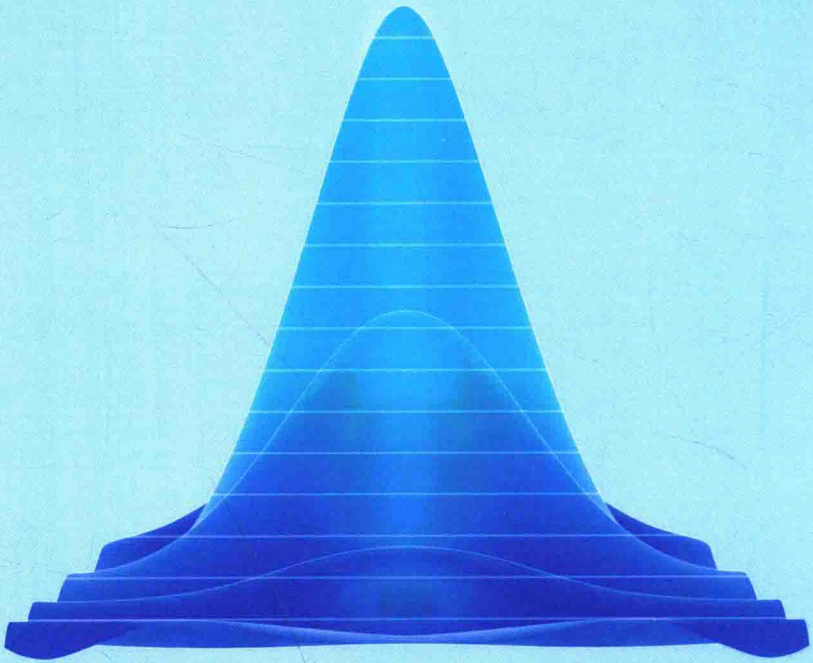
25位著名数据科学家的真知灼见

异步图书
www.epubit.com.cn

数据科学家访谈录

[美] 单研（Carl Shan） 陈子蔚（William Chen）
汪强明（Henry Wang） 宋迈思（Max Song） 著

田原 刘奕 译



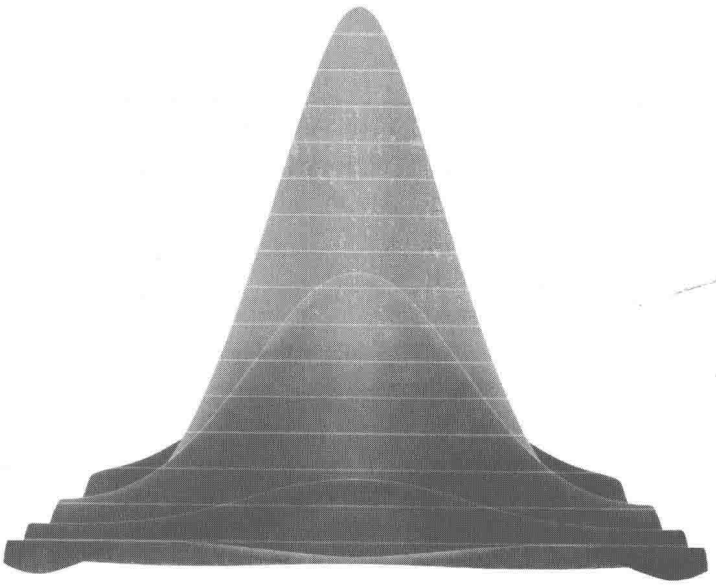
中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

数据科学家访谈录

[美] 单研（Carl Shan） 陈子蔚（William Chen）
汪强明（Henry Wang） 宋迈思（Max Song） 著

田原 刘奕 译



人 民 邮 电 出 版 社

北 京

图书在版编目（CIP）数据

数据科学家访谈录 / （美）单研（Carl Shan）等著 ；田原，刘奕译. -- 北京 ：人民邮电出版社，2018.2
ISBN 978-7-115-47091-1

Ⅰ. ①数… Ⅱ. ①单… ②田… ③刘… Ⅲ. ①数据管理－工程技术人员－访问记 Ⅳ. ①K816.16

中国版本图书馆CIP数据核字(2017)第301153号

版权声明

◆ 著 [美]单研（Carl Shan） 陈子蔚（William Chen）
汪强明（Henry Wang） 宋迈思（Max Song）
译 田 原 刘 奕
责任编辑 陈冀康
责任印制 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷

◆ 开本：720×960 1/16
印张：19.25
字数：411 千字 2018 年 2 月第 1 版
印数：1－2 400 册 2018 年 2 月河北第 1 次印刷
著作权合同登记号 图字：01-2016-5097 号

定价：69.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

内容提要

数据科学正在对商业、教育、能源、软件与互联网等各行各业产生深远的影响并贡献巨大的价值。作为21世纪最诱人的职业，数据科学家既有巨大市场需求的潜力，又面临着高难度的学习路径的挑战。

本书选取世界知名的25位数据科学家进行了深度的访谈，从不同的视角和维度，将他们的智慧、经验、指导和建议凝聚成册。每一篇访谈都是一次深度的交流，涵盖了这些数据科学家最初从菜鸟起步，运用各种知识武装和充实自己，一直到最终成为一名卓有成效的数据科学家的全过程。通过阅读本书中的访谈，读者可以形成对数据科学的宏观认识和了解，更深刻地认识和体验数据科学家的角色，并且从这些前辈的过往经历中学到宝贵的知识和经验以应用于自身的成长和事业中。

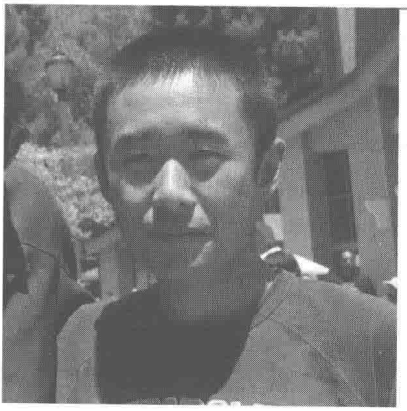
本书适合有志于成为数据科学家的人、正在从事数据科学相关工作的人、数据科学团队的领导者和企业家以及商业人士参考，也适合对数据感兴趣的普通读者阅读。

致亲爱的家人、朋友和导师们，

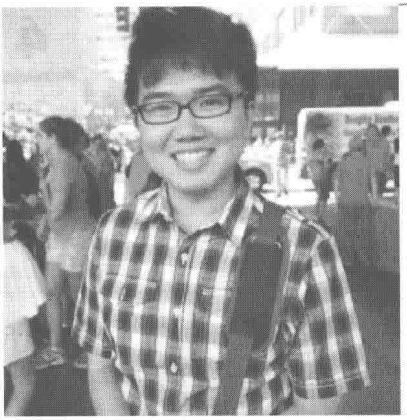
你们的支持与鼓励是我们生命之火的不竭动力。

作者简介

Carl Shan 于 2014 年在芝加哥大学 Eric & Wendy Schmidt 数据科学学会担任数据科学家，用数据模型协助非营利组织的工作。他与人合作撰写了一篇论文，将监督学习应用于公共政策问题。他以优异的成绩毕业于加州大学伯克利分校并获得了统计学学位。他目前在加州圣马特奥的 Nueva 学校教授机器学习和计算机科学。你可以通过 www.carlshan.com 了解关于他的更多信息。



Henry Wang 目前在伦敦，在一家专注于转型工作的金融公司工作 。在此之前，他曾在美国的一家可再生能源公司进行增长股权投资。在他的闲暇时间里，他喜欢参与诸如 Numer.ai 这样的数据科学竞赛，并且对基于随机梯度的机器学习优化算法很感兴趣。他拥有加州大学伯克利分校的统计学学位。你可以通过 www.henrywang7.com 了解关于他的更多信息。



William Chen 是 Quora 的数据科学经理，他在那里帮助公司发展壮大并与世界分享知识。他也是 Quora（https://www.quora.com/profile/William-Chen-6）上一个狂热的作家，在那里他回答各种关于数据科学、统计、机器学习、概率的问题。他参与本书的写作，分享了数据科学家的故事，以帮助那些想要进入这个行业的人。在闲暇时候，他的爱好是玩“密室逃脱”，他还开了一个专门用于分享这类“越狱经验”的博客。William 拥有哈佛大学的统计学学士和应用数学硕士学位。他的个人网站是 www.wzchen.com。

Max Song 曾在 Ayasdi 担任数据科学家，他也是 Neurocurious（后来被 Vium 收购）公司的联合创始人。他曾任奇点大学（Singularity University）的生物信息助教，从而接触人工智能的概念。他热爱学习、旅行和社区建设，并与其他人共同创立了“壹沙龙（onesalon.org）”。Max 拥有布朗大学 (Brown University) 的应用数学和生物学学士学位、清华大学苏世民学院（Schwarzman College）的硕士学位，他是苏世民学院的首届学生之一。他目前在香港的一家家族公司从事研究和投资。你可以通过 www.maxsong.io 了解关于他的更多信息。

序

在过去的 5 年里，数据科学差不多对人类所有重要的研究突破领域，都产生过深远的影响。从商业到教育界，再到能源领域，当然，也包括软件与互联网产业，在全球范围内，数据科学在这些形形色色的产业中产生了巨大的价值。实际上，在 2015 年年初，美国总统发布了白宫的一个新职位——首席数据科学家，并且任命 DJ Patil 担此重任，而 DJ Patil 正是本书中的受访者之一。

与世界上其他的发明创造如出一辙，数据科学产业的诞生同样归功于一小群积极踊跃的人。在过去的几年里，正是他们让数据分析这一理念可以走进任何领域，慢慢从无到有，发展壮大，并最终深入人心。在本书中，你将有机会遇见这些开拓者中的一部分，聆听他们一路走来的、精彩纷呈的第一手故事，并且了解他们对于数据科学未来的发展预见。

成为数据科学家的道路并不总是一帆风顺的。当我曾经试图从实验物理学领域转向这个领域时，和如今相比，那时的资源是如此的稀缺。实际上，虽然当时公司里确实已经存在数据科学方面的岗位需求了，但这一类人却连一个正式、统一的职位名称都没有。我曾经花费大量的时间自学这个领域的知识，也在不同的产业项目中磨砺过，到头来却发现我在学术圈的朋友遇到了和我同样的挑战。

我见过许多拥有极高天分及多年科研领域经验的研究人员，由于心仪数据科学领域而选择转向其中，愿意成为与数据为伍的人，但却挣扎多年不得要领。简而言之，他们不知道如何将自身惊人的数学功底、计算天赋以及数据分析技巧用在工业界。与此同时，我在硅谷工作的时候发现，相当多的科技公司其实都急需这方面的人才。

为了填补学术界与工业界之间的鸿沟，我于 2012 年创建了深入理解数据科学研究（Insight Data Science Fellows Program）社群。该项目旨在组建一个帮助计量相关领域的博士从学术界向工业界转职的训练团队。在过去的几年中，我们已经帮

助数百名项目成员，从诸如物理学、计算生物学、神经科学、数学以及工程学之类的科研背景转入工业界，在诸如 Facebook、Arbib、LinkIn、纽约时报公司、斯隆 - 凯特琳癌症中心以及其他上百家企业公司中担任重要的数据科学家职位。

在我的个人过往经历中，一方面，我自己成功走进了科技产业；另一方面，我也创造了一个让更多的人走上这条路的团队社区。在此过程中，我发现对我的事业给予重要帮助的一个资源就是：更多地与那些成功完成事业转型的人沟通交流。鉴于我创建并发展了数据科学社群，我有机会与硅谷的一些最好的数据科学家沟通交流，他们绝对是业内顶尖的大师：

Jonathan Goldman 创建了 LinkedIn 公司最初的一个数据产品，即“你可能认识的人（People You May Know）”，该产品直接促使公司改变了它的发展战略。DJ Patil 将 LinkedIn 内部的数据科学小分队发展壮大，最终发展成了该公司一个强大的部门，并且他也是“数据科学”这个术语最初的创造人之一。Riley Newman 在 Airbnb 公司内致力于产品开发与分析，该工作对于 Airbnb 的发展可谓举足轻重。Jace Kohlmeier 在可汗学院领导数据团队，致力于将上百万学子的网上学习最优化。

遗憾的是，想要与这些大师面对面交流是非常难的。在数据科学研究社群中，为了尽量争取与这些大师面对面交流高质量的内容，我们每年只会选择这样一群数据科学家以及工程师中的 3 位进行交流访谈。

本书把与这些大师的深度交流访谈整理出版，奉献给读者。

通过阅读本书中的访谈，你应该可以从这些前辈们的过往经历中学到一些知识并用于你自己的事业中，无论你现在身在何地，从事何业。每一篇访谈都是一次深度的交流，涵盖了这些科学家最初从菜鸟阶段起步，运用各种知识武装充实自己的经验，一直到最终成为数据科学家的事业全程。

并不只是早期的数据科学先驱们才有可能在这个领域做出卓越的贡献。这个领域源源不断地有新鲜血液注入，他们中的每一个人都有机会推动这个领域前进。在我遇到本书的作者们的时候，他们都曾只是梦想成为数据科学家的大学生，一个个急切地询问着那些每一个初入门道的人都想要了解的问题。

在 18 个月的努力学习过后，他们跑遍各地并寻访了全球的诸位顶尖数据科学家，探询了他们的观点、意见和指导。本书就是这些访谈的最终成果，将最出

类拔萃的一群数据科学家的 100 小时以上的智慧汇集整理成册（想象一下你去和奥巴马总统都要抢时间与之交谈的 DJ Patil 对话）。

通过阅读这些内容丰富且非正式的访谈，你将会坐在领域先驱 DJ Patil、Jonathan Goldman 和 Pete Skomoroch 对面，他们都是 LinkedIn 早期的员工，也是 LinkedIn 内部数据科学团队的核心成员。你将会遇到 Hilary Mason 与 Drew Conway，他们是声名远扬的纽约数据科学社区的主要发起人及推动人。你将会听到未来的数据科学领域先锋领袖（如 Diane Wu 和 Chris Moody）的建议，他们都曾是数据科学研究社群的成员，现在他们正分别在 MetaMinds 和 Stitch Fix 公司大放异彩。

你将会遇到那些在学术领域有巨大影响力的科学家，例如加州大学圣迭戈分校的 Bradley Voytek 和哈佛大学的 Joe Blitzstein。你也将见到初创公司里的数据科学家，例如 Mattermark 的 Clare Corthell 和 Bento Labs 的 Kunal Punera，他们会告诉你他们如何将数据科学作为让自己更有竞争力的武器来运用。

本书中提到过的科学家们与其他的千万同僚们一起，曾经创建了许多形形色色的对这个世界产生重大影响力的公司和企业。在本书里，他们主要讨论了那些促使他们厘清误区、不断开疆拓土的心路历程，并且分享了他们人生中那些有特别意义的挑战或成功的故事，以及他们对于自己的团队所需要的人才的想法。

我希望读者通过阅读此书，聆听他们所思，学习他们对于未来的数据科学世界的眼界，并最终找到适合自己的数据科学之路。祝愿你们在这条路上做出自己对于世界的贡献，甚至于推进这个领域的前沿发展。

深入理解数据科学研究社群、深入理解数据工程研究社群、深入理解健康数据科学研究社群的创始人 Jake Klamka

前言

欢迎阅读本书！

在本书此后的内容中，你将会看到针对25位卓越的数据科学家的深度采访。他们来自于不同的背景、职业以及产业。他们中的一些人，诸如DJ Patil和Hilary Mason，是曾经将这一领域从默默无闻推向全球皆知的伟大开拓者。也有一些刚刚开始数据科学家生涯的学者，例如Clare Corthell，她在这个领域内有自己独树一帜的贡献，即创造了开源数据科学导师课程，这是一套完全基于开源的互联网资源而建立的自学课程。

如何阅读本书

我们出版本书的目的，是创造一本可以历久弥香并且激发你对于数据科学的兴趣的图书，无论你的教育专业背景如何，希望你都能从中获益。我们每一次精心校对、编辑、推敲和拿捏，都是为了让本书成为你日后在不同的学习和事业阶段，可以不断回头翻阅，得以温故知新的一件礼物。

这里列出了本书中涵盖的知识点。尽管本书的每一篇访谈都是精彩绝伦的，并且涵盖了很广阔的知识领域，我们还是从中选择出了一些有助于你快速起步的访谈。

- 有志于成为数据科学家的读者：你可以从这些故事中得到如何转向数据科学领域的建议和经典案例。

  推荐阅读：William Chen、Clare Corthell和Diane Wu。

- 正在从事数据科学工作的读者：你可以从访谈中知道如何更高效地工作，以及如何更快地在职场中成长。

  推荐阅读：Josh Wills、Kunal Punera和Jace Kohlmeier。

- 数据科学团队领袖：你可以从访谈中知道如何招聘其他数据科学家，如何组建一个团队，以及如何与公司产品和工程部门通力协作等一系列历久弥新的经验。

  推荐阅读：Riley Newman、John Foreman和Kevin Novak。

- 企业家以及商业人士：你可以从中读到有关数据科学未来发展方向的灵感，从而拓展你的视野。

  推荐阅读：Sean Gourley、Jonathan Goldman和Luis Sanchez。

- 对数据感兴趣的普通读者：你可以通过阅读一些最早期的数据科学家的故事，来知道这个领域的来龙去脉与历史沿革。

  推荐阅读：DJ Patil、Hillary Mason、Drew Conway和Pete Skomoroch。

在收集、策划以及编纂这些访谈的时候，我们的重心一直是与这些科学家中的每一位都能有深度并且高质量的对话。这其中的很大一部分信息也同样是长久以来数据科学界众多周知的观点和故事。你将会听到他们每一个人独家的出身背景、宏观眼界、职场经历以及人生建议。

在本书后面的内容中，你将会看到这些数据科学家对于以下问题的观点和解答：

- 为什么数据科学对于今天的世界和经济如此重要？
- 如何同时掌握编程、统计以及领域知识，从而成为一名卓有成效的数据科学家？
- 如何从学术界或者其他领域，专职进入数据科学领域，并在其中找到一份工作。
- 数据科学家与统计学家、软件工程师有什么区别？他们如何协同工作？
- 如果你的公司有数据科学相关工作需求，你应该如何招聘员工？
- 如何建立一支出色的数据科学团队？
- 卓越的数据科学家与优秀的数据科学家相比，在心态、技术和能力等方面有什么区别？
- 数据科学的未来会是怎样的？

在你阅读这些访谈之后，我们希望你会发现，从不同的背景和领域转入数据科学领域，并最终成为数据科学家这一过程是非常多样化的。我们再次祝你一路好运，并且期待你与我们联系：contact@thedatasciencehandbook.com。

—— Carl、Henry、William 和 Max

目录

第1章　重要问题的取舍 …… 1
RelateIQ产品部副总裁DJ Patil
第2章　在成为成功的数据科学家之际 …… 14
Fast Forward Labs创始人Hillary Mason
第3章　无处不在的软件开始用数据重构这个世界 …… 25
Data Wrangling核心数据科学家Pete Skomoroch
第4章　学术期刊中的数据科学 …… 40
《纽约时报》数据科学家Mike Dewar
第5章　通过数据倾听你的客户 …… 50
Airbnb数据主管Riley Newman
第6章　建立你自己的数据科学课程表 …… 58
Mattermark数据主管Clare Corthell
第7章　均方误差根无法解决所有社会难题 …… 67
Project Florida数据主管Drew Conway
第8章　软件工匠学堂、软件工程及产品 …… 80
Uber数据科学主管Kevin Novak
第9章　从天体物理到数据科学 …… 89
Square数据科学家Chris Moody
第10章　数据科学中软件工程的重要性 …… 101
Facebook数据工程师Erich Owen

第11章 弥合领域的鸿沟：从生物信息到数据科学 …… 108
Ayasdi数据科学家Eithon Cadag
第12章 如何锻炼数据科学技能 …… 123
Intuit资深数据科学家&创新领袖George Roumeliotis
第13章 科学、工程和数据科学的交织 …… 132
Palantir数据科学家Diane Wu
第14章 从高频交易到驱动个性化教育 …… 140
Khan Academy 数据科学主管Jace Kohlmeier
第15章 针对数据科学与演讲能力的教育 …… 150
哈佛大学应用统计学教授Joe Blitzstein
第16章 数据科学不是Kaggle竞赛 …… 162
MailChimp首席科学家Jonh Foreman
第17章 数学、自谦以及成为更好的程序员 …… 182
Cloudera数据科学主任Josh Wills
第18章 数据科学和学术界 …… 195
UCSD计算神经科学教授，前Uber数据布道师Bradley Voytek
第19章 数据科学家的学术、量化金融与企业家之路 …… 205
ttwick创始人/数据科学家Luis Sanchez
第20章 美国总统竞选就像物理科学一样 …… 216
Civis Analytics资深数据科学家Michelangelo D'agostino
第21章 培养数据感觉的重要性 …… 226
LinkedIn数据科学家主任Michael Hochster
第22章 数据挖掘、数据产品与企业家精神 …… 240
Bento Labs联合创始人/CTO Kunal Punera
第23章 从战争建模到增强智能 …… 256

Quid联合创始人/CTO Sean Courley
第24章 如何创建新颖的数据产品和公司 …… 277
Intuit数据科学家主任Jonathan Goldman
第25章 从本科生到数据科学家 …… 284
Quora数据科学家William Chen

RelateIQ 产品部副总裁 DJ Patil

第 1 章　重要问题的取舍

DJ Patil 是“数据科学家”这个术语的创造者之一，也是哈佛商业周刊文章《数据科学家：21 世纪最诱人的工作》（*Data Scientist：Sexiest Job of the 21st Century*）的共同作者。

由于折服于数学的魅力，年轻时代的 DJ 在加利福尼亚大学圣地亚哥分校取得了数学学士学位，然后在马里兰州立大学取得应用数学博士学位。在攻读博士期间，他主要研究非线性动态过程、混沌理论以及复杂系统。在进入科技领域以前，他在气象领域做了将近十年的研究工作，并且为美国国防部和能源部提供咨询服务。在他的职业生涯中，DJ 曾在 eBay 担任首席架构师和研究科学家职位，然后在 LinkedIn 担任数据产品主管，正是在那段时光里，他与 Jeff Hammerbacher 一同创造了“数据科学家”这个术语，并且打造了一个出类拔萃的数据科学团队。他曾是 RelateIQ 公司产品部副总裁，RelateIQ 是新一代基于数据科学开发的客户关系管理软件（customer relationship management software）。近期，RelateIQ 公司因为其出众的数据科学技术而被 Salesforce.com 收购。

在对他的访谈中，DJ 将会谈论抓住时机的重要性，通过独立学习、团队工作，激发兴趣并回馈帮助过自己的社区，以此不断提高自己。

2015 年，DJ 被任命为美国历史上第一位首席数据科学家。