大数据/人工智能系列丛书

大数据/人工智能人才培养规划教材

| 数据采集 | 大数据存储 | 大数据分析 |
| 建立分析模型 | 大数据计算 | 模型优化 |

大数据分析导论

魏苗 陈述 吴禀雅 编著

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/人工智能系列丛书
大数据/人工智能人才培养规划教材

大数据分析导论

魏 苗 陈 述 吴禀雅 编著

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

内 容 简 介

本书主要介绍了大数据的相关背景、大数据基础知识、大数据下的相关编程语言、相关工具框架以及大数据下的其他相关技术等，另外，还对机器学习、神经网络和深度学习等内容进行了介绍，并且对常用的大数据可视化工具进行了阐述。由于大数据、云计算、人工智能有着密不可分的关系，本书既包含了大数据的基础知识，同时也涵盖了大数据下的人工智能以及可视化工具等相关内容，方便读者通过本书了解到大数据以及相关技术的基础知识。

本书可以作为本科院校、高等职业院校各专业学生学习大数据基础知识的教材，也可以满足对大数据分析感兴趣的广大读者的学习需求。

图书在版编目（CIP）数据

大数据分析导论 / 魏苗，陈述，吴禀雅编著. — 北京：电子工业出版社，2019.2
ISBN 978-7-121-36022-0
Ⅰ. ①大… Ⅱ. ①魏… ②陈… ③吴… Ⅲ. ①数据处理－高等学校－教材 Ⅳ. ①TP274
中国版本图书馆CIP数据核字（2019）第023012号

策划编辑：贺志洪
责任编辑：贺志洪
印　　刷：天津千鹤文化传播有限公司
装　　订：天津千鹤文化传播有限公司
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：787×1092　1/16　　印张：11.5　　字数：294.4千字
版　　次：2019年2月第1版
印　　次：2019年2月第1次印刷
定　　价：47.00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）88254609 或 hzh@phei.com.cn。

前言

PREFACE

现代社会中，人们每天都在使用不同的方式来收集数据，例如通过手机浏览短视频、通过各种自媒体来读取感兴趣的内容等。而每个人在收集和使用数据的同时，也都时时刻刻在产生着数据。透过查询搜索引擎的关键词趋势，人们可以清楚地看到，在中国，2011 年后“数据”这个词语的搜索率在不断地上升，这意味着人们越来越关注“数据”这个看不见摸不着而又确确实实是人们生活中必不可少的东西。从宏观的角度来看，一个人的数据似乎并不能产生什么特别大的价值，但是当海量的数据产生时，其中所蕴含的价值就能改变整个行业。大数据的出现，帮助了很多行业进行产业升级。无论是人们平时的生活还是商业行为甚至是政治活动，都在不知不觉中受到大数据的影响。例如当你每天打开电子商务网站，在将相关的推送商品添加进你的购物车中时，你就在享受大数据所带来的便利了。

大数据带来的不仅仅是购物车中物品的增加，同样也能改善人们的健康状况。透过大数据的技术，越来越多的遗传疾病可以在不久的将来被攻克。物联网、人工智能、云计算等更与大数据相辅相成，为信息化社会带来了新的技术革命。在这样的人工智能和数据为主的大时代中，国家更是将大数据作为战略之一进行重点发展。通过“大数据 + 民生”的方式来促进产业升级，强化民生服务，通过深度开发各类便民应用从而提升人民的幸福感。

为了更好地普及大数据以及相关知识，帮助读者了解基本的大数据以及相关技术的概念，本书主要从大数据的基础、云计算的基本概念、人工智能等方面着手，帮助读者构建一个大数据以及相关技术的框架，开拓读者的视野。

本书主要介绍了大数据、云计算、人工智能等基本概念。为了方便读者更好地理解进而对大数据产生兴趣，本书并未添加过多的数学逻辑内容，因此本书也可以作为通用教材被非计算机专业的学生所使用。

本书由来自业内知名企业的专家和来自教学一线的教师共同编写。其中第 1 章由浙江商业职业技术学院孔美云、宁波财经学院韦凝芳和浙江大学宁波理工学院唐云廷共同撰写，第 2 章由陈述魏编写，第 3 章至第 5 章由陈述、魏苗和吴熹雅共同撰写。本书在编写过程中也参考了许多优秀的中英文专著和文献，在此一并表示感谢。

大数据、人工智能以及相关技术作为新兴领域，技术日新月异，相关学术观点以及理解也存在争议，希望有兴趣的读者能继续加深了解。由于时间有限，书中难免存在疏漏和不妥之处，望读者雅正，以便本书后续的完善。更多关于神经网络以及机器学习等的具体算法以及相关实际案例将在第 2版中进行添加。

编著者

2018.12

目录

CONTENT

第1章 大数据导论

1.1 大数据的产生

所谓大数据，狭义上可以定义为：用现有的一般技术难以管理的大量数据的集合对大量数据进行分析，并从中获得有用观点。这种做法在一部分研究机构和大企业中，过去就已经存在了。现在的大数据和过去的相比，主要有三点区别：第一，随着社交媒体和物联网等的发展，在我们身边正产生出大量且多样的数据；第二，随着硬件和软件技术的发展，数据的存储、处理成本大幅下降；第三，随着云计算的兴起，大数据的存储、处理环境已经没有必要自行搭建。

所谓“用现有的一般技术难以管理”，可以是指用目前在企业数据库占据主流地位的关系型数据库无法进行管理的、具有复杂结构的数据，或者也可以说，是指由于数据量的增大，导致对数据的查询（Query）响应时间超出允许范围的庞大数据。

研究机构 Gartner 给出了这样的定义：“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的海量、高增长率和多样化的信息资产。

麦肯锡[①]说："大数据指的是所涉及的数据集规模已经超过了传统数据库软件获取、存储、管理和分析的能力，这是一个被故意设计成主观性的定义，并且是一个关于多大的数据集才能被认为是大数据的可变定义，即并不定义大于一个特定数字的TB才叫大数据。因为随着技术的不断发展，符合大数据标准的数据集容量也会增长；并且定义随不同的行业也有变化，这依赖于在一个特定行业通常使用何种软件和数据集有多大，因此，大数据在今天不同行业中的范围可以从几十TB到几PB。"

随着"大数据"的出现，数据仓库、数据安全、数据分析、数据挖掘等围绕大数据商业价值的应用正逐渐成为行业人士争相追捧的利润焦点，在全球引领了又一轮数据技术革新的浪潮。

大数据是一门专注于大量的、频繁产生于不同信息源的数据进行存储、处理和分析的学科。当传统的数据分析、处理和存储技术手段无法满足当前需求的时候，大数据的实践解决方案就显得非常重要。具体地说，大数据能满足许多不同的需求，例如，将多个没有联系的数据集结合在一起，或是处理大量非结构化的数据，抑或是从时间敏感的行为中获取隐藏的信息等。

虽然大数据看起来像是一门新兴的学科，但其已有多年的发展历史。对大型数据集的管理与分析是一个存在已久的问题——从利用劳动密集方法进行早期人口普查的工作，到计算保险收费背后的精算学科，都涉及这个方面的问题，大数据就由此发展起来。信息社会所带来的好处是显而易见的，每个人口袋里都有一部手机，每个办公桌上都放着一台计算机，每间办公室内都连接到局域网甚至互联网，半个世纪以来，随着计算机被全面和深度地带入社会生活，信息爆炸已经积累到一个开始引发变革的程度。它不仅使世界充斥着比以往更多的信息，而且其增长速度也在加快。信息总量的变化还导致了信息形态的变化——量变引起质变。

---

① 麦肯锡公司，世界级领先的全球管理咨询公司，自1926年成立以来，公司的使命就是帮助领先的金业机构实现显著持久的经营业绩改善，打造能够吸引、培育和激励杰出人才的优秀组织机构。麦肯锡在全球52个国家有94个分公司，在过去10年中，麦肯锡在大中华区完成了800多个项目，涉及公司整体与业务单元战略，企业金融、营销/销售、组织架构、制造/采购/供应链、技术、产品研发等领域。麦肯锡的经验是，关键是找那些企业的领导们，他们能够认识到公司必须不断变革以适应环境变化，并且愿意接受外部的建议，这些建议在帮助他们决定做何种变革和怎样变革方面大有裨益。

最先经历信息爆炸的学科，如天文学和基因学，创造出了“大数据”（Big Data）这个概念，现在“大数据”几乎应用到了所有人类致力于发展的领域中。

1.1.1 天文学——信息爆炸的起源

综合观察社会各个方面的变化趋势，我们就能真正意识到信息爆炸或者说大数据的时代已经到来。以天文学为例，2000年斯隆数字巡天项目[①]（如图1.1所示）启动的时候，位于新墨西哥州的望远镜在短短几周内收集到的数据，就比世界天文学历史上总共收集的数据还要多。

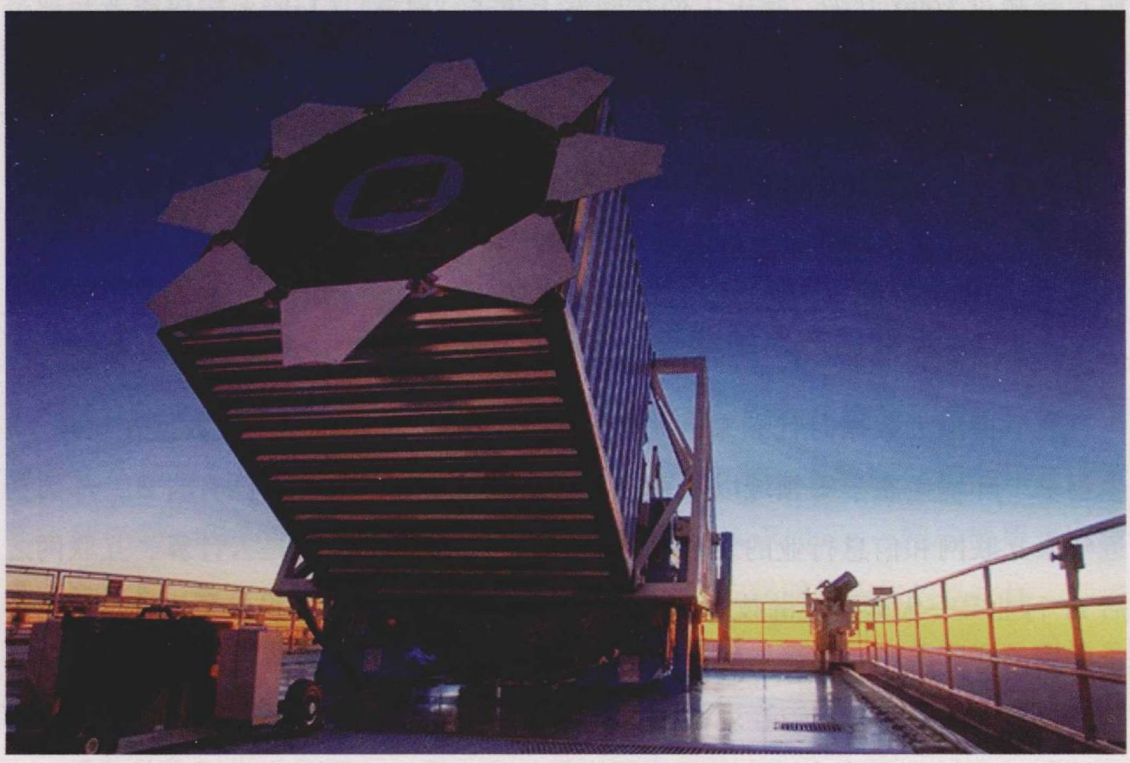

图 1.1 美国斯隆数字巡天望远镜

到了2010年，信息档案已经高达1.4×242GB。当时预计2016年在智利投入使用的大型视场全景巡天望远镜能在5天之内就获得同样多的信息。

① 斯隆数字巡天（Sloan Digital Sky Survey，SDSS）是使用位于新墨西哥州阿帕奇山顶天文台的2.5m口径望远镜进行的红移巡天项目，以阿尔弗雷德·斯隆的名字命名，计划观测25%的天空，获取超过一百万个天体的多色测光资料和光谱数据。2006年，斯隆数字巡天进入了名为SDSS-Ⅱ的新阶段，进一步探索银河系的结构和组成，而斯隆超新星巡天计划搜寻1a型超新星爆发，以测量宇宙学尺度上的距离。

天文学领域发生的变化在社会各个领域都在发生。2003 年，人类第一次破译人体基因密码的时候，辛苦工作了十年才完成了三十亿对碱基对的排序。大约十年之后，世界范围内的基因仪花 15 分钟就可以完成同样的工作。在金融领域，美国股市每天的成交量高达 70 亿股，而其中三分之二的交易都是由建立在数学模型和算法之上的计算机程序自动完成的，这些程序运用海量数据来预测利益和降低风险。

互联网公司更是被数据所淹没。谷歌公司每天要处理超过 24PB（2^{50}B，拍字节）的数据，这意味着其每天的数据处理量是美国国家图书馆所有纸质出版物所含数据量的上千倍。Facebook（脸书）这个创立不过十来年的公司，每天更新的照片量超过 1000 万张，每天人们在网站上单击“喜欢”（Like）按钮或者写评论大约有三十亿次，这就为 Facebook 公司挖掘用户喜好提供了大量的数据线索。与此同时，谷歌子公司 YouTube[①] 每月接待多达 8 亿的访客，平均每一秒钟就会有一段长度在一小时以上的视频被上传。推特（ Twitter）[②] 上的信息量几乎每年翻一番，每天都会发布超过 4 亿条推文。

1.1.2 大数据产生的背景

最早提出大数据时代到来的是全球知名咨询公司麦肯锡。大数据在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业存在已有时日，却因为近年来互联网和信息行业的发展而引起人们的关注。大数据是云计算、互联网之后 IT 行业又一大颠覆性的技术革命。云计算主要为数据资产提供了保管、访问的场所和渠道，而数据才是真正有价值的资产。

企业内部的经营信息、互联网世界中的商品物流信息、互联网世界中的人与人交互信息、位置信息等，其数据量将远远超越现有企业 IT 架构和基础设施的承载能力，实时性要求也将大大超越现有的计算能力。如何盘活这些数据资产，使

① YouTube是世界上最大的视频网站，注册于2005年2月15日，早期其总部位于加利福尼亚州的圣布鲁诺。2006 年 11 月，谷歌公司以 16.5 亿美元收购了 YouTube，并把其当作一间子公司来经营。

② Twitter（推特）是美国一家提供社交网络及微博客服务的网站，是全球互联网上访问量最大的 10 个网站之一。Twitter 是微博客的典型应用，其消息也被称作“推文（Tweet）”，被形容为“互联网的短信服务”。

其为国家治理、企业决策乃至个人生活服务，是大数据的核心议题，也是云计算内在的灵魂和必然的升级方向。

1. 产生背景——大数据时代来临

进入2012年，大数据（Dig Data）一词被越来越多地提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术的发展与创新。它已经上过《纽约时报》《华尔街日报》的专栏封面，进入美国白宫官网的新闻，现身在国内一些互联网主题的讲座沙龙中，甚至被嗅觉灵敏的国金证券、国泰君安、银河证券等写进了投资推荐报告。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然很多企业可能并没有意识到数据爆炸性增长带来的隐患，但是随着时间的推移，人们将越来越多地意识到数据对企业的重要性。正如《纽约时报》2012年2月的一篇专栏中所称，“大数据”时代已经降临，在商业、经济及其他领域，决策将愈发基于数据和分析而做出，而并非基于经验和直觉。哈佛大学社会学教授加里·金说：“这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。”

随着云时代的来临，大数据也吸引了越来越多的关注。分析师团队认为，大数据通常用来形容一个公司创造的大量非结构化和半结构化数据，这些数据在下载到关系型数据库用于分析时会花费过多时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像MapReduce一样的框架来向数十、数百甚至数千台计算机分配工作。“大数据”在互联网行业指的是这样一种现象：互联网公司在日常运营中生成和累积用户网络行为数据。这些数据的规模是如此庞大，以至于不能用G或T来衡量。大数据到底有多大？一组名为“互联网上一天”的数据告诉我们，一天之中，互联网产生的全部内容可以刻满1.68亿张DVD；发出的邮件有2940亿封之多（相当于美国两年的纸质信件数量）；发出的社区帖子达200万个（相当于《时代》杂志770年的文字量）；卖出的手机为37.8万部，高于全球每天出生的婴儿数量37.1万……截至2012年，数据量已经从TB（1024GB=1TB）级别跃升到PB（1024TB=1PB）、EB（1024PB=1EB）乃至ZB（1024EB=1ZB）级别。国际数据公司（IDC）的研究结果表明，2008年全球产生的数据量为0.49ZB，2009年的数据量为0.8ZB，2010年增长为1.2ZB，2011年的数量更是高达1.82ZB，相当于全球每人产生200GB以上的数据。而到2012年，

人类生产的所有印刷材料的数据量是200PB，全人类历史上说过的所有话的数据量大约是5EB。IBM的研究称，整个人类文明所获得的全部数据中，有90%是过去两年内产生的。而到了2020年，全世界所产生的数据规模将达到今天的44倍。每一天，全世界会上传超过5亿张图片，每分钟就有20小时时长的视频被分享。然而，即使是人们每天创造的全部信息——包括语音通话、电子邮件和信息在内的各种通信，以及上传的全部图片、视频与音乐，其信息量也无法企及每一天所创造出的关于人们自身的数字信息量。这样的趋势会持续下去。我们现在还处于所谓“物联网”的初级阶段，而随着技术日趋成熟，我们的设备、交通工具和迅速发展的“可穿戴”科技将能互相连接与沟通，届时产生的数据量更是不可想象。

2. 数据价值

现在，科技的进步已经使创造、信息收集和管理信息的成本降至2005年的六分之一，而从2005年起，用在硬件、软件、人才及服务之上的商业投资也增长了50%，达到了4000亿美元。一分钟内，推特上新发的数据量超过10万；社交网络“脸书”的浏览量超过600万……这些庞大的数字，意味着什么？它意味着，一种全新的致富手段也许就摆在你的面前，它的价值堪比石油和黄金。事实上，当你仍然在把微博等社交平台当作抒情或者发议论的工具时，华尔街的金融高手们却正在挖掘这些互联网的“数据财富”，先人一步用其预判市场走势，而且取得了不俗的收益。让我们一起来看看他们是怎么做的，这些数据都能干什么。数据具体有六大价值：

- 华尔街根据民众情绪抛售股票；
- 对冲基金依据购物网站的顾客评论，分析企业产品销售状况；
- 银行根据求职网站的岗位数量，推断就业率；
- 投资机构搜集并分析上市企业声明，从中寻找破产的蛛丝马迹；
- 美国疾病控制和预防中心依据网民搜索，分析全球范围内流感等病疫的传播状况；
- 时任美国总统奥巴马的竞选团队依据选民的Twitter，实时分析选民对总统竞选人的喜好。“数据是新的石油。”亚马逊前任首席科学家Andreas Weigend说。

大数据是如此的重要，以至于其获取、存储、搜索、共享、分析，乃至可视化地呈现，都成为了当前重要的研究课题。

【延伸阅读】

得数据者得天下

我们的衣食住行都与大数据有关，每天的生活都离不开大数据，每个人都被大数据裹挟着。大数据提高了我们的生活品质，为每个人提供创新平台和机会。

人们通过对大数据进行整合分析和深度挖掘，发现规律，创造价值，进而建立起物理世界到数字世界的无缝链接。大数据时代，线上与线下、虚拟与现实、软件与硬件，跨界融合，将重塑我们的认知和实践模式，开启一场新的产业突进与经济转型。

国家行政学院常务副院长马建堂说，大数据其实就是海量的、非结构化的、电子形态存在的数据，通过数据分析，能产生价值、带来商机的数据。

而《大数据时代》的作者维克多·迈尔·舍恩伯格这样定义大数据："大数据是人们在大规模数据的基础上可以做到的事情而这些事情在小规模数据的基础上无法完成。"

大数据是"21 世纪的石油和金矿"

工业和信息化部部长苗圩在为《大数据领导干部读本》作序时形容大数据为"21 世纪的石油和金矿"，是一个国家提升综合竞争力的又一关键资源。

而马建堂在致辞中也指出，大数据可以大幅提升人类认识和改造世界的能力，大数据正以前所未有的速度颠覆着人类探索世界的方法，焕发出变革经济社会的巨大力量。"得数据者得天下"已成为全球普遍共识。

"从资源的角度看，大数据是'未来的石油'；从国家治理的角度看，大数据可以提升治理效率、重构治理模式，将推动国家治理的变革；从经济增长角度看，大数据是全球经济低迷环境下的产业亮点；从国家安全角度看，大数据

能成为大国之间博弈和较量的利器。”马建堂在《大数据领导干部读本》序言中这样界定大数据的战略意义。

总之，国家竞争焦点因大数据而改变，国家间竞争将从资本、土地，人口、资源转向对大数据的争夺，全球竞争版图将分成数据强国和数据弱国两大新阵营。

苗圩在《大数据领导干部读本》序言中说，数据强国主要表现为国家拥有数据的规模、活跃程度及解释、处置、运用的能力，数字主权将成为继边防、海防、空防之后另一大国博弈的空间，谁掌握了数据的主动权和主导权，谁就能赢得未来。新一轮的大国竞争，并不只是在硝烟弥漫的战场，更是通过大数据增强对整个世界局势的影响力和主导权。

大数据可促进国家治理变革

专家们普遍认为，大数据的渗透力远超人们想象，它正改变甚至颠覆我们所处的时代，将对经济社会发展、企业经营和政府治理等方方面面产生深远影响。

的确，大数据不仅是一场技术革命，还是一场管理革命。它提升人们的认知能力，是进行国家治理变革的基础性力量，在国家治理领域，打造阳光政府、责任政府、智慧政府建设上都离不开大数据，大数据为解决以往的“疾”和“痛点”提供强大支撑；大数据还能将精准医疗、个性化教育、社会监管、舆情监测预警等以往无法实现的环节变得简单、可操作。

中国行政体制改革研究会副会长周文彰认同大数据是一场治理革命。他说：“大数据将通过全息数据呈现，使政府从主观又经验主义的模糊治理方式，迈向‘实事求是数据驱动’的精准治理方式，在大数据条件下，‘人在干、云在算、天在看’，数据驱动的精准治理体系、智慧决策体系、阳光权力平台，都将逐渐成为现实。”

马建堂在为《大数据领导干部读本》作序时也说，对于决策者而言，大数