



DATA
SCIENCE

人人可懂的数据科学

[爱尔兰] 约翰·D. 凯莱赫 (John D. Kelleher)
布伦丹·蒂尔尼 (Brendan Tierney) / 著

张世武 黄元勋 / 译

·浅出介绍数据科学基本要素
·基础直观认识数据科学系统



机械工业出版社
China Machine Press

Data Science

人人可懂的数据科学

[爱尔兰] 约翰·D. 凯莱赫 (John D. Kelleher)

布伦丹·蒂尔尼 (Brendan Tierney) 著

张世武 黄元勋 译



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

人人易懂的数据科学 / (爱尔兰) 约翰·D. 凯莱赫 (John D. Kelleher), (爱尔兰) 布伦丹·蒂尔尼 (Brendan Tierney) 著; 张世武, 黄元勋译. —北京: 机械工业出版社, 2019.9
书名原文: Data Science

ISBN 978-7-111-63726-4

I. 人… II. ① 约… ② 布… ③ 张… ④ 黄… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字 (2019) 第 207396 号

本书版权登记号: 图字 01-2018-4600

John D. Kelleher, Brendan Tierney: Data Science(ISBN 978-0-262-53543-4).

Original English language edition copyright © 2018 by Massachusetts Institute of Technology.

Simplified Chinese Translation Copyright © 2019 by China Machine Press.

Simplified Chinese translation rights arranged with MIT Press through Bardon-Chinese Media Agency.

No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system, without permission, in writing, from the publisher.

All rights reserved.

本书中文简体字版由 MIT Press 通过 Bardon-Chinese Media Agency 授权机械工业出版社在中华人民共和国境内 (不包括香港、澳门特别行政区及台湾地区) 独家出版发行。未经出版者书面许可, 不得以任何方式抄袭、复制或节录本书中的任何部分。

本书简要介绍了数据科学领域的发展、基础知识, 并阐释了数据科学项目的各个阶段。本书考虑了数据基础架构和集成多个数据源数据所面临的挑战, 介绍了机器学习基础, 探讨了如何应用机器学习专业技术解决现实问题。本书还综述了道德和法律问题、数据法规的发展以及保护隐私的计算方法。最后探讨了数据科学的未来影响, 并给出了数据科学项目成功的原则。

人人易懂的数据科学

出版发行: 机械工业出版社 (北京市西城区百万庄大街 22 号 邮政编码: 100037)

责任编辑: 赵 静

责任校对: 李秋荣

印 刷: 中国电影出版社印刷厂

版 次: 2019 年 10 月第 1 版第 1 次印刷

开 本: 147mm×210mm 1/32

印 张: 7

书 号: ISBN 978-7-111-63726-4

定 价: 59.00 元

客服电话: (010) 88361066 88379833 69326294

华章网站: www.hzbook.com

投稿热线: (010) 88379604

读者信箱: hzit@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版 本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

丨 内 容 简 介 丨

本书从数据科学发展演化史，数据科学定义，数据、数据集，数据科学生态系统，机器学习，数据科学标准任务，隐私与道德，发展趋势等角度，对数据科学展开了精彩的阐述。

本书精准界定了数据科学的术语、任务、生命周期，并介绍了主流的数据科学生态技术，以及决策树、回归分析、神经网络、深度学习等常见机器学习算法。同时也涵盖隐私、数据道德等方面的话题，介绍了数据科学可能引发的隐私泄露、人为歧视、不公平，以及欧盟、美国等组织或国家针对数据因素保护、数据道德的立法。

| 作者简介 |

约翰·D.凯莱赫(John D. Kelleher) 是都柏林理工学院计算机科学学院的教授以及信息、通信和娱乐研究所的学术负责人。他的研究得到了ADAPT中心的支持，该中心由爱尔兰科学基金会 (Grant 13 / RC / 2106) 资助，同时也接受欧洲区域发展基金的资助。他还是《Fundamentals of Machine Learning for Predictive Data Analytics》的作者之一。

布伦丹·蒂尔尼(Brendan Tierney) 是都柏林理工学院计算机科学学院的讲师，同时也是Oracle ACE 主任，还著有多本基于Oracle技术的数据挖掘类著作。

.. 译 者 序 ..

数据科学这个术语的出现可以追溯到 20 世纪 90 年代。事实上，该领域的历史更悠久。在古代人类就已经有收集数据和分析数据的传统，这些可视为数据科学的雏形。数据科学的目标是从大数据集中提取洞察力并基于它改进决策。数据科学与数据挖掘、机器学习紧密相关，但它的研究范围更广，人们在日常交流中通常会混淆这些概念。进入信息时代之后，数据科学逐步迈入应用阶段，但是真正被大众熟知则是在大数据时代。目前，大数据正在急剧改变着人们的工作、生活与思维模式，同时也对数据科学的学术研究及应用产生了深远影响。大数据技术日新月异的发展、可用数据的激增及计算能力的提升，为数据科学实践提供了肥沃的土壤，数据科学项目在各种规模的组织机构中如雨后春笋般涌现。

本书从数据科学发展演化史，数据科学定义，数据、数据集，数据科学生态系统，机器学习，数据科学标准任务，隐私与道德，发展趋势等角度，对数据科学展开了精彩的阐述。书

中精准界定了数据科学的术语、任务、生命周期，介绍了主流的数据科学生态技术，以及决策树（Decision Tree）、回归分析（Regression Analysis）、神经网络（Neural Network）、深度学习（Deep Learning）等常见机器学习算法。同时也涵盖了隐私、道德等方面的话题，介绍了数据科学可能引发的隐私泄露、人为歧视、不公平，以及美国、欧盟等国家或组织针对数据因素保护、数据道德的立法。本书深入浅出、案例丰富，适合各种类型的读者阅读。对数据科学感兴趣的非专业人士阅读本书正文内容即可获得对数据科学系统的、直观的认识；专业人士还可以阅读本书“延伸阅读”“参考文献”部分列举的文献。本书也对一些重要概念以及容易混淆的内容提供了注解，以帮助读者准确无误地掌握本书内容。

本书经过精心组织，结合了译者多年的数据科学研究、实践经验，并参考了微软、阿里巴巴、腾讯、百度等众多知名企业的业界专业人士的意见。本书翻译团队由拥有丰富经验的数据科学从业者组成。其中，张世武负责第1、2、4、5章的翻译和校对以及全书统稿，黄元勋负责第3、6、7章的翻译和校对。在本书翻译过程中，译者经过多次讨论、审校，力求信达雅。由于本书涉及很多新概念，业界尚无统一术语，另外由于译者水平有限，难免会出现一些问题，欢迎广大读者及业内同行批评指正。

最后，感谢家人的支持与宽容，为我们顺利交稿创造了宽松的环境。

.. 前 言 ..

数据科学的目标是通过将决策建立在从大数据集中提取洞察力的基础上来改进决策。作为一个活动领域，数据科学由一系列原理、问题定义、算法和过程组成，用于从大型数据集中提取有用但不显眼的模式。数据科学与数据挖掘和机器学习领域密切相关，但它涉及的范围更广。如今，数据科学推动了现代社会几乎所有领域决策的发展。数据科学可能影响人们日常生活的方方面面，如确定在线广告的呈现，向用户推荐电影、书籍、朋友，过滤垃圾邮件，用户续订手机合约时向他们提供合适的优惠套餐，降低医保成本，规划辖区内交通信号灯的布局及通行时间，药品设计，警力部署规划等。

大数据和社交媒体的出现、计算能力的加速、计算机内存成本的大幅降低以及更强大的数据分析和建模方法的发展推动了当代社会对数据科学需求的增长，其中典型的技术有深度学习。这些因素共同作用意味着组织收集、存储和处理数据将比以前简单。与此同时，这些技术创新和数据科学的广泛应用意

味着与数据使用和个人隐私相关的道德挑战从未如此迫切。本书的目的是提供数据科学的介绍，涵盖该领域的基本要素，并提供对该领域深刻的原则性见解。

本书第 1 章介绍了数据科学领域，简要回顾了数据科学的发展演化历史，还探讨了如今数据科学为什么那么重要，以及推动采用数据科学的一些因素。在这一章的最后，回顾并揭穿了与数据科学相关的一些神话。第 2 章介绍了与数据相关的基本概念，描述了数据科学项目的标准流程：业务理解、数据理解、数据准备、建模、评估和部署。第 3 章重点介绍了数据基础设施以及大数据和多源数据集成带来的挑战。数据基础设施的一个可能具有挑战性的典型方面是，数据库和数据仓库中的数据通常驻留在与用于数据分析的服务器不同的服务器上。因此，当处理大型数据集时，可能要花费大量时间在数据库或数据仓库所依赖的服务器与进行数据分析和机器学习处理的服务器之间移动数据。第 3 章首先描述组织中典型的数据科学基础设施，以及在数据科学基础设施中移动大型数据集的挑战的一些新兴解决方案，其中包括使用数据库内置机器学习算法，使用 **Hadoop** 进行数据存储和处理，以及混合数据库系统的开发，这些系统无缝地结合了传统的数据库软件和类似 **Hadoop** 的解决方案。这一章的最后强调了将整个组织的数据整合到适合机器学习的统一表示中的一些挑战。第 4 章介绍了机器学习领域，并解释了一些最流行的机器学习算法和模型，包括神经网络、

深度学习和决策树模型。第 5 章聚焦于通过审视一系列标准业务问题，描述了机器学习解决方案如何解决这些问题来将机器学习专业知识与现实问题联系起来。第 6 章回顾了数据科学的道德含义、数据监管的最新发展，以及在数据科学过程中保护个人隐私的一些新的计算方法。最后，第 7 章描述了数据科学在不久的将来会产生重大影响的一些领域，并列出了确定数据科学项目是否会成功的一些重要原则。

.. 致 谢 ..

John 和 **Brendan** 感谢 **Paul McElroy** 和 **Brian Leahy** 对早期草稿的阅读和评论。还要感谢两位匿名审稿人，感谢他们针对稿件给麻省理工学院出版社的工作人员提供了详细且有益的反馈意见，对本书的出版提供了强有力的支持和指导。

John 感谢他的家人和朋友在本书的准备过程中给予的支持和鼓励，并将这本书献给他的父亲——**John Bernard Kelleher**，以表示对他的爱 and 情谊。

Brendan 感谢 **Grace**、**Daniel** 及 **Eleanor** 在他撰写另一本著作（他的第 4 本著作）时，对他长期出差的容忍以及对日常工作的支持。

.. 作者简介 ..

约翰·D. 凯莱赫 (John D. Kelleher) 是都柏林理工学院计算机科学学院的教授以及信息、通信和娱乐研究所的学术负责人。他的研究得到了 ADAPT 中心的支持，该中心由爱尔兰科学基金会 (Grant 13 / RC / 2106) 资助，同时也接受欧洲区域发展基金的资助。他还是《Fundamentals of Machine Learning for Predictive Data Analytics》(麻省理工学院出版社) 的作者之一。

布伦丹·蒂尔尼 (Brendan Tierney) 是都柏林理工学院计算机科学学院的讲师，同时也是 Oracle ACE 主任，还著有多本基于 Oracle 技术的数据挖掘类著作。

.. 目 录 ..

译者序

前言

致谢

作者简介

第 1 章 什么是数据科学 1

1.1 数据科学简史 5

1.1.1 数据收集简史 5

1.1.2 数据分析简史 9

1.1.3 数据科学的产生与发展 14

1.2 数据科学用于何处 20

1.2.1 销售和营销中的数据科学 21

1.2.2 数据科学在政府中的应用 22

1.2.3 数据科学在竞技体育中的应用 23

1.3 为什么是现在 25

1.4 关于数据科学的神话 28

第 2 章 什么是数据，什么是数据集 31

- 2.1 关于数据的观点 38
- 2.2 数据可以积累，而智慧不能 43
- 2.3 CRISP-DM 45

第 3 章 数据科学生态系统 54

- 3.1 将算法迁移至数据 61
 - 3.1.1 传统数据库与现代的传统数据库 64
 - 3.1.2 大数据架构 67
 - 3.1.3 混合数据库世界 69
- 3.2 数据准备和集成 72

第 4 章 机器学习 77

- 4.1 有监督学习与无监督学习 78
- 4.2 学习预测模型 83
 - 4.2.1 相关性不等同于因果，但它有时非常有用 84
 - 4.2.2 线性回归 90
 - 4.2.3 神经网络与深度学习 96
 - 4.2.4 决策树 108
- 4.3 数据科学中的偏差 114
- 4.4 评估模型：泛化而不是记忆 116
- 4.5 摘要 119

第 5 章 标准的数据科学任务 121

- 5.1 谁是我们的目标客户（聚类） 122
- 5.2 这是欺诈吗（异常值检测） 128
- 5.3 你要配份炸薯条吗（关联规则挖掘） 131

5.4 流失还是不流失，这是一个问题（分类）…… 136

5.5 它价值几何（回归）…… 141

第 6 章 隐私与道德 …… 143

6.1 商业利益与个人隐私 …… 145

6.1.1 数据科学的道德启示：画像与歧视 …… 148

6.1.2 数据科学的道德含义：创建一个全景监狱 …… 154

6.2 隐私保护 …… 157

6.2.1 保护隐私的计算方法 …… 159

6.2.2 规范数据使用和保护隐私的法律框架 …… 161

6.3 通往道德的数据科学之路 …… 164

第 7 章 未来趋势与成功准则 …… 172

7.1 医疗数据科学 …… 172

7.2 智慧城市 …… 174

7.3 数据科学项目准则：为什么会成功或失败 …… 177

7.4 终极思考 …… 185

术语表 …… 188

延伸阅读 …… 201

参考文献 …… 203

第 1 章

什么是数据科学

数据科学 (data science) 由一系列原理、问题定义、算法及数据处理过程组成, 用于从大数据集中抽取不显眼但又非常有用的模式 (pattern)。数据科学中的很多元素由诸如机器学习 (machine learning)、数据挖掘 (data mining) 等相关学科中的元素进化而来。事实上, 很多场景中, 数据科学、机器学习和数据挖掘这些概念可以互换。它们有很多共同点, 主要集中在如何利用数据分析来改进决策。尽管数据科学从相关学科中借鉴了不少东西, 但是它的研究范围更广阔。机器学习更关注算法的设计与演化 (用于从数据中抽取有用的模式)。数据挖掘通常用来处理和分析结构化数据, 更侧重于商业应用。数据科学涵盖了机器学习与数据挖掘的所有内容, 此外还会应对很多其他的挑战, 如非结构化的社交媒体、Web 数据的捕获、清洗、转换, 利用大数据技术存储和处理海量非结构化数据集, 数据