O'REILLY

TURING 图灵程序设计丛书

Python 数据科学手册

Python Data Science Handbook

掌握用Scikit-Learn、NumPy等工具高效存储、处理和分析数据

[美] Jake VanderPlas 著

陶俊杰 陈小莉 译

powered by jupyter

中国工信出版集团

人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Python数据科学手册

Python Data Science Handbook

[美] Jake VanderPlas 著
陶俊杰 陈小莉 译

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®
O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目（CIP）数据

Python数据科学手册 /（美）杰克·万托布拉斯（Jake VanderPlas）著；陶俊杰，陈小莉译. -- 北京：人民邮电出版社，2018.2（2018.3重印）
（图灵程序设计丛书）
ISBN 978-7-115-47589-3

Ⅰ. ①P… Ⅱ. ①杰… ②陶… ③陈… Ⅲ. ①软件工具－程序设计－手册 Ⅳ. ①TP311.561-62

中国版本图书馆CIP数据核字(2017)第324296号

内 容 提 要

本书是对以数据深度需求为中心的科学、研究以及针对计算和统计方法的参考书。本书共五章，每章介绍一到两个 Python 数据科学中的重点工具包。首先从 IPython 和 Jupyter 开始，它们提供了数据科学家需要的计算环境；第 2 章讲解能提供 ndarray 对象的 NumPy，它可以用 Python 高效地存储和操作大型数组；第 3 章主要涉及提供 DataFrame 对象的 Pandas，它可以用 Python 高效地存储和操作带标签的 / 列式数据；第 4 章的主角是 Matplotlib，它为 Python 提供了许多数据可视化功能；第 5 章以 Scikit-Learn 为主，这个程序库为最重要的机器学习算法提供了高效整洁的 Python 版实现。

本书适合有编程背景，并打算将开源 Python 工具用作分析、操作、可视化以及学习数据的数据科学研究人员。

◆ 著 [美] Jake VanderPlas
译 陶俊杰 陈小莉
责任编辑 朱 巍
执行编辑 夏静文
责任印制 彭志环

◆ 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
北京圣夫亚美印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：29.25
字数：691千字 2018年 2 月第 1 版
印数：3 501－5 000册 2018年 3 月北京第 2 次印刷
著作权合同登记号 图字：01-2017-9018号

定价：109.00元

读者服务热线：(010)51095186转600 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

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

版权声明

© 2017 by Jake VanderPlas.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc.and Posts & Telecom Press, 2018. Authorized translation of the English edition, 2017 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版，2017。

简体中文版由人民邮电出版社出版，2018。英文原版的翻译得到 O'Reilly Media, Inc. 的授权。此简体中文版的出版和销售得到出版权和销售权的所有者 —— O'Reilly Media,Inc. 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

试读结束：需要全本请在线购买： www.ertongbook.com

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 *Make* 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过图书出版、在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——*Wired*

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——*Business 2.0*

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

译者序

本书主要介绍了 Python 在数据科学领域的基础工具，包括 IPython、Jupyter、NumPy、Pandas、Matplotlib 和 Scikit-Learn。当然，数据科学并非 Python 一家之“言”，Scala、Java、R、Julia 等编程语言在此领域都有各自不同的工具。至于要不要学 Python，我们认为没必要纠结，秉承李小龙的武术哲学即可——Absorb what is useful, discard what is not, and add what is uniquely your own（取其精华，去其糟粕，再加点自己的独创）。Python 的语法简洁直观、易学易用，是表现力最强的编程语言，学会它就可以让计算机跟随思想，快速完成许多有趣的事情。同时，它也是备受欢迎的胶水语言，许多由 Java、C/C++ 语言开发的工具都会提供 Python 接口，如 Spark、H2O、TensorFlow 等。2017 年 3 月 6 日，PyPI（https://pypi.python.org/pypi）网站上的程序包数量就已经达到 10 万，新的程序包还在不断地涌现，数据科学目前是 Python 星球最酷炫的风景之一。如果数据科学问题让你心有挂碍，那么 Python 这根数据科学的蛇杖（Asklēpiós，阿斯克勒庇俄斯之杖，医神手杖，医院的徽章）可以为你指点迷津。

本书书稿已经在 GitHub 上开源（https://github.com/jakevdp/PythonDataScienceHandbook）。由于本书的纸质版是黑白印刷的，因此作者在 GitHub 上建立了开源项目，以 Notebook 形式分享了本书的书稿，让读者可以看到彩色的可视化图。此外，作者也在博客（https://jakevdp.github.io/PythonDataScienceHandbook/）上发布了 Notebook 的 HTML 页面。除正文的部分内容外，Notebook 中的代码、注释与纸质版相同。由于 Notebook 是类 JSON 数据格式，因此也适合做版本管理，配合 GitHub 修复 bug 比较方便。配合本书同时开源的，还有作者编写的 Python 入门教程 *Whirlwind Tour of Python*，同样是使用 Notebook 撰写的。Notebook 是 IPython 的 Web 版，目前已经合并到 Jupyter（http://jupyter.org）项目中，是一款适合编程、写作、分享甚至教学（Jupyter/nbgrader）的开源工具，其基本功能将在本书第 1 章中介绍。Notebook 的操作十分简单，在浏览器上即可运行。它不仅可以在浏览器中直接编写代码、生成可视化图，还支持 Markdown 文本格式，能够在网页中快速插入常用的 Web 元素（标题、列表、链接、图像）乃至 Mathjax 数学公式，稍加调整便可以幻灯片形式播放内容，阅读体验一级棒。

看编程书的第一步是搭建开发环境，但这一步往往会吓退不少对编程感兴趣的读者。本书对应的开发环境可以通过三种方式实现。第一种方式是在线版 Notebook 编程环境，免安

装，有浏览器就可以学习编程知识，推荐想快速掌握知识的朋友使用。目前，有许多安装了 Python 编程环境的 Anaconda 发行版的网络平台（PaaS），支持 Jupyter Notebook 编程环境，可以免费使用，如 JupyterHub（https://tmpnb.org）、SageMathCloud（https://cloud.sagemath.com）、微软 Azure（https://notebooks.azure.com）在线编程环境。它们可以在线运行 Notebook 文件，编写调试运行代码，也支持文件的上传、下载、新建、删除，还可以运行 Terminal 工具。另外，基于 GitHub 代码仓库，有 nbviewer（https://nbviewer.jupyter.org）可以查看 GitHub 的 Notebook，还有 binder（http://mybinder.org）支持代码仓库一键部署，都是非常有趣的组合。类似的在线免费 Notebook 编程环境还有很多，特别推荐德国 Yves Hilpisch 博士的 The Python Quants Group 公司开发的 Python Quants Platform（http://tpq.io）。Yves 博士的三本 Python 金融学图书均使用该编程环境，读者可以免费注册使用，其硬件为 CPU Xeon 1231、16GB 内存，能够满足一般的学习与分析需要。Jupyter Notebook 支持许多编程语言（Python、R、Scala、Julia、Haskell、Ruby……），甚至支持 Kotlin（https://github.com/ligee/kotlin-jupyter）、Java 9 的 REPL 新功能 JShell（https://github.com/Bachmann1234/java9_kernel）。第二种方式是在电脑上安装 Anaconda 发行版。作者在本书前言中介绍了具体的安装方法，安装成功后即可创建 Notebook 编写代码。由于网络问题，建议国内的朋友使用清华大学 TUNA 镜像（https://mirror.tuna.tsinghua.edu.cn/help/anaconda/）下载和更新 Anaconda 集成开发环境。第三种方式适合了解 Docker（https://www.docker.com/）的朋友——可以直接使用 Jupyter 在 GitHub 上的 Docker 镜像（https://github.com/jupyter/docker-stacks），一键安装，省时省力。里面除了标准 Anaconda 开发环境，还支持 Spark、TensorFlow 的 Notebook 开发环境。

本书作者 Jake VanderPlus（GitHub 账号为 @jakevdp）目前是华盛顿大学 eScience 学院物理科学研究院院长。他既是一位天文学家，也是一位会议演讲达人，活跃于历年的 PyData 会议，尤其擅长 Python 科学计算与数据可视化。Jake 在数据可视化方面颇有建树，创建了 altair、mpld3、JSAnimation 可视化程序库，同时为 NumPy、Scikit-Learn、Scipy、Matplotlib、IPython 等著名 Python 程序库做了大量贡献。我在学习贝叶斯估计时，从他 2014 年的系列博文“Frequentism vs Bayesianism”（频率主义与贝叶斯主义）中获益颇多。2015 年，听说他要在 O'Reilly 出版《Python 数据科学手册》一书，一直持续关注，正式版终于在 2016 年年底发布。期间，他在 O'Reilly 做了一些 Python 数据科学教程（基于 O'Reilly 的 Atlas 平台创建 Notebook，代码可在线运行），介绍了 Pandas、Seaborn、Matplotlib 等工具。2017 年 2 月，他在 YouTube 发布了一组视频，通过美国西雅图市弗雷蒙特桥上穿行的自行车统计数据，演示了 Python 数据科学编程的最佳实践，包括在 Notebook 中编码、重构、测试、发布程序的技巧，可谓短小精悍。此次有幸能翻译大神的作品，与有荣焉。首先感谢图灵社区，尤其感谢朱巍老师的再次大力支持，夏静文老师、刘美英老师和岳新欣老师的细致审校。也要感谢一起合作过的小伙伴们，促使我们再次翻译数据科学的基础教程，让更多用 SQL、Excel、Matlab、SPSS 的分析师了解 Python 数据科学的工具，用数据更自由地表达，讲出更精彩的故事。

前言

什么是数据科学

这是一本介绍 Python 数据科学的书。可能话音未落，你脑海中便会浮现一个问题：什么是**数据科学**（data science）？要给这个术语下个定义其实很困难，尤其它现在还那么流行（自然也众口难调）。批评者们要么认为它是一个多余的标签（毕竟哪一门科学不需要数据呢），要么认为它是一个粉饰简历、吸引技术招聘者眼球的噱头。

我认为这些批评都没抓住重点。如果去掉浮华累赘的装饰，数据科学可能算是目前为止对跨学科技能的最佳称呼，在工业界和学术界的诸多应用中扮演着越来越重要的角色。跨学科是数据科学的关键；我认为，如今对数据科学最合理的定义，就是 Drew Conway 于 2010 年 9 月在自己的博客上首次发表的数据科学维恩图（如图 0-1 所示）。

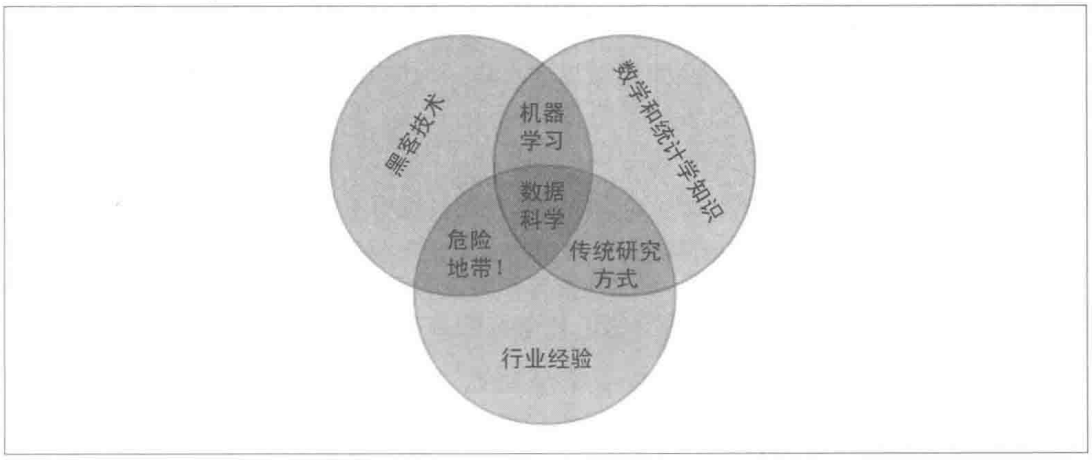

图 0-1：Drew Conway 的数据科学维恩图

虽然图中交错的标签看着跟开玩笑似的，但我还是认为这幅图道出了“数据科学”的真

谛：它是一个**跨学科**的课题。数据科学综合了三个领域的能力：**统计学家**的能力——能够建立模型和聚合（数据量正在不断增大的）数据；**计算机科学家**的能力——能够设计并使用算法对数据进行高效存储、分析和可视化；**领域专家**的能力——在细分领域中经过专业训练，既可以提出正确的问题，又可以作出专业的解答。

我希望你不要把数据科学看作一个新的知识领域，而要把它看成可以在自己熟悉的领域中运用的新能力。无论你是汇报竞选结果、预测股票收益、优化网络广告点击率、在显微镜下识别微生物、在太空中寻找新天体，还是在其他与数据相关的领域中工作，本书都会让你具备发现问题、解决问题的能力。

目标读者

无论是在华盛顿大学教书时，还是在各种科技会议上演讲时，经常有人问我这样一个问题："我应该怎样学习 Python 呢？"问这个问题的都是有技术能力的学生、程序员或科研人员，他们通常都具备很强的编程能力，善于使用计算机和数学工具。他们中的大多数人其实并不想学习 Python 本身，而是想把它作为数据密集型任务处理和计算机科学的工具来使用。虽然网上已经有很多教学视频、博客和教程，但是我一直觉得这个问题还缺少一个令我满意的答案——这就是创作本书的缘由。

这并不是一本介绍 Python 和编程基础知识的书。它假设读者已经熟悉 Python 的基本语法，包括定义函数、分配变量、调用对象方法、实现程序控制流等基本能力。这本书将帮助 Python 用户学习如何通过 Python 的数据科学栈——包括 IPython、NumPy、Pandas、Matplotlib、Scikit-Learn，以及其他相关的程序库——高效地存储、处理和分析数据。

为什么用Python

Python 作为科学计算的一流工具已经有几十年的历史了，它还被应用于大型数据集的分析和可视化。这可能会让 Python 早期的创导者感到惊奇，因为这门语言一开始并不是为数据分析和科学计算设计的。Python 之所以能在数据科学领域广泛应用，主要是因为它的第三方程序包拥有庞大而活跃的生态系统：NumPy 可以处理同类型（homogeneous）数组型数据、Pandas 可以处理多种类型（heterogeneous）带标签的数据、SciPy 可以解决常见的科学计算问题、Matplotlib 可以绘制可用于印刷的可视化图形、IPython 可以实现交互式编程和快速分享代码、Scikit-Learn 可以进行机器学习，还有其他很多工具将在后面的章节中介绍。

如果你需要一个 Python 入门教程，那么我推荐你阅读本书的姊妹篇 *A Whirlwind Tour of the Python Language*。这个简短的教程介绍了 Python 的基本特性，目的是让熟悉其他编程语言的数据科学家快速学习 Python。

Python 2与Python 3

本书使用 Python 3 的语法，其中包括了 Python 2.*x* 版本不兼容的语法技巧。虽然 Python 3.0 在 2008 年就发布了，但并没有被快速采用，尤其是在科学和 Web 开发领域。这主要是

因为许多第三方程序库和工具包需要时间来兼容 Python 的新版本。然而，从 2014 年初开始，数据科学领域最重要的工具的稳定版本都已经同时兼容 Python 2 和 Python 3，因此本书将使用新版本的 Python 3 语法，不过其中的大部分代码示例无须调整也可以在 Python 2 中运行。如果遇到了 Python 2 不兼容的地方，我会尽量详细说明。

内容概览

本书每一章都重点介绍一到两个程序包或工具，它们是 Python 数据科学的基础。

IPython 和 Jupyter（第 1 章）
: 这两个程序包为许多使用 Python 的数据科学家提供了计算环境。

NumPy（第 2 章）
: 这个程序库提供了 `ndarray` 对象，可以用 Python 高效地存储和操作大型数组。

Pandas（第 3 章）
: 这个程序库提供了 `DataFrame` 对象，可以用 Python 高效地存储和操作带标签的 / 列式数据。

Matplotlib（第 4 章）
: 这个程序库为 Python 提供了许多数据可视化功能。

Scikit-Learn（第 5 章）
: 这个程序库为最重要的机器学习算法提供了高效整洁的 Python 版实现。

Python 数据科学（PyData）世界里当然不只有这五个程序包；相反，情况是日新月异的。因此，我在每章结尾都列举了用 Python 实现的其他有趣的图书、项目和程序包的参考资料。不过这五个程序包是目前在 Python 数据科学领域中完成大部分工作的基础，即使生态系统在不断成长，我仍然觉得它们五个非常重要。

使用代码示例

本书的补充材料（代码示例、图像等）都可以在 https://github.com/jakevdp/PythonDataScienceHandbook 下载。本书是要帮你完成工作的。一般来说，如果本书提供了示例代码，你可以把它用在你的程序或文档中。除非你使用了很大一部分代码，否则无须联系我们获得许可。比如，用本书的几个代码片段写一个程序就无须获得许可，销售或分发 O'Reilly 图书的示例光盘则需要获得许可；引用本书中的示例代码回答问题无须获得许可，将书中大量的代码放到你的产品文档中则需要获得许可。

我们很希望但并不强制要求你在引用本书内容时加上引用说明。引用说明一般包括书名、作者、出版社和 ISBN，比如“*Python Data Science Handbook* by Jake VanderPlas (O'Reilly). Copyright 2017 Jake VanderPlas, 978-1-491-91205-8”。

如果你觉得自己对示例代码的用法超出了上述许可的范围，欢迎你通过 permissions@oreilly.com 与我们联系。

试读结束：需要全本请在线购买： www.ertongbook.com

软件安装注意事项

安装 Python 和科学计算程序库的方法其实很简单，下面列举一些在安装软件时的注意事项。

虽然安装 Python 的方法有很多，但是在数据科学方面，我推荐使用 Anaconda 发行版[1]，Windows、Linux 和 Mac OS X 操作系统的安装和使用方式类似。Anaconda 发行版有两种。

- Miniconda（http://conda.pydata.org/miniconda.html）只包含 Python 解释器和一个名为 conda 的命令行工具。conda 是一个跨平台的程序包管理器，可以管理各种 Python 程序包，类似于 Linux 用户熟悉的 apt 和 yum 程序包管理器。
- Anaconda（https://www.continuum.io/downloads）除了包含 Python 和 conda 之外，还同时绑定了四五百个科学计算程序包。由于预安装了许多包，因此安装它需要占用几个吉字节的存储空间。

如果安装了 Miniconda，所有程序包（包括 Anaconda）都可以手动安装。因此，我推荐先安装 Miniconda，其他包视情况安装。

首先，下载并安装 Miniconda 程序包（确认你选择的是适合 Python 3 的版本），然后安装本书的几个重要程序包。

```
[~]$ conda install numpy pandas scikit-learn matplotlib seaborn ipython-notebook
```

本书还会使用其他更专业的 Python 科学计算工具，安装方法同样很简单，就是 `conda install` *程序包名称*。关于 conda 的更多信息，包括 conda 虚拟环境（强烈推荐）的创建和使用，请参考 conda 在线文档（http://conda.pydata.org/docs/）。

排版约定

本书使用了下列排版约定。

- **黑体**
  表示新术语或重点强调的内容。
- 等宽字体（`constant width`）
  表示程序片段，以及正文中出现的变量、函数名、数据库、数据类型、环境变量、语句和关键字等。
- 加粗等宽字体（**`constant width bold`**）
  表示应该由用户输入的命令或其他文本。
- 等宽斜体（*`constant width italic`*）
  表示应该由用户输入的值或根据上下文确定的值替换的文本。

注 1：中国大陆用户请使用清华大学 TUNA 镜像（https://mirror.tuna.tsinghua.edu.cn/help/anaconda/）。——译者注

O'Reilly Safari

Safari（原来叫 Safari Books Online）是面向企业、政府、教育从业者和个人的会员制培训和参考咨询平台。

我们向会员开放成千上万本图书以及培训视频、学习路线、交互式教程和专业视频。这些资源来自 250 多家出版机构，其中包括 O'Reilly Media、Harvard Business Review、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Adobe、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett 和 Course Technology。

更多信息，请访问 http://oreilly.com/safari。

联系我们

请把对本书的评价和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

O'Reilly 的每一本书都有专属网页，你可以在那儿找到本书的相关信息，包括勘误表、示例代码以及其他信息。本书的网站地址是：

http://bit.ly/python-data-sci-handbook

对于本书的评论和技术性问题，请发送电子邮件到：bookquestions@oreilly.com

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：

http://www.oreilly.com

我们在 Facebook 的地址如下：

http://facebook.com/oreilly

请关注我们的 Twitter 动态：

http://twitter.com/oreillymedia

我们的 YouTube 视频地址如下：

http://www.youtube.com/oreillymedia

电子书

如需购买本书电子版，请扫描以下二维码。

目录

试读结束：需要全本请在线购买： www.ertongbook.com