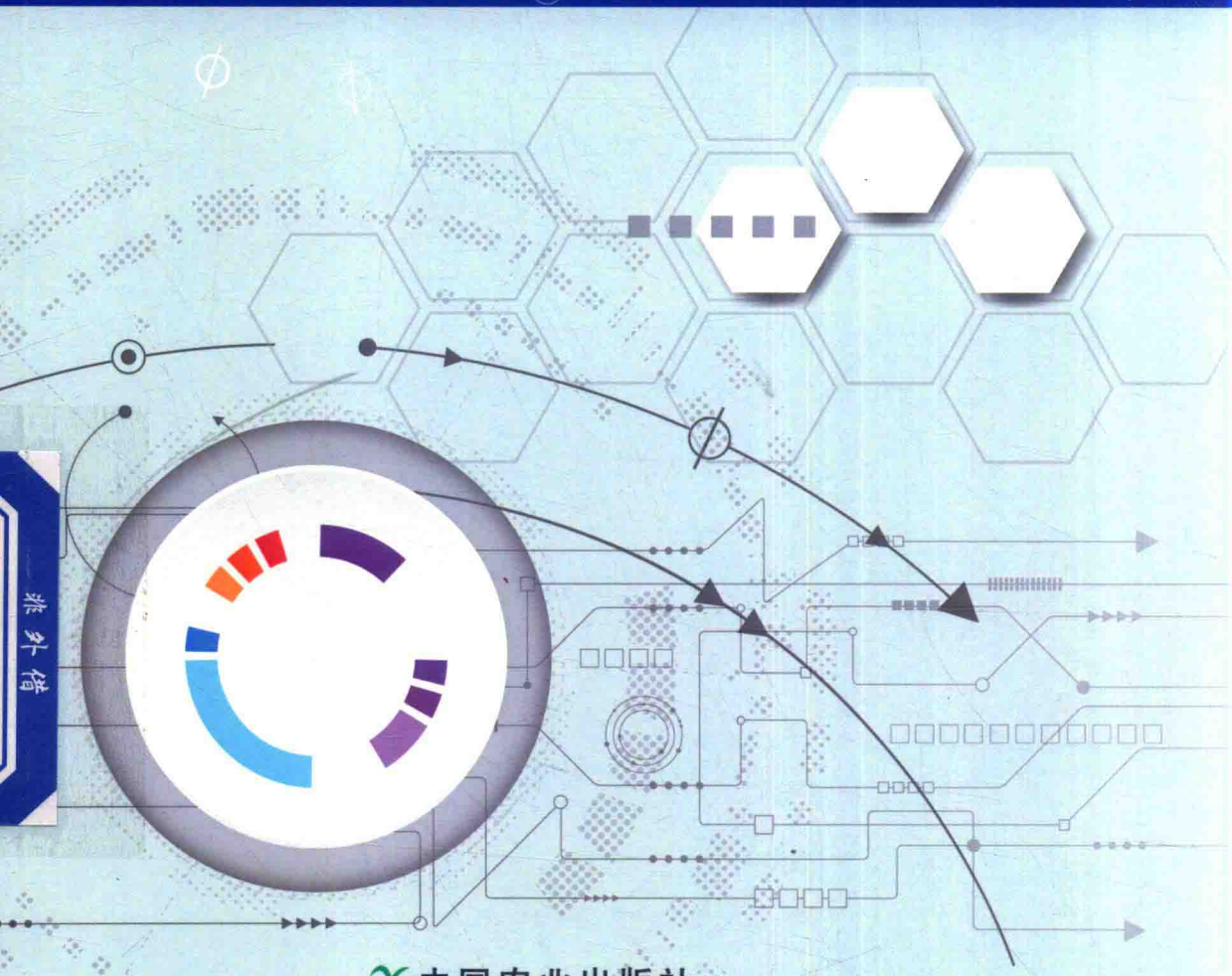
探索性空间数据分析及其应用

降 惠 著



非外借

中国农业出版社

探索性空间数据分析及其应用

降 惠 著

中国农业出版社

图书在版编目（CIP）数据

探索性空间数据分析及其应用 / 降惠著. —北京：中国农业出版社，2017.3

ISBN 978-7-109-21347-0

Ⅰ.①探… Ⅱ.①降… Ⅲ.①计算机应用 Ⅳ.①TP39

中国版本图书馆 CIP 数据核字（2017）第 038878 号

中国农业出版社出版

（北京市朝阳区麦子店街 18 号楼）

（邮政编码 100125）

策划编辑 魏明龙

文字编辑 魏明龙

北京印刷一厂印刷 新华书店北京发行所发行

2017 年 3 月第 1 版 2017 年 3 月北京第 1 次印刷

开本：720mm×960mm 1/16 印张：6.75

字数：125 千字

定价：14.00 元

（凡本版图书出现印刷、装订错误，请向出版社发行部调换）

前　言

FOREWORD

随着“大数据”时代的到来，“大数据”这个既陌生又熟悉的概念，迅速在世界各个国家战略层面得到认可和推广，成为当前世界各国竞相角逐、抢占先机的重要“砝码”。大数据涉及人类生活的方方面面。它广泛影响和促进着人类的政治、经济、社会、文化、生态、医疗、社交、教育、能源、交通等领域的发展和进步。然而，大数据对我们来讲，就像一把“双刃剑”，我们只有通过科学的采集、存储、整理、分析、处理等操作，才有可能挖掘出隐藏在大数据背后的含有巨大潜在价值的知识和规律，促进政府职能改革和公共服务更新，有力推动信息技术革命稳步向前发展。

为了让大数据爱好者、高校学生等更好地学习掌握大数据的理论知识，作者在查阅国内外大量资料的同时，结合10几年高校一线计算机教学和科研经验，尝试编写本书。本书按照数据分析的逻辑顺序和思维习惯，共分为四章。第一章大数据概述，本章具体介绍了大数据的起源、概念、产生、发展、运用，可以使读者对似曾“高冷”的“大数据”实现由浅入深、由表及里的初识和掌握。然而，在现实世界里数据又都具有空间特性。脱离了空间性质的数据分析是不科学和不准确的。本书第二章为空间数据与空间数据分析，从空间数据的概念、结构入手，深入介绍空间数据挖掘、分析的相关知识，让读者能在大数据基础上，进一步了解和掌握空间数据的理论知识。第三章探索性空间数据分析，本章从探索性空间数据分析的概念定义、研究现状、特点差异、数据模型、相关软件（GeoDa）等渐进式的介绍和讲解，能让读者系统全面地学习和掌握探索

性空间数据分析技术的总体知识框架，为更深入研究打下较好的理论基础。第四章医学数据空间探索性分析，本章重点结合医学“大数据”知识，介绍如何将探索性空间数据分析知识运用到医学领域，详细介绍了智慧医疗、健康城市等新兴概念，具体介绍了现有的医学数据源、医学数据空间探索性分析等理论，让读者能对空间数据分析与医学的结合运用有初步的了解和掌握，力求抛砖引玉，启迪思维。全书章节布局合理，各有重点，理论知识通俗易懂，循序渐进，各章节既相互独立，又彼此联系，理论与实践并重，严格遵循空间数据分析、医学等学科知识融合的规律和特点，能让读者相对轻松、高效地掌握本书知识，为进一步工作、学习提供帮助。

在本书的编写过程中，作者得到很多机关和部门，以及领导、同事、学生的帮助，感谢长治医学院各位领导以及计算机教学部的全体教师，给予我出国深造的机会，让我学到了更新的空间数据挖掘分析理论知识，感谢山西农业大学资源环境学院毕如田教授，在探索性空间数据挖掘方面给予的帮助和指导，感谢美国 Auburn University 计算机与软件工程学院 Kai H. Chang 教授在文献检索方面给予我的指导，感谢美国 West Virginia University Dr. Song 为我在医学数据源等方面提供的帮助。最后，我想引用鲁迅先生的一句话，“感谢命运，感谢人民，感谢一切我要感谢的人”。

鉴于作者的理论水平、实践经验、编著时间等非常有限，本书的内容难免有不足和纰漏，恳请广大读者批评指正，不胜感激。

降　惠

2016 年 12 月

目　录

CONTENTS

第一章　大数据概述

第一节　大数据的起源

自 1946 年 2 月 14 日，世界上第一台计算机在美国宾夕法尼亚大学诞生以来，经过 70 多年的发展，互联网信息技术发展日新月异，已经影响到人们生活的方方面面。目前，我们正处于一个“数据大爆炸”的时代，“数据宇宙”正以无法估计的速度膨胀。根据互联网数据中心（IDC）数据显示，2006 年全球产生的新数据总量为 180EB（1EB＝1024PB，1PB＝1024TB，1TB＝1024GB），个人用户数据使用量也进入了 TB 时代。到 2010 年，全球新增数据量高达 1.14ZB（1ZB=1024EB）以上，2011 年达到 1.8ZB，2012 年增长到 2.8ZB，2015 年已达到 8.61ZB [83]，依此估计，全球数据量将会以每两年翻一番的速度增长，到 2020 年预计会达到 2011 年数据总量的 50 倍 [82]。数据总量正以指数级速度持续猛增，它遍及世界每一个角落，甚至能到达月球、火星、更辽阔的宇宙。作为从事大数据分析的研究人员，我们必须熟悉了解和掌握“大数据”的产生和发展，并学会科学运用、分析大数据的规律和价值，从浩瀚的“大数据”海洋中，发掘出对我们人类社会有意义的价值。

一、溯　　源

大数据并非是近几年的新生事物，资料显示，早在 1981 年，著名的未来学者 Alvin Toffler 就在《第三次浪潮》中将神秘的“大数据”称作人类发展史上的“第三次浪潮”[72]。但由于当时信息化技术的滞后，人们对数据概念的模糊，大数据并没有引起人们足够的重视。随着近几年信息时代的到来，互联网、物联网、云计算等新兴技术的广泛使用，以及信息技术在人们生活、社交、购物、娱乐等方面的运用，人们才逐渐发现“大数据”所具有的潜在价值，大数据逐步受到学术界的重视和关注。2008 年 Lynch 在《Big data：How

do your data grow》一文中，从数据时代的产生、发展，以及互联网、广域网、局域网等各方面对大数据进行了介绍，客观展现了大数据对世界发展带来的冲击和影响［37］。2011 年 2 月《科学》杂志专门出版一期关于“数据处理”的专刊，针对人们如何科学采集和处理海量的“大数据”进行了探讨，并预估了即将到来的“大数据时代”，“大数据”对人类社会带来的革命性影响和变化。同一时期，商业界也发现了“大数据”的价值，并快速地将“大数据”分析技术运用到产品研发、市场分析、竞争预判等环节，有效降低了商业运行风险，提高了竞争力。如 2011 年，麦肯锡公司率先提出“大数据时代”这一概念，指出“数据”已经渗透到当今人类社会各行各业的每一个角落，并成为推动人类生产力发展的重要元素之一。2011 年 10 月，在 IOD（信息随需应变）大会上，IBM 高管 Steve Mills 指出了大数据时代用户将面临的四大挑战，即数据的快速生成、数据类型的多样化、数据量级的扩张、数据的低价值密度。2012 年 3 月，美国政府对“大数据研究和发展计划”投资 2 亿美元，充分表明对大数据的重视和关注。2013 年，“大数据”一词成为社会热词，已影响到社会生活的方方面面。2015 年 9 月，经李克强总理签批，我国国务院印发《促进大数据发展行动纲要》，系统部署了大数据发展工作。在短短几年间，政府机关、科研机构、教育、医疗、环保等各行各业的专家学者，都已感受到“大数据时代”的便利与价值，并对“大数据时代”即将带来的巨大变革深信不疑。

二、数据、信息、知识与智慧

数据表面看只是一些简单的符号，如数字、图像、声音、文字、表格等，它可以对客观事物进行直观描述，是很多未经加工原始信息的集合，是人们进行科学研究、探索发现、分析归纳的直接结果。数据通常并不能表现出明显的价值和规律，对我们准确认识客观世界的属性和特征并不能提供直观的、有价值的帮助。只有对数据进行科学有效的加工分析，将数据中有用的价值和规律发掘出来，才能成为指导我们生产、生活、决策的有用资源，充分体现数据的潜在价值。一直以来，人们对数据、信息、知识、智慧常常混淆不清，概念模糊。其实，数据、信息、知识、智慧各自具有不同的含义。

数据是专门对客观事物、各种现象的描述，它可以对客观世界进行准确描述和科学鉴别，并且对客观世界具有的性质、属性、状态及事物彼此间存在的相互关联性进行描述的物理或数学符号集合。它们具有可识别性、抽象性、集合性、概括性等特征。它们不仅仅是单个的数字、图像等符号信息，还兼具某些潜在的意义和价值，通过这些数字、字母、文字、图形符号的组合，可以直观展现出客观世界的数量、属性、状态、时空关系的抽象表示。例如，“0，1，

2，3，4，…”“东、南、西、北、中”“金、木、水、火、土”“医院患者的病例档案记录”“人民群众的个人健康体检档案”等都是数据。数据经过收集整理、探索分析、计算加工、归纳总结后就会成为有用的信息。针对计算机领域的数据而言，就是指那些能通过计算机输入设备输入到计算机中，经过计算机的精确处理，使得原本独立的数字、图形、声音体现出一定的属性和联系过程的符号或介质的集合。如今，能通过计算机进行输入、处理、分析的数据对象范围正在不断扩大，数据的类型、范围、属性也随之变得日益宽泛和复杂。

信息是某种客观事物、现象、属性及特点的集合。正是由于各种各样的集合，才使信息较数据具备更多更广泛的意义和价值。信息与数据既有区别，又有联系，数据可以通过数字、符号、视频等表现信息，并作为信息的载体。而信息则可以反映数据的内涵，信息依附于数据，解释表达数据的具体含义和价值。信息与数据相辅相成，密不可分，数据准确表达信息，信息依靠数据单元来组成。数据通常是物理或数学符号，具有独立性和单一性。信息是经过对一定规模数据进行分析处理后得出的具有一定规律和价值的数据集合，它可以指导人们做出正确决策，具备观念性、逻辑性。数据是信息的表现形式，信息是数据意义的体现结果，数据是信息的表现载体，信息是数据的内涵总结，是形式与本质的关系，原始数据没有过多的价值和意义，只有当数据或数据集合对客观世界产生影响和关联时才能成为广义上的信息。例如，一组数字“3.1415926……”，表面看就是无数个阿拉伯数字的简单集合，并不能看出其潜在的意义和价值，但是如果我们同“圆周率 π”联系起来，我们就会发现它背后隐藏的规律和价值；再如“长治医学院”，表面看这是五个汉字简单的组合，每个汉字各具词意，分开来看每个汉字都不能表达出这个组合的意义，但组合在一起，就构成了晋东南地区最大的医学类高等院校——长治医学院；被称为“信息论之父”的香农于 1948 年 10 月发表于《贝尔系统技术学报》上的论文《A Mathematical Theory of Communication》（通信的数学理论）给出了信息熵的定义，他认为：消息传递中的信息反映出事物确定性的增加，可以用概率方法进行定量描述，通过信息熵来度量数据包包含的信息量，将信息的传递过程作为一种统计学现象来分析，提出了估算通信信道容量的方法。把信息论看作是一门用数理统计方法来研究信息的度量、传递和变换规律的科学。它主要是研究解决信息在控制和通讯过程中体现出的共同规律以及研究解决信息的最佳传递、变换、度量、获限、储存等问题。总之，数据经过人们科学分析加工之后，将数据及数据组合潜在的价值和意义挖掘出来，就能指导人们实践，就成了真正意义上的信息。

知识是人类起源以来在实践中认识掌握客观世界（包括人类自身）的成果，它代表着人类探索发现物质世界和精神世界成果的总和，它是指运用某种结构化方式来表示现实世界客观事物的概念、特征、事件和过程。知识深刻地描述客观世界，它既能对客观世界进行抽象和提炼，也能对具体事物的起因条件和相互关联进行归纳总结，还能对事物的过去、现在、将来进行归纳、总结和预判，它涵盖客观事实、信息描述、实践技能，它始终符合人类文明的发展方向。知识兼具理论性与实践性，它是信息经过科学加工和准确分析后的成果。

知识按类型分类，可分为自然知识和社会知识、感性知识和理性知识、独有知识和共有知识、简单知识和复杂知识、显性知识和隐性知识、具体知识和抽象知识等。20 世纪 50 年代，世界著名的科学家 Michael Polanyi 发现了知识的隐性维度，引起了当时整个科学界的轰动，他认为："人类的知识有两种。人们广义上被描述为知识的，即通过书面文字、图表或数学公式加以表述的，称为显性知识。那些实际存在的未被表述的知识，像我们在生活实践中所拥有的知识，称为隐性知识。"依照波兰尼的理解，显性知识是指那些能够被人类以一定符号系统（最典型的是语言，也包括数学公式、各类图表、盲文、手势语、旗语等诸种符号形式）加以完整表述的知识。隐性知识是指那些我们知道但难以言述的知识。哲学范畴内对"知识"的研究被称为认识论，可以看作是构成社会文明和人类智慧最重要的因素。知识的获取过程要经过许多复杂的环节，感觉、沟通、推理、思考、总结等，知识具有普遍的真理性、公允性、一致性。至今，学术界对知识的定义仍然存在很多分歧。长期以来"什么是知识"吸引着很多著名思想家的兴趣，长久以来经过众多思想家、哲学家、数学家、天文学家激烈的争论，至今仍然没有形成一个普遍认可的定义。伟大思想家柏拉图曾提到一个经典的定义：一条陈述要称得上"知识"必须具备三个条件：一是它一定是被验证过的；二是它一定是正确的；三是它一定是被人们所相信的。这同时也是科学与非科学的本质区别。

智慧是指人类或生物具有的基于神经器官（物质基础）的某种高级的综合能力。包括知识、文化、记忆、情感、感觉、理解、现象、逻辑、识别、计算、分析、宽容、决策等多种能力。智慧可以让人类或生物深刻地理解人、事、情、景、物、自然界、世界、过去、现在和将来，具备思考、分析、探索规律真理的能力。智慧与智力、智能的意思并不相同，智慧包括智力而高于智力，智慧是智力器官的一种终极功能，与"形而上谓之道"有异曲同工之妙，智力、智能则可称作是"形而下谓之器"。智慧让我们正确地决策，科学地分析，精准地预测，我们常常说有智慧的人就是"智者"。广义的智慧

包括一系列知识体系、智力体系、技能和方法体系、非智力体系、思想与观念体系、评价和审美体系等，我们通常将人类或生物的遗传智慧和养成智慧、生理智慧和心理智慧、直观智慧和思维智慧、情感智慧和理性智慧、道德智慧和美感智慧、已有智慧和潜在智慧均作为构成智慧的主要要素。智慧按应用领域可以分为智慧医疗、智慧城市、智慧旅游、智慧金融等，各种各样的智慧影响和改变着人类生活的方方面面，为我们带来了幸福，提供了便利。

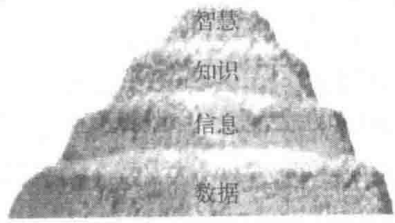


图 1-1　数据、信息、知识、智慧间的关系

数据、信息、知识、智慧四者之间的关系如图 1-1 所示，数据处于最低层，是大量原始记录的集合；信息高于数据，是对数据经过初级加工后的浓缩；知识高于信息，是对大量有用信息的理解、解释、提炼；智慧高于知识，是在知识的基础上，通过经验、阅历、见识的不断积累沉淀，形成对世界的深刻认知，体现为一种卓越的决策力和执行力。

第二节　大数据的产生

随着信息化时代的飞速发展，近几年“大数据”这一概念逐渐被人们所熟知，人们生产、生活、工作、学习无不被迅速猛增的“大数据”包围，很多人甚至感到措手不及，惊慌失措。“大数据”很快得到了人们的高度重视，很多专家学者开始对大数据产生的原因、发展趋势、优势和劣势进行分析研究，试图从“海量”的大数据中发现更多潜在的有用价值和规律。可想而知，大数据时代产生和发展符合社会生产力的发展，与目前科技、社会、生活、经济等因素休戚相关，具有社会发展的必然性。具体来讲，大数据产生和发展具备以下几方面原因和基础：

第一，计算机信息技术的快速发展，使得对各种“海量”数据的采集、存储、传递、分析、共享等成为可能。我们身边的世界从来不缺少数据，只是我们一直以来缺乏获取数据的有效途径。计算机信息化技术的进步和发展，使我们能够得到原来得不到的宝贵数据，如人类基因组数据。每个人都有自己独特的、大量的基因数据，如果我们仅仅记录碱基序列，不进行任何注释，那么需要存储的基因组数据容量就足足超过 3GB。但是由于医疗基因技术、大数据分析技术等技术的不足，我们的科研工作者迟迟不能完整地得到这些人体的宝贵数据。但随着医学基因测序技术的发展和突破，使我们获得人体全部基因数

据成为可能。自 1977 年开始，人类基因测序技术经过三代技术阶段的发展，在测量精度和效率方面实现了较大的突破。就芯片技术来说，2003 年一块 SNP 芯片只能整合 1 万个生物标记数据，至 2007 年，一块 SNP6.0 芯片就可以实现对 90 万（906600）个生物标记数据的存储，科学技术的发展为我们获取“海量”的大数据提供了有效的技术保障。

第二，随着传感器设备的改进和发展，人们对大数据采集的技术也得到突破。过去，人们对数据的采集大多依靠人工的输入和录入，效率低，成本高。如今，人们使用的手机、Pad、PC、汽车、智能水表、电表、煤气表、手表、眼镜等信息化设备中的传感器设备均能够实现自动智能地感知、采集、整理、存储、传递超大容量的数据。这些数据既涵盖了人们生产、生活、学习、工作中的思维、行为、感受、目的、数量、速度等传统数据，又涵盖有关地理位置、温度、湿度、震动、信号强弱等新型数据，很多传感器还可以根据人们的需求设立功能模块，实现对收集的实时数据内容进行智能调整、转化、分析等。据调查显示，目前全世界传感器的发明和运用正以 30%的速度迅速增长。我们可以相信，随着大数据智能化信息时代的到来，人们对智能设备会越来越依赖，传感器技术一定会被更加广泛地运用到医学、商业等各个领域，成为人们采集原始“大数据”的重要工具和渠道。

第三，数字化和网络化信息为大数据产生和发展提供了保证。在计算机还未普及的 20 世纪末，政府、企业都依靠人工在纸张上记录数据，数字化信息网络改变了传统的生产生活方式，为政府、企业的管理和决策提供了全新的平台和工具。如数字化网络管理系统可以为企业量身定做自动化数据信息收集系统；每个人的财务收入和支出都可以通过银行网络信息系统进行记录、存储、显示；每个人的身份信息都可以绑定身份证，存储在网络中；病人生病住院的医疗健康信息都可以通过医疗健康服务系统实现实时采集、存储、查阅、传递、共享等；数字化信息记录方式使得政府、企业、工厂、医院、学校等机构和组织彻底告别了过去堆积如山的纸质文件、治疗、账本等，使他们能够利用相应硬件和软件轻松地整理、采集、存储海量数据，加以分析利用，提高管理和决策效能，他们可以科学智能地管理这些数据，随时查阅、共享、传递更多需要的数据信息。正是大数据改变了世界，改变了人类，使得人们对数据的依赖越来越重，促进了“大数据”分析处理技术的发展和创新，并将掀起一场以“大数据”技术为核心的工业革命。

第四，计算机与互联网、物联网、云计算等技术改变了人们的传统生活方式。人们传统的生产、生活、交流、沟通、社交、娱乐，因互联网数字化技术而发生着深刻地变革，互联网数字化信息技术无时无刻不在记录和存储着人们

生活的点点滴滴。微信、QQ、陌陌、人人网、贴吧、论坛等网络工具和平台能够将人们社交活动中的每一句话，每一个标点，甚至一个表情长久完整地保存在网络中；京东、天猫、淘宝、亚马逊等电子商务平台能把人们每一次网购的记录分类保存在网络中，甚至还能智能化地为人们自动推荐自己感兴趣的商品和服务；马路、广场、商场、医院等公共场所的数字化监控采集设备，可以将人们的影像资料清晰完整地保存并传递到网络中；无数的信息系统、PC 软件、APP 应用，将人们的各种需求、活动、言语、行为、感受等信息转化为数据信息加以存储。毫不夸张地讲，在数字化互联网时代，每个人都是“透明人”，无时无刻无论身处何处都会被强大的数字化信息设备“监控”，每个人都是数据信息的制造者和传递者。大量的文字、图片、声音、震动、视频等数据信息被原始记录下来，通过数字化信息分析处理技术和工具加以分类整合，给我们的世界创造了难以想象的具有巨大价值的数据信息资源，导致了全球“大数据”时代的扑面而来。

第五，先进的数据分析处理硬件技术为大数据技术的发展提供了重要载体。大数据技术的发展和突破都离不开硬件载体，可靠先进的硬件为大数据技术提供了坚实基础。一方面，先进的数据存储硬件设备不断被设计开发，它们兼具体积小、成本低、容量大、性能稳定等特性，为“海量”的大数据提供了存储的“空间”。根据摩尔定律可知：当价格固定时，集成电路中可容纳的晶体管数目大约一年半可增加一倍，性能也能相应提升一倍。这意味着数据信息赖以存储的硬件载体，经过一年半到两年的周期，大小相同、价格相同的硬件，存储量就可能会增加一倍。我们可以巧合地发现摩尔定律中的 18 个月周期与硬件容量翻番的一年半到两年周期非常相近，这就清楚地说明数据信息的增长与硬件技术的发展息息相关，有空间才会有容量，有容量才能有数据。另一方面，随着处理器芯片技术和内存缓存技术的飞速发展，英特尔、速龙、联发科、高通等芯片公司研发能力的不断提升，四核、八核、十核 CPU 相继面市，2G、4G、8G 内存等硬件设备更是如雨后春笋般不断投入使用，使得人们对“海量”大数据的处理速度更快，分析效率更高，大大降低了过去对“海量”大数据分析的技术“门槛”，有力推动了“大数据”分析处理技术的普及和突破。

第三节　大数据的发展

自 2012 年以来，“第三次工业革命”的概念和观点逐步进入人们的视野。目前学术界对第三次工业革命的核心引领技术主要存在三种观点［75］。第一

种观点是美国著名经济学家 Jeremy Rifkin 认为，互联网技术和可再生能源技术两者融合发展产生了一种新的互联网模式——能源互联网，这种模式成为第三次工业革命的基础设施［50］。第二种观点是长尾理论创始人安德森在畅销书籍《创客：新工业革命（第二版）》中预测：互联网和制造业结合在一起即将引发的一场以制造业革命为主导的第三次工业革命［3］。他认为，在未来10年内，人类会越来越多地将网络智能运用于现实生活，未来的经济走向不仅属于虚拟的网络世界，还属于现实世界中高度发达的产业。他大胆预测，创客运动会成为“大数据”时代助推世界发生重大变革的主要动力，将在全球实现全民创造，引导新一轮的工业革命。第三种观点认为：“大数据”的探索挖掘和分析利用将成为第三次工业革命的标志。基于大数据为基础的物联网、云计算等技术将发挥出举足轻重的作用，成为全世界各国竞争角力的关键“砝码”，第三次工业革命将在信息化和数据化结合的基础上席卷而来。

纵观第一次工业革命和第二次工业革命，任何一种单一的技术都不足以掀起一场轰轰烈烈的工业革命。第一次工业革命中，珍妮纺纱机、蒸汽机、蒸汽机车、蒸汽船等一系列新技术和新发明共同推动了一场工业革命的到来。第二次工业革命中，电子计算机、电灯、航天科技、克隆技术等都是鲜明的标志。因此，人们判断新的工业革命能否发生常常从两方面来考虑：一方面是看是一项或几项新发明、新创造，还是一个科技创造集群，只有数量庞大、参与人数众多的科技发明创造集群才能引起工业革命的产生；另一方面是这种科技发明创造的技术群能否为人们生产、生活带来革命性的变革。总之，不论第三次工业革命的标志究竟是能源互联网，还是互联网与制造业结合，还是大数据探索分析技术，都不能单独推动第三次工业革命的到来。我们可以预测，第三次工业革命将是以“大数据”时代为背景，以绿色能源为主的新能源技术、以 3D 打印技术为主的数字化制造业、航天技术、数控技术等融合互通，协调发展，创新创造的共同体现。

第三次工业革命将具有以下几方面显著特征：

（1）**协作化**。在人们通过自身努力使得个人财富不断增加的同时，传统的竞争关系将逐步被协作所取代，销售者和购买者之间的矛盾关系将演化成供应方和使用方的协作互利关系。

（2）**个性化**。数字化打印技术及“互联网＋制造业”就为人们提供方便快捷的个性化定制服务，为了赢得竞争，企业会借助互联网大数据优势，为每一个顾客提供满足其个性化需求的产品，并为个性化产品提供长期有效的增值服务。

（3）**智能化**。智能化不仅包括设计、生产、销售、维修等环节的智能化，

还包括任何产品本身的智能化。智能化的技术和产品将改变传统制造业的生产模式，甚至改变整个国民经济结构，引领人类进入科技化、智能化、数字化时代。

（4）分散化。未来将以人们的个人需求为产品目标，企业会逐渐改变现在的“集中生产，统一销售，产品固定，等待购买”模式，转化为“因人而异，分散生产，各具特色，私人订制”模式，每个人都能成为产品设计者、生产者，根据自己的兴趣和喜好，方便快捷地生产出自己需要的产品。

毋庸置疑，第三次工业革命即将到来，“大数据”也一定会成为助推工业革命的重要核心技术之一，不论世界各国各地区，还是工业、农业、商业、服务业等各行业，谁掌握了“大数据”，谁就占得了第三次工业革命的先机，谁就能从容地应对和把握第三次工业革命带来的挑战和机遇，谁就能在工业革命的浪潮中勇立潮头，赢得胜利。

第四节　大数据的定义、特点、价值与意义

一、大数据的定义

“大数据”是信息化时代的一个热门词汇。对于“大数据”（Big data），研究机构 Gartner 指出“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来适应海量、高增长率和多样化的信息资产［79］；麦肯锡全球研究所指出，“大数据”是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［80］。但“大数据”的概念究竟是什么，目前，理论界众说纷纭，没有统一的标准和定义。本书将从以下视角对大数据的定义进行分类介绍，希望对读者能有所启发和帮助。

（1）从数据源方面定义。随着计算机技术和互联网的快速发展，大数据时代也随之到来，海量的数据扑面而来，迅速膨胀和传播，很多专家学者从大数据产生来源来判断和定义大数据。他们认为，基于互联网模式而产生的各种数据都可以称为大数据，这说明了大数据对互联网、物联网、云计算等技术的过分依赖。这种大数据的定义方式具有一定局限性，它忽视了化学、物理、数学、生物、航天等领域产生的大数据，仅把互联网产生的数据视为大数据。结合实际来看，目前我们认识和接触的大数据，大部分都是从互联网模式中收集的数据，大数据技术的发展也着重于互联网背景下的“海量”数据。

（2）从特征方面定义。一部分专家学者认为，IBM 公司提出的大数据 4V 特征具有代表性，如果数据具备容量够大、来源渠道丰富、数据产生速度快、

价值密度高四个特征，那就可以称之为大数据。另一种观点是从大数据的结构特征来判断是否属于大数据，如果数据属于数字化时代创造出的大量非结构化和半结构化数据，那么这些数据就符合大数据的结构特征，就可以称之为大数据。

(3) **从容量方面定义。**大数据的最直观特征就是容量“大”。传统数据库的最大工作数据容量一般为10～100TB。因此，我们习惯于将10～100TB作为是否成为大数据的容量标准。但是，从大数据本身的特征和属性来看，我们单单从容量这一方面进行定义难免有些偏颇，但是人们通常听到“大数据”这一概念时，脑海里第一印象就是容量“大”，容量的大小依旧是人们判断大数据的重要标准之一。

(4) **从大数据与传统数据的差别定义。**如今的“大数据”与传统的数据相比，具备两个显著的特点。一是大数据体量更大，形式更加丰富，数字化程度更高，内容更加全面；二是大数据之间存在更高的共享性、关联性、可操作性。大数据不再仅仅局限于某种特定的数据分析处理工作和方法，来自世界各地的不同数据不再是孤立存在的，它们之间具备相互关联性，可以实现方便快捷的传递共享，各种数据分析工具和手段都能对大数据进行处理，没有严格的界限和鸿沟。从这个视角来看，如果数据具备上述两个特点，就可以被称为大数据。

(5) **从分析处理技术可行度定义。**这是目前专家学者定义大数据的主要方法。大数据最直观的特征体现在通过当前存在的常见主流数据分析处理工具（Excel、VF、Access等）很难对庞大的数据进行有效的分析处理，而依赖于更新的数据分析硬件和软件来进行分析处理。麦肯锡公司对大数据作了定义，认为大数据是指那些不能在规定时间内用传统的数据分析处理软件进行有效的采集、存储、传递、分析、处理的数据集合。这种定义方式有三方面优势：一是大数据分析处理技术面临的困难和挑战，即传统计算机数据库分析处理软件用时过长、运算难度过大、硬件承载量过低；二是直观体现了大数据的4V特征；三是表明了大数据分析处理技术的发展方向和趋势。

关于大数据的定义和界定还有很多，在此我们不再一一列举，这些不同的定义和观点究竟哪种更能准确客观地解释“大数据”，有待于实践的检验和考证。《商学院》杂志2013年开展了一期以“如何定义大数据”为主题的市场调查活动，调查内容从与大数据有关的信息的范围、新型数据和分析、事实信息、源自新技术产生的数据、非传统形式的媒体、大量的数据、最新大数据流行语、社交媒体产生的大数据等几个方面，对95个国家和地区的1144名业务