前　言

近些年，各行各业聚集的“大数据”不仅对信息处理技术提出了挑战，而且深刻影响社会经济的各个方面。大数据时代的到来也催生一门新的学科：数据科学。数据科学是基于计算机科学、统计学、数学等学科的一门新兴的交叉学科，主要研究内容包括数据科学基础理论、数据预处理、数据计算和数据管理。作为一门新兴学科，很多学校开设了相关专业，也亟需教材讲授其核心理论体系和应用实践。本书顺应数据科学兴起的潮流，为计算机专业以及统计学，经济学等其他专业的学生，提供一本入门和导论性质的教材。

作者深入调研了现有的大数据教材和资料，结合十余年数据挖掘和机器学习等领域的科研实践和《计算机导论》等计算机专业基础课程的教学实践经验，以“建立知识体系、掌握基本原理、学会初级实践、了解前沿技术”为原则，精心设计编辑了《数据科学导论》教材内容，该教材具有如下特色：

(1)内容全面，重点突出。本书涵盖了数据科学的主要内容，包括：基础理论、数学基础、分析方法、应用前沿和处理技术。同时，作者也从数据挖掘的视角着重强调了数据分析的基本方法和技能。

(2)理论系统，实践丰富。本书比较系统介绍了数据科学紧密相关的基本理论和方法，并且配以丰富的示例进行讲解。作者以 Python 语言为例，配以大量实例详细讲解了数据分析的基本方法。

(3)模块设计，灵活组合。本书划分为三个模块：基础理论（第 1 和 2 章）、分析方法（第 3 至 6 章）、高级主题（第 7 和 8 章），三个模块相对独立，模块内部也是由浅入深。选择合适章节内容和讲授深度，可以支撑 2-6 学分的数据科学导论课程设置。

(4)深入浅出，可读性强。本书尽量介绍数据科学最相关的内容和最基本的概念，并配以实例介绍本质含义；此外，介绍了大量深入学习的扩展阅读材料。本书面向具有基础的计算机相关知识的学生和科技工作者，力争概念通俗易懂，方法便于上手。

全书内容分为 3 部分，共 8 章。第一部分是数据科学的基本理论和数学基础，由第 1 章和第 2 章组成。

第 1 章 “数据科学概论”是本书统领式的一章。主要介绍了数据科学的产生背景、基础知识、基本理论、以及数据科学家和数据科学的实践案例。通过串联数据和大数据的概念，阐述了人类社会的数据化进程；通过介绍数据科学的理论基础和应用实践引导学生在学习时应注重理论联系实际，学以致用。

第 2 章 “数学基础”介绍数据科学研究中广泛使用的数学工具。主要介绍了数据科学中需要用到的基础数学知识，分为四个部分：线性代数、概率论、优化理论和图论，并结合实例案例探讨他们的应用。

本书第二部分内容介绍数据科学中常用的数据分析方法，由第 3 至 6 章组成。

第 3 章 “Python 语言初步”介绍数据科学研究中主流的编程语言。全书的案例也都统一以 Python 语言讲解。该章涵盖 Python 的基本用法以及数据科学处理中重要库的使用。

第 4 章 “数据预处理”介绍数据科学处理中基本的数据预处理方法。该章节是整个数据处理中的前期核心步骤，包括数据清洗、数据集成、数据规约、数据变换等技术，并最后辅以一个实践案例具体阐述预处理各个步骤。

第 5 章 “分析方法初步”介绍数据科学研究中的基本机器学习模型。该章节介绍机器学习基本概念及主流的机器学习库，同时讲解回归、分类、神经网络等监督学习方法及聚类等无监督学习模型，每个模型均配有实例及代码演示。

第 6 章 “数据科学实践”以实战案例系统总结前面章节的数据处理技术。首先介绍数据分析流程，继而给出五个具体的案例，包括泰坦尼克号生存预测、时间序列预测等，每个案例从问题分析开始，阐述数据预处理、机器学习模型使用、结果分析等完整流程。

本书第三部分内容介绍数据科学的应用前沿和处理技术，由第 7 章和第 8 章组成。

第 7 章 “数据科学的重要研究领域”围绕非结构化数据，分别对文本数据、图像视频数据、图结构数据的分析与应用方法展开介绍。此外，还简要介绍了数据可视化分析技术、应用场景、常用的可视化分析工具。

第 8 章 “云计算与大数据处理工具”介绍了大数据处理的主流工具。主要介绍了云计算的相关概念和特点，核心技术虚拟化和多个商用的云计算平台；讨论了大数据处理工具 Hadoop 与 Spark 这两个框架的基本概念、核心算法以及生态环境。本章还提供了一个完整的搭建并使用 Hadoop 集群进行数据处理的应用案例。

本书可以作为高等学校信息类学生的数据科学和大数据分析等课程的入门教程，也可以作为科技工作者学习大数据分析的参考材料。作为大学教材使用，可以有短学时（2-3 学分）和长学时（4-6 学分）两种教学计划。针对短学时教学计划，可以选择第 1，3 至 6 章节讲授，其他章节选讲；针对长学时教学计划，可以讲授全部内容，并且增加上机实习环节。本书还提供了丰富的教学资料供教师教学参考和学生学习使用，包括教学幻灯片和所有示例源代码等资料。这些资料可以从 www.shichuan.org 下载使用。

全书框架由石川负责设计并统稿，并编写了第 1 章。王啸负责编写第 3-6 章，胡琳梅负责编写第 1、2、7、8 章；王柏对全书进行了校验。本书编写过程中得到了北京邮电大学计算机学院数据科学与服务中心的老师们的大力支持和帮助；也得到了许多研究生的支持，他们收集并整理了大量的资料。没有他们的帮助，本书很难在约定的时间内完成。在此，感谢他们对本书的编写过程中做出的巨大贡献。