Springer

Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications

Python 数据科学导论 概念、技术与应用

[西] 劳拉·伊瓜尔（Laura Igual）
桑蒂·塞　吉（Santi Seguí）　等著
章宗长　王艺深　等译

机械工业出版社
CHINA MACHINE PRESS

Springer

Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications

Python 数据科学导论

概念、技术与应用

[西] 劳拉·伊瓜尔（Laura Igual）
桑蒂·塞　吉（Santi Segu
章宗长　王艺深

机械工业出版社
CHINA MACHINE PRESS

本书通过理论与实践相结合的方式来阐述数据科学的一系列重要概念及算法，以使读者学会如何管理并利用数据。

本书共有11章，第1章概要地介绍了数据科学的现状并给出了一些使用本书的建议；第2章介绍了Python语言数据科学生态系统，涉及NumPy、SciPy和Pandas等热门第三方库；第3~7章着重讲解了统计学和机器学习的知识，涉及描述统计学、统计推断、监督学习、回归分析、无监督学习等主题；第8~10章详细介绍了数据科学的一些主要应用，如网络分析、推荐系统和情感分析；第11章介绍了并行计算及性能优化方法。

本书既可作为数据科学初学者的入门书籍，也可作为高等院校相关专业学生的参考书。

图书在版编目（CIP）数据

Python 数据科学导论：概念、技术与应用 /（西）劳拉·伊瓜尔（Laura Igual）等著；章宗长等译 . —北京：机械工业出版社，2018.8

书名原文：Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications

ISBN 978-7-111-60464-8

Ⅰ . ① P…　Ⅱ . ①劳… ②章…　Ⅲ . ①软件工具 – 程序设计　Ⅳ . ① TP311.561

中国版本图书馆 CIP 数据核字（2018）第 156654 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）
策划编辑：刘星宁　　责任编辑：刘星宁
责任校对：张　薇　　责任印制：常天培
北京富博印刷有限公司印刷厂印刷
2018年8月第1版第1次印刷
184mm × 240mm · 12 印张 · 4 插页 · 268 千字
0 001— 4000 册
标准书号：ISBN 978-7-111-60464-8
定价：59. 00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务		网络服务	
服务咨询热线：	010-88361066	机 工 官 网：	www.cmpbook.com
读者购书热线：	010-68326294	机 工 官 博：	weibo.com/cmp1952
	010-88379203	金　书　网：	www.golden-book.com
封面无防伪标均为盗版		教育服务网：	www.cmpedu.com

译者序

数据科学是一门新兴的学科，但它与我们的距离并非遥不可及，我们每时每刻都在使用数据科学产品。例如，微博通过网络分析向我们推荐可能感兴趣的人；淘宝通过推荐系统实现商品的精准展示；Facebook 通过情感分析来预测美国大选的结果等。这些应用都是基于数据科学来实现的。

人类社会正在步入智能时代，大数据是智能革命中不可或缺的驱动力。随着各行各业生成的数据越来越多，需要对海量数据进行有效的管理和利用，数据科学是在这种社会大背景下诞生的一门应用性学科。作者为了让更多人学会管理和利用数据而编写了本书。

在第 1 章，作者对数据科学的现状及本书的使用方法进行了简要阐述。学习数据科学需要一定的理论基础。如果读者此前没接触过统计学和机器学习方面的内容，也不必担心。本书在第 3、4 章及第 5~7 章分别对统计学和机器学习进行了概要介绍，涉及描述统计学、统计推断、监督学习、回归分析、无监督学习等主题，并结合实际案例来加深读者对这些知识的理解。

除了理论，学习数据科学离不开编程实践。本书的所有代码均使用 Python 语言编写。Python 语言简洁优美、功能强大、可读性强，对初学者非常友好。在第 2 章，作者对常用的 Python 语言数据科学工具箱进行了介绍，包括 NumPy、SciPy 和 Pandas 等。但是本书并没有讲解 Python 语言的基本语法，所以建议没有这方面编程经验的读者在学习本书之前，先阅读一本入门书籍。

数据科学有广泛的应用场景。本书在第 8~10 章分别介绍了网络分析、推荐系统和情感分析这些常见的数据科学应用。为满足数据科学对计算机性能的需求，第 11 章介绍了并行计算及性能优化方法。

本书既可作为数据科学初学者的入门书籍，也可作为高等院校相关专业学生的参考书。

本书的翻译工作主要由章宗长、王艺深完成，参与本书翻译工作的还有陈子璇、姜冲、陈浩然、沈永亮、王泽宇。此外，感谢机械工业出版社的刘星宁编辑在整个翻译工作中给予的支持和理解。

由于译者水平有限，书中难免会出现错漏之处，恳请读者朋友批评指正。意见和建议请发至 zzzhang@suda.edu.cn，我们不胜感激。

译　者

于苏州大学

原书前言

本书的主题范围

在这个时代，来自不同领域的大量信息被收集和存储，其分析和价值提取已成为公司和社会普遍关注的课题之一。需要多学科团队共同设计方案来解决数据带来的新问题。计算机科学家、统计学家、数学家、生物学家、记者和社会学家以及其他许多人现在一起工作，以便从数据中提供知识。这个新的跨学科领域被称为数据科学（data science）。

任何数据科学都涉及提出正确的问题、收集数据、清洗数据、生成假设、做出推断、可视化数据和评估解决方案等环节。

本书的组织和特点

本书是对数据科学的概念、技术和应用的介绍。内容侧重于数据分析，涵盖统计学和机器学习的概念，图像分析技术和并行编程技术以及推荐系统或情感分析等应用。

本书所有章节都通过使用真实数据的实际案例来阐述新概念。本书使用了欧盟统计局、不同的社交网络以及 MovieLens 等公共数据库。有关数据的具体问题在每章中都有提出。这些问题的解决方案是使用 Python 编程语言实现的，并在代码框中进行了恰当的展示。这使得读者可以通过解决问题来学习数据科学，做到举一反三。

本书不打算涵盖整套数据科学方法，也不提供完整的参考文献。目前，数据科学是一个日益增长的新兴领域，因此我们鼓励读者使用网络中的关键词来寻找具体的方法和文献。

目标读者

本书面向高年级本科生和一年级的工科研究生。此外，本书还面向参加继续教育短期课程的专业人员和来自不同领域的自学研究人员。

计算机科学、数学和统计学的基本知识是必需的。有 Python 代码编程背景学习起来会更轻松。但是，即使读者不熟悉 Python，也不是问题，因为在短时间内掌握 Python 的基础知识是可行的。

材料的先前用途

本书所提供材料的一部分已用于巴塞罗那大学“数据科学和大数据”（Data Science and Big Data）的研究生课程。本书所有的贡献者都参与了这门课程。

本书的使用建议

本书可被用于任何入门的数据科学课程。采用基于问题的方法来引入新概念对初学者

来说是有帮助的。针对不同问题实现的代码解决方案对学生来说是一种很好的练习。而且，当学生面对更大的项目时，这些代码可以作为基准。

配套资源

本书附带一套 IPython 笔记本，其中包含解决本书实际案例所需的所有代码。笔记本可以在以下 GitHub 库中找到：https://github.com/DataScienceUB/introduction-datascience-python-book。

致谢

我们感谢所有的贡献者：J. Vitrià、E. Puertas、P. Radeva、O. Pujol、S. Escalera，L. Garrido 和 F. Dantí。

Laura Igual
Santi Seguí
西班牙巴塞罗那

作者和贡献者简介

作者简介

Laura Igual 博士是巴塞罗那大学数学和计算机科学系的副教授。她于 2000 年获得西班牙瓦伦西亚大学的数学学位，并于 2006 年获得西班牙庞培法布拉大学的博士学位。她特别感兴趣的领域包括计算机视觉、医学成像、机器学习和数据科学。

Laura Igual 博士是第 3、6 和 8 章的合著者。

Santi Seguí 博士是巴塞罗那大学数学和计算机科学系的助理教授。自 2007 年起，他担任了西班牙巴塞罗那自治大学的计算机科学工程师。他于 2011 年获得西班牙巴塞罗那大学的博士学位。他特别感兴趣的领域包括计算机视觉、应用机器学习和数据科学。

Santi Seguí 博士是第 8~10 章的合著者。

贡献者简介

Francesc Dantí 是巴塞罗那大学数学和计算机科学系的兼职教授和系统管理员。他是西班牙加泰罗尼亚大学的计算机科学工程师。他特别感兴趣的领域是 HPC、网格计算、并行计算和网络安全。

Francesc Dantí 是第 2、11 章的合著者。

Sergio Escalera 博士是巴塞罗那大学数学和计算机科学系的副教授。自 2003 年起，他担任了西班牙巴塞罗那自治大学的计算机科学工程师。他于 2008 年获得该校的博士学位。他的研究兴趣包括统计模式识别、视觉对象识别，他特别关注人体姿态恢复和多模态数据的行为分析。

Sergio Escalera 博士是第 4、10 章的合著者。

Lluís Garrido 博士是巴塞罗那大学数学和计算机科学系的副教授。自 1996 年起，他担任了西班牙加泰罗尼亚理工大学的电信工程师。他于 2002 年获得该校的博士学位。他特别感兴趣的领域包括计算机视觉、图像处理、数值优化、并行计算和数据科学。

Lluís Garrido 博士是第 11 章的合著者。

Eloi Puertas 博士是巴塞罗那大学数学和计算机科学系的助理教授。自 2002 年起，他担任了西班牙巴塞罗那自治大学的计算机科学工程师。他于 2014 年获得西班牙巴塞罗那大学的博士学位。他特别感兴趣的领域包括人工智能、软件工程和数据科学。

Eloi Puertas 博士是第 2、9 章的合著者。

Oriol Pujol 博士是巴塞罗那大学数学和计算机科学系的终身副教授。由于在机器学习和计算机视觉方面的工作，他于 2004 年获得西班牙巴塞罗那自治大学的博士学位。他特别感兴趣的领域包括机器学习、计算机视觉和数据科学。

Oriol Pujol 博士是第 5、7 章的合著者。

Petia Radeva 博士是巴塞罗那大学的终身副教授和高级研究员。她在 1989 年毕业于保加利亚索非亚大学应用数学和计算机科学系，并于 1998 年在西班牙巴塞罗那自治大学获得医学影像计算机视觉博士学位。她是 2015 年 Icrea Academia 研究员，综合研究小组“巴塞罗那大学计算机视觉”的负责人以及计算机视觉中心 MiLab 的负责人。她目前的研究兴趣是开发基于学习的计算机视觉、深度学习、自我中心视觉、生命记录和数据科学方法。

Petia Radeva 博士是第 3、5 和 7 章的合著者。

Jordi Vitrià 博士是巴塞罗那大学数学和计算机科学系的教授。他于 1990 年获得西班牙巴塞罗那自治大学的博士学位。Jordi Vitrià 博士在 SCI 索引期刊上已发表了 100 多篇论文，在计算机视觉、人工智能以及它们在多个领域的应用等方面拥有超过 25 年的经验。他现在是“UB 数据科学集团”的领导者。“UB 数据科学集团”是一家技术转移部门，负责巴塞罗那大学和私人公司之间的合作研究项目。

Jordi Vitrià 博士是第 1、4 和 6 章的合著者。

目 录

第 1 章
数据科学概述

1.1 什么是数据科学

毫无疑问，你已经通过许多方式接触过数据科学。当你使用搜索引擎在网上查找资料或向手机询问方向时，你都是在与数据科学产品进行交互。近些年来，数据科学一直在幕后解决一些我们最常见的日常任务。

大多数为数据科学助力的科学方法并不是新鲜事物，它们已存在很长时间，正等待应用被开发出来。统计学是一门古老的科学，它站在 18 世纪的巨人，诸如皮埃尔·西蒙·拉普拉斯（1749—1827）和托马斯·贝叶斯（1701—1761）等的肩膀上。机器学习更年轻些，但它已经走出婴儿期，可以被认为是一门完善的学科。计算机科学在几十年前改变了我们的生活，且这样的改变仍会继续下去；但是它不能被认为是新的科学。

那么，为什么数据科学在商业评论、技术博客和学术会议中被视为一种新的趋势呢？

数据科学的新颖性并不根植于最新的科学知识，而是根植于在我们社会中由技术演化引起的一个颠覆性的变化：数据化（datification）。数据化是把世界上以前从未被量化的方面转化成数据的过程。在个人层面，商业网络、我们的书单、我们喜欢的电影、我们吃的食物、我们的体育活动、我们的购物清单、我们的驾驶行为等数据化的（datified）概念列表是很长的，且仍在增长。甚至当我们把想法公布到自己喜欢的社交网络上时，它们便被数据化了；在不远的将来，你的所见所得也能通过可穿戴视觉记录设备数据化。在商业层面，公司正在数据化以前被舍弃的半结构化数据，如网络活动日志、计算机网络活动、机械信号等。书面报告、电子邮件、声音记录等非结构化数据正在被存储，目的不仅是为了存档，更是为了分析。

然而，数据化并不是数据科学革命的唯一要素。另一要素是数据分析的民主化。当数据科学还未兴起之时，Google、Yahoo、IBM、SAS 等大公司是该领域仅有的玩家。在本世纪初，那些公司庞大的计算资源使他们能够使用分析技术来利用数据化的优势以开发出创新性的产品，甚至为他们的业务提供决策。今天，那些公司和世界其他地方（公司和人员）之间的分析能力的差距正在缩小。云计算允许任何个人在短的时间内分析海量的数据。数据分析的知识是免费的，提供解决方案所需的大多数关键算法都可以被找到，因为开源开发是这个领域的常态。因此，使用丰富数据进行循证决策的大门实际上是对任何个人或公司敞开的。

数据科学常被看成是一种能从数据中推断出可行见解的方法论。相比于以前的数据分析方法，如商业智能或探索性统计，这是一个微妙但非常重要的差异。研究数据科学是一个有着雄心勃勃目标的任务：从数据中产生信念（belief），并将其作为决策制定的基础。在缺乏数据的情况下，信念是不可知的，在最好情况下的决策是基于最佳实践或直觉的。通过丰富的数据来表示复杂的环境，开辟了应用所有我们所拥有的、关于如何从数据中推断知识的科学知识的可能性。

总体而言，数据科学允许采用四种不同的策略来使用数据探索世界：

1. 探查现实。数据可以通过被动或主动的方法来收集。在后一种情况下，数据代表了世界对我们行动的响应。当我们对后续行动进行决策时，那些响应的分析是非常有价值的。这种策略最好的例子之一是使用A / B测试进行网站开发：按钮最合适的大小和颜色是什么？最好的答案只能通过探查世界来找到。

2. 模式发现。分治法是一种古老的启发式方法，它被用来解决复杂的问题；但如何将这种常识应用于问题中并不总是那么容易。数据化后的问题可以被自动分析以发现有用的模式和自然的群集，这可以大大简化它们的解决方案。使用这种技术来分析用户是今天在程序化广告或数字营销等重要领域的关键组成部分。

3. 预测未来事件。自早期的统计学开始，最重要的科学问题之一就是如何构建强大的数据模型，使之能够预测未来的数据样本。预测分析可以做出响应未来事件的决策，而不仅仅是对其的反应。当然，在任何环境下预测未来是不可能的，总会有不可预测的事件出现；但可预测事件的辨识代表了有价值的知识。例如，通过分析天气、历史销售量、交通状况等数据，预测分析可用来优化零售店员工在接下来一周内的任务。

4. 了解人和世界。目前这个目标已经超出了大多数公司和人员的能力范畴，但是大公司和政府正在把大量的资金投入到自然语言理解、计算机视觉、心理学和神经科学等研究领域。对于数据科学来说，科学地理解这些领域非常重要，因为为了最后做出最优决策，有必要了解驱使人们做出决策和进行行为选择的实际过程。深度学习方法在自然语言理解和视觉对象识别的发展就是这方面研究的一个很好的例子。

1.2 关于本书

数据科学绝对是一门很酷、很时髦的学科，它经常出现在那些非常重要的报纸和电视台的头版头条。从这些报道中可以看出，数据专家是非常稀缺珍贵的资源。在这种情况下，数据科学被看作是一门复杂的、令人畏惧的学科，只有那些为大公司工作的天才们才能掌握。本书的主要目的是通过讲述一系列工具和技术，使拥有基本计算机科学、数学和统计学技能的人员能够完成与数据科学相关的任务，从而揭开数据科学的神秘面纱。

为此，本书是基于以下假设编写的：

• 数据科学是一个复杂的、多维度的领域，它可以从多个角度来研究：道德、方法论、商业模式、处理大数据的方法、数据工程、数据治理等。每个角度都值得进行漫长而有趣的探讨，但本书侧重于分析技术，因为这些技术构成每个数据专家的核心工具箱，是预测未来事件、发现有用模式和探查世界的关键组成部分。

• 需要有一些 Python 编程经历。基于这个因素，本书不会对 Python 语言作介绍。但即使你刚接触 Python，这也不是个问题。在阅读本书之前，你可以从任一 Python 网络教程开始学习。精通 Python 不容易，但对任何人来说，在短时间内掌握 Python 的基本知识是一项可行的任务。

• 数据科学是基于事实讲故事的过程，这种过程需要适当的工具。Python 数据科学工具箱是研究数据科学最成熟的环境之一。通过使用 Anaconda[⊖]，你可以很容易地安装所有你需要的东西。Anaconda 是一款免费的产品，它包含了编程语言 Python，开发和演示数据科学项目的交互式环境 Jupyter 笔记本（Jupyter Notebook），以及进行数据分析所需要的大部分工具箱。

• 边做边学是学习数据科学最好的方法。本书所有的示例代码和数据均可从 https://github.com/DataScienceUB/introduction-datascience-python-book 下载。

• 数据科学能够解决现实世界的问题。所以本书中的所有章节均包含并讨论了使用真实数据的实际案例。

本书包含三类不同的章节。第一类章节介绍了 Python 的扩展。Python 最初的设计中仅有最少量的数据对象（如 int、float、string 等），但在处理数据时，有必要将原始集扩展为更复杂的对象，如数值数组（numpy）或数据框架（pandas）。第二类章节包括了进行统计分析和机器学习的技术和模块。最后的一些章节介绍了数据科学在搭建推荐系统、进行情感分析等几个方面的应用。所选的这些章节提供了数据科学领域的全景图，但我们鼓励读者更深入研究这些主题，并探索本书没有涉及的话题，如大数据分析、深度学习技术和计算机代数、贝叶斯统计等更高级的数学、统计学方法。

致谢　本章和 Jordi Vitrià 共同编写。

⊖ https://www.continuum.io/downloads.

第 2 章
数据专家的工具箱

2.1 引言

本章首先介绍数据专家使用的一些工具。像任何类型的程序员一样，工具箱对于任何数据专家来说都是成功和提升绩效的基本要素。选择合适的工具可以节省大量的时间，从而使我们能够专注于数据分析。

需要决定的最基本的工具是我们将使用哪种编程语言。许多人一生只使用一种编程语言：他们学习的第一种语言，也是唯一的一种。对许多人来说，学习一门新语言是一项艰巨的任务，如果可能，这项任务应该只承担一次。问题在于某些语言是为开发高效的或高产出的代码而设计的，如 C、C++ 或 Java，而其他语言则更侧重于原型代码，其中最著名的就是所谓的脚本语言，如 Ruby、Perl 和 Python。所以，根据你学习的第一种语言来完成（至少）某些任务将会相当的枯燥。使用单一语言的主要问题是，许多基本工具将无法在其中使用，最终你将要么必须重新实现它们，要么必须创建一个桥梁来使用其他语言去完成特定的任务。

总而言之，你必须准备好为每项任务切换到最合适的语言，然后将结果粘合在一起，或者选择一种非常灵活的、有丰富生态系统（如第三方开源代码库）的语言。在本书中，选择 Python 作为编程语言。

2.2 为什么选择 Python

Python[⊖] 是一门成熟的编程语言，对于编程新手来说，它也有着优良特性，这使它成为那些以前从未接触过编程的人的理想选择。其中一些最值得一提的特性包括：代码易读，禁止非强制性的分隔符，使用动态类型和动态内存。Python 是一种解释型语言，所以代码可以立即在 Python 控制台中执行，而不需要编译成机器语言。除了自带的 Python 控制台外，也可以使用其他的交互式控制台，如 IPython[⊜]，它提供了一个更丰富的 Python 代码执行环境。

目前，Python 是最灵活的编程语言之一。使它如此灵活的主要特性之一是它可被视为一种多范型语言。这对于那些使用其他语言编程过的人来说尤其有用，因为他们可以以同

⊖ https://www.python.org/downloads/.

⊜ http://ipython.org/install.html.

样的方式快速地使用 Python 进行编程。例如，Java 程序员使用 Python 会感到很亲切，因为 Python 支持面向对象的范型，或者 C 程序员能使用 Cython 来混合 Python 和 C 的代码。此外，针对有函数式语言（如 Haskell 或 Lisp）编程背景的人，Python 在其核心库中也提供了函数式编程的基本声明。

本书决定使用 Python 语言，因为正如之前所解释的那样，它是一种成熟的编程语言，对于新手来说很容易，并且它能够作为数据专家的特定平台。以上这些，都要归功于其庞大的科学类库生态系统和它充满活力的社区。除了 Python，数据专家还较常使用 R 和 MATLAB/Octave。

2.3 数据专家的基本 Python 库

Python 社区拥有大量成熟的工具箱，是最活跃的编程社区之一。数据专家最常用的 Python 工具箱是 NumPy、SciPy、Pandas 和 Scikit-learn。

2.3.1 数值和科学计算：NumPy 和 SciPy

NumPy[㊀] 是 Python 科学计算的基础工具箱。NumPy 除了能支持多维数组及其基本操作外，还提供了有用的线性代数函数。许多工具箱使用 NumPy 数组表示作为一种有效的基本数据结构。同时，SciPy 提供了一个囊括许多数值算法和信号处理、优化、统计等特定领域工具箱的集合。SciPy 中的另一个核心工具箱是绘图库 Matplotlib。这个工具箱中有许多数据可视化工具。

2.3.2 Scikit-learn：Python 中的机器学习库

Scikit-learn[㊁] 是一个基于 NumPy、SciPy 和 Matplotlib 的机器学习库。Scikit-learn 为数据分析中的分类、回归、聚类、降维、模型选择和预处理等常见任务提供了简单高效的工具。

2.3.3 Pandas：Python 数据分析库

Pandas[㊂] 提供了高性能的数据结构和数据分析工具。Pandas 的关键特征是有一个快速且高效的 DataFrame 对象，其能够通过索引高效地对数据进行处理。DataFrame 结构可被视为一个电子表格，它提供了非常灵活的工作方式。你可以通过重塑（reshaping）、添加（或删除）列或行等你想要的方式来改变数据集。它也提供了能实现聚合、合并、连接数据集等功能的高效函数。Pandas 也有工具来导入或导出不同格式的数据，如逗号分隔值（comma-separated value, CSV）、文本文件、微软 Excel、SQL 数据库和快速的 HDF5 格式文件。在许多情况下，上述格式的数据是不完整的或不是完全结构化的。对于这些情况，Pandas 提供了丢失数据处理和智能数据对齐的功能。此外，Pandas 提供了一个方便的 Matplotlib 接口。

㊀ http://www.scipy.org/scipylib/download.html.

㊁ http://www.scipy.org/scipylib/download.html.

㊂ http://pandas.pydata.org/getpandas.html.

2.4 数据科学生态系统的安装

在开始解决面向数据的（data-oriented）问题之前，需要安装编程环境。需要回答的第一个问题是关于 Python 语言自身。目前，有两个不同版本的 Python，即 Python 2.X 和 Python 3.X。不同版本有重要差异，因而不同版本的代码间没有兼容性，即用 Python 2.X 编写的代码在 Python 3.X 中无法运行，反之亦然。Python 3.X 是在 2008 年末推出的。那时，基于 Python 2.X（Python 2.0 最初在 2000 年推出）的大量代码和工具箱已完成部署。因此，很多科学社区没有立即升级为 Python 3.0，而是继续使用 Python 2.7。到目前为止，几乎所有的库都已经开始支持 Python 3.0，但是支持 Python 2.7 的版本仍在被维护，所以其中任一版本均能选用。但是，仍然有大量的 Python 2.X 的代码不兼容 Python 3.X。在本书的例子中，将使用 Python 2.7。

一旦选好 Python 的版本，接下来要解决的是如何安装 Python 数据科学生态系统，是安装一个个工具箱还是捆绑安装所有（甚至更多）需要的工具箱。对于新手，推荐第二个选项。如果选择第一个选项，那么就必须按照正确的顺序来安装前面章节中所有提到过的工具箱。

然而，如果选择捆绑安装，那么 Anaconda Python 发行版[㊀] 就会是一个好的选择。Anaconda 发行版将把我们所需的所有 Python 工具箱和应用程序集成到一个目录中，不会对已安装在机器上的其他 Python 工具箱产生干扰。它不仅包含了 NumPy、Pandas、SciPy、Matplotlib、Scikit-learn、IPython、Spyder 等核心工具箱和应用程序，还有能完成数据可视化、代码优化和大数据处理等其他相关任务的特定工具。

2.5 集成开发环境

对于程序员和数据专家来说，集成开发环境（integrated development environment, IDE）是一个基础工具。IDE 旨在最大化程序员的产出。因此，这类软件多年来的发展就是为了使编程任务更加容易。所以，选择合适的 IDE 是至关重要的，不幸的是，没有通用的编程环境。最好的解决方案是尝试社区中最受欢迎的 IDE，并在每种情况下选择最合适的。

一般来说，任何 IDE 都有三个基本部件：编辑器、编译器（或解释器）和调试器。一些IDE通过安装特定的语言插件来支持多种编程语言，如Netbeans[㊁] 或Eclipse[㊂] 。其他IDE则只支持某一种语言，甚至是只针对特定的编程任务。无论是在商业领域（PyCharm[㊃] 、WingIDE[㊄] ……），还是在开源领域，都有很多支持 Python 的 IDE。开源社区帮助了 IDE 发展，这样任何人都可以定制他们自己的环境，并与社区的其他成员共享。例如，Spyder[㊅]（科学 Python 开发环境）是根据数据专家的需求而配置的 IDE。

㊀ http://continuum.io/downloads.
㊁ https://netbeans.org/downloads/.
㊂ https://eclipse.org/downloads/.
㊃ https://www.jetbrains.com/pycharm/.
㊄ https://wingware.com/.
㊅ https://github.com/spyder-ide/spyder.