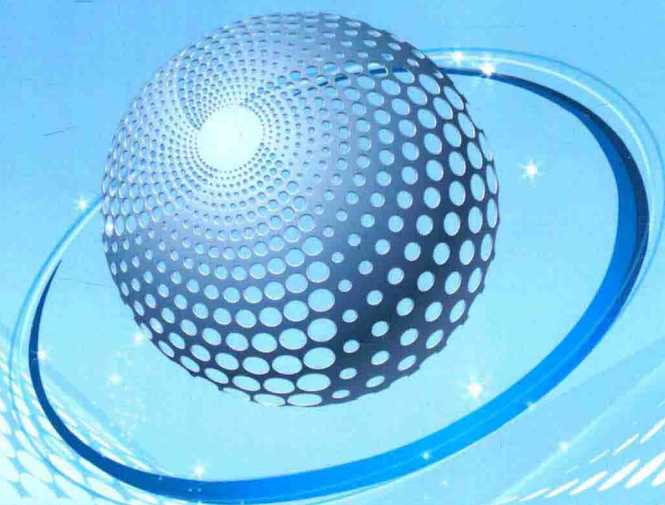


中国科协新一代信息技术系列丛书

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

网络服务

服务咨询热线: 010-88379833

机工官网: www.cmpbook.com

读者购书热线: 010-88379649

机工官博: weibo.com/cmp1952

教育服务网: www.cmpedu.com

封面无防伪标均为盗版

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

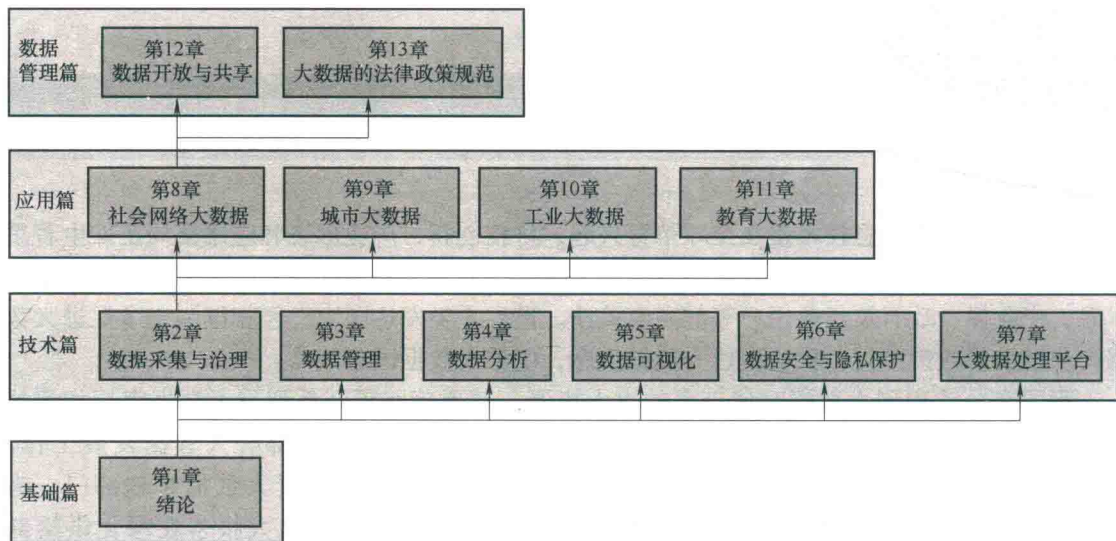
当前，新一代信息技术正在全球孕育兴起，科技创新、产业形态和应用格局正发生着重大的变革。随着数据获取和计算技术的进步，大数据已成为一种新的国家战略资源，引起了学术界、产业界、政府及行业用户等的高度关注。世界主要发达国家已经相继制定了促进大数据产业发展的政策法规，积极构建大数据生态，实施大数据国家战略。

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、任良全博士、韩喬博士、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

前 言

基 础 篇

第1章 绪论	3
1.1 概述	3
1.1.1 数据	4
1.1.2 数据中蕴含的价值	6
1.1.3 获取数据中蕴含的价值	8
1.2 大数据的内涵和外延	9
1.2.1 大数据时代的驱动力	9
1.2.2 大数据的概念和特征	10
1.2.3 大数据带来的思维模式改变	12
1.2.4 大数据的作用和意义	13
1.3 大数据的技术挑战和科学意义	15
1.3.1 数据处理的一般过程	16
1.3.2 大数据计算面临的挑战	17
1.3.3 大数据计算的特点	18
1.3.4 大数据计算平台	19
1.3.5 大数据与云计算、人工智能的关系	20
1.4 数据科学	22
1.4.1 数据科学的提出	23
1.4.2 数据科学的范畴	23
1.4.3 数据科学对学科发展的影响	24
习题	25
参考文献及扩展阅读资料	26

技 术 篇

第2章 数据采集与治理	29
2.1 概述	30
2.2 大数据的来源与多源数据的采集方式	30

2.2.1	大数据的来源	30
2.2.2	多源数据的采集	31
2.2.3	数据离散化	33
2.3	数据集成和跨界应用的数据集成方法	34
2.3.1	数据集成的定义与形式	34
2.3.2	传统数据集成	35
2.3.3	跨界数据集成	38
2.4	数据的预处理	40
2.4.1	数据变换	40
2.4.2	数据质量的检验与提升	41
	习题	44
	参考文献及扩展阅读资料	44
第3章	数据管理	46
3.1	概述	46
3.2	关系数据库	47
3.2.1	关系数据模型	48
3.2.2	结构化查询语言	51
3.2.3	数据库事务	52
3.2.4	关系数据库管理系统	53
3.3	分布式文件系统	54
3.3.1	Hadoop	55
3.3.2	Ceph	57
3.3.3	GlusterFS	59
3.3.4	分布式文件系统对比	60
3.4	新型数据管理与查询系统	61
3.4.1	NoSQL 数据库	61
3.4.2	SQL on Hadoop 系统	65
	习题	68
	参考文献及扩展阅读资料	68
第4章	数据分析	69
4.1	概述	69
4.2	统计数据分析	71
4.2.1	数据描述性分析	71
4.2.2	回归分析	74
4.3	基于机器学习的数据分析	76
4.3.1	非监督学习方法	77
4.3.2	监督学习方法	77

4.4	图的数据分析	84
4.4.1	图的基本概念	85
4.4.2	中心性和相似性分析	86
4.4.3	社交网络上的算法	89
4.5	自然语言中的数据分析	92
4.5.1	词表示分析	92
4.5.2	语言模型	94
4.5.3	话题模型	95
	习题	96
	参考文献及扩展阅读资料	96
第5章	数据可视化	98
5.1	概述	98
5.2	数据可视化主要技术	101
5.2.1	高维数据可视化	102
5.2.2	网络数据可视化	106
5.2.3	层次结构数据可视化	109
5.2.4	时空数据可视化	112
5.2.5	文本数据可视化	115
5.3	高可扩展可视化技术	117
5.3.1	科学可视化中的高可扩展性	117
5.3.2	支持数据高效的存储和检索的可视化	121
5.3.3	支持可扩展可视化的交互手段	123
5.4	大数据可视化与可视分析案例	125
5.4.1	VAST Challenge 2017 的可视分析案例	125
5.4.2	车辆轨迹数据的可视分析案例	128
5.5	可视化工具和软件	131
5.5.1	高维数据可视化工具	131
5.5.2	文本可视化工具	132
5.5.3	网络可视化工具	132
5.5.4	可视化编程工具	132
	习题	132
	参考文献及扩展阅读资料	132
第6章	数据安全与隐私保护	135
6.1	概述	135
6.1.1	数据安全与传统信息安全的共异点	136
6.1.2	数据采集及传输中的安全与隐私	136
6.1.3	数据存储的安全与隐私	138

6.1.4	数据分析挖掘及处理的安全与隐私	138
6.1.5	数据交互、共享与服务的安全与隐私	139
6.2	数据安全及隐私保护支撑技术	140
6.2.1	密码学基础及关键技术	140
6.2.2	公钥基础设施	146
6.2.3	授权管理基础设施	147
6.2.4	PKI 与 PMI 协同工作原理	148
6.2.5	秘密分割与共享管理技术	149
6.3	数据脱敏技术与实践	150
6.3.1	数据交互安全与脱敏技术	150
6.3.2	静态数据脱敏技术	150
6.3.3	动态数据脱敏技术	150
6.3.4	数据脱敏实例	151
6.4	数据生命周期安全的防护及管理体系	151
6.4.1	数据安全防护体系	151
6.4.2	数据安全标准	153
6.4.3	数据生命周期安全实施方案与数据安全管	154
	习题	155
	参考文献及扩展阅读资料	155
第 7 章	大数据处理平台	157
7.1	概述	157
7.2	大数据处理平台架构	158
7.2.1	技术架构	158
7.2.2	开源平台	159
7.3	批量大数据计算	161
7.3.1	基本概念	161
7.3.2	典型批量计算系统	162
7.3.3	实例：微博用户群体年度热词统计	164
7.4	流式大数据计算	166
7.4.1	基本概念	166
7.4.2	典型流式计算系统	168
7.4.3	实例：微博用户群体实时热门话题分析	169
7.5	大规模图数据计算	170
7.5.1	基本概念	170
7.5.2	典型图计算系统	172
7.5.3	实例：微博用户影响力排名	174
	习题	175
	参考文献及扩展阅读资料	176

应用篇

第8章 社会网络大数据	179
8.1 概述	179
8.2 社会网络大数据面临的挑战	181
8.3 社会网络中的用户影响力	182
8.3.1 影响力检测实验	183
8.3.2 影响力传播模型	185
8.3.3 影响力度量算法	186
8.3.4 社会影响力应用	186
8.4 在线社会媒体中信息传播的建模与预测	187
8.4.1 网络信息传播模型	187
8.4.2 传播网络推断	188
8.4.3 影响力最大化	188
8.4.4 信息传播预测	189
习题	192
参考文献及扩展阅读资料	192
第9章 城市大数据	194
9.1 概述	194
9.1.1 城市数据的分类	195
9.1.2 城市数据的特点	195
9.2 智慧城市	197
9.2.1 智慧城市的概念	197
9.2.2 智慧城市的发展现状	198
9.2.3 智慧城市的未来趋势	199
9.3 智慧城市的技术体系框架	200
9.3.1 智慧城市的技术框架	200
9.3.2 以数据为中心的智慧城市特点	201
9.3.3 智慧城市中的典型应用与服务	203
9.4 城市大数据应用案例	205
9.4.1 交通大数据的来源与种类	206
9.4.2 交通大数据的分析与处理	206
9.4.3 交通大数据的应用成果	207
9.5 城市大数据未来展望	208
习题	209
参考文献及扩展阅读资料	210

第 10 章 工业大数据	211
10.1 概述	211
10.1.1 工业大数据的内涵	212
10.1.2 工业大数据的特点	213
10.2 工业大数据典型应用场景	218
10.2.1 现有业务优化	218
10.2.2 促进企业升级转型	219
10.3 工业大数据分析技术	220
10.3.1 工业大数据分析工作准备	221
10.3.2 工业大数据分析工作实施	222
10.3.3 工业大数据分析关键技术	225
10.4 工业大数据分析案例	226
10.4.1 大唐集团工业大数据应用实践	226
10.4.2 中联重科工业大数据应用实践	229
习题.....	233
参考文献及扩展阅读资料.....	233
第 11 章 教育大数据	234
11.1 概述	234
11.2 教育大数据的采集与应用场景	235
11.2.1 信息化校园	236
11.2.2 智能辅导系统和在线题库	238
11.2.3 大规模开放式在线课程	239
11.3 认知诊断分析	241
11.3.1 认知诊断任务描述	242
11.3.2 经典认知诊断方法	242
11.3.3 基于大数据的协同认知诊断	244
11.4 知识跟踪分析	245
11.4.1 知识跟踪任务描述	246
11.4.2 经典知识跟踪方法	246
11.4.3 联合知识跟踪	247
11.5 习题资源分析与挖掘	249
11.5.1 相似习题判定任务描述	249
11.5.2 相似习题判定技术	249
11.5.3 其他习题分析与挖掘应用	250
11.6 MOOC 平台活跃度预测	251
11.6.1 活跃度预测任务描述	251
11.6.2 活跃度预测分析方法	251

11.7 教育大数据应用案例	252
11.7.1 基于大数据分析的学生“隐形补助”体系	252
11.7.2 基于大数据技术的个性化学习	255
习题	258
参考文献及扩展阅读资料	258

数据管理篇

第12章 数据开放与共享

263

12.1 概述	263
12.1.1 数据开放与共享的概念	263
12.1.2 数据开放与共享的发展历程	264
12.2 数据开放与共享的原则与政策	266
12.2.1 数据开放与共享原则	266
12.2.2 国外数据开放与共享政策	266
12.2.3 中国数据开放与共享政策	269
12.2.4 数据开放与共享实施指南	270
12.3 数据开放与共享分类	270
12.3.1 政府数据开放与共享	271
12.3.2 公共财政资助产生的科学数据开放与共享	271
12.3.3 企业数据开放与共享	272
12.3.4 个人数据开放与共享	272
12.4 数据开放与共享平台	273
12.4.1 数据开放与共享综合平台	273
12.4.2 数据开放与共享领域平台	275
12.4.3 数据开放与共享平台的基本功能	279
12.4.4 数据开放与共享平台的产权保护	281
习题	282
参考文献及扩展阅读资料	282

第13章 大数据的法律政策规范

284

13.1 大数据政策法规指引	284
13.1.1 大数据政策法规发展过程	284
13.1.2 中国的数据保护监管机构	286
13.2 数据主权与数据权利	287
13.2.1 数据主权	288
13.2.2 数据权利	288
13.2.3 数据权利主体和其他利益相关主体	288
13.3 个人数据立法保护	289

13.3.1	国外个人数据保护制度	289
13.3.2	中国个人数据保护制度	292
13.4	数据跨境流动监管法律机制	295
13.4.1	国外数据跨境及数据本地化立法	296
13.4.2	中国数据跨境流动法律制度	298
13.4.3	数据跨境流动法律制度设计	299
13.5	大数据伦理	301
	习题	302
	参考文献及扩展阅读资料	302

基础篇

