DATA SCIENTIST

The Definitive Guide to Becoming a Data Scientist

数据科学家

修炼之道

[美] Zacharias Voulgaris 著

吴文磊 田原 译

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

DATA SCIENTIST

The Definitive Guide to Becoming a Data Scientist

数据科学家

修炼之道

[美] Zacharias Voulgaris 著

吴文磊 田原 译

人 民 邮 电 出 版 社

北 京

图书在版编目（CIP）数据

数据科学家修炼之道 / （美）弗格里斯(Voulgaris,Z.）著 ; 吴文磊, 田原译. -- 北京 : 人民邮电出版社, 2016.4
ISBN 978-7-115-41824-1

Ⅰ. ①数… Ⅱ. ①弗… ②吴… ③田… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第047210号

版权声明

◆ 著　　[美] Zacharias Voulgaris
译　　吴文磊　田　原
责任编辑　陈冀康
责任印制　张佳莹　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
固安县铭成印刷有限公司印刷

◆ 开本：720×960　1/16
印张：15.5
字数：196 千字　　2016 年 4 月第 1 版
印数：1 – 3 000 册　　2016 年 4 月河北第 1 次印刷
著作权合同登记号　图字：01-2014-5848 号

定价：49.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

数据科学家是指采用科学方法、运用数据挖掘工具寻找新的数据洞察力的工程师，他们往往集技术专家和数据分析师的角色于一身。在IT行业中，数据科学家将在创造力、沟通能力以及与商业世界的联系方面得到更多的锻炼机会，是当前非常有发展潜力的新兴职位。

本书全面介绍了成为数据科学家应当了解的各类知识。全书共分18章，首先介绍了数据科学与大数据、数据科学的重要性，接着介绍了数据科学家的类型、思维体系、技术资质、经验、社交圈、所用的软件、学习新知和解决问题，另外还介绍了机器学习与R语言平台、数据科学的处理流程、所需的具体技能，最后介绍了数据科学求职、自我展示，并提供了一些有关职业数据科学家和资深数据科学家的案例学习。

本书内容全面、轻松易读，非常适合从事数据科学相关工作的读者阅读，是一本可以帮助读者应聘数据科学家职位的求职指南。

前言

在一年半以前，对数据科学家这个角色，我既没有清晰的认识，也不明白它的重要价值。那时候，我陷没在一个毫无希望的网络营销公司，已经开始忘记我曾艰辛数年所学的知识。不知道是什么让我下定决心去研究这个主题（当时根本没有什么像样的书，更不用奢谈有谁能来指导我），但我内心非常清楚，那段日子让我的身心得到了放松。当然，在这个过程里，我碰到了许多既没学过，也不知道如何去学的问题，而这还是在我每周花费 50 小时的前提下，而且在我生活的地方根本没有一个能深入教导数据科学的课程。但我努力坚持了下来，我深信我所付出的一切都是值得的，更不用说这个过程让我感到愉悦。就算我在这条道路上失败了，至少在这个过程里我学到了许多有用的技能。

这本书是写给那些同样渴望探索这个领域的人们。当我开始探索数据科学的时候，我不得不用一种艰难的方式学习，并通过尝试和犯错，以及艰难地通过文章、视频和网络上其他的资源来研究数据科学。好在，这些对你来说将会容易得多。而这正是我写这本书的原因，也就是让你有这样一本手册，可以在这个充满挑战的变革中获得指导。

数据科学是具有高回报的领域，它处理的是数据世界中激动人心的新大陆——大数据。数据科学是一个由许多有趣的挑战所构成的领域，而这些挑战是因为我们还尚不能直接有效地利用大数据。但这也为创造力的发挥提供了空间，也使你能够以数据科学家的身份去探索和开拓更多的可能性。

另外，通过数据科学家这个角色，你会获得锻炼自己能力的机会。在

IT 领域里，没有第二个角色能够赋予你这样的机会：创造能力、沟通能力与商业世界的直接联系，等等。通过精妙地使用你能获取的数据，你将为你的组织（例如公司、政府机构，甚至是慈善组织）提供价值。这些数据与你平常在漂亮的数据库里碰到的数据完全不同，它们体量巨大，类型多样，而且很散乱。于是，这催生了大数据与数据科学家的诞生，他们就是以科学的富有创造力且易于理解的方式来处理大数据的专业人士。但是，到底什么是大数据呢？大数据跟传统数据有什么差别？一个简略的回答是“没办法只用一台电脑去处理的数据就是大数据。”尽管这通常是因为数据的体量巨大，但还是有些别的原因。我们一般总结为 4 个特征，通常被称为大数据的 4 个 V。

- **体量（Volume）**：与我们常见的“一般”数据不同，大数据要大得多得多，例如几个“太”字节（TB）到几个“泽”字节（ZB）。ZB 是 TB 的 100 万倍，是 GB 的 10000 亿倍。非常非常大！在 2010 年，整个世界的数据量达到了差不多 1ZB，相当于是 125000000 个 8GB 媒体播放器的存储容量。更夸张的是，这个数字在过去几年里以极快的速度增长着，并且没有任何想要停下来的意思，这么巨量的数据塑造了大数据的身形。这样巨大的体量与无法被单机处理的特性（即便是一台超级计算机）带来了并行计算的方法（一组计算机通过网络连接一起工作）。在大数据项目里，这种计算方式被大量采用。
- **多样（Variety）**：由于数据源既可以是传统的，也可以是非传统的，这造就了大数据的多样性。我们平常所处理的数据通常来说是结构化的数据，这种数据常见于数据库中，严格用行和列表示。我们非常清楚它的类型和大小，而且我们大概也能猜到数据是什么样子。然而，大数据还囊括了非结构化数据和半结构化数据。非结构化数据没有提前定义下层数据的行列结构（例如说，Facebook 的贴子、

微博和电话呼叫记录）。而半结构化数据则是介于结构化与非结构化特征的数据结构（如机器的日志和电子邮件的地址格式）之间。

- **高速（Velocity）**：大数据另一个重要的特征就是它的速度，或者说，是数据传到企业以及被处理的速率。传统数据被认为是低速的，或是相对静止的，这包括从它的形成方式以及它产生的位置到它被处理的位置的传送速度可以体现。而大数据却是一直处于移动状态，而且是高速移动（尽管有时候也会有些例外）。这就意味着，大数据需要快速地处理，如果可能的话，最好是实时处理，以便于激发它的潜力。例如说，金融服务公司每秒需要以延迟不超过 30 微秒（微秒即百万分之秒）的速度分析 500 万条市场讯息。
- **精确（Veracity）**：由于这是一个新近加入的特征，所以有许多关于大数据的参考资料上仍写着 3V 属性。大数据同样也应是精确的，这是一个与数据的质量（即是否值得相信）息息相关的属性。正如大家所料，只要是数据就会掺有杂讯或者称之为噪音。有效地使用大数据就意味着要能够识别出这些藏在信号之后的噪音。这项挑战需要高超的分析技巧，如果不仔细，那么得出的统计结论就不是由真实的数据所支撑的，还可能导致我们做出不准确的决定。

有时候还能够看到另外两个 V，即多变性和可见性，但不是所有人都认同这两个特征。

不难发现，要让大数据有效地运作起来往往充满挑战。由于信息所蕴藏的潜在价值越来越明显，而驾驭大数据的手段也正持续增加着，大数据对产业界来说已经不容忽视了。想想亚马逊（Amazon）和奈飞（Netflix），他们对大数据精妙的使用为它们带来了竞争优势，同时也为行业开辟了一条新路。如果你身处在线购物行业，举个例子，你的大量客户群将带来大量数据，想象一下，你能获得他们的购物习惯、年龄构成、男女比例等特征，以及相应而来的通过分析数据所获得的大量机会。

基于这些你获得的科技新知，你可以再往前多走一步：也就是设计一个小工具或者 App 来利用你通过数据所洞察到的结果，或者是帮助你的用户获得类似的洞察力来提升他们的使用体验（如在线商店）。这也是亚马逊（Amazon）变得如此成功的一个原因。这就不止是向用户提供品类多样的产品，而是可以通过一些有益的特性，如推荐系统使整个购物体验变得更容易且更愉悦。许多类似的基于大数据智能分析的小程序被称为数据产品，它们集合了大量数据科学家的成果。而说到数据科学家，他们其实并不直接参与到产品的具体制造，而是以工程师的方式来帮助其他的数据科学家。所以即使是在一些具体工作中，数据科学家们由于各自技能的不同，所处领域也是千差万别的。

所以问题并不在于是否要跳上大数据的快车，而是怎么跳上去。这就得问数据科学家了。数据科学家在业内是一个崭新的角色，而自从他被引入求职市场以来，越来越受到人们的欢迎。这个职位涉及数据尤其是大数据处理的方方面面，并以充满智慧且有条不紊的方式创造着有用的产品（前文述及的数据产品）。这种以小工具或是 App 应用的形式出现的产品往往能给出连用户都还不知道的有价值的信息（后一点是由一位非常成功且富有经验的数据科学家 John Foreman 强调的）。大数据带来了数据处理和图像化的新法则，这种强有力的工具只有具备了独特的思维与技能的数据科学家才能驾驭。

许多人总是把科学家和数据分析师混淆在一起。其实他们的差别很大，这种差别很像航天飞机和普通飞机的差别。数据分析师会一些数据处理的技巧，尽管这些数据可能极为接近大数据的范畴，但相比于数据科学家来说，这种处理方式低效而且缺乏灵活性。数据分析师依靠一系列被预定义好的模型来导出有用的信息并制成报表交给业务人员过目，而数据科学家则是自己构建模型，或是在他的分析过程中使用完全由数据驱动的分析方法，这通常会推出一些能为他人所用的成果，不仅限于自己公司的业务人

员使用。数据分析师在他的报告中会提出一些靠直觉产生的图表，而数据科学家则会创造一个可以交互的仪表盘，表盘上可以看到所有必要的实时信息。

换句话说，数据分析是一种非常有用的工具，但如果有人想利用我们今天遍及四处的数据，那么他就不仅需要高效的数据分析技能，而且需要本书在后文所述及的各种其他实用知识。作为一个数据分析师是很棒的，但这也会把你困在结构化数据里的某个单一数据类型里面，而且，这些数据集还会把你变成一个只能处理小数据集的人。如果你想尝试更大而且更复杂的数据，你就需要学习数据科学家的处理思路。

数据科学家不能仅仅知道该怎么做，尽管对于那些对此感兴趣的人，这是一件极具愉悦感而且令人向往的工作。由于新技术的发展，他[1]所处的领域无时无刻不在发生变化，这使得这个领域充满活力。在科技的最前沿，与十分有趣的人群沟通，帮助其中一些人去驱动变革。数据科学是一个交叉学科领域，所以数据科学家们需要通过学习以及整合各学科的知识，用更系统性的方式去扩展自己的世界观。最重要的是，他们处理问题和数据的方式是十分有创造力的。

数据科学家同样是一个很好的职业。举例来说，这是一个能够给组织带来战略性优势的新角色（同时，不是很多人都会接受相应的训练）。根据 Indeed.com 的数据，对于具有相同工作年限的求职者，数据科学家所能获得的薪水相当丰厚，在一般情况下都好于其他的 IT 职业。此外，数据科学家还有机会发展自己各方面的技能，使自己成为一个多面手，能够有机会与行业或是科技界的各种各样的人打交道。尤其在金融风暴时期，技术专才求职会变得特别困难，而经受过数据科学的训练就凸显出其价值了。

这本书由 18 章组成，涵盖了数据科学的多项基础知识。在最开始的几

---

[1] 尽管我使用“他”在全书中指代数据科学家，但这个从事这个职位的既可以是“他”也可以是“她”。

章里，你会学到更多关于大数据领域的细节（数据科学和大数据是什么；为什么数据科学非常重要，尤其是现在；以及数据科学家的各种不同类型）。接着，你会有机会学习到如何成为一名数据科学家（数据科学家的思维体系，他们的技能资质，这个角色所需要的经验以及一些社交网络的内容）。之后你会有机会了解数据科学家的日常生活（他用什么软件；在这份工作里学习新知的重要性，以及数据科学流程的主要环节）。后面的章节里，你会看到各种从现在的工作转职到数据科学家的路径图（如果你是一个名程序员，你应该学什么，如果你是一个统计员或是机器学习的从业者，如果你是一个与数据相关的职业，或者你是一个学生）。然后，我会向你提供一些如何找到你的第一份数据科学工作的切实可行的建议（到哪里去找，如何以一个“未来的”数据科学家表达自己，以及如果你想做自由职业者，你应该去思考哪些问题）。最后，你会读到一些真实的数据科学家故事，他们的经历、观点，还会有一些真实的数据科学家岗位的任职要求。书的最后是一个术语表以及 3 个附录，还有一些有用的网站、网上相关的文章和线下扩展阅读资源以及一些综合索引。

这本书用高山鹦鹉（Kea Bird）来形容数据科学家。这种鹦鹉以它的聪慧、创新、好奇而闻名，并且高山鹦鹉也是这个族群中最为稀少的一个品种，这些特点使高山鹦鹉与众不同，同样，这些特点也为数据科学从业者所共有。

我真诚地希望这本书能为大家带来帮助，也希望能让你从中获得快乐。尽管转型本身的要求十分苛刻（尤其是如果你还在职业生涯的初始阶段），但这份经历是迷人且具有高回报的。当你最后成功转型成为数据科学家之后，这个领域的精彩程度仍丝毫不减。这不是一个给轻言放弃的人设置的角色，成为数据科学家在各个层面上都是精彩的经历，而且会是一段使人着迷的旅程。那么，你准备好启程了吗？

Zacharias Voulgaris 博士

目录

第 1 章
数据科学与大数据

我们今天面临着诸多来自大数据和其他数据分析带来的困难，而数据科学正是对这些挑战的回应。在介绍中，我们简要剖析了一下大数据，但那仅仅是“冰山的一角”。事实上，围绕着大数据，能说的太多了，单凭这一章仍无法得其全貌。但是，你能够通过本章认识到大数据在今时今日的重要性。更重要的是，这一章能让你拨开大数据的迷雾（过去几年里日益弥散的炒作），让你明白数据科学的重要性。

大数据是当今商业的基础资产。大数据以及大数据相关的技术能够得到这么广泛地利用绝不是巧合，现今的诸多行业要么正用着大数据，要么准备要去用大数据。尽管关于大数据的各种炒作甚嚣尘上，但大数据并不是昙花一现。对这些资源善加利用会带来诸多优势，而目前这种资源的日益丰富也是值得关注的信号，不仅要用，而且要快！也许在某些行业里，大数据还不能带来价值，因为这些行业的数据非常混乱，甚至不存在数据。而那些拥有数据并对数据善加利用的人，会在当今竞争激烈的经济环境下占得先机并立于不败之地。

1.1 深挖大数据

大数据含有与我们身边的业务难题息息相关的丰富信息。举例来说，如果你是一个电商公司的经理，你就可以在你公司网站上收集到关于你客户和访客的丰富信息，若能对此善加利用，你就能够增加公司的销售额、

提升网站设计并改善客户服务，它还能为你提供市场策略和提升公司的整体策略的建议。这些都是由居住在你的服务器中的 0 和 1 实现的。你只需要从你的资源中分出一小部分，并从这些数据中间提炼出信息。这当然不是一桩赔本买卖，我们稍后会再回到这个例子。尽管有些网络数据披着大数据的外衣，但并不是每一种数据融合都可以叫做大数据。这主要是因为大数据的 4 个 V 特性。[1]

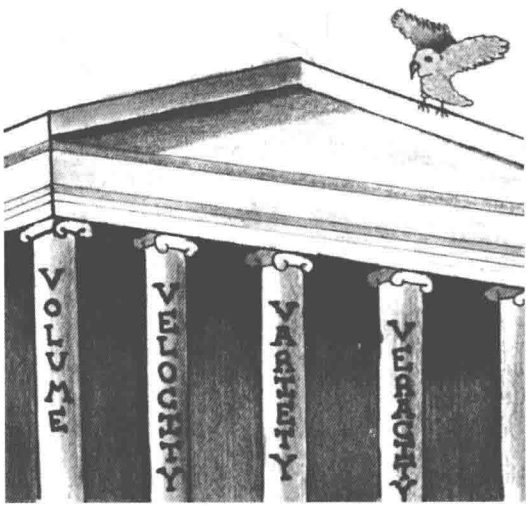

图 1.1 大数据的 4 个 V 特性

如我们之前所看到的，它们有如下几个特性。

- **体量**（Volume）：大数据由大量数据组成，从几个 TB 到几个 ZB。这些数据可能会分布在许多地方，通常是在一些连入因特网的计算网络中。一般来说，凡是满足大数据的几个 V 的条件的数据都会因为太大而无法被单独的计算机处理。单单这一个问题就需要一种不同的数据处理思路，这也使得并行计算技术（例如 MapReduce）得

[1] 事实上，有些人还会提及另外两个 V：多变性（variability）和可见性（visibility），这说的是大数据通常会随着时间变化，而且对于用户来说，很难洞察其中的变化。