Broadview®
www.broadview.com.cn

# Python绝技

## 运用Python成为顶级数据工程师

黄文青◎编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

# Python绝技

## 运用Python成为

# 顶级数据工程师

黄文青◎编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

## 内 容 简 介

Python 已成为广受数据科学领域欢迎的开发语言。本书契合这一趋势，结合具体的业务场景，从数据思维的角度出发，剖析各业务环节中数据处理的策略、算法，并运用 Python 代码呈现翔实的案例，构建出一个完整的数据分析体系。

在内容的组织和安排上，本书层次分明、详略得当：针对简单的数据分析工作，读者可以先浏览第 1 章至第 3 章；专职从事数据分析的工程师可以通篇阅读本书，以构建数据处理工程的完整知识框架；本书的最后一章针对从事大数据分析的工程师提供了一些常见问题的解决思路和方法。

本书既适合刚接触数据工程的从业人员作为入门参考，也可以帮助具有一定经验的数据工程师搭建知识体系，洞悉业务场景中的数据奥秘，得心应手地运用数据指导业务。



图书在版编目（CIP）数据

Python 绝技：运用 Python 成为顶级数据工程师 / 黄文青编著. —北京：电子工业出版社，2018.6
ISBN 978-7-121-33654-6

Ⅰ. ①P… Ⅱ. ①黄… 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(2018)第 024601 号

责任编辑：刘　皎
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：720×1000　1/16　印张：13.5　字数：232 千字
版　　次：2018 年 6 月第 1 版
印　　次：2018 年 6 月第 1 次印刷
定　　价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

# 好评袭来

数据工程师必备三大技能：数据工程能力、数据分析能力、业务能力，三者相辅相承，缺一不可。本书从这三个方面出发，以简单易懂的 Python 为基础工具，介绍了很多基础知识和工程案例，读起来非常痛快！

——路人甲，公众号“一个程序员的日常”

基于开源的第三方库和活跃的社区支持，Python 数据科学生态体系得到了快速的发展，越来越多的数据工程师选择 Python 作为开发语言。然而，在实际工作中，许多工程师往往侧重于需求实现而忽视对业务的理解。本书针对这一盲区，根据不同的业务场景，从数据的角度梳理、思考问题，并有针对性地阐述了不同的策略、算法和案例。

在跟随本书学习的过程中，我们可以从全局上深入理解数据分析的精髓，并融会贯通——这对于初学者和初级数据工程师的能力提升尤为重要。

——阿橙，“Python 中文社区”公众号主编

数据分析是近年来的热点。几乎所有的互联网公司在产品上都告别了“拍脑袋”做决定的方式，而选择“用数据说话”。因此，也有越来越多的人投入到相关领域当中。Python 作为数据分析的重要语言，受到了广泛关注。然而，对于想要成为数据工程师的人来说，仅完成编程语言的学习是远远不够的。本书恰恰为这一阶段的学习者提供了很好的帮助：从数据分析的基本理论，到业内实践中的分析流程和常用工具，本书均做了较为完整的梳理。

除了理论讲解外，书中还附带了不少分析实例，便于读者理解和演练；此外，作者的行业经验保证了本书的实用性，为入行者指出了清晰的学习路径。

——Crossin，公众号“Crossin 的编程教室”作者、码课创始人

Python 语言继在 Web 大潮之下成为网站快速开发、服务端运维的明星语言之后，随着人工智能技术的飞速发展又迎来了新的一波高潮，成为人工智能领域的首选编程语言。

Python 语言易学易用，有丰富的数据处理包，社区也相当成熟，在数据工程师群体中是非常流行的语言。作为中国最早一批使用 Python 的人之一，看见 Python 逐步从一门小众语言变成推动技术进步的主流语言，很是欣慰。希望此书能够帮助有志于成为顶级数据工程师的朋友更好地掌握这门优秀的语言。

——洪强宁，爱因互动创始人兼 CTO

人工智能是当下最热门的技术领域之一，各大厂商紧锣密鼓进行战略布局：自动驾驶、个人助手、医疗健康、电商零售、金融、教育……如果把人工智能比喻成火箭，那么数据就是燃料。不管你是从事人工智能、机器学习，还是数据分析，都离不开数据，由此诞生了数据工程师的职业。

本书从数据分析、数据挖掘、深度学习等方面介绍了一名数据工程师应该掌握的数据工程的方法和数据分析的思路，书中总结的数学公式和代码实践让原来枯燥的概念变得有滋有味。有志于成为数据工程师的你，细细“品尝”本书，必有收获！

——刘志军，公众号“Python 之禅”

本书内容由浅入深，分别介绍了数据分析的常用工具、Python 在数据分析方面常用的包、如何运用 Python 做基础的统计分析和如何运用 Python 做数据建模……读完以后令人有一种从侏罗纪时代穿梭到未来时代的感觉，信息量很大。

更难得的是作者拥有工业界的背景，这使他可以从实践操作的角度，手把手教您打造一把数据分析的利剑。

一言以概之，本书没有繁杂的数学公式，只有挤不出水的干货。

——挖数，公众号“Washu66”

# 前　言

数据分析、数据挖掘、深度学习及云计算，是当前最热门的技术领域。1830年前后，Gauss、Legendre 等数学家奠基了数据分析的基础理论；1943 年，心理学家 Warren McCulloch 和数理逻辑学家 Walter Pitts 首次提出神经网络；19 世纪80 年代，Hinton、Yann LeCun 等人提出 BP 算法及卷积神经网络；2006 年，深度置信网络研究成果发表。至此，数据建模理论研究的宏观大厦已初见雏形。

历史是如此的巧合，正当需要海量数据集和工程技术方案来处理数据时，云计算应运而生。2003 年，谷歌发表关于 Google File System、Google Bigtable 及 MapReduce 三篇论文，让大数据处理技术风靡全球。以此为基础，2010 年前后，整个云计算的概念及技术体系已经非常完善了。

数据理论的完善、工程技术的发展与无数创意的结合，使得 2010 年以后，整个人类社会进入了“数据时代”。无论是精细化运营，还是人工智能产品，对数据的应用无处不在；无论是政府机构，还是私有的大、中、小型企业，使用数据的热情都达到空前的高度。

2014 年，我加入百度公司，从事大数据处理及数据建模等相关工作。回首过

往，在该领域的几年中，我经历了云计算从雾里看花到如今的方兴未艾；人工智能的初现端倪到如今的高潮迭起。作为一名前线的数据工程师， 我深刻认识到，对我及大多数工程师而言，既无法像Jeff Dean等一样提出经典的大数据计算模型；也无法像Hinton、Yann LeCun一样提出具有深远影响的建模算法。我们所要做的，就是学习与汲取当前的理论与技术，结合应用领域，实现工程应用。这也是我写本书的初衷，希望能从宏观框架上梳理已有的数据分析理论与工程实施技术，并搭建相对系统的知识体系；同时，阐述工作实践中遇到的问题及解决的思路。

Python简洁易懂的语法、丰富的类库、与大数据组件的无缝集成等诸多特点，使其成为数据工程师的首选编程语言。当然，只是掌握Python还完全不足以成为顶级数据工程师，因此，本书介绍数据处理知识体系，并以Python实现相关代码示例，力求让读者能使用Python完成数据处理的各个环节。

本书的第1章和第2章，简要说明了数据处理领域的基本概念，旨在让读者对数据处理工作有宏观的了解。第3章~第5章，主要讲述数据分析理论。笔者按照难易程度，将其划分成三个部分，即基础分析、数据挖掘和深度学习。第6章针对大数据分析，介绍了在工程实施过程中需要用到的工程组件和架构模式，并以一个具体的案例说明整个数据工程的实施流程。

本书适合以下读者阅读：① 对人工智能和云计算感兴趣的读者；② 刚进入数据处理领域的 IT 工程师；③ 希望从宏观上梳理数据处理知识体系的读者；④ 用Excel、SPSS、Python做过数据分析的数据分析师；⑤ 应用过HDFS、Kafka等大数据组件的IT工程师。

本书能够完稿，得益于外界诸多的帮助与指导。感谢数据领域的先驱者Geoffrey Hinton、Yann LeCun、Jeff Dean等，他们的著作是数据时代最重要的理论依据；感谢在百度工作中遇到杨振宇、李华青、王珉然、陈合等许多优秀的同事和领导，在和他们一起试错、交流的过程中，让我取得巨大的进步；感谢本书的编辑刘皎，在她不厌其烦地督促下，本书才从凌乱的只言片语中编辑成书；特别感谢女友孙万兴，在本书的撰写过程中给予的谅解与支持。

# 目　录


1　概述 ..... 1

1.1　何为数据工程师 ..... 1

1.2　数据分析的流程 ..... 3

1.3　数据分析的工具 ..... 11

1.4　大数据的思与辨 ..... 14

2　关于 Python ..... 17

2.1　为什么是 Python ..... 17

2.2　常用基础库 ..... 19

2.2.1　Numpy ..... 19

2.2.2　Pandas ..... 26

2.2.3　Scipy ..... 37

2.2.4　Matplotlib ..... 38

3　基础分析 ..... 43

3.1　场景分析与建模策略 ..... 43

3.1.1　统计量 ..... 43

3.1.2　概率分布 ..... 48

3.2 实例讲解 ..... 55
3.2.1 谁的成绩更优秀 ..... 55
3.2.2 应该库存多少水果 ..... 57
4 数据挖掘 ..... 60
4.1 场景分析与建模策略 ..... 60
4.1.1 分类 ..... 61
4.1.2 聚类 ..... 76
4.1.3 回归 ..... 86
4.1.4 关联规则 ..... 90
4.2 数据挖掘的重要概念 ..... 93
4.2.1 数据预处理 ..... 93
4.2.2 评估与验证 ..... 97
4.2.3 Bagging 与 Adaboost ..... 99
4.2.4 梯度下降与牛顿法 ..... 102
4.3 实例讲解 ..... 105
4.3.1 信用卡欺诈监测 ..... 105
4.3.2 员工离职预判 ..... 110
5 深度学习 ..... 114
5.1 场景分析与建模策略 ..... 115
5.1.1 感知机 ..... 115
5.1.2 自编码器 ..... 119
5.1.3 限制玻尔兹曼机 ..... 123
5.1.4 深度信念神经网络 ..... 127
5.1.5 卷积神经网络 ..... 129
5.2 人工智能应用概况 ..... 137
5.2.1 深度学习的历史 ..... 137
5.2.2 人工智能的杰作 ..... 140
5.3 实例讲解 ..... 146
5.3.1 学习识别手写数字 ..... 146
5.3.2 让机器认识一只猫 ..... 151

6 大数据分析 ..... 160
6.1 常用组件介绍 ..... 160
6.1.1 数据传输 ..... 160
6.1.2 数据存储 ..... 165
6.1.3 数据计算 ..... 174
6.1.4 数据展示 ..... 180
6.2 大数据处理架构 ..... 188
6.2.1 Lambda 架构 ..... 189
6.2.2 Kappa 架构 ..... 192
6.2.3 ELK 架构 ..... 193
6.3 项目设计 ..... 194

参考文献 ..... 202

# 1 概述

首先，我们会从“软实力”与“硬实力”两个方面，介绍一名数据工程师应该具备的能力，并以“能力图谱”的方式列出数据工程领域的知识体系。本书正是围绕这些知识点，逐层细化讲解的。其次，本书会总结数据处理的一般流程：明确目标、确定方案、数据整理、建模分析、结果验证、总结展现，继而对各环节的具体工作进行详尽说明；并从易用性、适用领域等多个维度，介绍工作中常用的数据处理工具。最后，阐述笔者做大数据处理与分析中的一些思考，旨在让读者对大数据有更进一步的认识。

## 1.1 何为数据工程师

数据工程师无疑是大数据时代最热的名词之一。只要是从事数据相关工作的人员，都可以划分到该范畴。笔者认为“数据分析”与“数据工程”这两项能力，是数据工程师的核心能力。

“数据分析”能力，从实践的角度来看，就是工程师能根据数据给出分析建模的策略，解决业务中遇到的问题的能力。制定有效的建模策略，要求数据工程师

必须至少具备以下三类知识储备：统计分析、数据挖掘以及深度学习。本书的第4~6章分别讲述了这三个领域的相关知识。

“数据工程”能力，就是灵活运用数据处理组件及相关技术，实现数据分析中拟定策略的能力。数据处理的工程实施包含了“数据搜集”“数据传输”“数据存储”“数据计算”四个部分。这一系列流程的完成，要求数据工程师必须具备以下工程知识：消息队列、数据库技术、数据仓库、分布式文件系统和分布式计算平台等。在本书的最后一章，会对此做详细的介绍。

“数据分析”以及“数据工程”两项能力是数据工程师的硬实力。相比之下，对业务的理解则是与硬实力相辅相成的软实力。现实的工作中，一切技术手段最终都是为了业务的发展，理解这一点对数据工程师尤为重要。从初级数据工程师晋级为高级数据工程师，也是一个从被动实现到主动创新的过程。通常情况下，当面对上级或者同事提出的具体数据需求时，初级数据工程师一般处于被动实现的境地，这在初级阶段是无可厚非的；但是，对于高级数据工程师而言，则不能满足于被动地去实现需求，而要更进一步做到以下两点：① 应该主动去理解战略目标、产品方向和营销意图，围绕产品发展的各个阶段，综合当前业界已有的方法和思路，建立一套完善的数据支持体系；② 善于从数据的角度思考问题，保持数据敏感度。“啤酒-尿布”的故事正说明只有对数据敏感的人，才有可能从这种角度来发现数据间的关联。总之，一名优秀的数据工程师，不但要专心锤炼技术，更要着眼于现实的业务。

实践中，数据工程师不但要专注能力图谱（如图 1.1 所示）中的某几个点，还要在各个领域都有所涉猎。比如，合理地评估对海量数据集实施建模策略的可行性，就要求数据工程师不仅要了解模型构建，还必须具有工程实施的经验。所以，只有具备广而深的知识体系，才能掌控全局，更有效地提炼数据价值。

不可否认，数据工程师是一个更加需要经验的岗位。随着业务的深入发展，数据诉求会时刻发生变化；同时，不同业务之间的数据诉求也是天差地别的。笔者认为，拥有完善的分析体系和工程实施策略，并能针对具体场景给出完整解决方案的工程师，才能称为顶级数据工程师。

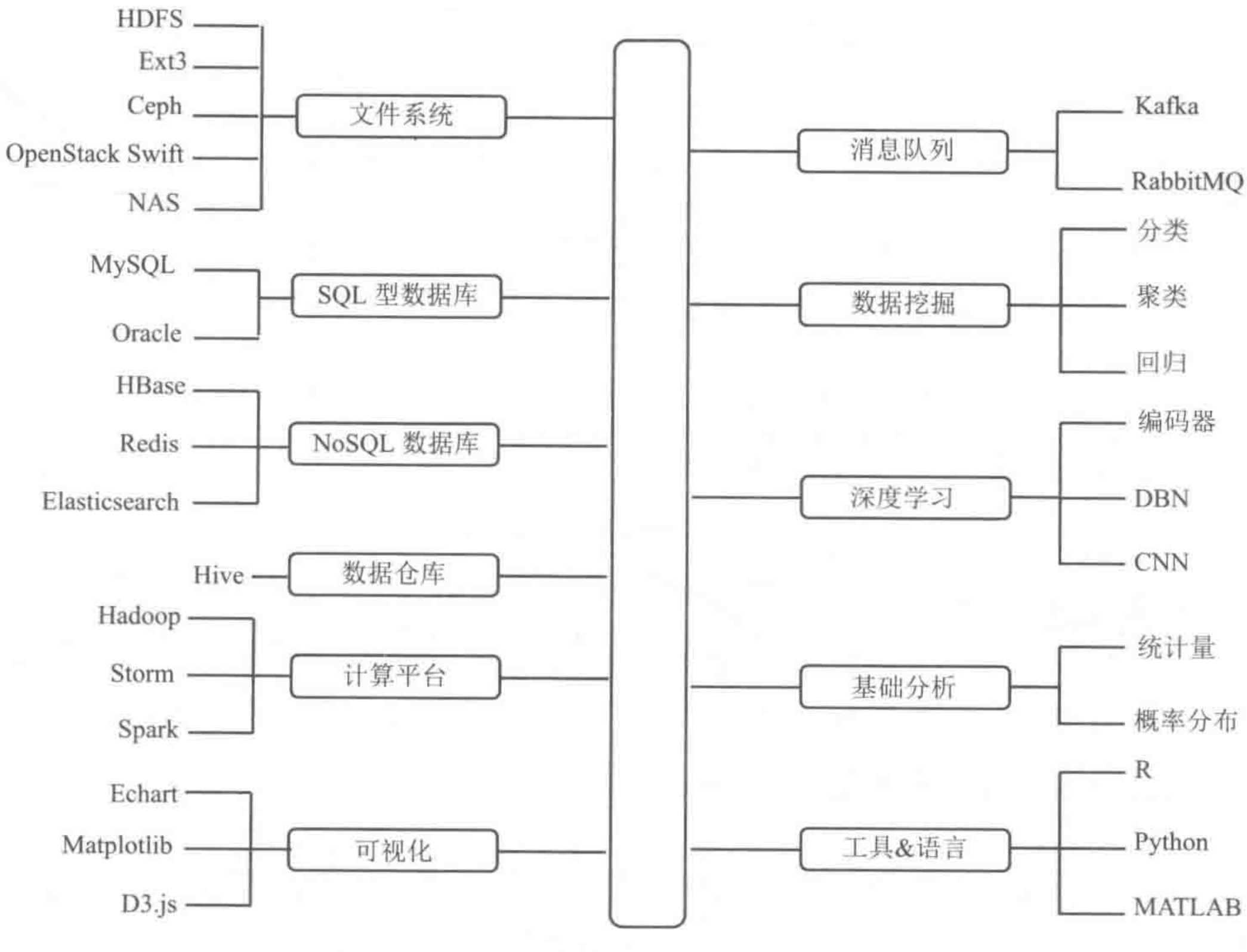


图 1.1 数据工程师能力图谱

## 1.2 数据分析的流程

与解决其他问题一样，数据分析也存在一般的模式化流程，通常可以划分为六个阶段。即明确目标、确定方案、数据整理、实施建模、结果验证和总结展现。

图 1.2 展示了数据分析的各个阶段。以下将详细说明各个阶段的具体实施思路，以及由“结果验证”阶段回退到“确定方案”阶段的原因。

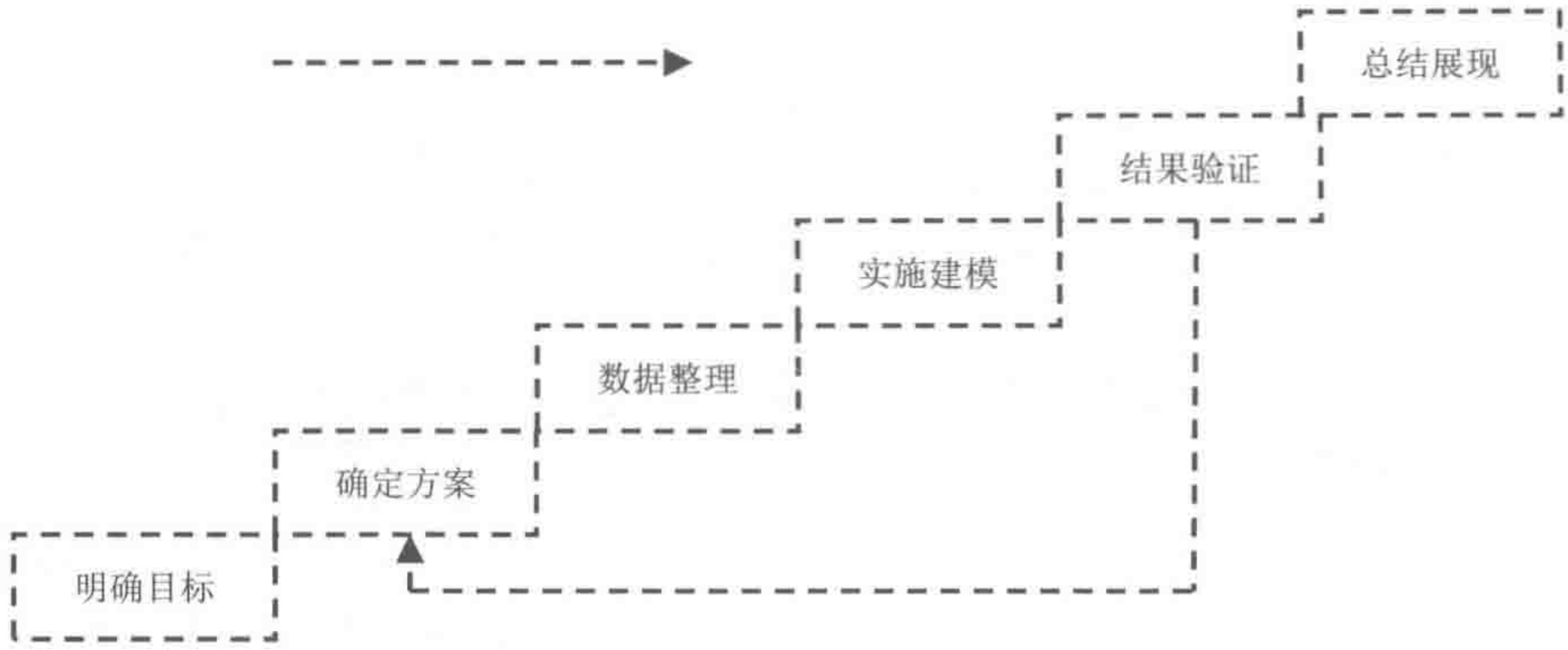


图 1.2　数据处理流程图

### 1. 明确目标

明确且细化分析的目标，是数据分析中极为重要的一点，直接关系到全部工作是否能有效展开、业务是否能有收益。

从“分析目标”的维度，对数据分析做抽象归纳，可以把它分为三种类型，如图 1.3 所示：① 验证型分析；② 描述型分析；③ 预测型分析。

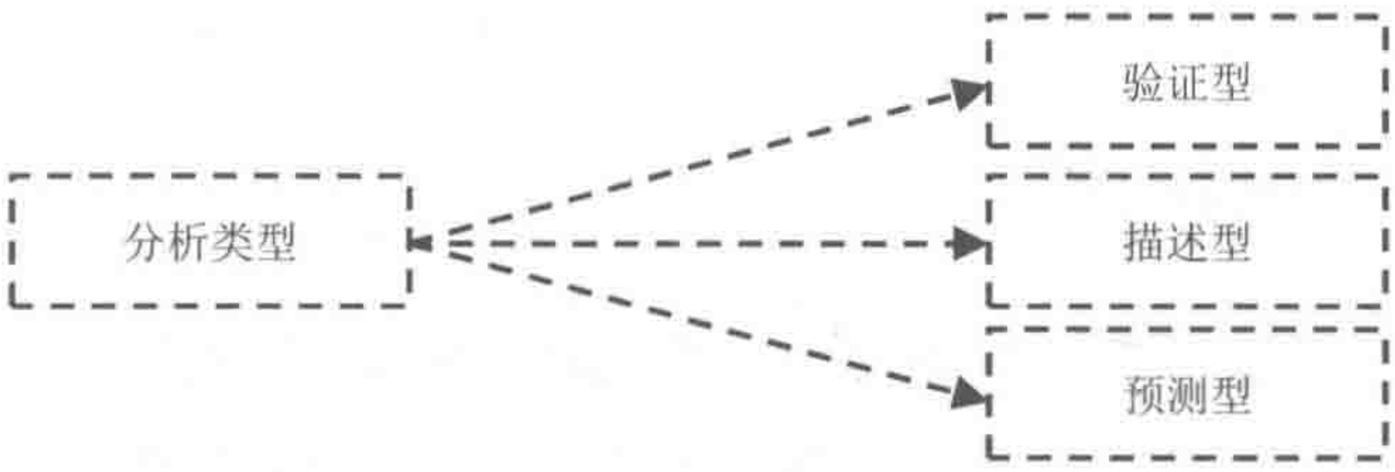


图 1.3　数据分析的三种类型

验证型分析，主要指对提出的问题进行数据验证。比如，某公司第二季度的销售额相比去年同期是否有所下滑；或者相比于选手 B，选手 A 的水平是否更加稳定等。提出问题的同时，用数据对其进行合理的验证，就是验证型的数据分析，也是日常分析中最常用的数据分析手段。

描述型分析，主要是从数据的角度说明现状或者问题。比如网站的运营状况可以通过网站的 PV、UV 或者留存率、转化率等数据指标进行描述。具体选择哪

些数据维度或者评价指标来描述数据？就笔者的经验而言，不同领域、不同事物的描述分析的维度虽然千头万绪，但都存在一定的、可以套用的模式。网站运营较为常用的指标有 PV、UV、二跳率、转化率、留存率等；网站运维常用的指标有页面平均响应时长、故障率、缓存命中率等；营销常用的有 STP 理论、SWOT 等。这些模式和领域密切相关，是业界同行经验的积累和总结，完全可以吸纳和应用，不必闭门造车。

预测型分析，主要指根据历史数据或者其他的数据信息，对可能发生或者即将发生的事情做出数据上的合理推测。比如，根据某地上年同期的降雨量预测今年同期的降雨量；或者通过 ARMA[1]模型预测股票大盘的走向；数据挖掘中回归决策树[2]对数据的分类，也是一种预测型数据分析。

### 2. 确定方案

确定方案有三个步骤（参见图 1.4）：① 确认能否获取相关数据；② 选择可行的分析建模以及实施方法；③ 制定结果的校验准则。

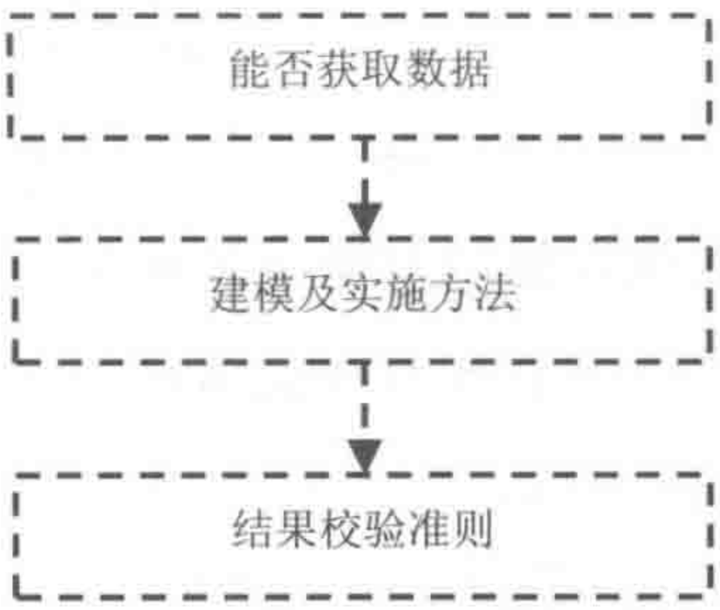


图 1.4　确定方案的三个步骤

能否获得数据决定了分析是否能够进行。在工程实践中，往往有很多可行的分析方法，然而能否获取相应的数据是一切工作的前提。一定要在已有数据的基础上，选择与拟定建模方法，以防止后续工作的徒劳无功。

建模的方法，由简单到复杂可以划分为三类（参见图 1.5）：① 基础分析 ② 数据挖掘 ③ 深度学习。这也是本书后续章节的叙述组织方式，虽不能大而全地呈现数据分析的细节，但是基本描绘了整体的知识框架。

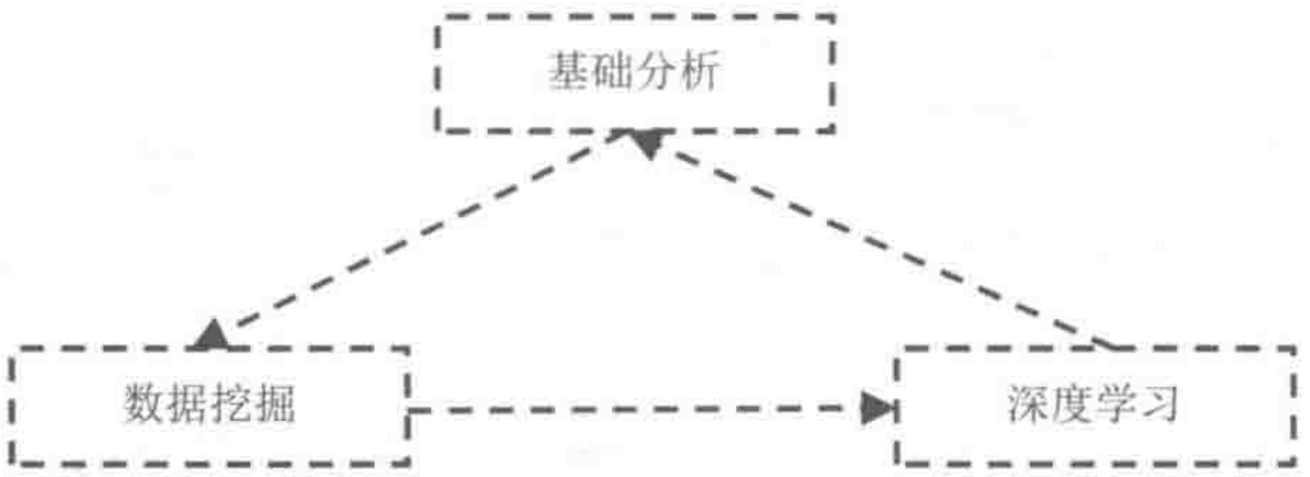


图 1.5　建模的方法

基础分析主要解答事物的统计特征，以及概率的相关问题。它首先研究是否可以通过均值、方差等简单的统计量来说明问题；其次，分析数据是否符合某种分布，如果能给出数据的有效分布，就可以合理地计算事物的概率。这些简单的分析在某些场景下具有事半功倍的效果。数据挖掘主要解决分类、聚类、关联的相关问题。如果在分析中遇到这几类问题，可以尝试从数据挖掘的角度来寻求方法。深度学习和数据挖掘类似，也是用来研究分类和识别的问题。它们之间的最大区别是：前者能自动提取数据的特征，并对非线性数据集具有良好的效果，有些文献把数据挖掘称为浅层学习。深度学习常用在图像识别以及声音识别的场景中。本书的第 5 章和第 6 章将分别对两者进行更为详细的讲述。

一般情况下，应由易到难地选择建模方式，解决实际问题。比如针对一组数据，我们首先要考虑基本的统计量以及概率分布是否能达到数据分析的目标；其次，思考能否运用数据挖掘的方法来对数据做进一步的分析；最后探讨深度学习的思路能否更好地解决问题。总之，兵无常势，水无常形，在具体问题的基础上，只有灵活运用多种建模手段，才能更好地达到分析的目标。

最后，制定结果的校验准则对数据分析尤为重要。在实践中，统计数据的对错往往很难被发现和评估；同时，错误的统计结果或者分析结论在某些时候可能会造成巨大的损失。因此，制定完善的数据校验策略来验证数据分析结果的可信度是极其重要的。本书会在数据校验部分对做校验的一些经验性方法进行简单的阐述。