微课
视频版

作者20多年软件开发、大数据实战、企业应用**经验积累**，10万忠实粉丝**读者群**，
继《**Python编程从零基础到项目实战**》之后的**Python进阶力作!**

Python编程
从数据分析到机器学习实践

刘瑜 ⊕ 著

视频讲解 + 示例案例 + 源代码 + 习题实验 + 在线服务

随“三酷猫”一起学Python，去探索机器学习的乐趣

视频讲解：608分钟微视频讲解，实战专家带你畅游Python。
示例案例：“三酷猫”带你进入Numpy、Scipy、Pandas、Matplotlib、Scikit-learn库学习科学计算、数据分析与机器学习。
源码解析：本书提供实例解析的源代码，方便读者学习与操练。
习题实验：每章均配有习题及实验，便于巩固与提高。
在线服务：QQ群、公众号在线服务，让Python学习无后顾之忧。

中国水利水电出版社
www.waterpub.com.cn

Python 编程从数据分析到机器学习实践

（微课视频版）

刘　瑜　著

中国水利水电出版社
www.waterpub.com.cn
·北京·

内 容 提 要

《Python 编程从数据分析到机器学习实践（微课视频版）》是一本基于 Python 语言进行数据分析和机器学习的入门与应用类图书，也是一本兼顾实战要求的视频教程。具体内容包括：Jupyter Notebook 应用，Numpy 科学计算、矩阵、线性代数和高级技术，Matplotlib 基础知识和高级应用，Scipy 基础知识和高级应用，Pandas 基础知识、数据处理和基于时间应用，Scikit-learn 基础知识与应用等。本书突出了代码编写的实战要求，为每一章提供了生动有趣的实践内容，包含了文字处理、图像识别、音频编辑、数据分析及预测等实际案例。本书的编写基于 Python 3.7 的最新版本。另外，本书配备了 608 分钟的微视频讲解、提供完整的源代码及 PPT 课件下载。具体下载方法见“前言”中的相关介绍。

《Python 编程从数据分析到机器学习实践（微课视频版）》适合具有 Python 编程基础的 IT 编程工程师、计算机相关专业的学生、专业科学研究人员、数据工程师、高校老师等使用。本书可作为高校、相关培训机构的教材使用。

图书在版编目（CIP）数据

Python 编程从数据分析到机器学习实践 : 微课视频版 / 刘瑜著. -- 北京 : 中国水利水电出版社, 2020.2

ISBN 978-7-5170-8152-4

Ⅰ. ①P… Ⅱ. ①刘… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.561

中国版本图书馆 CIP 数据核字（2019）第 248847 号

书　　名	Python 编程从数据分析到机器学习实践（微课视频版） Python BIANCHENG CONG SHUJU FENXI DAO JIQI XUEXI SHIJIAN
作　　者	刘瑜　著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路 1 号 D 座　100038） 网址：www.waterpub.com.cn E-mail：zhiboshangshu@163.com 电话：（010）62572966-2205/2266/2201（营销中心）
经　　售	北京科水图书销售中心（零售） 电话：（010）88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	北京智博尚书文化传媒有限公司
印　　刷	三河市龙大印装有限公司
规　　格	190mm×235mm　16 开本　28.25 印张　697 千字
版　　次	2020 年 2 月第 1 版　2020 年 2 月第 1 次印刷
印　　数	0001—5000 册
定　　价	99.80 元

凡购买我社图书，如有缺页、倒页、脱页的，本社营销中心负责调换

前　言

Preface

进入21世纪后，随着大数据、人工智能、物联网、云平台等新技术的发展，利用各种各样的数据进行科学分析，成为可能。而这里最引人关注的是基于Python语言的各种分析工具，这也是本书将要介绍的主要工具，它们可以轻而易举地解决网页、表格、图片、音频、视频等数据的处理和应用，不再受关系数据和非关系数据的制约，并且，大多数情况下在普通计算机上就可以执行。这是一件非常令人兴奋的事情！我们可以借助这些工具做数据工程师、数据分析及研究科学家、人工智能专家所做的工作。这在以前是不可想象的，总觉得那是一件非常遥远或高不可攀的事情，如今一切皆有可能。

一、本书主要涉及工具

1．Jupyter Notebook

该工具的前身Ipython是科学家科学计算可视化操作的专业工具，如今被普遍应用到数据处理及分析当中。本书主要介绍其后续项目Jupyter Notebook的使用功能。

2．Numpy

Numpy是Python技术体系下科学计算的基础工具，具有基石的作用，其他很多工具都是基于该工具进行功能扩展的。

3．Scipy

Scipy是在Numpy基础上开发而成的，是一种高级科学计算工具，除了继承Numpy大量的功能外，还拥有一系列高级计算功能。

4．Pandas

Pandas 是进行数据处理和分析的主流技术工具，尤其擅长二维表格的数据处理，业内人士认为Pandas将二维数据处理技术用到了极致。

5．Matplotlib

对数据进行图像可视化处理的首选工具，其提供了丰富多彩的二维、三维处理功能，可以看作Numpy、Scipy、Matplotlib的组合，可以替代MATLAB的相关功能。

6．Scikit-learn

Scikit-learn 是人工智能入门的首选工具，可以借助它深入理解人工智能的一些基础知识和原理，而且该工具相对简单易学，所提供的功能强大、案例丰富。

当然，上述工具都是免费的、开源的，用户可以直接进行商业应用。

二、本书构建思路

1．让读者相对容易地入门

本书可以使读者方便地入门，并能持续地深入学习。基于此，本书对基础知识进行了详细讲解，并通过图、表、提示等技巧使读者更加容易接受；另外，本书遵循由浅入深、层层推进的原则对知识点进行了深入浅出的分析；本书还突出了代码示例、案例的作用。

2．让读者有相对清晰的知识结构

由于科学计算涉及工具众多，如何解释清楚它们之间的关系，并说明工具本身的知识范围，是一件非常具有挑战意义的事情。本书在这方面做了很多工作，可以让读者在一步步掌握知识点的同时，也能很清晰地掌握相关知识结构。

3．让读者能具备一定的实战能力

学完本书的读者可以具备基本扎实的数据处理能力，这也是作为数据工程师所必需具备的。读者在掌握了基本数据处理能力并有一定实操经验后可以比较顺利地进入互联网企业、大数据中心等进行就业，当然，也可以进一步深入学习和研究。

4．让读者觉得本书有趣味性

为了增加趣味性，本书引入“三酷猫”的故事，“三酷猫”是电影《九条命》主题曲 *Three Cool Cats* 的中文翻译。因为作者和作者的孩子都喜欢那只可爱的猫咪。

三、本书适合读者对象

（1）高校学生。本书主体定位于科学计算入门教材，主要面向高校中已具有 Python 编程基础，并立志于从事数据分析、人工智能研究的在校学生。

（2）IT 行业编程人员。本书还适合已经掌握 Python 语言，并且想在数据工程、人工智能应用方面有所了解或发展的技术人员。

（3）相关培训机构的教师和学员。本书可作为相关培训机构的培训教材使用。

（4）专业科研机构的研究人员、高校教师。

（5）编程爱好者。

四、本书相关资源及获取方式

（1）为方便读者学习，本书提供视频源文件，方便读者下载后在电脑上学习观看。

（2）为方便读者巩固知识，本书免费提供配套电子版《习题及答案》。

（3）为方便读者实战学习，本书附赠所有章节的案例源代码。

（4）为方便教师教学，本书制作了教学课件 PPT。

资源下载方式

（1）读者通过扫描下面的二维码，关注微信公众号后，输入“Py524”即可获得本书资源下载链接。

（2）读者可以加入本书 QQ 交流群：797965584（若群满，会创建新群，请注意加群时的提示），按群公告中的提示获取本书资源下载链接。

五、作者介绍

刘瑜，具有 20 多年 C、ASP、Basic、Foxbase、Delphi、Java、C#、Python 等编程经验，著有《战神——软件项目管理深度实战》《NoSQL 数据库入门与实践》《Python 编程从零基础到项目实战》，作者是高级信息系统项目管理师、软件工程硕士、CIO、硕士企业导师、协会理事。

六、编写内容约定

（1）虽然本书主体代码是在 Windows 操作系统下通过的测试，但是在 Linux、Mac 操作系统下也可以实现（书中主要通过“说明”给予提示）。

（2）书中代码行首出现的“$”若不作特殊说明，指 Linux 下的命令提示符。

（3）本书主要内容在是 Python 3.7.2 基础上结合最新第三方库来实现的。

（4）编排格式及阅读方式提示。本书所有代码在 Notebook 里执行并通过，采用如下编排风格：

```
s1=[100,100,50]                    #列表对象 s1
s1                                 #执行 s1
[100, 100, 50]                     #列表执行结果
sum(s1)                            #求列表元素和
250                                #求和结果
```

背景颜色，浅灰色为输入并执行的代码，深灰色为执行结果输出。本书对代码进行了注释，除了方便读者阅读和理解外，也是为了尽量利用纸张的空间，在有限的空间提供尽量多的知识内容。所以提醒读者，代码中的注释内容与正文内容同等重要，必须严格阅读和掌握。在涉及数据使用时，尽量在一章节的开始，先实现数据存储及展现的过程，后续同一节的内容，避免反复展现相同的数据内容，这要求读者在同一节能前后对照着阅读代码执行内容。如 s1 数组变量，可以用于同一节的后续不同代码案例，而不用反复定义 s1。

七、习题及实验使用说明

《习题及答案》主要是为高校学生提供知识巩固测验之用。作者将为购买本书作为教材的学校提供习题标准答案（可以通过 QQ 群或微信公众号联系获取）。

实验是针对所有读者的，无论在校学生，还是编程从业人员，均应该认真完成每章所提供的实验任务，以切实掌握每章的核心编程内容（对于实验结果，学校可以从作者处获得标准答案；编程从业人员可以参与 QQ 群讨论和咨询）。

八、致谢

这里先要致谢视频支持老师——屈晓渊，榆林学院副教授，为本书录制了配套视频讲座。

同时感谢其他参与本书编写的人员——裴英尚、董树南、阚伟、刘勇。

本书受作者知识水平局限，虽然尽心尽力确保书中内容的质量，但难免存在疏漏之处。欢迎广大读者提出宝贵意见，谢谢！

编　者

目 录

Contents

第1章 入门准备

Python 语言作为“胶水”语言，“粘连”了大量的第三方库，这些第三方库在科学计算、数据分析、人工智能等方面都发挥了强大的应用功能，其中，包括了 Jupyter Notebook、Numpy、Scipy、Pandas、Matplotlib、Scikit-learn 等。对于已经有 Python 语言基础，想从事数据分析、科学计算、人工智能等工作的读者，本章提供了基础入门知识。

学习内容

- 自然智能、人工智能、机器学习、深度学习等概念
- 工具安装
- 数据
- 对读者的建议
- 公共约定
- 习题及实验

扫一扫，看视频

1.1 基本知识概述

本节介绍目前 IT 行业内比较流行的且跟本书紧密相关的一些背景知识，同时把 Jupyter Notebook、Numpy、Scipy、Pandas、Matplotlib、Scikit-learn 的基本情况及特点进行整体介绍，而本书的主体内容也是围绕上述六大库进行系统讲解和应用的。

1.1.1 背景知识

进入 21 世纪，Python 语言得到业内越来越多人的青睐，在全球范围内程序软件使用排名一路上升。截至 2019 年 2 月，其在 TIOBE 网[①]上排名居于前 3，且仍具有强烈的上升势头。Python 的火爆主要基于其两个优点，一是简单易学，赢得了大量编程初学者的认可；二是其粘连了大量的第三方库，如 Jupyter Notebook、Numpy、Scipy、Pandas、Matplotlib、Scikit-learn 等[②]。借助这些第三方库，Python 可以做（大）数据分析、科学计算、人工智能等相关的工作，而且这些第三方库都是免费的、开源的，这给了很多技术人员、研究人员更多的机会，他们可以更加轻松地学习，并通过上述工具，来解决看起来非常“高、大、上”的事情。

2003 年，两个美国科学家为了解决从网上爬取的以 10 亿页计的海量数据，开始研究大数据技术，并在 2008 年获得了成功[③]。而最近几年，我国各省市也陆续建立了省级或国家级的数据中心，累计的数据达到 TB、PB 甚至 EB 级别。另外，不同企业，甚至个人计算机上也都积累了数字、文字、网页、表格、图片、音频、视频等形式的大量数据。这些堆积的数据需要有专业的工具进行加工分析，从而产生对用户有价值的信息，而这正是 Python 所支持的第三方库所擅长的。

2016 年和 2017 年，阿尔法狗（AlphaGo）连续击败了世界围棋冠军李世石和排名第一的世界围棋冠军柯洁，这也宣告了目前人工智能在该领域彻底击败人类[④]。而阿尔法狗的工作原理是人工智能技术（主要是深度学习技术），要想在人工智能技术方面有所作为，利用基于 Python 语言的第三方库学习，是一个很好的选择。

另外，Python 语言在第三方库的支持下，还实现了强大的科学计算功能，如通过 Python 与 Numpy、Scipy、Matplotlib 的结合，可以替代 MATLAB 的部分功能，这对工程技术人员、科学研究人员来说，是一个极具诱惑的选择。

数学家说这个世界是数字的，今天我们距离可操控的数字世界是如此之近。

自从冯·诺依曼设计并制造出第一台通用电子计算机以来，“0”和“1”展现了这个世界越来越多的信息。从财务表格、数学公式、图片、音频、视频，到机器人智能博弈、音频图像识别、机器翻译、医疗诊断、自动驾驶、指纹和人脸识别等，计算机越来越多地替代人类的各种智能活动。让我们来看看董树南老师亲手研究并制作的“基于数字驱动三维碰撞模型”，如图 1.1 所

① TIOBE 网地址 https://www.tiobe.com/tiobe-index/。

② PyPi 网提供了超过 17 万个基于 Python 的项目（库），其地址为 https://pypi.org/。

③ 刘瑜、刘胜松《NoSQL 数据库入门与实践》(基于 MongoDB、Redis)，2018 年 2 月，第 4 页。

④ 刘瑜《Python 编程从零基础到项目实战》第 375 页。

示，整个模型用数字搭建，再现了数字在三维世界里的仿真功能。

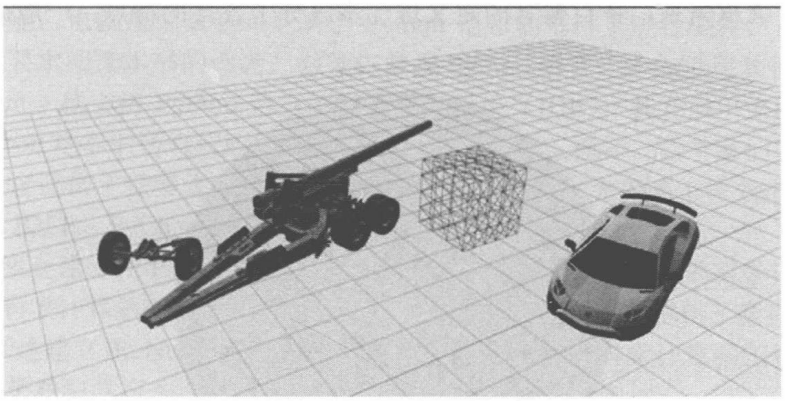

图 1.1　数字驱动三维动态仿真

动态的、随着时间变化的、碰撞的仿真，意味着什么呢？意味着整个世界都可以通过类似数字模型来展示。当读者耐心学完这本书时，大脑里会充满各种数字坐标和公式，也就是说这个世界可以用数字来表达，并可以通过二维、三维视觉来展现。

1.1.2　智能概述

在介绍相关工具前，需要先理清一些概念及关系，方便读者准确理解本书的知识。

自然智能（Natural Intelligence，NI），指人通过大脑的运算和决策产生有价值的行为。这些行为包括了人的大脑思考及决策、耳朵听力及判断、眼睛视觉及判断、鼻子嗅觉及判断、皮肤触觉及判断等，体现在人行为的方方面面。

人工智能（Artificial Intelligence，AI），通过机器替代人，实现人具有的智能行为。这里的机器主要指计算机、数据、相关软件，也可以包括相关的智能终端设备。目前人工智能应用比较成熟的技术方向包括机器博弈（智能机器人）、声音识别、图像图片识别（文字、指纹、人脸等）、传感器等提供数据的分析与预测。人工智能研究的主要学科涵盖计算机科学、信息论、控制论、自动化、仿生学、生物学、心理学、数理逻辑、语言学、医学和哲学等。

机器学习（Machine Learning，ML）是算法和统计模型的科学研究，计算机系统使用它来有效地执行特定任务，无须使用明确的指令，而是依赖于模式和推理。它被视为人工智能的一个子集，也是人工智能的核心。[①]机器学习必须借助数据进行“学习”。机器学习的形式可以分为监督学习、半监督学习、无监督学习、强化学习。

深度学习（Deep Learning，DL）（也称为深度结构化学习或分层学习）是基于学习数据表示的机器学习方法系列的一部分，而不是特定于任务的算法。深度学习受生物神经系统中信息处理和通信模式的启发，但与生物大脑的结构和功能存在差异。目前，深度学习架构，如深度神经网络、深度置信网络和递归神经网络，已应用于计算机视觉、语音识别、自然语言处理、音频识别、社交网络过滤、机器翻译、生物信息学、药物设计和医学图像分析等领域。

[①] Machine Learning, https://en.wikipedia.org/wiki/Machine_learning#cite_note-1。

图 1.2 直观地体现了自然智能、人工智能、机器学习、深度学习之间的关系，它们是包含关系，同时也是部分继承关系，各自概念的定义进一步区分了各自的特点，一般读者对此有印象、会简单区分即可。

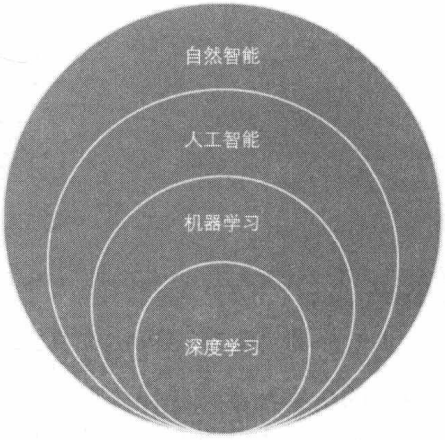

图 1.2 NI、AI、ML、DL 的关系

1.1.3 主要库功能

Python、Jupyter、Numpy、Scipy、Pandas、Scikit-learn、Matplotlib 等主要分别实现什么功能呢？它们之间是什么关系呢？如图 1.3 所示。

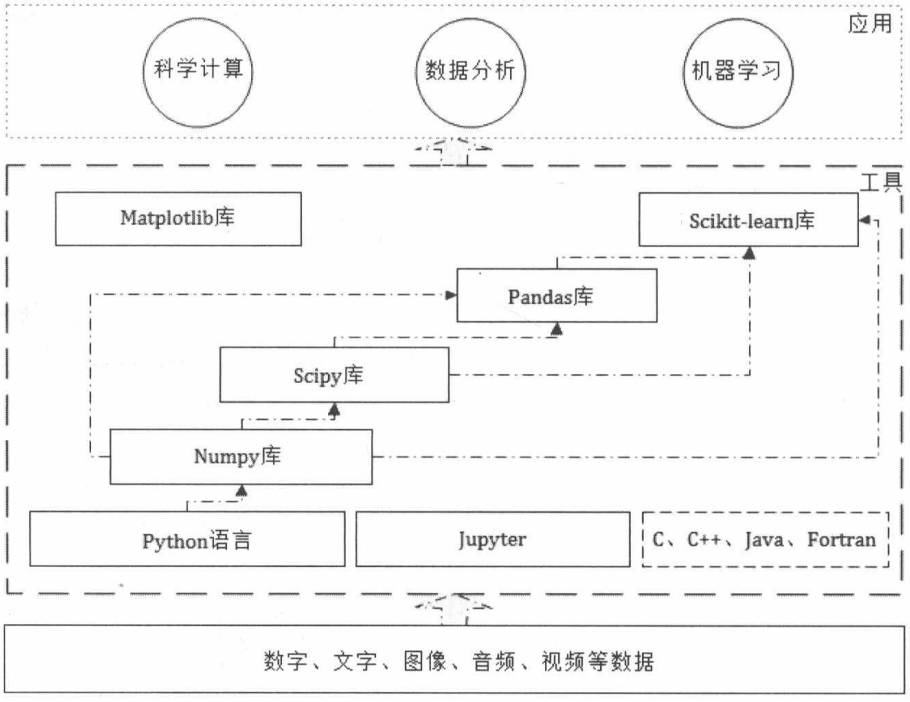

图 1.3 主要功能及相关关系