



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

金海系列

# 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

## 内 容 简 介

本书系统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。

版权专有 侵权必究

---

### 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

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

---

图书出现印装质量问题, 请拨打售后服务热线, 本社负责调换

# 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 1 章 引论	003
1.1 序言	003
1.2 数据科学简述	003
1.2.1 数据科学的定义	003
1.2.2 数据科学的由来	004
1.2.3 数据科学的研究范畴	005
1.2.4 数据科学的学习意义	006
1.3 本书结构	006

### 第二部分 大数据及其产生根源

第 2 章 数据	011
2.1 数据的定义	011
2.1.1 数据的定义	011
2.1.2 其他相关概念	011
2.2 数据简史	012
第 3 章 大数据的概念和特征	019
3.1 大数据的概念	019
3.2 大数据的 4 V 特性	020
3.2.1 体量 (Volume)	020
3.2.2 多样性 (Variety)	022
3.2.3 价值/真实性 (Value/Veracity)	023
3.2.4 速度 (Velocity)	024
3.2.5 对 4 V 特性的体会	024
第 4 章 大数据的产生根源	025
4.1 大数据的产生根源	025
4.1.1 大数据时代出现的技术基础	025

4.1.2 大数据时代出现的数据基础	025
4.2 大数据简史	026
4.3 大数据时代的挑战	028
4.3.1 数据规模	030
4.3.2 数据的多样性和异构性	031
4.3.3 数据的不可靠问题	031
4.3.4 数据的实时性要求	031
4.3.5 数据隐私问题	031
4.3.6 人机协作问题	031
4.3.7 数据的访问与共享	031
4.3.8 数据运用的合理性	032
4.3.9 小结	032

### 第三部分 大数据研究的重要性

第5章 大数据研究的现状	035
5.1 政府篇	035
5.1.1 联合国的大数据研究	035
5.1.2 美国的大数据研究	036
5.1.3 欧盟的大数据研究	037
5.2 企业篇	039
5.2.1 谷歌	039
5.2.2 IBM	039
5.2.3 百度	041
5.2.4 阿里巴巴	041
5.2.5 腾讯	042
第6章 关联分析	044
6.1 啤酒与尿布	044
6.1.1 案例详析	044
6.1.2 购物篮分析法	045
6.1.3 商品间相关性分析	047
6.1.4 外界因素的影响	051
6.1.5 思维启示	052
6.2 亚马逊的个性化推荐	054
6.2.1 案例详析	054
6.2.2 亚马逊的推荐方式	055
6.2.3 推荐算法	057
6.3 潘多拉音乐组计划	061
6.3.1 案例详析	061

6.3.2 标签的运用	064
6.4 塔吉特的大数据营销	067
6.4.1 案例详析	067
6.4.2 思维启示——数据应用已经渗入生活的方方面面	068
<b>第7章 趋势预测</b>	<b>070</b>
7.1 “搜索+比价”	070
7.1.1 Farecast 案例详析	070
7.1.2 Decide 案例详析	072
7.1.3 思维启示	075
7.2 Twitter 与对冲基金	075
7.2.1 案例详析	075
7.2.2 思维启示：数据可以预测趋势与规律	077
7.3 疾病预测	077
7.3.1 谷歌流感趋势	077
7.3.2 其他案例	081
7.3.3 思维启示	081
7.4 电影票房预测	083
7.4.1 案例详析	083
7.4.2 工作模式	084
7.4.3 思维启示：简单的就是最好的	087
7.5 奥斯卡预测	088
7.5.1 案例详析	088
7.5.2 思维启示：大数据可以做预测	090
<b>第8章 决策支持</b>	<b>092</b>
8.1 《纸牌屋》	092
8.1.1 案例详析	092
8.1.2 大数据的运用方式	093
8.1.3 思维启示	093
8.2 美国总统大选	096
8.2.1 案例详析	096
8.2.2 大数据的运用方式	096
8.2.3 思维启示	099
<b>第9章 模式创新</b>	<b>100</b>
9.1 大数据与反恐	100
9.1.1 美国“棱镜”计划	100
9.1.2 加拿大的“棱镜门”	102
9.1.3 思维启示	103
9.2 利用大数据打击犯罪	105
9.2.1 “先知”系统	105

9.2.2 “犯罪数据分析和趋势预测系统”	106
9.3 大数据与破案	107
9.3.1 《源代码》	107
9.4 大数据的其他运用方式	108
9.4.1 大数据与纽约沙井盖维护	108
9.4.2 大数据帮助寻根问祖	109

## 第四部分 数据科学的研究方式

第10章 数据密集型研究方法	117
10.1 范式和范式的演化过程	117
10.1.1 范式的定义	117
10.1.2 范式的演变过程	118
10.2 第四范式兴起的根源	120
10.2.1 数据洪流的到来	120
10.2.2 科学界对海量数据的关注	121
10.2.3 关联数据运动	122
10.2.4 政府数据开放运动	123
10.3 对第四范式的分析	124
10.3.1 科学数据与科学研究的问题	124
10.3.2 解决方案	124
10.4 数据科学研究的一般流程	125
第11章 数据的获取和预处理	127
11.1 数据的获取	127
11.1.1 数据的类型	127
11.1.2 网络爬虫技术	129
11.2 数据预处理的目的是	136
11.3 数据清洗	137
11.3.1 填补空缺值	137
11.3.2 平滑噪声数据	138
11.4 数据集成	142
11.4.1 多信息源的匹配	142
11.4.2 冗余数据的处理	143
11.5 数据变换	145
11.5.1 数据规范化	145
11.6 数据归约	146
11.6.1 数据立方体聚集	147
11.6.2 维归约	148



11.6.3	特征值归约 .....	150
<b>第 12 章</b>	<b>数据的存储与管理</b> .....	<b>151</b>
12.1	数据的存储 .....	151
12.1.1	数据存储的发展 .....	151
12.1.2	大数据对存储带来的挑战 .....	155
12.1.3	云存储方式 .....	156
12.2	数据的管理 .....	157
12.2.1	数据管理的发展阶段 .....	157
12.2.2	大数据时代数据管理的特点 .....	160
12.2.3	非关系型数据 .....	161
12.2.4	开源的 NoSQL 数据库软件 .....	161
<b>第 13 章</b>	<b>数据的处理</b> .....	<b>165</b>
13.1	Hadoop .....	165
13.1.1	Hadoop 的起源 .....	165
13.1.2	优点 .....	165
13.1.3	架构 .....	166
13.1.4	MapReduce 流程 .....	167
13.2	Spark .....	168
13.2.1	概述 .....	168
13.2.2	Spark 的特点 .....	169
13.2.3	编程模型 .....	169
13.2.4	运行和调度 .....	170
<b>第 14 章</b>	<b>数据的可视化</b> .....	<b>173</b>
14.1	概述 .....	173
14.2	可视化工具 .....	174
14.2.1	Excel .....	174
14.2.2	Raphaël .....	174
14.2.3	Visual.ly .....	175
14.2.4	Crossfilter .....	175
14.2.5	Polymaps .....	175
14.2.6	Kartograph .....	176
14.2.7	Processing .....	176
14.2.8	R .....	177
14.2.9	Weka .....	177
14.2.10	Gephi .....	178

## 第五部分 数据与未来

<b>第 15 章 大数据与智慧城市</b> .....	183
15.1 概述.....	183
15.1.1 智慧城市的定义.....	183
15.1.2 智慧城市产生背景.....	184
15.2 大数据与智慧城市.....	186
15.2.1 智慧城市的基本特征与层次构成.....	186
15.2.2 智慧城市建设中所应用的数据科学技术.....	188
15.3 智慧城市案例.....	190
15.3.1 韩国.....	191
15.3.2 日本.....	192
15.3.3 美国.....	192
15.3.4 爱沙尼亚.....	196
15.3.5 荷兰.....	198
15.3.6 英国.....	199
15.3.7 巴西.....	200
<b>第 16 章 大数据与智慧医疗</b> .....	205
16.1 概述.....	205
16.2 智慧医疗的范畴.....	205
16.2.1 临床操作.....	205
16.2.2 付款/定价.....	207
16.2.3 研发.....	208
16.2.4 新的商业模式.....	209
16.2.5 公众健康.....	209
16.2.6 给我们的思维模式启示.....	210
16.3 大数据与智慧医疗.....	210
16.3.1 大数据服务心脏病患者.....	210
16.3.2 “魔毯”病人的监控.....	211
16.3.3 大数据监测脑外伤病人恢复.....	211
16.3.4 大数据帮助实现个性化用药和诊断.....	212
16.4 可穿戴技术.....	213
16.4.1 可穿戴技术的概念.....	213
16.4.2 可穿戴设备简析.....	214
16.4.3 可穿戴设备与智慧医疗.....	218
16.4.4 思维启示——可穿戴设备的缺陷.....	218
<b>第 17 章 大数据与未来生活</b> .....	221
17.1 数据科学家.....	221

17.1.1 数据科学家的定义 .....	221
17.1.2 数据科学家的从业前景 .....	221
17.2 对未来数据科学发展的探讨 .....	224
17.2.1 数据不是万能 .....	224
17.2.2 提防进入数据误区 .....	225

# 图 目 录

图 1-1	科学体系	004
图 1-2	彼得·诺尔（前图灵奖得主，丹麦人）	004
图 1-3	吴建福（国际知名统计学家，美国国家工程院院士）	005
图 1-4	数据科学的研究范畴	005
图 1-5	数据科学与其他学科的关系	006
图 1-6	数据科学家——未来最性感的职业	007
图 2-1	穴居壁画——最古老的数据记录形式	012
图 2-2	结绳记事	013
图 2-3	古埃及人用莎草记录数据	013
图 2-4	造纸流程	014
图 2-5	最早的留声机	014
图 2-6	最古老的照相机	014
图 2-7	早期的摄影机	014
图 2-8	世界上第一台电子计算机	015
图 2-9	当前世界上最快的计算机——天河 2 号	015
图 2-10	网络与数据	016
图 2-11	物联网	017
图 2-12	用数据产生智慧之花	017
图 2-13	用数据来一窥未来	018
图 3-1	国际数据公司对大数据的定义	019
图 3-2	大数据的 4 V 特性	020
图 3-3	大数据与数据仓库对比	021
图 3-4	大数据的类型繁多	022
图 3-5	寻找数据价值	023
图 4-1	移动互联网的飞速发展	026
图 4-2	物联网天生就是大数据	026
图 4-3	穿孔卡片	027
图 4-4	IBM 沃森计算机系统	028
图 4-5	大数据时代来临组图一	028
图 4-6	大数据时代来临组图二	029
图 4-7	大数据时代来临组图三	029
图 4-8	大数据带来新的洞察力	030
图 4-9	分布式计算	030

图 4-10	大数据时代的隐私保护	031
图 4-11	大数据要求复杂多领域人机协作	032
图 5-1	全球脉搏计划	035
图 5-2	data.gov 网站	037
图 5-3	西班牙桑坦德	038
图 5-4	谷歌的大数据布局	039
图 5-5	IBM 的大数据平台和应用程序框架	040
图 5-6	百度的大数据引擎	041
图 5-7	阿里巴巴大数据竞赛	042
图 5-8	腾讯大数据平台	043
图 6-1	啤酒与尿布	044
图 6-2	美式购物篮分析法的代表——沃尔玛	045
图 6-3	日式购物篮分析法的代表——7-11 便利店	046
图 6-4	研究数据不应停留于表面	054
图 6-5	亚马逊的个性化推荐系统	055
图 6-6	鼓励用户参与投票、书评等主观性活动	056
图 6-7	推荐系统	057
图 6-8	基于人口统计学的推荐	058
图 6-9	基于内容的推荐	059
图 6-10	潘多拉音乐盒子	061
图 6-11	潘多拉网络电台	062
图 6-12	私人定制的电台音乐	063
图 6-13	潘多拉电台 App	063
图 6-14	推荐的方式	064
图 6-15	Delicious 网站	065
图 6-16	Lastfm 网站	066
图 6-17	CiteULike	066
图 6-18	Hulu	067
图 6-19	塔吉特预测怀孕	068
图 7-1	Farecast 机票预测	071
图 7-2	Decide 电商比价网站	073
图 7-3	Decide 的比价无远弗届	073
图 7-4	对电子产品系列更替的把握	074
图 7-5	分数化评价	074
图 7-6	“平静”指数和道琼斯指数对比	077
图 7-7	谷歌预测美国流感趋势	078
图 7-8	谷歌预测数据与真实数据的对比	079
图 7-9	检索词个数的选取	080
图 7-10	谷歌流感趋势预测结果	080

图 7-11	大数据傲慢	082
图 7-12	2012 年票房收入与搜索量的曲线	085
图 7-13	2012 年票房收入和两类搜索量的曲线	085
图 7-14	搜索量与首周票房收入之间的关系	086
图 7-15	提前一周预测票房的效果	086
图 7-16	提前一个月预测票房的效果	087
图 7-17	微软研究院的戴维德·罗斯柴尔德	088
图 7-18	微软研究院的戴维德·罗斯柴尔德博士预测奥斯卡获奖名单	088
图 7-19	PredictWise 网站	089
图 7-20	奥斯卡投票预测器	090
图 8-1	纸牌屋	092
图 8-2	大数据帮助大选	098
图 9-1	美国“棱镜门”风波	100
图 9-2	谁是“棱镜”计划的帮凶	101
图 9-3	美国“棱镜”计划	101
图 9-4	大数据时代的个人隐私安全	103
图 9-5	《少数派报告》	106
图 9-6	《源代码》海报	107
图 9-7	纽约沙井盖	109
图 9-8	Ancestry.com	110
图 9-9	家谱网站帮助寻根问祖	111
图 10-1	关联数据运动	122
图 10-2	数据科学的研究流程	126
图 11-1	订单	128
图 11-2	订单数据的数据类型区分	129
图 11-3	通用的网络爬虫框架	129
图 11-4	互联网网页的划分	130
图 11-5	网页拓扑结构示例	131
图 11-6	聚类抽样策略	132
图 11-7	分布式抓取系统结构	133
图 11-8	主从式基本结构	133
图 11-9	对等式基本结构	134
图 11-10	一致性哈希法确定服务器分工	135
图 11-11	数据清洗	137
图 11-12	聚类平滑噪声数据	142
图 11-13	回归方法平滑噪声数据	142
图 11-14	数据集成	142
图 11-15	客户基本情况	143
图 11-16	客户交易数据	143

图 11-17	数据表	144
图 11-18	写入期望值	144
图 11-19	数据立方体的示例	147
图 11-20	数据立方体的聚集	148
图 11-21	维归约	149
图 12-1	布乔的构想	151
图 12-2	自动提花编织机	151
图 12-3	打孔纸卡	152
图 12-4	穿孔纸带	152
图 12-5	计数电子管	152
图 12-6	盘式磁带	153
图 12-7	盒式录音磁带	153
图 12-8	磁鼓	154
图 12-9	软磁盘	154
图 12-10	硬盘机	154
图 12-11	日立 Deskstar 7K 500 硬盘	154
图 12-12	光盘	155
图 12-13	CD 光盘	155
图 12-14	DVD 光盘	155
图 12-15	云存储	157
图 12-16	互联网数据呈爆炸式增长	160
图 13-1	Spark 工作空间	170
图 13-2	Spark 程序运行示意图	171
图 13-3	窄依赖和宽依赖	171
图 13-4	Stage 的划分	172
图 14-1	约翰·图基	173
图 14-2	Anscombe 的四重奏	173
图 14-3	Anscombe 的四重奏的可视化呈现	174
图 14-4	Google Chart API	175
图 14-5	Raphaël	175
图 14-6	Visual.ly	175
图 14-7	Crossfilter	176
图 14-8	Polymaps	176
图 14-9	Kartograph	176
图 14-10	Processing 编程环境	177
图 14-11	R 语言编程	178
图 14-12	Weka 编程环境	178
图 14-13	利用 Gephi 做数据可视化	179
图 15-1	智慧城市	183

图 15-2	城市建设所面临的社会基础设施问题	186
图 15-3	智慧城市的基本特征	186
图 15-4	智慧城市组件示意图	187
图 15-5	智慧城市建设必需的数据分析技术的发展趋势	189
图 15-6	松岛新城	191
图 15-7	韩国政府斥巨资修建的松岛新城	192
图 15-8	美国俄亥俄州的哥伦布市	194
图 15-9	My Columbu 移动应用程序	194
图 15-10	爱沙尼亚的塔林市	197
图 15-11	阿姆斯特丹的 ASC 计划	198
图 15-12	里约热内卢的城市运营中心	201
图 15-13	里约热内卢的智慧城市建设	201
图 16-1	科学家利用大数据监控脑伤病人的恢复情况	212
图 16-2	建立个性化的用药模型	213
图 16-3	可穿戴设备	213
图 16-4	Maxim 生命体征测量 T 恤	214
图 16-5	TI 公司的 Health Tech	215
图 16-6	Valencell: 可随身穿戴的微型生理监测模块	216
图 16-7	Google Glass	216
图 16-8	苹果的 iWatch	217
图 16-9	BrainLink 意念头箍	217
图 17-1	麦肯锡对数据科学方面人才需求空缺的预测	222
图 17-2	数据科学从业人员的未来成长性	224
图 17-3	明确数据的优势和不足	225



# 第一部分 引论

