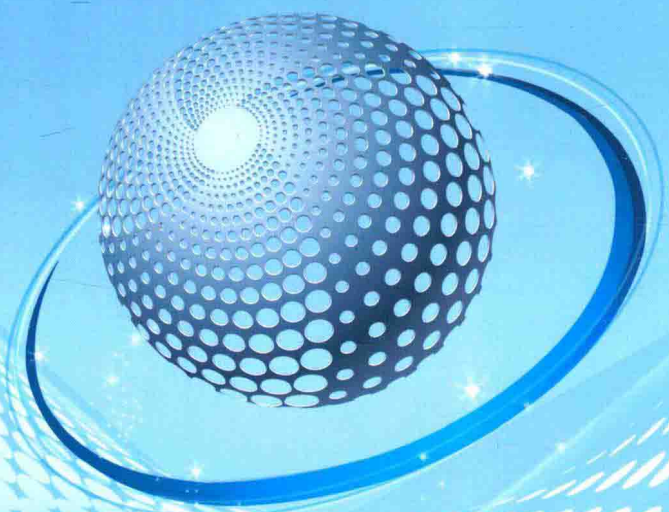
中国科协新一代信息技术系列丛书

大数据导论

Introduction to Big Data

张尧学 主编 / 胡春明 执行主编

中国电子学会 组编



机械工业出版社
CHINA MACHINE PRESS

中国科协新一代信息技术系列丛书

大数据导论

主　　编　张尧学
执行主编　胡春明
参　　编　王宏志　唐　杰　王建民
袁晓如　朱跃生　吴中海
吕金虎　王　晨　陈恩红
刘　闯　王德庆　马民虎

中国电子学会　组编

机械工业出版社

本书是中国科协新一代信息技术系列丛书之一。

本书重点阐述大数据的基本原理、技术、平台和不同领域的应用案例。全书共分13章，第1章为绪论；第2～7章为技术章节，介绍了数据采集与治理、数据管理、数据分析、数据可视化、数据安全与隐私保护和大数据处理平台；第8～11章为大数据在不同领域的应用案例，包括社会网络大数据、城市大数据、工业大数据和教育大数据；第12、13章为数据管理章节，包括数据开放与共享和大数据的法律政策规范。

本书主要面向大学非计算机类的工科专业的高年级学生与研究生，亦可作为大数据爱好者的科普读物。

本书配有免费的电子课件，欢迎选用本书作教材的老师登录www. cmpedu. com注册下载。

图书在版编目（CIP）数据

大数据导论/张尧学主编. —北京：机械工业出版社，2018. 8
（中国科协新一代信息技术系列丛书）
ISBN 978-7-111-60767-0

Ⅰ. ①大… Ⅱ. ①张… Ⅲ. ①数据处理－研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2018）第195535号

机械工业出版社（北京市百万庄大街22号 邮政编码100037）
策划编辑：路乙达 责任编辑：路乙达 王雅新
责任校对：黄兴伟 樊钟英 封面设计：张 静
责任印制：张 博
三河市宏达印刷有限公司印刷
2018年8月第1版第1次印刷
184mm×260mm · 20印张 · 484千字
标准书号：ISBN 978-7-111-60767-0
定价：49. 80元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务
服务咨询热线：010-88379833
读者购书热线：010-88379649

网络服务
机 工 官 网：www. cmpbook. com
机 工 官 博：weibo. com/cmp1952
教育服务网：www. cmpedu. com
金 书 网：www. golden-book. com

《大数据导论》编写组

顾问：

李德毅　中国工程院院士
梅　宏　中国科学院院士
王海峰　百度高级副总裁

主编：

张尧学　中国工程院院士

执行主编：

胡春明　北京航空航天大学

参编：

王宏志　哈尔滨工业大学
唐　杰　清华大学
王建民　清华大学
袁晓如　北京大学
朱跃生　北京大学
吴中海　北京大学
吕金虎　北京航空航天大学
王　晨　清华大学
陈恩红　中国科学技术大学
刘　闯　中国科学院
王德庆　北京航空航天大学
马民虎　西安交通大学

前　言

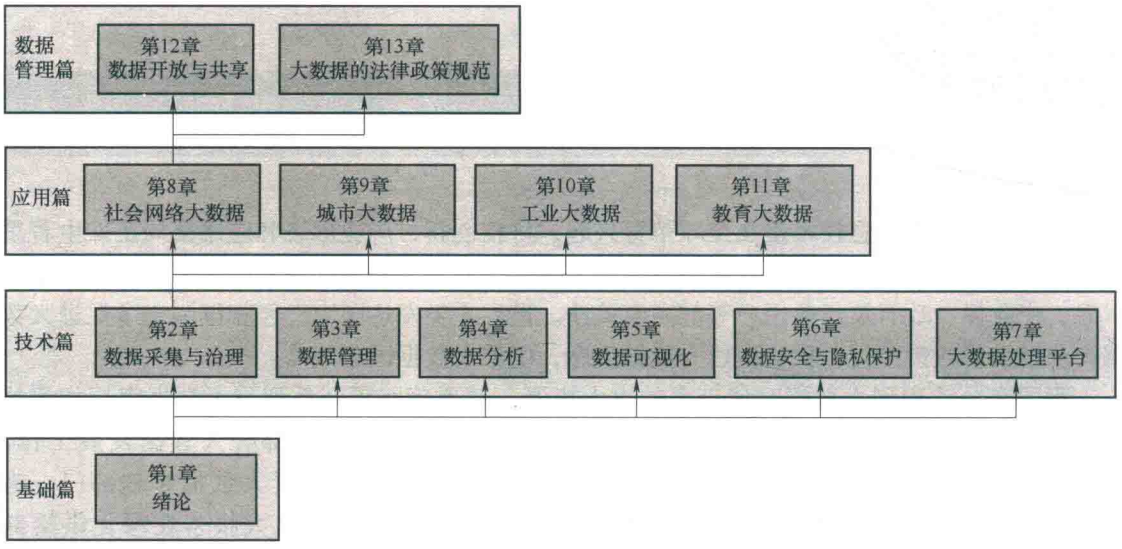
当前，新一代信息技术正在全球孕育兴起，科技创新、产业形态和应用格局正发生着重大变革。随着数据获取和计算技术的进步，大数据已成为一种新的国家战略资源，引起了学术界、产业界、政府及行业用户等的高度关注。世界主要发达国家已经相继制定了促进大数据产业发展的政策法规，积极构建大数据生态，实施大数据国家战略。

我国充分认识到大数据时代带来的重大机遇，部署落实了一系列与大数据密切相关的规划。2015 年，国务院印发《促进大数据发展行动纲要》，系统部署大数据发展工作。目前，多个省市已经出台大数据相关政策，一些地方政府专门设置大数据管理部门，为大数据基础设施、技术创新、产业发展营造了良好环境。党的十九大报告在深化供给侧结构性改革中指出："加快建设制造强国，加快发展先进制造业，推动互联网、大数据、人工智能和实体经济深度融合，在中高端消费、创新引领、绿色低碳、共享经济、现代供应链、人力资本服务等领域培育新增长点、形成新动能。"更加明确大数据应与各个行业深度融合。

为落实国家战略，加速新一代信息技术人才培养，满足数字经济发展的人才需求，为实现经济高质量发展提供人才支撑，中国科学技术协会策划并组织编写以云计算、大数据、人工智能等为代表的新一代信息技术系列丛书，成立了中国科协新一代信息技术系列丛书编制委员会，聘请梅宏院士为编委会主任，李培根院士、李德毅院士、李伯虎院士、张尧学院士、李骏院士、谭铁牛院士、赵春江院士为编委会委员，统筹丛书编制工作。本书是该系列丛书之一。

本书主要面向大学非计算机类的工科专业的高年级学生与研究生，目的是帮助学生掌握大数据的基本原理和基本知识，熟悉大数据技术在多个行业应用中"能与不能"的边界，培养学生在本专业应用大数据的能力。同时，对于计算机相关专业的学生，本书也可作为大数据专业课程的导论课教材。本书注重知识结构的基础性与完整性，确保技术内容的通用性、普适性与先进性，遵循教育规律，加强能力培养，同时，精选行业真实案例，开阔学生视野，启发创新思维。本书期望为跨学科研究者提供学科方法论和技术概述，满足新一代信息技术人才的要求。

本书可分为四部分。基础篇（第 1 章）介绍大数据的发展历程、内涵和外延、价值与意义；技术篇（第 2 ~ 7 章）以数据采集与治理、数据管理、数据分析与可视化的典型大数据应用生命周期作为主线，对大数据关键技术进行讲解，进而阐述数据安全与大数据平台的关键技术，努力呈现技术的逻辑性和严密的科学思维；应用篇（第 8 ~ 11 章）使学生熟悉大数据的典型应用领域，从社会网络大数据、城市大数据、工业大数据和教育大数据方面进行案例剖析，满足多个学科的教学需求；数据管理篇（第 12 ~ 13 章）则让学生了解数据开放与共享和法律政策规范等方面的现状。本书的具体结构如下图所示：



本书采用模块化教学思维进行编写，授课老师和学生可以根据专业和现有知识结构，选取不同的教学方案。本书教学建议为 32 ~ 48 学时，基础篇和技术篇建议 16 ~ 20 学时，应用篇建议 14 ~ 24 学时；数据管理篇建议 2 ~ 4 个学时或由学生自学。根据不同专业的实际情况，老师可以根据学时安排选择 1 ~ 2 个不同行业应用进行重点讲解，满足教学需求。

本书的编写汇集了多位专家学者的智慧。本书主编张尧学院士带领编写组全体成员，从教学理论和学术研究等多角度系统地进行顶层设计和撰写工作。本书第 1 章由胡春明编写，第 2 章由王宏志编写，第 3 章由王建民编写，第 4 章和第 8 章由唐杰编写，第 5 章由袁晓如编写，第 6 章由朱跃生编写，第 7 章由吴中海编写，第 9 章由吕金虎编写，第 10 章由王晨编写，第 11 章由陈恩红编写，第 12 章由刘闯、王德庆编写，第 13 章由马民虎编写。全书由胡春明统稿。

本书邀请了李德毅院士、梅宏院士和百度公司王海峰博士担任顾问专家，他们对本书的学术观点、技术方向以及内容组织都提供了极具价值的意见和建议。在此对各位领导和专家表示深深的敬意和感谢。

此外，中国科协领导多次协调，确保了丛书编制和推广工作的顺利进行。中国科学技术协会学术部对丛书的撰写、出版、推广全过程提供了大力支持和具体指导。中国科协智能制造学会联合体承担了丛书的前期调研、组织协调和推广宣传工作。中国电子学会承担了本书编写的全部组织工作，学会副理事长兼秘书长徐晓兰对本书高度重视，布置相关工作。中国电子学会的林润华副秘书长组织并指导了本书的编撰工作。中国电子学会的宁慧聪博士以及团队王娟副主任、王海涛老师和张玲老师在本书写作过程中精心组织，扎实推进此项工作。机械工业出版社的全力支持和悉心编校，让这本书的付梓成为可能，感谢他们的辛勤工作。本书编写还得到了多个单位和专家的支持，他们是北京航空航天大学王磊副教授、王静远副教授、邓婷博士，清华大学宋韶旭老师、任艮全博士、韩裔博士、丁铭博士，北京大学刘宏志副教授、莫同副教授，洪帆博士、赖楚凡、张江、李国政、陈帅、岳成磊、韩云、蒋瑞珂、林丽静。同时也感谢北京大学深圳研究生院“数据科学与智能计算”学科负责人白志强教授，中国科学技术大学刘淇副教授，西安交通大学党家玉博士，中国社科院陈新欣研究

员，人民大学陈跃国副教授，武汉大学彭煜玮副教授，百度大数据总监刘丽萍，科大讯飞大数据研究院执行院长谭昶，工程师于俊等多位专家学者。感谢大数据专家委员会顾问李德毅院士、主任委员梅宏院士，副主任委员张尧学院士，秘书长林润华、副秘书长胡春明，委员朱跃生、吕金虎对本书的大力支持。同时感谢本书被引用和作为参考文献的作者和机构。

大数据是一个新兴领域，处于飞速发展的阶段，且与多领域、多学科紧密结合，还有很大发展空间。由于时间、精力、知识结构有限，书中难免存在错误和不妥之处，恳请广大读者批评指正，以便编写组对本书的进一步完善。

《大数据导论》主编张尧学和编写组全体成员

2018 年 8 月

目 录

应 用 篇

数据管理篇

基础篇