个问题，即大数据背景下的相关技术如何显著地区别于传统的数据处理技术？对于这个问题的基本理解，我们首先从对大数据的“量化”入手，即给出一个精确的定义。事实上，很多学者试图定义或描述了什么是“大数据”。

第一个使用“大数据”这个术语的记录出现在 1997 年，由美国航空航天局(NASA)的科学家撰写的文章 *Application-controlled demand paging for out-of-core visualization* 中。他们描述了什么是“大数据”问题，即数据集一般相当大，消耗主内存、本地磁盘、甚至远程磁盘的能力。同时指出，当数据集不再适合主内存，甚至当它们不适合本地磁盘时，最常见的解决方案是获取更多的资源。[9]

在 2001 年的 *3-D data management: controlling data volume, velocity and variety* 一文中，工业分析师 Doug Laney 提出了 3V 特性，作为企业“数据管理挑战”的关键，并提出大数据是具有 3V 特性(即数量、速度和种类)的信息资产，需要信息处理的创新形式，用以提高洞察力、决策和自动处理的能力。[10]

在 2008 年，一些杰出的美国计算机科学家普及了这个术语。他们预测“大数据计算”将改变企业、科学研究人员、医疗从业人员，以及我们国家的国防和情报行动的活动，但是，“大数据计算”这个术语在文献中并没有定义[11]。

“大数据”定义的另一个权威来源是 Viktor Mayer-Schönberger 和 Kenneth Cukier 的专著 *Big Data: A Revolution That Will Transform How We Live, Work and Think*。书中讨论了可以用数据来做什么，以及数据量规模的重要性。他们认为社会正以全新的方式利用信息，产生有用的见解或产生重要价值的商品和服务，以及在提取新的观点或创造新价值形式的目标下，基于大规模数据可以完成的事情，绝大多数不可能基于较小的数据集完成。

全球各大知名的企业组织纷纷加入“大数据”产业化的角逐中。

(1) 引用次数较多的定义是由 Gartner 在 2001 年的报告给出的。Gartner 的报告没有太多提及“大数据”这个名词，而是预估了当前的趋势。迄今为止，该报告被普遍认为是“大数据”的关键定义之一。Gartner 提出的三重定义不仅涵盖了 3V——数量(volume)、速度(velocity)、种类(variety)，同时讨论了数据量的增加、数据产生速度的增长，以及数据格式和表达形式的增加等问题。

虽然作为大数据领域的常见文献，Gartner 提出的定义佐证不足，没有大数据的数值量化。NIST [12]和 Gartner 在 2012 年对这个定义进行了重申[13]，并由 IBM 进行了扩充[14]，包括提出第四个 V——准确性(veracity)。准确性包括关于数据以及数据的分析结果的信任度和不确定性的问题。

(2) Oracle 避免采用任何一个 V 进行定义。相反，Oracle 声称，大数据是从业务决策驱动的传统关系数据库中推导出的“价值”的派生词，另外，他们还补充了新的来源——非结构化数据[15]。这种新的来源包括博客、社会媒体、传感器网络、图像数据和其他形式的数据，这些非结构化数据的大小、结构、格式等在不停地变化。

他们认为，大数据包含额外的数据源，以增加现有的操作。值得注意的是，Oracle 定义的重点是基础设施。与其他定义不同，Oracle 强调一整套技术，如 Na-

sal、Hadoop、HDFS、R和关系数据库。在提出大数据定义的同时，他们也给出了大数据的解决方案。虽然这个定义比较容易应用，但它同样缺乏量化。Oracle的定义仍然没有明确给出什么时候可以应用大数据技术，而仅仅告诉了我们什么是大数据——“当你看到它，你就知道它”。

(3) 麦肯锡(McKinsey)在2011年关于大数据的研究报告中提到，数据集的大小超出了典型的数据库软件获取、存储、管理和分析数据的能力。麦肯锡的研究人员承认这个定义是相对主观的，并给出了另外一个定义，即数据集应该具有什么规模，才能被认为是大数据。不一定大于某一数量级(TB)的数据就是大数据。可以想象，随着科技的进步及时间的推移，被称为“大数据”的数据集的大小也将随之增加。同样地，随着部门的不同，大数据的定义可能会有所不同，这取决于在某个特定的行业，通常应用什么样的软件工具，处理多大规模的数据集。今天，在许多行业里，大数据的范围通常从几十TB到几百PB[16]。

因此，所有关于“大数据”的定量研究结果，包括加利福尼亚大学伯克利分校的数据更新(预估每年企业和用户存储了多少新数据)，都只与数据量的数值相关。例如，没人试图评估企业存储的多少数据(或“数据集”)是所谓的“大数据”。

(4) Intel是为数不多的在文献中提供具体数字的公司。Intel将大数据链接到“每周产生的数据中位数为300 TB”的公司[17]。不同于上述组织提供的定义，Intel通过量化其业务伙伴描述了大数据。Intel指出，被调查的公司广泛使用非结构化数据，并对产生速度大于每星期500TB的数据进行分析。Intel认为，数据分析中最常见的数据类型是存储在关系型数据库中的业务交易数据，其次是文件、电子邮件、传感器数据、博客和社交媒体。

(5) Microsoft提供了一个非常简洁的定义：大数据是这样一个术语，它被越来越多地用来描述运用重要的计算能力(如最新的机器学习和人工智能)来处理超大规模和高度复杂度的信息集[18]。这个定义表达了一种含义，即大数据需要显著的计算能力。这在以前的定义中虽然有所提起，但从未直接说。此外，这个定义介绍了两类技术，即机器学习和人工智能，这是以前的定义所忽视的。因此，这一概念引入了一组相关技术，组成该定义的关键部分。

(6) IBM的定义：“每天，我们创造2.5EB(10^{18}B)的数据资料，数据量如此之多，而世界上90%的数据是在过去的两年里创建的。数据的来源无处不在：传感器收集的气候信息、社会媒体网站数据、数字图像和视频、交易记录，以及手机全球定位系统(GPS)的信息等，这些数据就是大数据。”[14]

(7) Google Trends提供了有关大数据的以下各类术语[19]，从最常见的开始列举：数据分析、Hadoop、NoSQL、Google、IBM、Oracle。通过这些术语，我们可以看出：首先，大数据本质上与数据分析有关，旨在从数据中观察和发现；其次，大数据与一些技术相关，如NoSQL和Apache Hadoop；最后，有许多组织，特别是工业