帮你构筑数据科学的逻辑思维
帮你找到数据科学的入门钥匙

Broadview
www.broadview.com.cn

Data Scientist

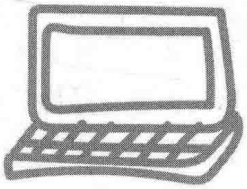
数据科学家养成手册

高 扬◎编著

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

数据科学家养成手册

高 扬◎编著



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

内 容 简 介

作为认知科学的延伸，数据科学一方面应该越来越引起广大大数据工作者的重视，另一方面也要撩开自己的神秘面纱，以最为亲民的姿态和每位大数据工作者成为亲密无间的战友，为用科学的思维方式进行工作做好理论准备。本书从众多先贤及科学家的轶事讲起，以逐步归纳和递进的脉络总结出科学及数据科学应关注的要点，然后在生产的各个环节中对这些要点逐一进行讨论与落实，从更高、更广的视角回看科学及数据科学在各个生产环节的缩影。本书并不以高深的数学理论研究作为目的，也不以某一种计算机语言编程作为主线脉络，而是在一个个看似孤立的故事与工程中不断拾遗，并试着从中悟出一些道理。

本书适合大数据从业人员和对大数据相关知识感兴趣的人，初级和中级程序员、架构师及希望通过对数据的感知改进工作的人，产品经理、运营经理、数据分析师、数据库开发工程师等对数据分析工作敏感的人，以及所有对数据科学感兴趣并希望逐步深入了解数据科学知识体系的人阅读。

图书在版编目（CIP）数据

数据科学家养成手册 / 高扬编著. —北京：电子工业出版社，2017.5

ISBN 978-7-121-31304-2

Ⅰ. ①数… Ⅱ. ①高… Ⅲ. ①数据管理—手册 Ⅳ. ①TP274-62

中国版本图书馆 CIP 数据核字（2017）第 071293 号

策划编辑：潘 昕

责任编辑：潘 昕

印 刷：三河市良远印务有限公司

装 订：三河市良远印务有限公司

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：23 彩插：2 字数：500 千字

版 次：2017 年 5 月第 1 版

印 次：2017 年 5 月第 1 次印刷

定 价：79.00 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。

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。

图1-3

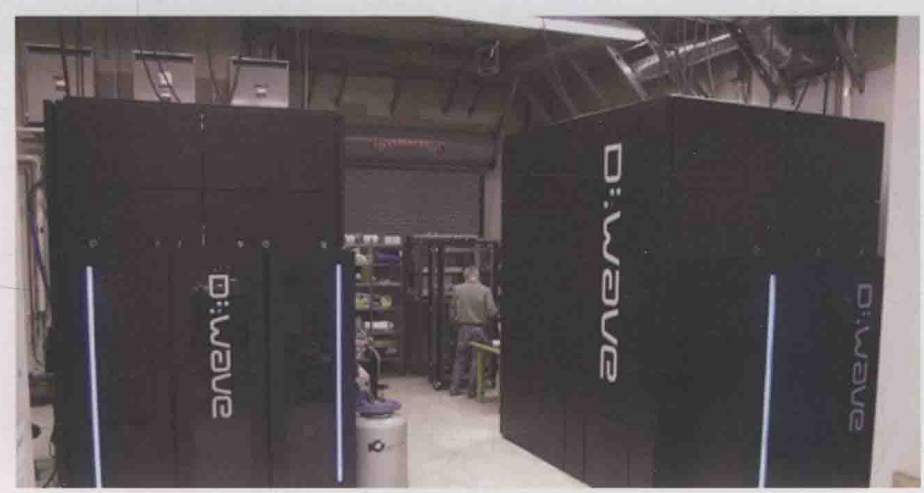


图4-12

图7-7

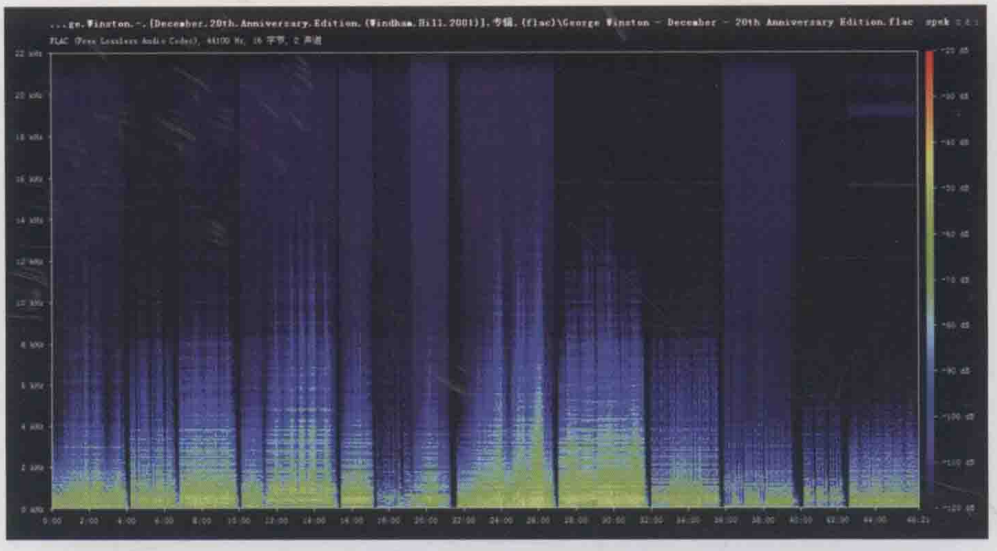


图9-10

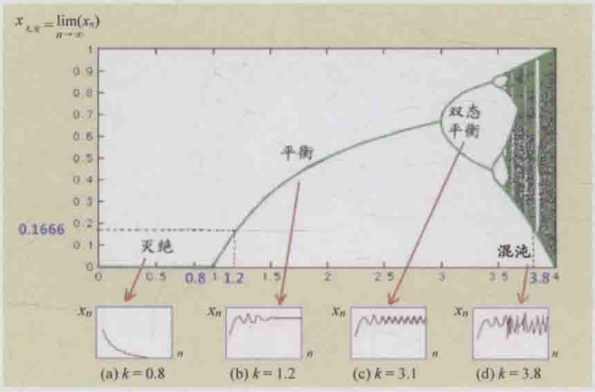


图10-4

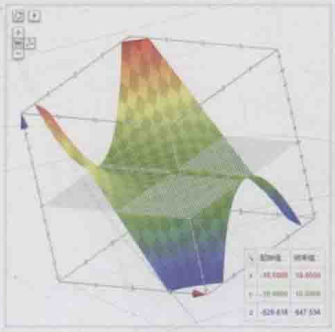


图10-5

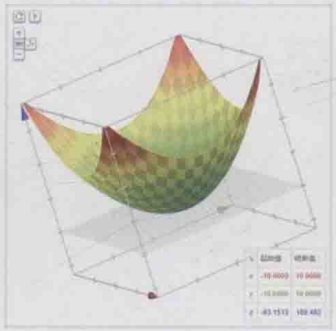


图11-18

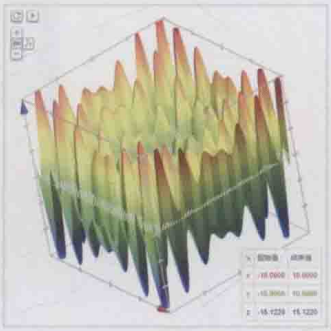


图11-20

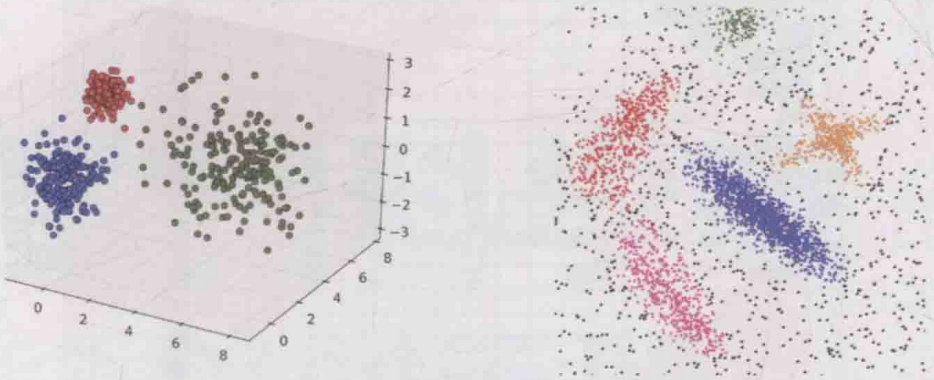


图11-22

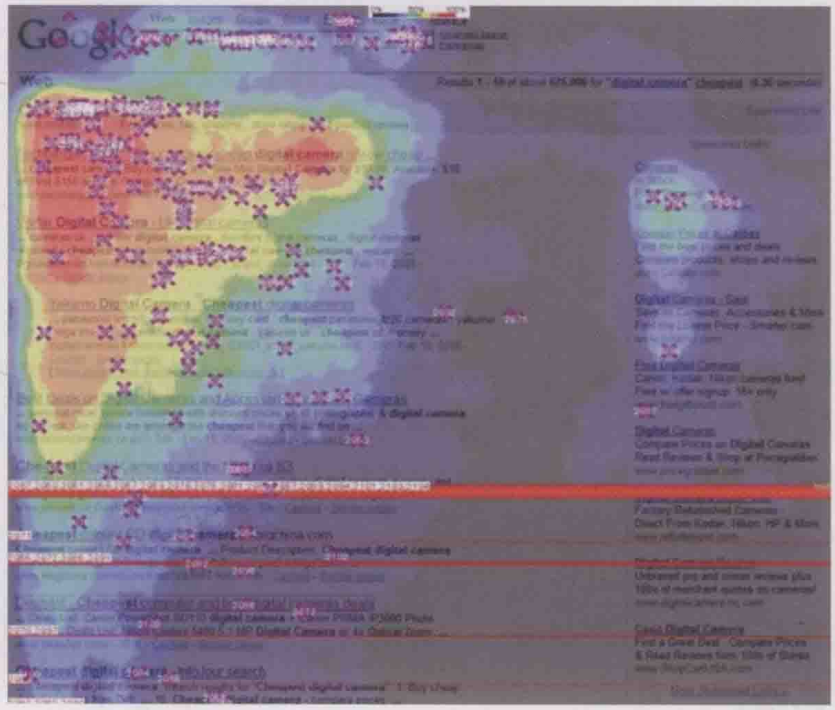


图16-6

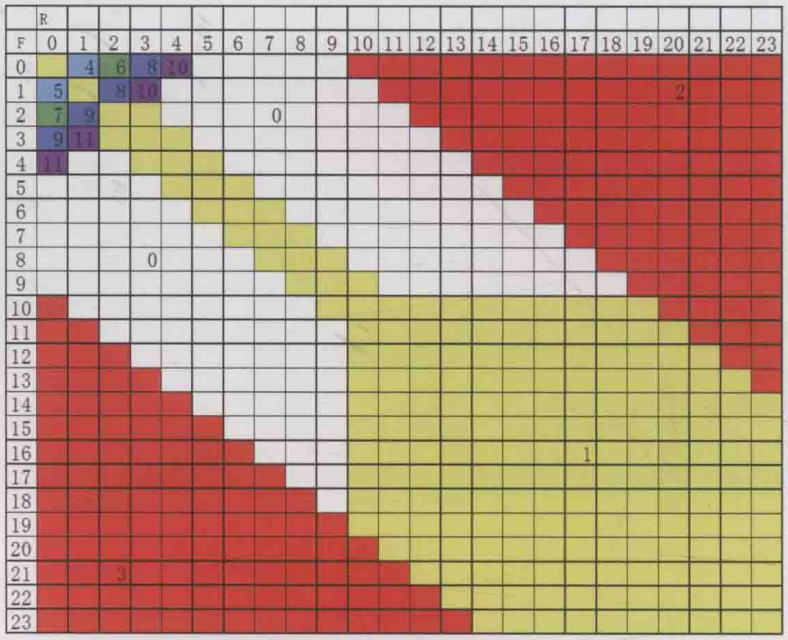


图18-21

序

十几二十年前，读书是学习新技术的不二法门。当时如果要学习一门技术，都需要买上几本“砖头书”，一边阅读，一边动手，一页一页“啃”下来。很多在今天叱咤风云的高手，当年都是用这种方式打下基础的。

最近几年，技术学习的方式发生了深刻的变化，大量的在线视频课程、交互式学习环境、开箱即用的工具箱，使技术学习的效率大幅度提升。特别是在动手能力方面，培训效率有了质的飞跃。最近一年，受人工智能领域突破性进展的鼓舞，机器学习和数据科学成为技术圈中的显学，而在线学习成为主流的学习方式。在这种情况下，大批学习者仅仅看过一些视频教程，按要求在 Jupyter Notebook 中做过一些练习，就基本具备动手解决问题的能力，可以上岗了。

这当然很好。但是，倘若你想在某一个领域取得真知，读书仍然是不可或缺的手段。中国信息安全领域的领军人物冯登国院士曾经说，以他的经验，想要真的搞懂某一个领域，非得深入“啃”至少一本书不可。读书的效率相对于听课、看视频要低得多，而多维的知识体系通过单维的文字表达出来，也给理解带来了挑战。然而，唯其有这种挑战，才需要读者进入深度思考状态，使读书成为一个推敲、琢磨、设问和破解的过程。不经过这个过程，我们所学到的知识一般来说只能是浮于表面的，很难达到“知其然知其所以然”的高度。正因如此，我们已经开始发现，仅通过在线视频和动手练习的学习者，对于相关领域的理论掌握经常是肤浅的。可以说，到目前为止，读书作为一种学习手段，依然是其他方式无法取代的。

机器学习和数据科学领域有几本非常重要的著作，每一个有野心的学习者都应该选择至少一本深入研究。Christopher Bishop 于 2006 年出版的 *Pattern Recognition and Machine Learning*，Kevin Murphy 2012 年的巨著 *Machine Learning: A Probabilistic Perspective*，斯坦福大学两位机器学习泰斗 Trevor Hastie 和 Robert Tibshirani 及其学生合著的 *An Introduction to Statistical Learning*，当然还有 Ian Goodfellow 和 Yoshua Bengio 最近出版的 *Deep Learning*——称这几本书为这个领域的“四书五经”，应该没有争议。

但是，这几本书有一个共同的问题——都是按照教材的体例编写的，所以都是尽全力系统化地介绍知识，对这个领域丰富多彩的应用、历史、人文和故事却很少展开论述。而要成为一名数据科学家，仅有知识和动手能力是不够的，还需要有相应的素养，这包括特有的思维方式、

价值观，对相关历史背景和掌故的了解，以及对数据科学社区的认知和互动——这恰恰是本书的最大价值。

作者把数据科学放在一个更广阔的背景之中，从数学、统计学、方法论甚至认知论的层面出发，讨论数据科学的内涵和外延，内容丰富，旁征博引，语言生动，灵活有趣，帮助读者站在一个更丰富的势场中认识数据科学，理解数据科学的基本思想。尤为令人欣喜的是，作者将信息论、混沌理论纳入讨论之中，表明作者敏锐地注意到数据科学与系统科学和认识论的深层联系，这是难能可贵的。从这个角度来解说数据科学的书，应该说在中国是第一本，即使在全球范围内也是独具特色的。为此，我们愿意向读者推荐本书，并相信读者一定能从中获得非常有价值的启发。

CSDN、AI100 创始人　蒋　涛

AI100 合伙人　孟　岩

前　言

为什么要写这本书

随着计算机科学和数据科学的发展，越来越多的人开始把目光投向其中最为耀眼的互联网、物联网、大数据、人工智能等高新技术领域，并且有相当多的高级技术人才已经在这些领域获得了令人瞩目的成就。

在追逐信息技术发展浪潮的过程中，数据科学成为人们在信息技术海洋中遨游所沉淀下来的理论与科学基础。我们都渴望通过对数据科学的理解来对生产工作进行指导和改善，这种工作的意义与其他各种在信息技术产业一线工作所创造价值的意义一样非比寻常。它给我们更广的辩证思考的空间，更高的观察事物的眼界，更多的自新的维度与动力。它是那么神秘且有趣。

今天，数据科学已经渗透到我们每个人的工作和生活之中。在你早上起来赶公车或者地铁的时候，你其实正在享受由数据科学辅助进行的精确调度服务；在你阅读工作报表的时候，你其实正在享受由数据科学辅助进行的大数据统计服务；在你吃午餐的时候，你其实正在享受由数据科学辅助进行的外卖快餐数据分发或食堂菜品改良服务；当你晚上回到家，在网上尽情购物的时候，你其实正在享受由数据科学辅助进行的高效电子商务和智能推荐服务。驾车出行有智能导航，就医问药有分诊机器人……也许你的家人或者朋友现在就在自己的工作岗位上，作为一名普通的销售人员、产品经理、人力资源师、售后服务人员、商务代表等，通过数据决策系统、数据库甚至电子表格来观察数据，作出判断，开展工作。数据科学给我们带来的红利已经紧紧把我们包围。

这本书绝无说教的想法，而是希望以书为媒，用谈天说地的方式，以激发每个人的思考为主要手段，归纳总结数据科学的实质及成就一位数据科学家所需要的基本素养。

遗憾的是，越是基础性、本源性的学科，与变成现实利益的距离就越远，让人觉得似乎不够实惠，不够亲近。至少读完这本书没办法帮你直接在第二天变出米饭、房子和汽车。不过我认为，楼房再高再漂亮，也需要人们看不到的深厚地基来支持；花儿再芬芳再娇艳，也需要在土壤之下吮吸养分的丰富根系来供能。这些看不到的东西，往往起着我们无法想象的巨大作用，

而这才是我希望与你一同讨论并思考的。

我们热爱生活，我们热爱所做的工作，我们希望在不断的攀登中看到更深更远的世界并去伪存真。那就让我们在点点滴滴的知识片段中一起开始慢慢思索、细细揣摩这一养成过程吧。

本书特色

本书从众多先贤及科学家的轶事开始讲起，以逐步归纳和递进的脉络总结出科学及数据科学应关注的要点，然后在生产的各个环节中对这些要点逐一进行讨论与落实，将这本书变成一本具有一定思维升华价值的参考书，从更高、更广的视角回看科学及数据科学在各个生产环节的缩影。

本书并不以高深的数学理论研究作为目的，也不以某一种计算机语言编程作为主线脉络，而是在一个个看似孤立的故事与工程中不断拾遗，并试着从中悟出一些道理。

简洁与深刻并重是本书的另一大特点。作为认知科学的延伸，数据科学一方面应该越来越引起广大大数据工作者的重视，另一方面也要撩开自己的神秘面纱，以最为亲民的姿态和每位大数据工作者成为亲密无间的战友，为用科学的思维方式进行工作做好理论准备。

读者对象

（1）大数据从业人员和对大数据相关知识感兴趣的人。

（2）初级和中级程序员、架构师，以及希望通过对数据的感知改进工作的人。

（3）产品经理、运营经理、数据分析师、数据库开发工程师等对数据分析工作敏感的人。

（4）希望在思维方式领域进行拓展的高校毕业生和希望接触并了解数据科学的社会人员。

（5）所有对数据科学感兴趣并希望逐步深入了解数据科学知识体系的人。

如何阅读本书

本书分为 3 篇，分别是认知篇、分化篇和实践篇。

认知篇（第 1 章 ~ 第 7 章）

归纳了什么是科学，数据科学的范围、定义与实践价值，以及辩证思维、哲学和实验的关系等问题。这些是认知观点的基石。

分化篇（第 8 章 ~ 第 11 章）

重点介绍了数据科学中与现代社会各行业联系最为紧密的统计学、信息论、算法学，另外把混沌论作为一个知识点进行了补充。这些是认知观点在不同细分学科中所形成的一些具体解决问题的思维方式和科学观点。

实践篇（第 12 章 ~ 第 19 章）

沿着数据生命周期进行演进。任何行业的数据生命周期都是按照采集、存储、统计与建模、算法、可视化与分析、决策支持的沿革来进行的，本篇对各个环节的注意事项和思维方式都做了详细的讨论，并在第 18 章介绍了两个具体的数据分析案例。

在本书的最后，补充了过去与同行们讨论过的，并在会议演讲及日常分享的过程中总结出来的一些精彩问答。

如果你希望读完这本书后能够在数学方面有很大的提升，在工程代码能力方面有巨大的进步，这本书恐怕帮不上什么大忙。但我相信，在读完这本书后，你会在一些以前并不熟知的领域有所了解和感悟，并逐步完善理解和分析问题的视角。如果你不是数据研究人员，也可以把这本书当成一个休闲读本。这本书里既没有太多的公式，也没有太过高深的理论，有的只是我在和你攀谈的过程中与你一起发现的新视角。

特别致谢

感谢绘麟社相辉先生和李晓林女士对本书的插画助力。

参加本书编写工作的有高扬、卫峥、左妍、尹会生、杨艺、陈钢、肖力。

勘误和支持

由于作者的水平有限，编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。如果您有更多的宝贵意见，欢迎扫描本页的二维码，关注“奇点大数据”微信公众号与我们进行互动讨论。本书后续的代码上传及勘误等相关更新内容都会在这个微信公众号发布。关注大数据尖端技术发展，关注奇点大数据。

同时，您也可以通过邮箱 77232517@qq.com 与我联系，期待能够得到您的真挚反馈，在技术之路上互勉共进。

高　扬

2017 年 1 月于珠海

反侵权盗版声明

目　录

认知篇

分化篇

实践篇