普通高等院校数据科学与大数据技术专业“十三五”规划教材

数据科学与大数据技术导论

SHUJU KEXUE 与 DASHUJU JISHU DAOLUN

张祖平 ◉ 编著

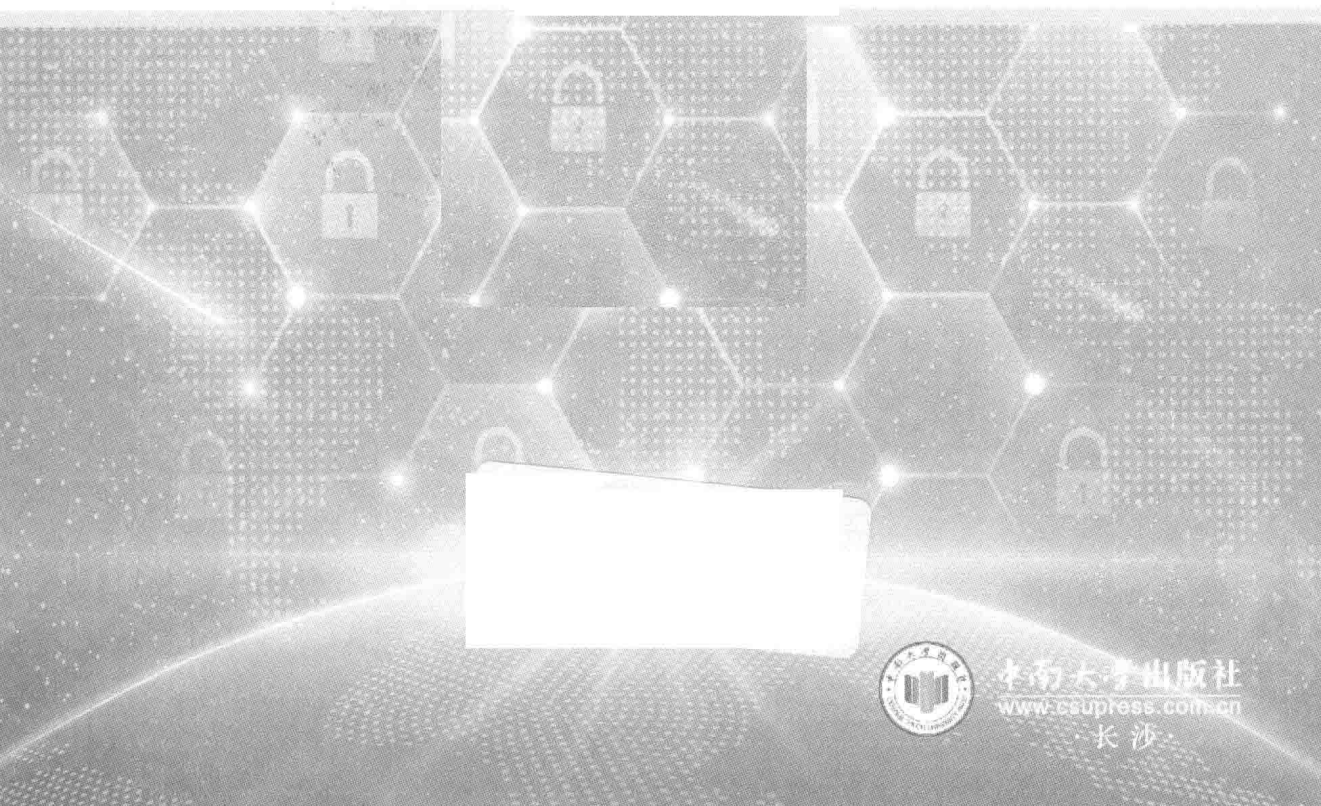
中南大学出版社
www.csupress.com.cn

普通高等院校数据科学与大数据技术专业“十三五”规划教材

数据科学与大数据技术导论

SHUJU KEXUE 与 DASHUJU JISHU DAOLUN

张祖平 ◉ 编著



中南大学出版社
www.csupress.com.cn
长沙

图书在版编目（CIP）数据

数据科学与大数据技术导论／张祖平编著. --长沙：中南大学出版社，2018.12

ISBN 978-7-5487-3374-4

Ⅰ.①数… Ⅱ.①张… 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(2018)第 202688 号

数据科学与大数据技术导论

张祖平　编著

□责任编辑　韩　雪
□责任印制　易建国
□出版发行　中南大学出版社
社址：长沙市麓山南路　　邮编：410083
发行科电话：0731-88876770　　传真：0731-88710482
□印　　装　长沙印通印刷有限公司

□开　　本　787×1092　1/16　□印张 14.5　□字数 366 千字
□版　　次　2018 年 12 月第 1 版　□2018 年 12 月第 1 次印刷
□书　　号　ISBN 978-7-5487-3374-4
□定　　价　38.00 元

普通高等院校数据科学与大数据技术专业“十三五”规划教材

编委会

总序

Preface

随着移动互联网的兴起，全球数据呈爆炸性增长，目前90%以上的数据是近年产生的，数据规模大约每两年翻一番；而随着人工智能下物联网生态圈的形成，数据的采集、存储及分析处理、融合共享等技术需求都能得到响应，各行各业都在体验大数据带来的革命，“大数据时代”真正来临。这是一个产生大数据的时代，更是需要大数据力量的时代。

大数据具有体量巨大、速度极快、类型众多、价值巨大的特点，对数据从产生、分析到利用提出了前所未有的新要求。高等教育只有转变观念，更新方法与手段，寻求变革与突破，才能在大数据与人工智能的信息大潮面前立于不败之地。据预测，中国近年来大数据相关人才缺口达200万人，全世界相关人才缺口更超过1000万人之多。我国教育部门为了响应社会发展需要，率先于2016年开始正式开设“数据科学与大数据技术”本科专业及“大数据技术与应用”专科专业，近几年，全国形成了申报与建设大数据相关专业的热潮。随着专业建设的深入，大家发现一个共同的难题：没有成系列的大数据相关教材。

中南大学作为首批申报大数据专业的学校，2015年在我校计算机科学与技术专业设立大数据方向时，信息科学与工程学院院领导便意识到系列教材缺失的严重问题，因此院领导规划由课程团队在教学的同时积累素材，形成面向大数据专业知识体系与能力体系、老师自己愿意用、同学觉得买得值、关联性强的系列教材。经过两年的准备，针对2017年《教育部办公厅关于推荐新工科研究与实践项目的通知》的精神，中南大学出版社组织对系列教材文稿进行相应的打磨，最终于2018年底出版“高等院校数据科学与大数据技术专业‘十三五’规划教材”。

该套系列教材具有如下特点：

1. 本套教材主要参照“数据科学与大数据技术”本科专业的培养方案，综合考虑专业的来源，如从计算机类专业、数学统计类专业以及经济类专业发展而来；同时适当兼顾了专科类偏向实际应用的特点。

2. 注重理论联系实际，注重能力培养。该系列教材中既有理论教材也有配套的实践教程。力图通过理论或原理教学、案例教学、课堂讨论、课程实验与实训实习等多个环节，训练学生掌握知识、运用知识分析并解决实际问题的能力，以满足学生今后就业或科研的需求；同时兼顾“全国工程教育专业认证”对学生基本能力的培养要求与复杂问题求解能力的

要求。

3. 在规范教材编写体例的同时，注重写作风格的灵活性。本套系列教材中每本书的内容都由教学目的、本章小结、思考题或练习题、实验要求等组成。每本教材都配有 PPT 电子教案及相关的电子资源，如实验要求及 DEMO、配套的实验资源管理与服务平台等。本套系列教材的文本层次分明、逻辑性强、概念清晰、图文并茂、表达准确、可读性强，同时相关配套电子资源与教材的相关性强，形成了新媒体式的立体型系列教材。

4. 响应了教育部“新工科”研究与实践项目的要求。本套教材从专业导论课开始设立相关的实验环节，作为知识主线与技术主线把相关课程串接起来，力争让学生尽早具有培养自己动手能力的意识、综合利用各种技术与平台的能力。同时为了避免新技术发展太快、教材纸质文字内容容易过时的问题，在相关技术及平台的叙述与实践中，融合了网络电子资源容易更新的特点，使新技术保持时效性。

5. 本套丛书配有丰富的多媒体教学资源，将扩展知识、习题解析思路等内容做成二维码放在书中，丰富了教材内容，增强了教学互动，增加了学生的学习积极性与主动性。

本套丛书吸纳了数据科学与大数据技术教育工作者多年的教学与科研成果，凝聚了作者们的辛勤劳动，同时也得到了中南大学等院校领导和专家的大力支持。我相信本套教材的出版，对我国数据科学与大数据技术专业本科、专科教学质量的提高将有很好的促进作用。

桂卫华

2018 年 11 月

前言

Foreword

数据科学与大数据技术导论是一门面向本专业的导论性课程，旨在让学生在大学入学最初阶段对本专业的发展历史、知识结构、培养目标与要求及数据科学与大数据技术相关的基础知识、典型技术、具体应用等有一个直观的认识，区别于新生课的普识性介绍，相关内容偏专业，目标是让学生对本专业的知识及培养要求有一个相对全面而直观的了解，同时也会概述性地介绍有关计算机学科相关内容及典型人物，以激发学生的学习兴趣，促进学生进一步了解设置新专业的历史背景与总体要求。

数据科学与大数据技术导论课程的基本要求包括：

知识：较好掌握数据科学与大数据技术的发展历史及相关典型概念，如与数据相关的基本概念，与数据特征相关的测度及与大数据相关的5V特性等；了解典型的大数据分析环境所包括的技术体系，如Hadoop；了解计算机典型的基础概念如数据、算法；了解专业所需要掌握的知识体系及课程要求；对大数据技术的典型应用有相对直接的了解并能联想到生活中的大数据技术应用场景。

能力：主要培养学生对本专业课程体系的区别与选择能力，对典型的大数据分析环境的技术体系有一定的判别与选择的能力，对应用系统是否要用到大数据平台有一定的判别能力，对整个专业的知识体系有一定提前的预判与认知。

素质：对数据科学与大数据技术专业的相关基础知识有相对全面的了解，逐渐形成采用数据分析的思维解决实际系统需求的意识。能够通过网络搜索平台找到大数据分析平台所需要的典型开源性工具软件，尝试通过网上教学视频进行安装与调试，逐步形成直观认识与一定的学习、操练兴趣。通过课外导学的模式，从网上大量相关的实例中得到启发，从而提升自主学习和终身学习的意识，形成不断学习和适应发展的素质。

为了适应这一要求，笔者组织编写了这本教材。

本教材既包括数据科学与大数据技术专业的发展历程及专业知识要求与技能基本要求等的分析，同时也包括了有关数据科学的基本概念、数据挖掘基本方法及大数据分析主要技术

等，将大数据分析的各流程环节中采用的关键技术及核心技术进行了梳理，对主要的大数据技术生态体系进行了介绍，最后基于实际项目，介绍医疗大数据与智能城市交通大数据，旨在为学生既提供基本的内容，又介绍实际应用的技术与高层次平台或项目申报所需要表达的大数据相关内容，寄希望于同学们能从教材中感受到“抬头看天，放眼未来；低头看地，把握当今”的意境，通过此教材的引导，通过大学四年的专业学习，能形成“格局宏大，布局精细”的个人特质。

本书区别于传统的导论课教材，书中包含20个实验，综合考虑了数据科学与大数据专业需要较好动手能力的特点，同时也顺应了教育部有关“新工科”的要求，培养数据科学与大数据专业学生的动手能力。经过导论课的学习，希望学生能对本专业的知识体系有感性认识，走入社会时，能找到与自己专业相关强的社会岗位，并能尽快适应、快速成长。

本书在编写过程中得到了广泛的支持与帮助。中南大学为数据科学与大数据专业设立了教材出版专项；中南大学出版社与中南大学信息科学与工程学院的相关领导也高度重视，成立了系列教材编写委员会，多次组织专题讨论会，并带领编委会成员多次外出学习、访问；邀请了厦门大学林子雨老师参加编委会教材专题讨论。在此，对支持、帮助及关注本书的各位同仁表示感谢。

本书在正式出版之前，作者将书稿交由数据科学与大数据专业的同学先行试用，部份同学在课程学习的同时认真阅读了书稿，并提出了一些意见或建议，为本书的进一步完善作出了贡献，在此表示感谢。

由于编者水平有限，书中难免有不足之处，恳请读者批评指正。

编　者

2018年8月

目录

Contents

第1章　专业概论

本章主要介绍数据科学与大数据技术专业的产生背景与发展历史、专业的特点与综合要求、专业相关的完整知识体系与技能体系；介绍了与本专业密切相关的专业如计算机科学与技术、统计学等的关联关系。同时，本章还对专业的出路与就业进行了简述。

1.1　专业发展历史

1.1.1　专业产生背景

随着移动互联网的崛起，全球数据正呈爆炸性增长。据统计，目前全球90%以上的数据是最近几年产生的，数据规模大约每两年翻一番。现有数据不仅指人们在互联网上发布的海量信息，还包括各种设备、建筑、系统、人员、业务、场景等产生的各种结构化、半结构化与非结构化数据，这些数据随时测量和传递着有关对象的各种状态及变化。这是一个产生大数据的时代，更是需要大数据力量的时代。

数据统计、分析和应用这一专业历史悠久，但是传统的相关专业已经难以适应大数据时代的新要求。大数据具有体量巨大、速度极快、类型众多、价值巨大的特点，对数据采集、存储、处理、传输和应用提出了前所未有的新要求。高等教育只有转变观念，更新方法和手段，寻求变革与突破，才能在信息大潮面前立于不败之地。

开设数据科学与大数据技术专业正是实现上述变革与突破的重要举措。

1. 行业发展与人才需求

大数据(big data)或称巨量信息，指的是所涉及的信息量规模巨大，以至无法通过目前主流软件工具在合理时间内实现采集、管理、处理，并成为帮助企业经营决策以达到更积极目的的数据。大数据这个术语最早期的引用可追溯到Apache基金会的开源项目Nutch。当时大数据用来描述为更新网络搜索索引需要同时进行批量处理或分析的大量数据集。随着谷歌Map Reduce、Google File System(GFS)以及Hadoop的发布，大数据不再仅用来描述大量的数据，还涵盖了处理数据的速度、数据的阶段和准确性及数据的复杂性等。从某种程度上说，大数据是数据分析的前沿技术，从各种各样类型的数据中快速获得有价值信息的能力就是大数据技术。全球知名咨询公司麦肯锡指出："数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用预示着新一轮生产率增长和消费者盈余浪潮的到来。"

大数据不是一个片断，也不是简单的一项技能，而是综合性的科学与技术，从理念层面

延伸到技术、科学和管理等层面。只有在真实的应用场景中才能让企业对大数据的价值有一个直观的感受，而应用场景的建立需要从企业战略本身出发。目前来看，大数据主要有五个方面的应用场景，分别是：

(1)利用大数据实现庞大知识库：客户服务、保险、汽车、维修、医药等行业需要巨大储备规模的知识库。

(2)利用大数据实现客户交互改进：电信、零售、旅游、金融服务和汽车等行业将“快速抓取客户信息从而了解客户需求”列为首要任务。

(3)利用大数据实现运营分析优化：制造、能源、公共事业、电信、旅行和运输等行业需要时刻关注突发事件，通过监控提升运营效率并预测潜在风险。

(4)利用大数据实现 IT 效率和规模效益：企业需要增强现有数据仓库基础架构，实现大数据传输、低时延和查询的需求，确保有效利用预测分析和商业智能实现性能的扩展。

(5)利用大数据实现智能安全防范：政府、保险等行业亟待利用大数据技术补充和加强传统的安全解决方案。

尽管大数据行业刚刚开始进入发展期，但市场竞争已相当激烈。企业要想在竞争中保持领先优势，仅仅是收集大量的数据显然是不够的。那些已经成功实施了大数据策略的企业，如百度、腾讯、阿里巴巴等，在大数据战略上都具有以下特点：第一，收集一切数据，并进行集中式存储，之后再决定是否需要这些数据；第二，使用数据驱动的产品，确保可以收集到可用的数据；第三，保持不断追求技术创新的动力；第四，对所有收集的重要数据信息进行正确的分析，建立信息中心文化；第五，聘请专家，注重培养大数据专业人才。

IDC 发布的报告显示，2012—2016 年全球大数据技术及服务市场复合年增长率(CAGR)将达到 31.7%，未来 4 年累计增长超过 200%，到 2016 年，大数据行业的收入将达 238～500 亿美元，其增速约为信息通信技术市场整体增速的 7 倍，其中，中国大数据技术和服务市场未来 5 年的复合增长率将达 51.4%。

毫无疑问，大数据的市场前景广阔，对各行各业的贡献也将是巨大的。目前来看，未来大数据技术能否达到预期的效果，关键是在于能否找到适合信息社会需求的应用模式以及是否能够建立起配套的教育培训体系，为大数据行业的发展输送合适的人才，使大数据产业保持创新能力，并具有长期的可持续发展性。

从传统架构到大数据时代应用程序架构的转变往往都会遇到一些问题和挑战。在对计算框架门槛的调查中，非专业人士难于入手这一难题的比例达到了 46.5%，这对企业人才的培训提出了迫切的要求。专业开发者期望从技术培训中获取的知识是什么？据调查，第一是计算框架，如 Map Reduce、Google File System(GFS)以及 Hadoop 等，占 63%；第二是面向大数据处理的数据库系统，如 NoSQL 等，占 37%；第三是云计算解决方案，占 35%；第四是编程语言，占 25%。

综上所述，大数据技术在企业界中有广泛的需求，未来大数据技术的需求者不仅仅是大企业，还包含大量的中小企业，其中的人才缺口是可观和长期的。而目前对大数据技术已经掌握并运用的企业数量不足 3 成，后发企业迫切需要对现有 IT 人员进行大数据技术培训，并招揽具有大数据技术背景的应届毕业生。

2. 专业人才缺口与求知需求

首先，从理论上看，由于社会生活与生产已经被大数据与云计算所笼罩，随之而来的数

据仓库、数据安全、数据分析、数据挖掘、数据可视化等技术，正在为大数据与云计算行业带来大量的商业价值，逐渐成为行业人士争相追捧的利润焦点。因此，与之相关的职业需求也必然呈爆发式增长，如很多企事业单位包括像互联网公司如阿里巴巴、腾讯等，同时也包括银行、大型制造型企业及商务型单位等都设有专门的数据科学家、数据分析员、大数据中心管理员、大数据系统架构师等与大数据密切相关的岗位，而现实情况是大数据职业的相关人才严重匮乏，人才缺口非常大。国际知名咨询公司盖特纳曾分析，2014 年，大数据与云计算专业将为全球带来 440 万个 IT 新岗位和上千万个非 IT 岗位，而实际情况是，相关岗位需求远不止如此。大数据开发工程师缺口大，预计近几年，大数据工作者人才缺口达到 150 万。

其次，从教育界的动向看，国际国内一些高校已经开始举办大数据相关的专业，这也反映了教育界对大数据专业人才需求的共识。国际上，美国北卡罗来纳州立大学、耶鲁大学、哈佛大学等开设了应用统计专业的成熟院校，早就开始关注大数据课程设置。2013 年起，美国纽约大学、英国邓迪大学等知名高校也设立了数据科学硕士学位。国内，香港中文大学、西安交通大学、浙江大学、厦门大学、中南大学、中山大学等高校设立了数据科学与大数据相关研究中心或研究院，开始培养具备大数据思维和创新能力的复合型人才。中国人民大学、北京航空航天大学、厦门大学等举办了专业教学班，推出了包括 Hadoop、HBase 技术等在内的系列课程。尤为引人注目的是，北京大学和清华大学于 2014 年秋开始培养第一批大数据硕士研究生。清华大学招收的第一批大数据硕士研究生分为五个方向，分别是数据科学与工程、商务分析、大数据与国家治理、社会数据、互联网金融。而北京大学等五院校的第一期大数据分析硕士实验班于 2014 年秋季开班，约有 100 多位教师参与到 50 名研究生的培养中。这一大数据分析硕士培养协同创新平台由中国人民大学、北京大学、中国科学院大学、中央财经大学和首都经济贸易大学五所院校，联合新华社、人民日报、中央电视台、中国移动、中国联通、中国电信等业界大数据应用单位共同成立。目前，该协同创新平台开发出 6 门必修课程，必修课将采用联合授课的方式在同一地点授课，计入各校学分体系。

再次，从一些企业人才高级管理人员、高校专业负责人发表的言谈中，能够明显感受到对大数据人才的期待。例如，戴尔全球副总裁、中国区大型企业及公共事业部总经理容永康曾表示，国内现在懂得在 Hadoop 上进行开发的专业技术人员非常少，而一些金融行业的用户虽然很想现在就部署大数据解决方案，但是苦于找不到既懂数据分析技术，又懂金融业务的专业人才。北京航空航天大学互联网营销专业带头人姜旭平教授认为，随着互联网一代的成长，企业的营销主战场越来越转移到互联网上，也可以说谁掌握了互联网，谁就掌握了未来，因此，对互联网营销人才需求将十分迫切。

最后，也是最能反映大数据人才需求趋势的事实，就是目前大学生求职招聘市场上的信息。在 2014 年，全国高校毕业生数量继续增加，727 万名大学毕业生拥入就业市场，再创历史新高，再加上往年没有找到工作的毕业生，就业人数突破 800 万，被称为“史上最难就业季”。但是，就是在这样的严峻形势下，IT 产业作为知识密集、技术密集的产业，就业形势却十分可观。“前程无忧”网站最新发布的无忧指数显示，全国 IT 招聘市场人才需求继续向上攀升，全国 IT 类(计算机、互联网、通信、电子)职能的 4 月份网上发布职位数将近 50 万个，环比涨幅达到 11%，同比涨幅高达 39%，成为招聘需求最热门行业，位居榜首。特别值得注意的是，互联网、电子商务、网络游戏和数据分析专业，涨幅高达 60%。很多公司指名招聘 Hadoop、HBase、Map Reduce 开发工程师。此外，计算机软件网上发布职位数同比涨幅均超

过20%，计算机硬件行业的网上发布职位数同比涨幅也达到了15%。

1.1.2 专业创办与申报

数据科学与大数据技术专业的创办最早可以追溯到2013年以培养大数据专业人才为目标，为计算机、数学等本科高年级专业学生或相关学科的研究生开设并设立相关学位的数据科学或大数据技术课程，如美国的纽约大学、英国的邓迪大学等知名高校设立了数据科学硕士学位。国内著名大学如北京大学、清华大学、浙江大学、中南大学、厦门大学、中山大学、西安交通大学及香港中文大学等相继在2014年前后成立数据科学研究中心或信息安全与大数据研究院，并且成立大数据相关的国家工程实验室、协同创新中心等平台，将培养具备大数据思维和创新能力的复合型人才作为主要任务之一，同时国家开始考虑设立大数据相关的新专业。在新专业正式招生之前，部分院校在计算机专业设立大数据方向，如中南大学在计算机科学与技术专业2015级设立的大数据方向即计算机科学与技术专业（大数据方向）正式掀开了大数据专业建设序幕。

2015年7月申报、2016年2月获批，北京大学、中南大学、对外经济贸易大学三所院校首次成功申请“数据科学与大数据技术”（专业代码：080910T）本科新专业。2017年3月，第二批32所高校获批。2018年3月，教育部最新公布的高校新增专业名单中，有248所学校获批，至此共有283所高校获批建设“数据科学与大数据技术”本科专业。

1.2 专业特点要求

1.2.1 专业培养定位

数据科学与大数据技术专业，瞄准社会各领域对大数据专业人才的需求，面向数据科学基本理论、计算机科学基础知识与核心技术以及大数据技术体系与大数据分析应用等多个层面，培养具有扎实的数学、信息科学、数据科学、计算机科学等知识，熟练掌握大数据的采集、预处理、存储、分析及应用等核心技术，具有大数据思维，能够承担企事业单位、政府机关、社会团体等单位与数据密切相关的有关系统分析、设计及开发应用工作，具有大数据系统相关技能的专业技术人才。该专业培养出来的合格学生，能掌握大数据应用中的各种典型问题的解决办法，能将知识领域与计算机技术和大数据技术进行融合、创新的能力，具有解决实际问题的能力，同时还能够从事大数据相关的研究。

1.2.2 交叉型与复合型

本专业强调培养具有多学科交叉能力的大数据人才。该专业重点培养具有以下三方面素质的人才：一是理论性，主要是对数据科学中模型的理解和运用；二是实践性，主要是处理实际数据的能力；三是应用性，主要是利用大数据的方法解决具体行业应用问题的能力。

强调学生具有学科交叉、知识融合，能解决具体行业应用问题的能力。