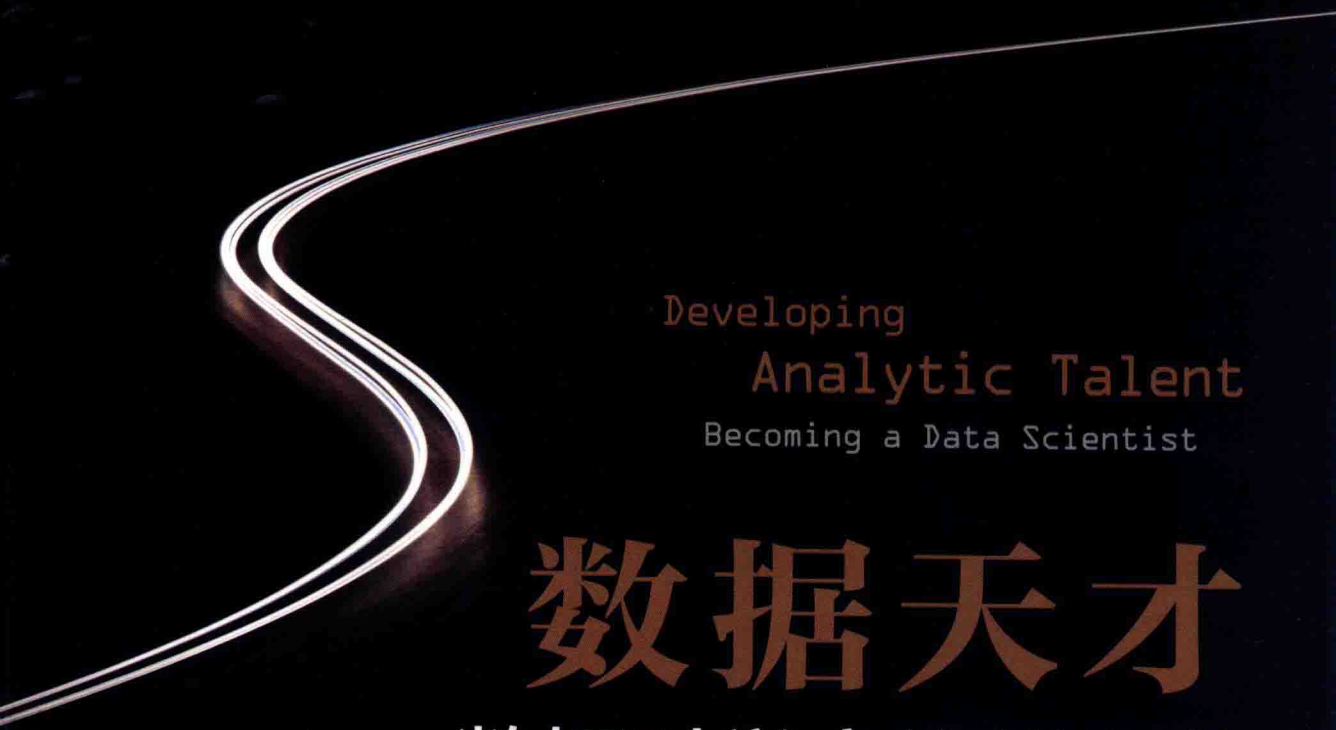
Broadview
www.broadview.com.cn

WILEY



Developing
Analytic Talent
Becoming a Data Scientist

数据天才

数据科学家修炼之道

【美】Vincent Granville 著
吴博 张晓峰 季春霖 译

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Developing Analytic Talent

Becoming a Data Scientist

数据天才

数据科学家修炼之道

【美】Vincent Granville 著

吴博 张晓峰 季春霖 译

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

这是一本跟数据科学和数据科学家有关的“手册”，它还包含传统统计学、编程或计算机科学教科书中所没有的信息。

本书有3个组成部分：一是多层次地讨论数据科学是什么，以及数据科学涉及哪些其他学科；二是数据科学的技术应用层面，包括教程和案例研究；三是给正在从业和有抱负的数据科学家介绍一些职业资源。本书中有很多职业和培训相关资源（如数据集、网络爬虫源代码、数据视频和如何编写API），所以借助本书，你现在就可以开始数据科学实践，并快速地提升你的职业水平。

本书是写给数据科学家和相关专业人士的（如业务分析师、计算机科学家、软件工程师、数据工程师和统计学家），也适合有兴趣转投大数据科学事业的人阅读。

图书在版编目（CIP）数据

数据天才：数据科学家修炼之道 /（美）文森特·格兰维尔（Vincent Granville）著；吴博，张晓峰，季春霖译. —北京：电子工业出版社，2017.5
书名原文：Developing Analytic Talent: Becoming a Data Scientist
ISBN 978-7-121-30883-3
Ⅰ. ①数… Ⅱ. ①文… ②吴… ③张… ④季… Ⅲ.①数据管理 Ⅳ. ①TP274
中国版本图书馆 CIP 数据核字(2017)第 021480 号

责任编辑：付　睿
印　　刷：北京天宇星印刷厂
装　　订：北京天宇星印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：22.25　字数：385 千字
版　　次：2017 年 5 月第 1 版
印　　次：2017 年 5 月第 1 次印刷
定　　价：85.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

专家推荐

数据科学家是商业分析、统计学和计算机科学等领域的通才，成为数据科学家正成为年轻人的新梦想。光启高等理工研究院季春霖副院长推荐我读这本他参与翻译的书之后，我一开始期望这是一本数学味、计算机味很浓的书籍。后来，完全出乎意料，这本书行文竟然如此清晰明白，原理与观点娓娓道来，并剖析了大量实际而有趣的案例，读起来丝毫没有教科书式的刻板感。通过本书，你可以了解一个数据科学家所需的完善的知识体系，包括商业分析、数据库、统计模型、蒙特卡洛模拟、机器学习、Hadoop、MapReduce、哈希连接等。本书显然为有梦想的你在通往数据科学家的道路上铺就了阶梯，我相信你只要拾阶而上，到达目的地就是确定无疑的。

——王磊　国家统计局高级统计师
中国国际经济交流中心金融学博士后
北京大学肿瘤医院核医学科客座教授

2017 年大数据行业已经从上半场开始挺进下半场，数据在不知不觉中影响着我们的生产、生活、娱乐等方方面面。我们深耕在行业，深知目前国内从行业角度真正缺乏的是有着商业精神的数据科学家，本书从场景出发给我们展示了如何成为数据天才。我与吴博、晓峰、春霖交流很多，他们有深厚的学术素养，但仍实实在在地做着数据商业，恰恰这一点也是目前国内缺乏的，我一直认为在中国不缺数据技术人才，但缺乏的是真正懂商业的数据天才、数据科学家。希望大家能从本书中汲取知识，真正走向数据科学的商业之路。

——汪祥斌　DataEye 创始人、CEO

数据科学家是“21 世纪最性感的职位”，全球到 2018 年对数据科学家有上千万的职位空缺，仅中国就稀缺上百万这样的人才。这本《数据天才：数据科学家修炼之道》是成为数据科学家的必备宝典。书中对数据科学有着翔实的介绍，并针对数据科学家日常工作中所需的技能进行了深度的剖析，辅以大量的实用案例分析，有助于快速提升大家对数据科学的理解和应用。本书势必会成为继维克托·迈尔-舍恩伯格的《大数据时代》后的又一经典大作！

——刘金玲　中国大数据产业第一媒体“36 大数据”创始人

大数据是近年来媒体的热点话题，大数据时代在科学领域里的表现就是数据科学的兴起。那么人们不禁会问：什么是数据科学以及如何成为数据科学家？作者通过本书及时地为读者用一种全景式的方式给出了答案。本书以通俗易懂的语言风格和众多的真实案例，讲活了大数据与数据科学，全面而又深入浅出地阐明了数据科学的实质与内涵，揭示了数据科学家的修炼秘笈。相信不同读者一定都能从书中得到启发，了解价值，找到灵感，更好地以全新的视角审视自己的专业领域以及汲取更多的新理念、新思想。

——谌东宇　教授　深圳云数通科技有限公司总裁
前海云游数据运营（深圳）有限公司首席数据官
西南交通大学数学学院客座教授

人生的关键决策只有几个，择业就是其中之一。良好的职业决定和素质准备来自于对未来的场景有清晰而且正确的认知。吴博的这本译著，不仅能够帮助我们认识未来几十年社会、商业和技术场景中的数据行业，数据科学家的是和不是，更重要的是提供了修炼自己的宝鉴。本书横跨中美视野、结合生活事件的描述，使得我们带着轻松、开心的心情完成对数据科学的认知、体悟，让人有一种跃跃欲试和大展宏图的感觉。实在是 4.0 时代必备的一本书！

——郑立新　德摩资本董事长
2017 年 3 月 14 日于深圳

献给我挚爱的妻子 Paris、可爱的女儿 Capri 和儿子 Zinal，感谢他们长久以来的支持。献给我怀念的父亲 Roger，感谢他在我孩提时代引领我接触到数学。

译者序

本书最适合有志于在大数据与数据科学领域从业的人学习。格拉德威尔在《异类》一书中强调，“若要成为行业专家，离不开十万小时的刻意学习（deliberate learning）”，这跟中国俗语里“板凳要坐十年冷”有些类似。但要实现刻意学习，就不能一味依赖通识科普书籍。在大数据与数据科学领域，市面上已不缺通识性的科普书籍，唯缺这类烧脑、有专业性、适合进行刻意学习的数据科学书籍。

本书不失专业性，但也不是令人生畏的大学教材。它处处体现理论与实践的结合，还兼顾技术与商业的平衡。这要归功于原作者 Vincent 是学术、技术、商业三栖高手。比如书中对于星空双星的估算、陨石撞地球的建模推算，让作者在数学奥赛方面的天分展现得淋漓尽致；在垃圾邮件、水印加密、点击欺诈等案例中，作者又分享了诸多为大公司实施数据项目的经验；在方案选择、股市预测等场景中，作者更侧重商业视角，帮读者提升对数据科学方法投入/产出比及适用性的敏感度。

本书虽然专业度高，但也因为案例翔实、讲求实际，适合其他行业或领域的人士阅读。特别建议业务跟数据息息相关的企业负责人或高管，或者对数据相关项目感兴趣的投资者品读。毕竟数据科学家这一高层职位，跟企业负责人及高管的对接较多。虽说好的数据科学家，应具备与非技术人士沟通的能力，但作为数据科学家的领导，一旦多懂一些数据科学的思考模式及流程，便会对数据科学家有更多理解，

也会对数据化的决策有更深的认识。

本书也传递出对行业热词的审慎态度。比如本书就对“大数据”的缘起、演变、更替、历史、迷思和幻象，着墨不少。就像书中所说，大数据领域许多看似新的方法，可以追溯到二三十年前，如今的不少创新，实乃新瓶旧酒。想必读者从 Gartner 的成熟度曲线里，可以看到大数据一词已渡过巅峰、渐趋理性，与之相随的，是跟数据科学息息相关的人工智能（AI）重新崛起。若理解本书的立场和价值取向，就知道人工智能 60 多年来几起几落，不少如今大放异彩的方法，也可找到前身。透过现象看本质，人工智能多少因为数据体量更大、数据分析更细、计算能力更强，才成为行业焦点。忽视基础理论盲目追随人工智能热点无异于舍本逐末，认真和刻意学习数据科学及人工智能的基础理论和实践，方是正途。

正因为这本书内容如此之好，能满足读者所需，于是我痛快答应电子工业出版社付睿编辑的邀约来翻译本书。但这个小想法变成最终成品，却耗费不少人的时间和精力，对他们的感谢和亏欠不能尽录。我最要鸣谢翻译合作者光启研究院的副院长季春霖博士，还有在哈工大深圳研究生院任教的张晓峰博士，两位的研究和管理任务都很繁重，面对译书这种流程漫长、成效滞后的工作，他们展现了学界出身的坚韧素养，而在翻译校对本书的过程中，又处处体现出手不凡的专业功力。同时，也要感谢配合翻译校对本书的助手和出版社工作人员，他们对我有莫大的包容和支持。本书准备期间，也正是我的一对小孩——泰学和雅学——从孕育到出生的过程，所以要感谢我的太太熊瑛，容许我为本书挤出不少本来可以陪伴家人的时间。

最后，我还要代表季春霖博士感谢广东省自然科学杰出青年基金项目（No.S20120011253）和深圳市数据科学与建模技术重点实验室的资助。也要感谢我所在的宜远智能团队，他们在将本书中许多数据科学方法实践到医疗健康领域时，提出了诸多宝贵的翻译修正补充建议。当然，对专业内容的翻译，难在对作者见识的理解和原意的把握，所以总有力有不逮、不甚精确之处，请各位读者和专家对此海涵，提出宝贵的建议。

本书译者　吴博

关于作者

Vincent Granville 博士是一名富有远见的数据科学家，有 15 年大数据、预测建模、数字分析和业务分析的经验。Vincent 在评分技术、欺诈检测和网络流量优化及增长等领域，是举世公认的权威专家。在过去的 10 年中，他曾与 Visa 一起研究实时信用卡欺诈检测，与 CNET 一起研究广告组合优化，与 Microsoft（微软公司）一起研究“改变点检测”，与 Wells Fargo（富国银行）一起研究在线用户体验，与 InfoSpace 一起研究搜索智能，与 eBay 一起研究自动竞价，与各大搜索引擎、广告网络和大型广告客户一起研究点击欺诈检测。Vincent 也管理着 LinkedIn 上最大的“大数据及分析数据科学家”小组，该小组拥有超过 100 000 名成员。

最近，Vincent 推出了数据科学中心（Data Science Center）这个大数据、业务分析和数据科学界的领先社区。Vincent 曾是剑桥大学和美国国家统计科学学院的博士后。他曾入围沃顿商业计划竞赛和比利时数学奥林匹克的决赛。Vincent 已经在统计期刊上发表了 40 篇论文，并且是许多国际会议的受邀演讲嘉宾。他还开发了一种新的数据挖掘技术，被称为隐性决策树，他还拥有多项专利，是发表数据科学书籍的第一人，并筹集了 600 万美元的创业启动资金。根据福布斯的排名，Vincent 是大数据领域前 20 位有影响力的人物之一，被 VentureBeat、MarketWatch 和美国有线新闻网（CNN）专门报道。Vincent 的 Twitter 账号为`@Analyticbridge`。

关于技术编辑

Joni Ngai 是一位数字化传播者，在财富 500 强企业任职高管，负责数字视觉开发，以及利用技术和数据来吸引大量当今互联世界里的客户。她拥有带领机构和客户在数字业务、客户关系管理（CRM）、网络媒体、分析和技术发展领域进行新尝试的丰富经验。从 2000 年开始，Joni 在纽约的 Razorfish 公司开始了她的数据咨询职业生涯。此后，她在许多顶尖的数字化机构里工作过，如 MRM Worldwide 和 Havas Digital。还有亚太地区的一些全球品牌，如 Intel（英特尔公司）、Microsoft（微软公司）和 P&G（宝洁公司）。她曾是 I-COM China 的副主席，这是一个由工业界所支持的全球数字度量论坛，该论坛通过促进在线指标度量标准化的发展，帮助相关行业成长。

Joni 毕业于滑铁卢大学，主修电气工程，并辅修管理学。她还获得了西北大学凯洛格管理学院和香港科技大学的行政工商管理硕士学位。她也在香港中文大学传授新媒体理科硕士课程。

前言

这是一本跟数据科学和数据科学家有关的“手册”，它还包含传统统计学、编程或计算机科学教科书中所没有的信息。凭借作者在数据科学领域 20 多年的领导者地位，他在本书中收集了他认为对从事数据科学职业最重要的一些信息。在过去 3 年里，本书中的很多内容首先被发表在 Data Science Central 官网上，被数百万的网站用户所阅读。本书介绍了数据科学与其他相关领域的差异，以及使用大数据能给组织带来的价值。

本书有 3 个组成部分：一是多层次地讨论数据科学是什么，以及数据科学涉及哪些其他学科；二是数据科学的技术应用层面，包括教程和案例研究；三是给正在从业和有抱负的数据科学家介绍一些职业资源。本书中有很多职业和培训相关资源（如数据集、网络爬虫源代码、数据视频和如何编写 API），所以借助本书，你现在就可以开始数据科学实践，并快速地提升你的职业水平。如果你是一位决策者，你会在本书中找到一些信息，来帮助你建立更好的分析团队，以及决定是否需要及何时需要专业的解决方案，以及哪些方案最为恰当。

这本书是写给谁的

这本书是写给数据科学家和相关专业人士的（如业务分析师、计算机科学家、软件工程师、数据工程师和统计学家），以及有兴趣转投大数据科学事业的人。本书也是为学习定量课程、想成为数据科学家的大学生所准备的。最后，本书也可供数据科学家的上级领导、想创建数据科学初创公司开展业务或提供数据科学咨询的人阅读。

这些读者将在本书中找到有价值的信息，特别是在以下几章中。

- 第 2、4、5、6 章对**数据科学工作者**特别有价值，因为它们包含大数据技术内容（如聚类和分类技术），以及前沿数据科学技术，如组合特征选择、隐性决策树、分析类 API、判断 MapReduce 何时有用等。这些章节里很多案例研究（如欺诈检测、数字分析、股票市场策略和其他更多）的说明非常详细，详细到可以让读者在实际工作中面临类似数据时，能沿用这些案例的分析方法。然而，它们的文字描述都很简单，高层管理人员不用花太多时间在细节、代码或公式上，也能阅读下来。
- 修读计算机科学、数据科学或工商管理硕士课程的**学生**，会在第 2、4、5、6 章中找到对他们有用的信息。特别是在第 2、4、5 章，他们能从中找到进阶内容，如实际的数据科学方法和原则，这些在一般的教科书或典型的大学课程里都没有。第 6 章还介绍了现实生活应用和案例研究，并包含更深入的技术细节。
- **求职者**将会在第 3 章中找到有关数据科学的培训和课程资源。第 7、8 章为求职者提供了大量的资源，包括面试问题、简历模板、招聘广告样板，经常招聘数据科学家的公司的清单，以及薪资调查等。
- 对于想要创建一个数据科学创业公司或顾问公司的**企业家**，在第 3 章中会找到商业计划书样板、创业公司点子和针对顾问职位的薪酬调查。同时，在本书中，数据顾问会了解如何提高数据科学工作沟通效率，掌握数据科学项目的生命周期，并得到相关书籍、会议参考和许多其他资源。
- 对于试图评估数据科学的价值和它们对企业项目的益处，以及评估

MapReduce 架构何时有用的**高管**们，会在第 1、2、6（案例部分）、8 章（招聘广告样板、简历、薪金调查）中找到有价值的信息。这些章节的重点通常不是技术。顶多会在第 2 章和第 6 章介绍一些新的分析技术。

这本书涵盖了什么

本书的技术部分包括数据科学的核心内容，比如：

- 将大数据和传统的算法应用到大数据时的挑战（例如在进行大数据聚类或分类时的解决方案）。
- 一种统计科学上新颖、简化、对数据科学友好的方法，重点在于它是一种健壮的无模型方法。
- 顶尖的机器学习方法（隐性决策树和组合特征选择）。
- 新型数据的新指标（综合指标、预测能力、波动系数）。
- 创建快速算法所需的计算机科学要素。
- MapReduce 和 Hadoop，以及 Hadoop 进行计算时的数值稳定性。

重点还是最新的技术。在本书中你不会找到关于旧技术的资料介绍，如线性回归（除非在引文里涉及），因为这些在经典书籍里已经讨论了很多。在本书中，对逻辑回归类的知识讨论不多。我们只是将逻辑回归与其他分类器混合，提出一种数值稳定的近似算法（近似的解决方案往往和精确模型一样有效，毕竟没有任何数据完全符合理论模型）。

除了技术，本书还提供了有用的工作资源，包括工作面试的相关问题、简历模板和招聘广告样板。本书的另一个重要组成部分是案例研究。本书的案例研究，有些带有统计或机器学习的意味，有些则跟商业或决策科学或运筹学有关，有些则关乎数据工程。大多数时候，我喜欢 Data Science Central（这是个数据科学家的领先社区）上最新发表和非常热门的主题，而不是我特别重视的话题。

本书是如何架构的

本书由三大主题构成。

- 数据科学和大数据是什么和不是什么，以及与其他学科的区别（第 1、2、3 章）。
- 职业和培训资源（第 3 章和第 8 章）。
- 用作教程的技术材料（第 4 章和第 5 章，以及第 2 章中关于大规模数据集聚类和分类的内容，第 8 章中关于 Hadoop 的新变化和大数据的内容），以及案例研究（第 6 章和第 7 章）。

本书为潜在的和现有的数据科学家和相关专业人员（以及他们的管理者和老板）提供了宝贵的职业资源。宽泛而言，本书适用于所有处理更大、更复杂、更新、频率更快的数据的专业人士。本书还提供一些数据科学的秘诀、技巧、概念（其中许多是原创和首次公开的）、带实施方法和技术的案例研究，以及已经在不同领域，不论是手动还是自动，能成功分析现代数据的技术。

阅读本书你需要什么知识

这本书包含了少量的 R 或 Perl 示例代码。你可以在 `http://www.activestate.com/activeperl/downloads` 下载 Perl，在 `http://cran.r-project.org/bin/windows/base/` 下载 R。如果你使用 Windows 计算机，首先需要安装一个 Linux 式环境：Cygwin。你可以在 `http://cygwin.com/install.html` 上下载 Cygwin 软件。Python 也是开源的，且有一个有用的、被称为 Pandas 的库。

如果你有一两年大学基本定量课程的知识基础，就足以理解书中大多数内容。本书不需要微积分或高等数学的相关知识——事实上，它几乎不包含任何数学公式或符号。

然而，本书也包含一些高度概括性的进阶材料。本书中的一些技术讲义，是针对那些对数学更有倾向和有兴趣深入挖掘的读者。有两年大学微积分、统计学和矩阵理论知识的读者，将能更好地理解这些技术细节。本书提供了一些源代码（R、Perl）和数据集，但本书的重点不是编码。

本书通过多种技术水平混合的介绍方式，让你不用具备高级数学知识，也有机会深度探索数据科学（这有点像 Carl Sagan 向主流公众介绍天文学的方式）。

惯例标记

为了帮助你从本书中学到最多的东西，而不是一头雾水，我们将在本书中使用惯例标记。

注意 本书中的注意、提示、交叉参考，以及对当前讨论的辅助说明，将像这个注意的方式显示。

至于文本的样式标记如下。

- 当我们介绍术语和重要的词时，我们会用楷体突出它们。
- 快捷键用这种方式表示：Ctrl+A。
- 我们在书中显示文件名、链接和代码的格式如下。

  `persistence.properties`
- 我们介绍代码的格式如下。

  对于大多数代码， 我们使用 `Courier New` 字体，不加粗。

致谢

我要感谢来自 Wiley 的 Chris Haviland 和 Carol Long，他们对本书的出版有很大的贡献，承担了不少风险，他们把我很多有价值、分散未经组织的在线文章，整合成一本连贯、全面和有用的书。从许多方面来看，这个复杂的过程类似于将非结构化数据转化为结构化数据，这是许多数据科学家经常面对的常规挑战，而这本书也

正好提供了将非结构化数据转化为结构化数据的解决方案。同时，我要感谢我的商业伙伴和共同创始人 Tim Matteson，他帮助 Data Science Central 这个网站成为数据科学社区的领导者，还变成了一个现代的、专注于产生价值的创业项目。最后，我要感谢我们社区的所有成员，感谢他们的评论和支持。如果没有他们的帮忙，本书也无法出版。

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/30883

目录