普通高等教育“十三五”规划教材

数据科学导论

（第2版）

Introduction to Data Science（2nd Edition）

杨 旭 汤海京 丁刚毅 ◎ 编著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

普通高等教育“十三五”规划教材

数据科学导论

（第2版）

Introduction to Data Science（2nd Edition）

杨 旭 汤海京 丁刚毅 ◎ 编著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

本书系统性地讲述了与数据科学相关的各方面知识，着重培养数据工程师所需要的技能与思维。本书将从与数据科学相关的概念出发，通过丰富、翔实的案例，从各方面展示数据科学的运用方式，并且在其中穿插数据科学研究方式下新的思维模式的讲解，让读者有一个更为直观的认识，也可以从中感受到运用数据科学处理各个领域问题的方法和流程。本书还从工程概论的流程角度来讲述数据科学的工程体系架构，并展望数据科学的未来发展。同时本书特意加强了对于数据预处理的理论和技术的讲解。本书可作为计算机相关专业本科生教材，也可供相关专业技术人员阅读参考。

图书在版编目（CIP）数据

数据科学导论 / 杨旭，汤海京，丁刚毅编著. —2 版. —北京：北京理工大学出版社，2017.1

ISBN 978-7-5682-3115-2

Ⅰ. ①数…　Ⅱ. ①杨…　②汤…　③丁…　Ⅲ. ①数据管理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 222723 号

出版发行 / 北京理工大学出版社有限责任公司
社　　址 / 北京市海淀区中关村南大街 5 号
邮　　编 / 100081
电　　话 / （010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68948351（其他图书服务热线）
网　　址 / http://www.bitpress.com.cn
经　　销 / 全国各地新华书店
印　　刷 / 三河市华骏印务包装有限公司
开　　本 / 787 毫米×1092 毫米　1/16
印　　张 / 15.25
字　　数 / 384 千字
版　　次 / 2017 年 1 月第 2 版　2017 年 1 月第 1 次印刷
定　　价 / 58.00 元

责任编辑 / 王晓莉
文案编辑 / 王晓莉
责任校对 / 周瑞红
责任印制 / 王美丽

PREFACE

前言

我们已经处于一个数据爆炸的时代，在新的时代背景下，需要运用新的科学研究方式去应对新的挑战。数据科学，作为一门正在蓬勃发展的新学科，所关注的正是如何在大数据时代的背景下，运用各种与数据相关的技术和理论，服务社会，让人们可以更好地利用身边的数据，使生活变得更加美好。

本书系统性地讲述了与数据科学相关的各方面知识，着重培养数据工程师所需要的技能与思维。本书从与数据科学相关的概念出发，通过丰富、翔实的案例，从各方面展示数据科学的运用方式，并且在其中穿插对数据科学研究方式下新的思维模式的讲解，使读者有一个更为直观的认识，读者也可以从中感受到运用数据科学处理各个领域问题的方法和流程。本书还从工程概论的流程角度来讲述数据科学的工程体系架构，并展望数据科学的未来发展。

此次改版，结合了之前教学实践中的诸多收获，对本书的组织结构做了优化，对案例做了更为科学的划分和归纳。综观数据处理全过程，数据预处理的任务量最大，而且对数据处理的效果也有深远的影响。因此在本书的编写过程中我们特意加强了对数据预处理的理论和技术的讲解。

本书由北京理工大学软件学院“数据科学与技术”课题组的杨旭老师、汤海京老师，以及北京理工大学软件学院丁刚毅教授编著。其中第三、四部分由杨旭老师负责；第五部分由汤海京老师负责；第一、二部分由丁刚毅教授负责。全书由丁刚毅教授统稿并审校。书中存在的错误及不妥之处，恳请各位读者、同行批评指正。

作　者

目 录

CONTENTS

第一部分 引 论

第二部分 大数据及其产生根源

第三部分 大数据研究的重要性

第四部分 数据科学的研究方式

第五部分　数据与未来

图 目 录

第一部分　引论