新时代 • 技术新未来

深入浅出 Python 数据分析

张维元　编著

清華大學出版社
北　京

内 容 简 介

数据时代的来临带动了新一波的智能革命，数据与算法驱动了各个领域的改变。在几个市场热门的讨论议题中，都可以看到数据应用扮演的角色。在面对真实世界的数据时，有许许多多的事情需要考虑。本书试图从最务实的角度开始，结合理论与实践去探索数据科学的真实世界，帮助读者一步一步地培养数据时代下的思维与技术。本书将从基础的 Python 编程开始，以数据分析的流程为主轴一步一步地解析，然后展开介绍数据收集、数据前处理、特征工程、探索式分析等。本书系统性地从函数库开始学习，并拓展到不同的应用场景。

本书实用性强，提供数据分析所必需的编程技能的培训，以及常见第三方软件和库的使用方法；以数据科学家、数据分析师等数据应用工作的实践经验作为培养目标，适合对 Python 与数据分析有兴趣的人阅读。

图书在版编目(CIP)数据

深入浅出 Python 数据分析 / 张维元编著 . —北京：清华大学出版社，2022.3
（新时代 • 技术新未来）
ISBN 978-7-302-57453-8

Ⅰ . ①深… Ⅱ . ①张… Ⅲ . ①软件工具－程序设计－教材 Ⅳ . ① TP311.561

中国版本图书馆 CIP 数据核字 (2021) 第 022582 号

责任编辑：刘 洋
封面设计：徐 超
版式设计：方加青
责任校对：宋玉莲
责任印制：丛怀宇

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 **邮 编：**100084
社 总 机：010-83470000 **邮 购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市国英印务有限公司
经 销：全国新华书店
开 本：187mm×235mm **印 张：**14.25 **字 数：**259 千字
版 次：2022 年 4 月第 1 版 **印 次：**2022 年 4 月第 1 次印刷
定 价：89.00 元

产品编号：087611-01

前　言

为什么要写这本书?

数据时代下，数据将驱动很多领域产生有趣的新进展。数据的使用也变成了一个实用的技能，不再仅限于计算机或统计学行业。在这个技术的推动之下，任何领域的人或多或少都应该要培养数据的思考与使用能力。本书将以浅显易懂的内容与实务场景，逐步培养数据开发者的相应技能。

本书采用 Python 作为主要的程序语言，Python 语言拥有简单、易用、易上手、社区资源丰富等优点，特别在数据分析这个领域，它有很多优秀的第三方套件，能够帮助开发者专注项目本身。本书与其他图书的主要区别是，先系统分析几个数据分析中的主流套件，再进一步将场景拉回实际应用。本书以数据分析的流程为主轴一步一步解析各个环节，包括数据收集、数据前处理、特征工程、探索式分析等，让读者全面、深入、透彻地理解 Python 的数据分析套件，并将其用于实际应用。

本书有何特色?

1. 涵盖Python用于数据分析的主流工具

本书涵盖了数据收集的 Request、BeautifulSoup、Seleium 套件，以及高效能的数学运算工具 NumPy、串起数据与程序分析的 Pandas，还有用于视觉化呈现数据的 Matplotlib。

2. 解析与深入探讨数据分析的步骤

本书将套件与工具应用到不同的使用情境，对数据收集、数据前处理、特征工程、探索式分析等每个环节的实践内容进行深入探讨。

3. 大量的范例与实用代码

本书在每个章节都提供大量的范例作为参考，代码都来自真实的项目。通过对每一段代码的详细了解，读者可以充分理解其作用，并且能够重复地将这些代码应用于项目中。

4. 真实的案例解析

本书最后一章提供了 3 个实战案例。读者可以将本书所介绍的思考方法与实操代码用于真实项目中，从零开始思考解法。

5. 提供完善的技术支持和售后服务

本书提供了技术支持邮箱：v123582@gmail.com。读者在阅读本书的过程中有任何疑问都可以通过该邮箱获得帮助。

本书内容及知识体系

第1篇　数据分析与Python程序语言（第1～2章）

本书第 1 章从数据分析的发展说起，从早期的统计分析到现今的大数据与人工智能发展，介绍计算机科学的演进如何带动数据时代的到来；接着阐述数据项目分析流程应如何制定，以及 Python 与数据分析的关系；最后介绍数据科学家必备的知识与技能。第 2 章介绍与 Python 相关的基础知识，为后续深入学习 Python 打下基础。

第2篇　数据的存取与使用（第3～4章）

第 3 章介绍常见的数据来源与获取方式，归纳成几种常见的形式，即文件、API 与网页爬虫。第 4 章深入讨论网络爬虫的实操技术，从认识 HTTP 网站框开始到爬虫应用，全方位解析网络爬虫相关内容。

第3篇　常见数据分析工具（第5章）

第 5 章介绍 3 个将 Python 用于数据分析的主流套件，分别是高效能的数学运算工具 NumPy、串起数据与程序设计工具 Pandas 和可视化呈现数据工具 Matplotlib，并系统性地介绍这 3 个主流套件的使用方法与其核心目标。

第4篇　数据分析流程（第6～9章）

第 6 ～ 9 章，依照数据分析的流程——“定义问题与观察数据”“数据清理与类型转换”“数据探索与可视化”“特征工程”4 个环节，解析如何使用 Python 与搭配适当的工具进行数据分析。

第5篇　数据分析流程示例应用（第10章）

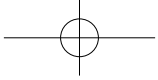
本书第 10 章提供了 3 个项目实战案例，利用几个真实的数据集实践本书前面讨论的各种方法。

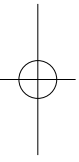
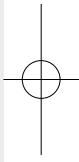
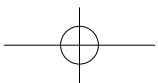
适合阅读本书的读者

- 需要全面学习 Python 数据分析的人员。
- 对数据分析、人工智能有兴趣的人员。
- 希望能够从零开始完成数据分析项目的人员。
- 即将成为与数据分析相关的从业人员。
- 需要一本数据分析训练手册的人员。

阅读本书的建议

- 没有Python基础的读者，建议从第2章开始阅读，先培养基础的程序能力。
- 有一定Python程序基础的读者，可以根据实际情况有重点地选择阅读各个模块和项目案例。
- 对于每一个使用情境和范例代码，自己先思考一下再阅读，并且在每个案例后都尝试其他的优化方式，能够达到最佳的学习效果。
- 可以了解书中的每一个案例，然后套用不同的数据集实现类似的过程。

目　录

第 3 章 数据来源与获取

第 6 章 定义问题与观察数据

第 7 章 数据清理与类型转换

第 8 章 数据探索与可视化

第 9 章 特征工程

第 10 章 示例应用

第 1 章　数据分析与 Python

本章从介绍数据分析的兴起与发展的时代背景说起，介绍计算机技术的演进如何带动数据时代的到来，简要介绍数据项目相关基础知识，并介绍 Python 与数据分析的关系。

本章主要涉及的知识点：

- 数据分析的发展；
- 数据项目团队的组成与数据项目分析流程；
- Python 与数据分析的关系；
- 数据分析人员的学习地图。

1.1 数据分析概述

本节将以数据分析的兴起与发展为重点，介绍数据时代的发展背景与数据分析的几个流派。本节涉及几个重要的关键名词，如大数据、人工智能、机器学习与深度学习等。

1.1.1 数据分析兴起与发展的时代背景

第一次工业革命开始于 18 世纪 60 年代，是一场以大规模工厂化生产取代个体工场手工生产的革命。工业革命以蒸汽机、煤、铁和钢为主要因素，将传统生产模式升级为新的机器制造生产模式，全面地改变了人们的生活。19 世纪 60 年代后期，第二次工业革命开始，人类进入电气时代。第三次工业革命开始于 20 世纪四五十年代，核心技术是电子计算机技术。

电子计算机技术依靠其运算速度快、处理数据量大的优势代替了人类的部分脑力劳动或体力劳动，改变了人类社会的信息处理方式，进而改变了现代社会的运作结构。电子计算机技术的快速发展带动了一大批高新技术的演进。过去几次工业革命都是站在新技术革新的转折点，如今，我们也站在一个新时代——数据时代的浪尖上。

随着计算机技术的发展，数据量快速增长，数据储存成本进一步下降，云端环境逐渐成熟。计算机的计算能力大幅提升，带来的是数据量的快速增长，因此造就了数据分析的新时代思维。具体而言，过去人们使用演绎方法研究科学发展，根据推论求得规律，随着问题的复杂化，人们通过演绎方法解决问题面临瓶颈，于是形成了通过归纳方法来解决问题的观点。因此，人们将数据分析与巨量数据推上了显学。巨量数据分析不同于传统统计抽样方法，它考虑的是数据母体，利用比实证研究更耗费计算成本的数据驱动方法，对数据中挖掘出的数据背后的关系进行全面分析。当前，我们正处于人类有史以来发展最快的时代。基于“数据”与“分析”，我们将迎来一场新的变革。技术驱动的演进，促进经济的结构性改革，我们正在走向一个充满变化的未来。最重要的是，我们必须把握创新的机会，而且是技术驱动创新的机会。

数据时代席卷的不只是信息界。巨量数据带来的是各个领域的改变。例如，Fintech（金融 + 科技）、Growth Hacking（营销 + 科技）、Health Care（医学 + 科技）等都是数据时代跨领域整合的趋势。换句话说，巨量数据 / 数据思维，需要的是一种跨域的宏观视野。从以上这几个逐渐兴起的热门领域，我们就可以看到数据分析的重要性。

1.1.2　什么是数据分析

数据分析（Data Analysis）往往又称数据科学（Data Science），其目标是在数据中找到有价值的规律或特征，是一门利用数据学习的科学。它结合了各种不同的领域，如数学、统计、机器学习、数据可视化、数据库、云计算等。非专业人士能够利用数据分析来理解问题，通过数据的解读与分析来正确地处理数据。数据分析能够用于不同的领域，如教育、金融或商业。

简单来说，数据分析就是“从数据中找出洞见”的一种技术、一种方法。“数据分析”这个名词兴起于 2012—2013 年。随着计算机技术的发展，计算机的存储技术与运算效率都有了巨大的提升，进而带出了“云计算”→“大数据”→“物联网”→“机器学习”→“深度学习”→“人工智能”这一系列技术新浪潮（见图 1.1），而数据分析也是跟着出场的一个新名词。

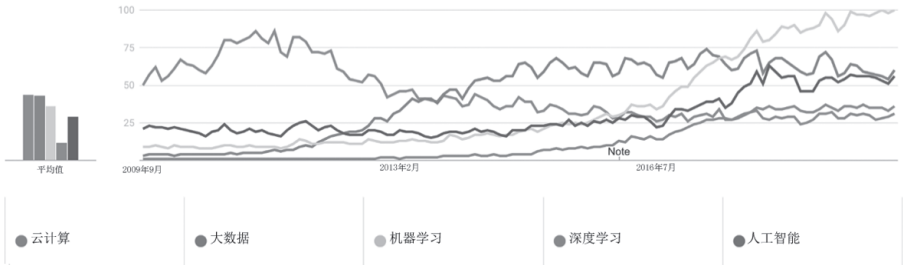


图 1.1　几个数据相关名词的搜寻量变化

事实上，数据分析并不是一个新技术，过去传统的科学研究其实都算是广义的数据分析，但是受限于硬件与计算资源的不足，多半只是统计学上的量化研究。现今的数据分析只是一个升级版，是融合了统计、计算机与数据发展的数据分析。

1.1.3　数据分析的发展方向

数据分析的发展方向有如下两个：

- 由“问题导向”的推论统计和假设检验（Hypothesis Testing）。
- 由“数据驱动”的数据挖掘（Data Mining）或知识发现（Knowledge Discovery）。

推论统计其实就是统计学中的量化研究方法，即人们根据观察或专业知识对一个问题提出虚无假设与对立假设，先证明虚无假设正确，再依照对立假设进行推论。

t- 检验、*Z*- 检验、卡方检验都属于假设检验。假设检验是一种由上而下的研究方法，换句话说，必须先有假设，才能有检验。在真实世界中，提出假设本身是一件困难的工作。

另一个困难点在于很多假设是由具有专业知识背景的科学家提出的，难免会掺杂主观的想法，具有一定的不可控性。假设检验是问题导向的，人们可以尝试去证实或举反证来验证预设的想法。

数据挖掘是另一种由下而上（由数据反过来观察结果）的数据驱动方法。在没有任何假设的情况下，人们可以直接通过数据观察归纳出某些重要的特性。不同于必须要先假设的推论统计，数据挖掘仅通过数据由下而上得到结果。数据挖掘不需要过多的事前假设，也不会有主观意念的影响。

不过数据挖掘就像是大海捞针一样，人们需要在茫茫的数据中找寻特性。可想而知，这种方法需要大量的计算与储存资源。这也是数据挖掘过去一直无法成为主流研究方向的主要原因，但随着计算机科学的发展，更快的计算资源与更大的储存空间让数据挖掘逐渐受到重视。数据挖掘是数据驱动的，人们可以从现有的数据中分析出一些未知的事情。机器学习是数据挖掘的一种方法，这两个名词现在经常混用。

- 统计分析：利用数学模型学习数据，找出一组参数来“描述”数据，目标是找出数据背后的规律，解释数据间的关系。
- 机器学习：通过抽象模型学习拟合数据，着重在学习模型的最佳化过程，目标是达到最好的预测效果。
- 数据挖掘：强调演算方法或步骤，目标是找出数据背后的价值。人们通常会根据所需要的数据选择适合的方法。

数据挖掘与统计分析这两种方法的目标是相近的，只是使用背景有所不同。数据挖掘是计算机领域发展的议题；统计分析是统计学所探讨的领域。无论是数据挖掘，还是统计分析，它们都有一个共同的目标——从数据中学习。这两种方法的目的都是使人们通过处理数据的过程，对数据有更进一步的了解与认识。数据挖掘、大数据、统计学三者的关系如图 1.2 所示。

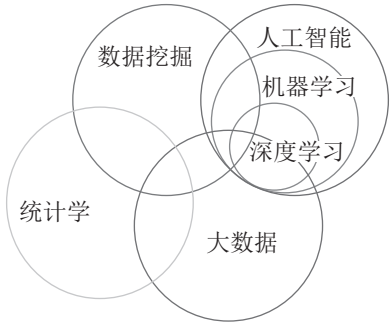


图 1.2　数据挖掘、大数据、统计学三者的关系

统计方法是人们利用方程描述分类问题，为数据找出一个分割线，将结果分成两类的方法。然而，人们利用机器学习的方法找出来的是一圈一圈的等曲线，看起来似乎可以得到更广泛的结果，而不只是简单的分类问题。机器学习是由人工智慧发展而来的领域，通过非规则的方法学习数据分布的关系。统计模型是统计学中描述自变量（特征栏位）与因变量（目标栏位）的关系的模型。统计模型是基于严格的假说限制进行统计检验的（称为假设检验）。假设检验与机器学习方法的不同之处在于机器学习方法是在无假说的情况下对数据进行计算的算法。

基于假设检验的发展，统计模型能找出更贴近现有数据的趋势。然而，预测的目的是找出“未来数据”或所有数据，但假设会使得数据太贴近现有数据（在机器学习中称为过拟和）。严格的假设是统计学习的一把“双刃剑”，就像数据分析中流传的一句话所说的那样：预测模型中较小的假设，预测能力较强（The lesser assumptions in a predictive model，higher will be the predictive power）。

总的来说，数据分析的前身其实就是统计学，随着数据累积才有了大数据，带动了演算法的发展，也就是现在的机器学习与深度学习。现今，数据分析技术正在发展的浪潮上，数据分析的终极目标是利用数据与算法打造一个更智慧的系统，即人工智能。

1.1.4　大数据与厚数据

无论是统计分析还是数据挖掘，数据都扮演着决定性的角色。数据量越大，其所支持的分析模型越完善。如果数据的可用性太低，那么模型再厉害也无法充分发挥作用。所以，数据有两种指标：量与质。

我们把巨量的数据称为大数据，简单的定义如下：当抽样的数量大到接近“母体”时，这类数据就可以称为大数据，带来的效益是大幅降低因为抽样产生的误差。大数据具备 Volume（数据量）、Variety（多元性）、Velocity（即时性）的 3V 特性。

为什么巨量数据是一件重要的事情？迈尔 • 舍恩伯格在《大数据》一书中这样说明：“通过更完整的数据分析，通过接近母体的数据量，可以大幅降低传统抽样所产生的统计误差。”换言之，实现巨量数据需要付出更多、更快的运算机器，所以巨量数据与计算机技术的进步是相辅相成的。不过，数据分析也不尽然要盲目地追求“巨量”这件事。大企业能享有巨量数据的规模优势，但小团队也有成本及创新上的优势，因为速度够快、灵活度高，就算维持小规模，还是能够蓬勃发展的。重要的是，能否掌握数据时代的思维与创新。

从数据可用性角度来看数据，数据分析领域还有另一个值得关注的名词——厚数据。

厚数据由美国社会学者克利福德•格尔茨提出，是指利用人类学定性研究法来定义的数据，数据隐含大量感性的内容。少量的数据能够记载更多的意义，也就是说数据本身具有较大的信息量。厚数据不同于大数据的量化，更多的是数据的质性。

1.1.5 数据挖掘、机器学习与深度学习

1. 数据挖掘

数据挖掘的英文是 Data Mining，其主要的意思是 Mining From Data，即从数据中挖掘金矿。另外，KDD（Knowledge Discovery in Databases）是数据挖掘的另一个常见的同义词。Data Mining 是在 20 世纪 90 年代从数据库领域发展而来的，所以一开始通常用 KDD 这个名称，在知名的学术论坛也称为 SIGKDD。

第一届 SIGKDD 会议讨论了这个问题，即沿用 KDD 还是改名为 Data Mining。会议最终决定这两个名字都保留，KDD 有其科学研究上的含义，而 Data Mining 也适用于产业界。数据挖掘方法主要分为 3 种：关联（Association）法、分类（Classification）法和聚类（Clustering）法。

提到数据挖掘，一定会提到“啤酒尿布”这样的案例。该案例涉及一个经典的数据挖掘算法——关联规则（Association Rule）。因其常用在商品数据上，所以也被称为购物篮数据分析（Basket Data Analysis）。关联规则通过数据间的关系，找出怎样的组合是比较常出现的。关联规则与传统统计的相关性差异在于关联法则更重视关联性。

分类法是数据挖掘与机器学习中的重要算法。分类法主要用于区分数据，判断数据属于哪一个类别，即从原有的已知类别的数据集进行学习，以判断新进的未知类别数据。因为是用已知类别的数据集进行学习，所以分类法也被称为监督式学习（Supervised Learning）。

分类法的用法有两种：分析与预测。

分析：解释模型形成的原因，以了解数据本身的特性及应用。

预测：根据数据的特征及模型预测未来新的数据走向。

分类法可应用在多个领域，如银行用来判断是否发放贷款，医生用来判断某人是否患病等。

聚类法又称丛集法，是相对于分类法的另一种数据挖掘方法。聚类法也是用来区分数据的，它与分类法的差别在于原本的数据都是未经类别区分的。因为是对未知类别的数据集进行区分，所以聚类法也被称为非监督式学习（Unsupervised Learning）。

聚类法通常用于分组。举例来说，一家营销公司想要对不同的用户投放广告，就可

以利用聚类法先对其进行初步的分组。聚类法可以用在市场研究、图形识别等领域。因为数据是由不同的属性所组成的向量，会呈现一个多维的对象，所以人们通常利用“距离”的概念表示相似程度。两笔数据会被表示为两个点，两点之间的距离越大，代表两笔数据越相似，反之越不相似。

当然，随着数据样式的变化，许多进阶用法不断出现，如时间序列分析（Time Series Analysis）和序列模式分析（Sequential Pattern Analysis）。

2. 机器学习

机器学习是从人工智能这门学科延伸出来的分支，主要是通过演算法试图从数据中“学习”到数据的规律，从而预测数据的特性。机器学习、数据挖掘与统计分析是用不同的观点看待“数据”的技术。随着技术的演进，这些技术所涵盖的方法与技术越来越相近。《大演算》一书从不同的思维角度将机器学习流派分成 5 种。

- 符号理论学派：归纳法——从数据反向推导出结论的方法。
- 演化论学派：遗传算法——通过程序模拟遗传演化产出最后的结果。
- 类神经网络学派：通过多层的节点模拟脑神经传导的思考。
- 贝氏定理学派：根据统计学及概率的理论产生模型。
- 类比推理学派：基于相似度判断进行推论学习。

3. 深度学习

深度学习是机器学习的一个支派，也称为进阶的方法，以前也称为类神经网络。目前业界使用较多的是深度学习这个名称。1980 年，多层类神经网络失败，浅层机器学习方法（SVM 等）兴起。直到 2006 年辛顿成功训练出多层神经网络，带动了新一波的深度学习发展。几个数据相关名词的搜寻量变化如图 1.3 所示。

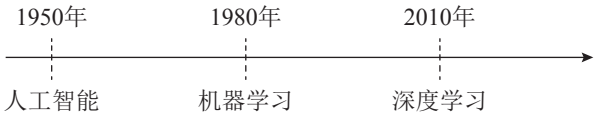


图 1.3　几个数据相关名词的搜寻量变化

1.2　数据项目

本节首先介绍了如何定义一个数据项目，然后介绍了如何构建数据项目团队，最后介绍了一个完整的数据项目的分析流程。

1.2.1 定义数据项目

数据项目的核心在于数据。要解决好问题，相关人员必然要先了解有哪些常见的方法与技术可以应用在数据分析上。下面我们先来快速了解一下数据分析模型。

根据要解决的目标，数据分析模型可分成 3 种类型：监督式学习、非监督学习与半监督学习（Semi-Supervised Learning）。监督式学习指的是数据有一个明确的栏位，用来做预测或分类的目标变量。例如，人们可以利用过去的天气数据，包含“有没有下雨”这个栏位，来预测明天“会不会下雨”。此时，就可以称“下雨与否”为目标变量或统计学上的反应变量。简单来说，就是从过去数据中的其他栏位，找出与“有没有下雨”这个栏位之间的关系，并将其关系套用到一组未知数据“会不会下雨”的其他栏位中，得出“会不会下雨”的预测值。以上这个例子也是监督式学习的典型案例。监督式学习可以想象成根据目标找关系，有一个明确学习的栏位，因此被称为监督式学习。

数据驱动（Data Driven）的方法论是数据分析的一个概念。对于初学者而言，可以先聚焦在特定的问题上讨论，再在一个最小可解上进行优化；当熟悉各种方法论之后，再试着进行更泛化的数据驱动。

1.2.2 数据项目团队的组成

数据分析是一个跨领域的方法论，涉及计算机科学、数学、神经学、心理学、经济学、统计学等领域。换句话说，数据分析并不是单一领域的学科。要完成一个好的数据项目，一个合作无间的数据项目团队必不可少，并且数据项目团队的人员必须同时掌握不同领域的知识，也需要有跨领域合作的思维。数据思维是一种跨领域宏观视野下的思维模式。

另外，跨领域的整合也是一个重要的数据应用关键。无论数据多寡，数据项目都建立在信息、统计、可视化等不同的领域专业上。不过从现实层面上来说，很难有人可以同时具备那么多能力，因此数据项目更需要团队合作。

一个完整的数据项目团队，除了要有特定领域的专家之外，还需要以下 3 种角色：数据科学家（Data Scientist）、数据分析师（Data Analyst）及数据工程师（Data Engineer）。

数据科学家是一个数据项目团队的核心，需要具备综合统筹的能力，包括观察数据、发现问题、组织整个数据团队，可以视为数据项目团队的组长，拥有相关领域的各种技能，哪里需要就往哪里去，能独立实现从分析数据、处理数据到实践应用直到最终产生价值的过程。简单来说，数据科学家就是“用数据解决真实问题的人”。也正因为如此，

数据科学家须具有多元化的能力包括与其他角色沟通的能力，从处理数据的工程到分析数据的建模都需要涉猎，还要拥有洞察力。听起来好像数据科学家什么都要会，不过实际上很难有人可以样样精通，所以团队才显得更为重要。一个好的数据科学家，必须能够驾驭一个数据项目团队。

数据科学家的主要工作是观察数据，从中发现有趣的和需要解决的问题（通常这个过程被称为数据驱动）；然后和工程师商量如何从数据库中建立分析架构；最终，与统计学家用统计模型 / 数据挖掘 / 机器学习的技术进一步分析数据，同时产生一份数据报告。数据科学家可以视为数据分析师的“进阶版”，解决数据分析师难以解决的复杂问题，终极目标是找出藏在数据背后的信息，并根据这些信息预测未来趋势。

数据科学家需要涉猎不同的领域，如基本的数学理论、大数据、程序设计、统计、机器学习与数据可视化等。简单来说，数据科学家需具备一定的综合能力。

数据分析师通常是指对数据进行解释的工作者。其工作步骤是“搜集数据—整理数据—分析数据—产生结果”，最常见的技能是利用常见的商业统计软件（如 SQL、R、SAS、Excel）得出统计报告，并对统计报告进行解释。数据分析师所做的一切都是为了回答问题 [通常这个过程被称为问题驱动（Problem Driven）]。

数据分析师在数据工程师提供的数据基础之上对数据进行探索性分析，目的是找到问题的正确答案，主要工作通常是例行性任务，定期出一个报告来分析季度数据，供管理层决策参考。数据分析师需要具有操作统计软件的基本技能，往往对数字及数据有一定的敏感性。

数据工程师的主要任务是进行数据的架构设计，专注于环境与平台的架设。其所做的一切都是为了让数据可以容易地被使用，负责建立和维持公司数据储存的技术基准，策划硬体和软件的结构，确保数据储存系统可以支持未来的数据量和分析需求，最终目标是把数据整理好，达到降低储存成本、提高查询效率的目的。

随着巨量数据的需求，现在的数据通常存在很多的噪声及干扰，相关人员需要花更多的精力在数据清理上。数据项目团队的主要工作包括收集数据、管理数据，设计一个好的架构以便存取数据，针对用户需求设计产出的数据集，需要具备数据爬虫、数据库架构、数据预处理（数据清理、转换）、数据建模、分散式系统等相关专业知识和技能。

1.2.3　数据项目的分析流程

数据项目的分析流程是：从数据开始，通过一连串的过程发现隐藏在数据中的规

则，利用这些规则完成一些有趣的应用，大致概括为取得数据—数据预处理—数据转换—数据分析—数据解释—发现知识。

图 1.4 所示为乌萨马·菲亚德在 *The KDD Process for Extracting Useful Knowledge from Volumes of Data* 中提到的数据项目的分析流程。这个看似单一的流程，其实需要相关人员不断重复地尝试，一层一层探索，最终才能找到真正具有价值的数据。

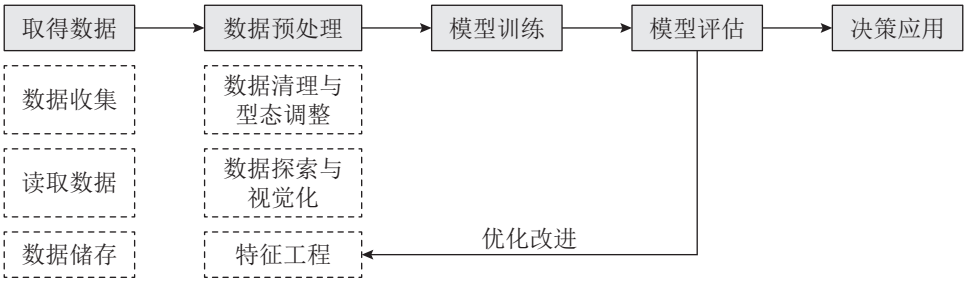


图 1.4　数据项目的分析流程

取得数据是指从原始数据到决定存放数据库的过程，一般来说会涉及数据获取、数据爬虫、数据管理、数据仓储等内容。

数据预处理是指根据规则（API、SQL）从数据库中取出数据集，进行数据清理，处理数据中的噪声或错误信息，或进行多个数据集的整合。

数据转换是指在取得数据集之后，我们经常需要针对分析的具体用法进行调整，将原始数据转换成适合分析模型的格式，如筛选栏位、长宽表转置等。

数据分析可以分为两个阶段，即探索性数据分析（Exploratory Data Analysis）与数据挖掘 / 机器学习。我们可以把探索性数据分析视为一种前期的观察，再经由数据挖掘进行进一步挖掘。

数据解释指人们通常会通过数据可视化的方式及图表方式呈现前述的结果，运用一些可能的原因对数据进行解释，然后把这一整套数据联系起来。

人们一般在数据分析的范畴中把数据清理和特征工程放在数据预处理环节一起讨论，但是在 kaggle 竞赛中，通常会把数据清理视为“处理遗失值”这个动作，也把特征工程视为一个独立过程。常见的特征工程包括特征编码（Categorical Encoding）、特征选取（Feature Selection）、特征降维（Dimensionality Reduction）、正规化（Normalization）/ 标准化（Standardization），如图 1.5 所示。

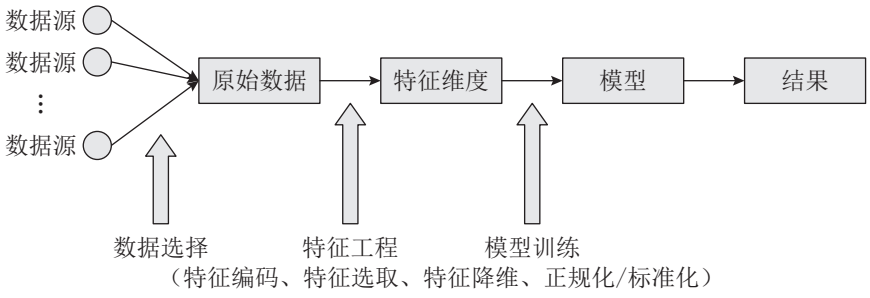


图 1.5　特征工程

1.3　Python 与数据分析的关系

本书采用 Python 作为数据分析的程序语言。本节将介绍为什么选用 Python 进行数据分析及 Python 的数据分析系统。

1.3.1　为什么要用Python进行数据分析

Python 具有简单易用、社区资源丰富两个主要优点。特别是在数据分析这个领域，Python 有很多优秀的第三方库，能够帮助开发者专注于项目本身。本节将从 Python 的基础讲起，运用大量范例说明 Python 的语法与操作。

1.3.2　Python的数据分析系统

Python 拥有完善的数据分析系统，简单分为数据收集、数据预处理、数据可视化、数据模型训练、深度学习、自然语言与文本数据处理。

Request、Beautifulsoup、Scrapy 用于数据收集与网页爬虫，NumPy 与 Pandas 提供了更贴近数据分析的数据结构，SciPy 能够做更复杂的科学计算，Matplotlib 是数据可视化的核心，Seaborn 用于优化样式，Bokeh 和 Plotly 提供了交互的图表。

在模型方面，相关工具有专注于统计的 Statsmodels 和专注于机器学习的 SciKit-Learn，此外也有 xgboost 提供复杂的进阶模型。在深度学习方面，相关工具有 TensorFlow(Theano)、Pytorch、Keras，各自都有拥护者。NLTK、Gensim 用于自然语言与文本数据处理。Python 的数据分析系统如图 1.6 所示。

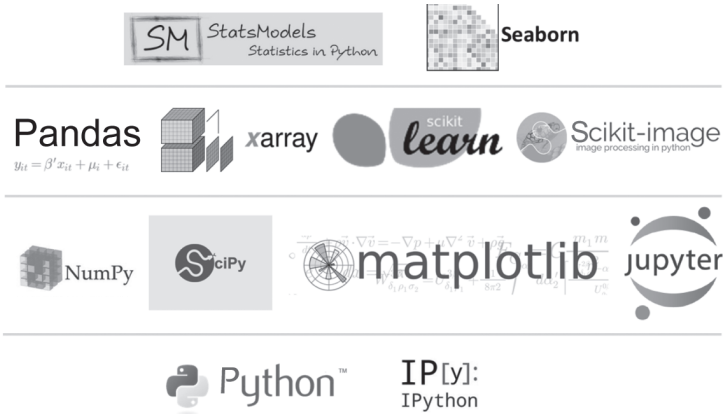


图 1.6　Python 的数据分析系统

1.4　数据分析人员的学习地图

数据分析是一个跨领域的学科，而不是单一领域的学科。数据分析人员必须同时掌握不同领域的知识，需要有跨领域合作的思维。

1.4.1　怎样成为数据分析人员

要完成一个好的数据项目，靠的不能只是一个厉害的强者，而是需要一支合作无间的数据团队。换句话说，只要能够找到一个在团队中的位置，人人都有机会参与数据项目。不过，找到这个位置也不是那么容易的，相关人员需要具备跨领域的复合技能与沟通合作的硬实力。

数据分析技能可以分为 3 种：程序技术、理论分析与专业应用。

- 程序技术：Python、数据清理、数据工程。
- 理论分析：统计分析、数据挖掘、机器学习、深度学习。
- 专业应用：数据分析、数据爬虫、人工智慧。

程序技术员指的是擅长程序开发的人，有比较扎实的工程背景，适合往数据工程方向发展。数理分析能力比较强的人一般具有较好的理论分析能力，其可能具有数学统计或信息背景，可以深入研究数据分析领域或机器学习分析领域。如果一个人写不好程序、也不擅长数学，那么他是不是就难以入门数据分析呢？答案是否定的。拥有某一个领域专业背景的人，也可以往专业应用的方向发展。在擅长领域中积累知识、找出数据分析

可以发挥的空间也是一件很重要的事情。这类人需要的是“相信数据分析的信念”与跨领域沟通的能力。数据分析人员如图 1.7 所示。

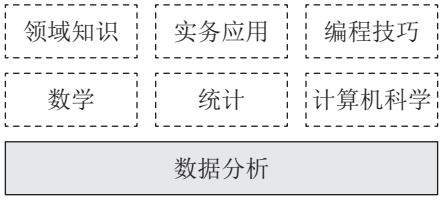


图 1.7　数据分析人员应具备的学科知识

1.4.2　技能树养成之路

数据分析技能那么多，那么技能该怎么学，该从何学起呢？在不同的数据分析教材或课程中，学习地图或课程规划都不太相同，这意味着学习数据分析其实并没有一条绝对的道路。对于新手，建议其首先学好一个程序语言，其次学习相关的系统工具，然后把一个基本的分析过程从头到尾研究透彻，最后就可以摸索自己适合在数据项目团队中的角色了。在学习过程中，数据分析人员应培养与不同角色沟通合作的能力，逐步学习各种数据分析技能，最终成为一个独立的数据分析人员。简单来说，数据分析人员应先学会基本技能，再通过大量的项目掌握完整技能。

那么如何开始数据分析呢？首先挑选一个自己感兴趣的数据集，找出一个可以回答的问题，然后根据这个问题找到一个最基本的原型解（Prototype Solution）来检验这个问题是否可解，通常就是选用最简单的模型当作基础线（Baseline）；接着从基础线开始对解进行优化。一般来说，我们可以从以下两种角度进行优化：更好分的数据和更厉害的模型。

（1）更好分的数据：从数据下手，对数据进行转换与重组，称为“特征工程”。

（2）更厉害的模型：利用复杂的模型，如集成式或深度学习的模型。

除了对模型的准确度进行优化之外，速度与代码质量也是重要的优化指标。

我们可以先利用原型解建立一个基础线的工作流，将预处理与模型比较分为不同的模组；持续从不同的角度进行调整，去观察做哪些动作会造成怎样的优化，最终慢慢提炼出适合数据的手法；建立数据工作流与优化模组之后，就可以快速地将其迁移到类似的数据与问题上；通过反复练习，从每次的调整中让自己更从容地查看数据。