数据科学与大数据技术丛书

DATA

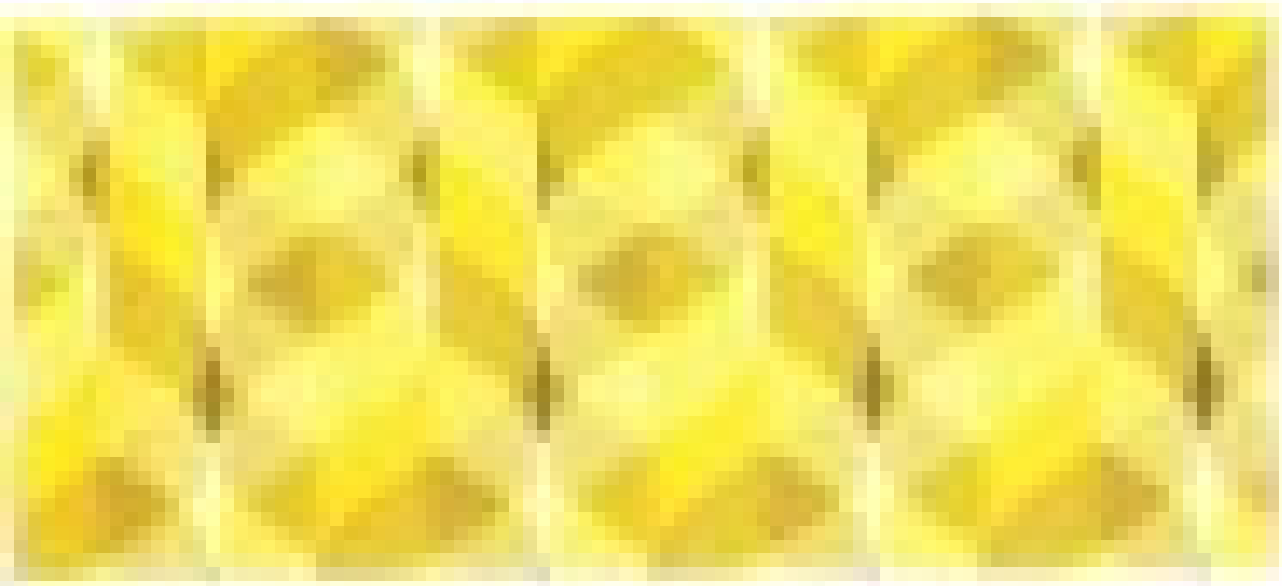
数据科学概论

Introduction to Data Science

覃雄派 陈跃国 杜小勇◎编著

DATA SCIENCE

中国人民大学出版社



数据科学概论

Introduction to Data Science

DATA SCIENCE

数据科学与大数据技术丛书

数据科学概论

Introduction to Data Science

覃雄派 陈跃国 杜小勇◎编著

中国人民大学出版社

• 北京 •

图书在版编目（CIP）数据

数据科学概论/覃雄派等著．—北京：中国人民大学出版社，2018.1
（数据科学与大数据技术丛书）
ISBN 978-7-300-25292-6

Ⅰ.①数… Ⅱ.①覃… Ⅲ.①数据管理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2017）第 313409 号

数据科学与大数据技术丛书
数据科学概论
覃雄派　陈跃国　杜小勇　编著
Shuju Kexue Gailun

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	邮政编码	100080
电　　话	010－62511242（总编室）	010－62511770（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京溢漾印刷有限公司		
规　　格	185 mm×260 mm　16 开本	版　　次	2018 年 1 月第 1 版
印　　张	36 插页 1	印　　次	2018 年 1 月第 1 次印刷
字　　数	880 000	定　　价	68.00 元

推荐序

大数据时代已经来临，数据的价值逐渐为人们所认知，大数据的研究和应用方兴未艾，然而大数据人才无论是在中国还是在世界上都非常缺乏。为了应对大数据人才的培养问题，教育部设立了“数据科学与大数据技术”本科新专业。到目前为止，总共有 35 所高校获教育部批准开设数据科学与大数据技术专业，中国人民大学是其中一所。

本人认为，数据科学作为一个学科似乎尚未成熟，很多理论问题尚未搞清楚，其知识体系应该是怎样的，莫衷一是，但社会对于大数据人才的需求是明确和巨大的，时不我待，为培养人才计，可以先把专业建设起来。

开设新专业需要设置成体系的课程和编写成系列的教材。中国人民大学信息学院计算机系正在为数据科学与大数据技术专业重新规划课程体系，拟建设系统课程群、算法课程群、数据科学课程群三大课程群。其中，数据科学课程群将由一系列课程构成，包括数据科学概论、数据库、数据挖掘、统计分析、深度学习、商务智能等，数据科学概论定位为该课程群的入门和导论性质课程。

杜小勇老师领导的数据库与信息检索实验室（隶属于教育部数据工程与知识工程重点实验室）较早开展了大数据系统及应用的相关研究工作，他们编写的这本《数据科学概论》全面论述了大数据相关的主要数据类型和主流的数据管理、分析技术。在内容上，从简单的数据管理和分析、多维分析和结构化数据分析，到复杂的数据挖掘和机器学习，由浅入深渐次展开，同时，对文本、社交网络、时间序列、轨迹等数据分析技术分别进行介绍，完成了内容的宽度展开，这是本书比较突出的一个特点。数据科学是一门对实践动手能力要求很高的学科。本书还提供了大量的基于 Python 的数据分析实例，读者可以直接运行或进一步修改，加深对技术原理的理解，真正掌握这些数据分析技术。

《数据科学概论》在论述的风格上深入浅出，通过一些案例和可视化技术辅助原理的讲述，使得读者容易理解和把握。作为数据科学与大数据技术的入门教材，本书是不错的选择，谨向大家推荐。

文继荣
中国人民大学信息学院教授、院长

作者自序

数据科学是基于计算机科学（数据库、数据挖掘、机器学习等）、统计学、数学等学科的一门新兴的交叉学科，它研究数据的各种类型、状态、属性及其变化规律，研究如何对数据进行分析，从而揭示自然界和人类行为等现象背后的规律。

本书顺应大数据时代的到来以及数据科学兴起的潮流，为计算机专业以及统计学、经济学、金融学等其他专业的学生，提供一本入门和导论性质的教材。

数据科学家是大数据时代最急需的人才，他们具有宽广的视野，同时具有扎实的理论和技术功底。《数据科学概论》为数据科学概论课程而设计，具有入门和统领的作用，为学生学习后续课程打下坚实的基础，有利于培养新一代数据科学家，为各行各业的数据处理提供急需的人才。

本教材对数据科学的核心问题，即对数据进行分析、挖掘并提取其价值，获得对事物的洞察的各种技术手段，进行全面论述，把读者引进数据科学的大门，帮助读者建立数据科学的知识体系。

在本教材中，大量使用通俗易懂的实例，配合技术原理的讲解。本教材力图体现学科交叉的特点，让不同学科背景的读者，感受到数据科学的魅力，数据分析技术是如此有趣和有价值，能够解决各行各业的实际问题，使得学习过程充满趣味而不是枯燥无味。

在本教材最后，我们面向金融领域的量化交易应用，从数据采集、模型训练、预测、评价，到可视化等环节，带领读者完成数据处理和分析的实践，打通整个流程，锤炼读者的编程能力，使其深刻体会运用数据科学方法解决实际问题的乐趣。

本教材适合本科生使用。

由于编者的水平有限，书中难免有错漏之处，敬请读者和同行批评指正。

覃雄派　陈跃国　杜小勇

前　言

一、本教材的四大模块

本教材的内容分为四大模块，分别是：

（1）数据科学基础（Fundamentals）：讲述数据科学的基本概念和原则。

（2）数据和数据上的计算（Data and Computing on Data）：讲述不同的数据类型及其分析方法，数据类型包括结构化数据、非结构化数据、半结构化数据，分析方法包括统计学方法、数据挖掘和机器学习方法等。

（3）数据处理基础设施、平台和工具（Infrastructure，Platforms and Tools）：讲述云平台、数据库、大数据平台和工具以及编程语言 Python。

（4）数据科学案例和实践（Applications and Practice）：讲述大数据应用的案例，并且面向金融领域的量化交易应用，从数据采集、模型训练、预测、评价到可视化等环节，带领读者完成数据分析处理的实践。

二、各章内容介绍及其相互联系

第 1 章“数据科学概论”是本书统领式的一章。本章介绍数据科学的概念、数据科学的原则、数据的价值、数据的类型、数据处理的流程、批处理/实时处理/交互式处理以及 Lambda 系统架构等。

我们把数据处理系统提供的功能（或者对数据的操作）分为数据服务和数据分析两大类。

数据服务（Data Serving）是面向大量操作型用户，提供对细节层面数据的存储、检索等服务，从而支撑业务系统的运行。对数据的操作主要是增加、删除、修改、查询少量的记录，或者计算一些简单的统计信息。由于要处理大众化的、简单的数据处理任务，每个任务必须能够被快速地处理掉，因此，单个任务一般不会很复杂，直接处理的数据量不会太大。

传统的关系数据库（也称为 SQL 数据库，因为关系数据库管理系统一般提供 SQL 查询语言的支持）具备事务处理能力，提供关键任务（Mission-Critical）的数据服务，比如银行核心业务系统，提供了存钱、取钱、转账、查明细等功能。随着移动互联网时代的到来，很多互联网特有的现象级应用给事务处理带来了不小的挑战。两个较为典型的例子是

“双十一”和12306网站春节期间的购票。高并发的事务处理请求，给现有数据库系统带来了巨大的挑战，研究人员尝试使用新硬件提升数据库系统性能，以及在软件层面重新构建数据库系统，这是对关系数据库的改造升级或者再造，产生了一批扩展性良好的新型事务处理系统，统称NewSQL数据库。

在某些应用系统中（比如电商的购物框、互联网用户画像等），数据处理逻辑相对简单，主要的操作是快速查找数据和修改数据，对数据的一致性要求没有那么强，但是要求极高的吞吐能力。传统的关系数据库扩展性不足，无法胜任这样的数据处理任务。NoSQL数据库（不是某个数据库，而是一类数据库的统称）由于弱化了数据的一致性要求（采用最终一致性），数据操作简单，适合并行化处理。在一定规模的集群下，能够达到较高的数据读/写吞吐率（每秒百万级操作），满足这类应用。典型的例子是采用键值对数据模型的NoSQL数据库，比如Dynamo等，对数据的主要操作包括根据某个Key查找到其Value，或者根据某个Key修改其Value等。

互联网公司在为用户提供服务的过程中，收集到很多用户行为数据，可以利用数据分析技术构建用户画像，以便提供更好的（个性化的）服务。为了精细地刻画用户的特征，经常使用成千上万个属性。对特定用户的属性进行检索和修改，就是典型的数据服务。在用户登录时，快速提取用户的属性，利用这些属性和一些特有的业务规则，对用户进行个性化的界面显示，也是典型的数据服务。整个服务过程在几毫秒到几十毫秒的时间内处理完，要求很高的性能。根据用户的属性，计算用户的相似度，对用户进行聚类，则是复杂的数据分析。

数据分析（Data Analysis）是面向分析人员（或者管理层），通过对大量数据使用复杂的计算模型进行分析，从而发现数据中隐藏的模式或规律。数据分析结果的信息量通常很大，适用于宏观决策，为中层和高层管理人员所使用。数据分析分为简单分析和复杂分析。

简单分析是指简单汇总和报表等。联机分析处理（Online Analytic Processing，OLAP）采用星型模型或者数据立方体组织数据，保存在数据仓库中，然后进行多维度的聚集查询处理。在OLAP应用中，在事实表中记录一系列的事件，在维表中记录事件的维度，然后从维表中选取一些维度对事实表进行汇总操作，这些汇总操作包括计数、求和、求平均值、求最大值、求最小值等。比如，某连锁企业汇集了全国各个地区的门店的销售数据，就可以从时间、产品、地区等维度对销售数据进行汇总。在一个维度上，还可以选择不同的汇总层次，比如在产品维度上，可以查看各个产品大类的销售额，或者更加深入细致地查看某个产品大类底下各个产品小类的销售额等。对销售额最低的门店的详细销售数据进行查看是一种下钻操作，比如，具体到每个商品每天的销售情况，可以帮助管理者找到销售不佳的原因。联机分析处理从分析的复杂度上来讲属于简单分析。为了把数据从联机事务处理（Online Transaction Processing，OLTP）/数据服务系统转移到数据仓库系统中，需要对数据进行抽取、转换和装载，对于其中有错的数据，需要进行必要的清洗，提高数据质量。如果数据来源多，还需要对数据进行集成。

复杂分析一般需要利用统计分析方法、数据挖掘与机器学习方法，对数据进行深入分析，提取有用信息并形成结论，辅助人们决策。机器学习是一个活跃的研究领域。机器学习从数据中训练一个模型，用于预测与分类。比如，电商网站的推荐系统，根据大规模用

户的购买或浏览行为数据，使用机器学习算法，得到一个推荐模型。当用户在电商网站选购商品时，网站会利用先前学习到的推荐模型，结合用户近期的浏览和购物行为，为用户推荐商品。

近年来，通过和大数据的结合，机器学习取得了令人振奋的进展。最为典型的是深度学习（深度神经网络）技术已经在计算机视觉、语音、自然语言处理、游戏博弈等领域取得了巨大的突破。交通标志检测是无人驾驶技术中一项非常具有挑战性的任务。交通标志的正确识别，对辅助定位和导航具有决定性的作用。交通标志的种类繁多，其大小和角度不一致，受天气、光照等环境因素影响大，这使得对交通标志的检测非常困难，但是我们有一个优势，即很容易获得大量真实场景下的图像数据，用于机器学习模型的训练和测试。2016 年中国计算机学会大数据与计算智能大赛中的一个题目是自动驾驶场景中的交通标志识别，获得该赛题一等奖的团队采用的正是深度神经网络技术，他们还通过将关键部位的图像进行放大增强等措施，提高了识别的准确率。

在用户和数据处理系统进行一次交互的过程中，用户使用到的功能可能囊括数据服务、数据的简单分析和复杂分析等不同的数据处理功能。比如，在用户给搜索引擎提交关键字进行查询的过程中，搜索引擎利用信息检索技术提供查询结果。信息检索（Information Retrieval）是指从大规模的数据集（大量文档构成的文档集）中快速查找满足用户需求的数据（少量文档）的过程。

用户通过关键字（或自然语言语句）表达信息需求，为了快速得到查询结果，信息检索系统一般预先离线式地建立索引结构（如倒排表）。检索出一系列的文档后，搜索引擎还需要根据结果跟查询的相关性，对结果列表进行排序。索引的创建、结果的排序都是典型的数据分析操作，索引创建是离线进行的（Offline），结果排序是在线进行的（Online）。用户根据搜索引擎提供的网址列表，点击某个网址以后，搜索引擎提供该网址所指向的网页，则是典型的数据服务操作，即从数据库中提取网页返回给用户。在为大量互联网用户提供海量 Web 数据的信息检索服务的基础上，搜索引擎公司能够获得大规模的用户行为数据，通过对这些数据进行分析，它们能够为自身提升信息检索服务质量、拓宽广告服务等增值业务奠定基础。

数据服务的主要目的是支撑业务的运行，数据分析才能真正发挥数据的内在价值，为决策服务。还是以搜索引擎为例，之前人们一直认为信息检索的核心是排序模型，并投入了大量精力改进排序模型，以求提升信息检索的精度。然而，随着越来越多的用户使用搜索引擎，搜索引擎公司逐渐意识到，用户对结果的点击行为是一种非常好的反馈。利用海量用户的点击数据，研究人员使用排序学习的方法，可以大幅提升信息检索的精度。这是搜索引擎公司对其收集的大数据的一种重要的价值发现过程。数据分析（特别是复杂分析）是发挥数据价值的关键。

本书第 2 章“OLTP 与数据服务”介绍联机事务处理（OLTP）与数据服务的关键技术与实际系统。第 3 章“OLAP 与结构化数据分析”介绍联机分析处理（OLAP）与结构化数据分析的关键技术与实际系统。

第 5 章“数据的深度分析”介绍主流的统计分析方法、数据挖掘与机器学习方法。从 OLAP 与结构化数据分析到数据的深度分析，分析的复杂度提高了。

到这里为止，我们主要讨论了结构化数据、半结构化数据的分析处理。在互联网时

代，有几类数据的分析尤为重要，包括文本数据、图数据（社交网络）、知识图谱等。第7章“文本分析”介绍文本分析的意义、方法和工具。第8章“社交网络分析”介绍社交网络分析的应用、社交网络分析方法、社交网络分析工具等。社交网络的发展，让图分析技术发挥越来越重要的作用。图数据分析（社交网络分析）的目的在于分析图上节点（包括边）的影响关系、发现图的模式等。比如，在论文数据库DBLP之上，可以获得作者之间的合作关系，评估作者之间的相互影响程度。对于一个特定的作者来说，可以分析出哪些作者对其影响最大，还可以分领域（用关键字表达）、按时间阶段等展示这种影响。第9章“语义网与知识图谱”介绍语义网概念、关键技术以及知识图谱的创建、挖掘和应用。我们认为，时间序列和轨迹数据两类数据的分析是数据科学的重要内容。由于篇幅所限，我们将在本书的未来版本中将其包含进来。

在数据科学的实践中，数据的可视化是非常重要的一个环节。第10章介绍数据可视化的作用、过程、原则、实例等，并介绍可视分析以及可视化的工具。此外，由于探索式数据分析往往需要数据可视化的帮助，我们在这一章中也介绍了探索式数据分析。

在已经介绍的内容中，涉及的数据处理模式主要有批处理（比如训练机器学习模型）和交互式处理（OLAP分析）。第6章“流数据处理”介绍第三种数据处理模式，包括流数据处理的概念、技术与实际系统。

在学习了以上内容之后，读者应该熟悉各种数据分析的技术和方法，但是这些方法要落到实处，需要数据处理的基础设施、平台和工具的支撑。

第11章“云计算平台”介绍云计算的概念、技术与主流厂商的实际系统。我们希望需要数据时能够迅速利用它，当数据规模大到一定程度时，传统计算架构已经不再适用，云计算是为大数据而生的计算模式。云计算与大数据相辅相成，两者之间互相推动与促进。

在介绍数据分析的技术和方法的内容中，我们穿插介绍了大量的工具，包括关系数据库管理系统（RDBMS）、NoSQL系统、NewSQL系统、流数据处理系统等，以及数据挖掘、机器学习、文本分析、图数据分析的相关工具。

接下来，我们将介绍两大主流大数据处理平台和生态系统——Hadoop和Spark。第12章“Hadoop及其生态系统”介绍Hadoop大数据处理平台及其生态系统，包括Hadoop1.0与Hadoop2.0。第13章“Spark及其生态系统”介绍Spark大数据处理平台及其生态系统。Hadoop和Spark都具有数据集线器（Hub）的功能，它们把各个来源的数据集中到一个地方，进行后续的分析。

当用户面临各种类型的数据需要管理，大量可用的工具可供选择时，他们需要某种手段做出抉择。第15章“评测基准”介绍数据处理和分析系统的各类评测基准。评测基准不仅帮助最终用户进行系统选择，而且可以帮助厂家提高产品的性能，以及帮助研究人员验证新的技术和思想。

我们还介绍了主流的数据科学编程语言Python。第14章“Python与数据科学”介绍了Python语言、工具库，并给出大量的数据分析实例，帮助读者掌握Python编程。

本书最后的部分跟实际应用有关。第16章“数据科学案例”介绍数据科学在各个行业的成功案例。第17章“数据科学实践”则介绍数据科学在金融领域的一个重要应用，即量化交易，该章介绍了量化交易系统的基本概念、系统设计、实现以及评估等方面的

内容。

全书各章的内容及其联系参见图 1。

图 1　各章内容及其联系

三、本教材的特点

市面上关于数据科学的教材，有的偏重理论，有的偏重实践。偏重理论的教材，重视严谨性和数学的推导；偏重实践的教材，则用大量的篇幅讲某个（些）软件的使用和编程。本教材兼顾原理和编程，采用案例讲解技术原理以及问题解决策略。

为了帮助学生（读者）把握技术原理，并且能够开始运用这些原理，对复杂的工程问题进行求解，我们从两个方面进行案例式讲解。一个方面是针对各个知识点，给出实例。通过简单的实例，讲解每个技术的原理，使得学生迅速把握其本质，而不是陷入艰难的数学推导和绝望当中。这并不是说不需要数学推导，而是本书作为一门导论性质的课程的教材，更为重要的是让学生把握技术的原理和思想，艰难的但是必要的、深入的数学推导过程，可以在后续的课程中介绍。换言之，本教材不会陷入数学公式的复杂推导过程（必要的数学知识是需要的），而是采用浅显易懂的语言，结合实例，讲清楚各种技术的基本原

理。不仅方便计算机专业的学生阅读，其他专业的学生在理解上也不会有太大的困难。

另一个方面，我们从问题出发，展示问题的分析和解决策略及其实现过程，这是传统意义上的综合实例，是面向实际应用的综合实例。我们选择的综合案例是金融领域的量化交易应用。通过案例，带领读者完成数据采集、模型训练、预测、评价以及可视化等重要环节。锤炼读者的编程能力，使其深刻体会运用数据科学方法解决实际问题的乐趣。

四、教学计划

根据每周的课时数我们给出两个教学计划。

如果每周两次课，每次 2 课时，每周的课时数是 4 个课时，有上机。按照一个学期 16 周（不算考试周）计算，总课时数是 64 个课时，本书的内容可以全部讲授。每周的两次课，可以分别讲授基本原理（第 1～10 章）和实践技术（第 11～17 章，可以先讲 Python），按照两条线索对内容进行讲授。这样学生就可以在每周上完课以后开始上机了。

如果每周一次课，每次 2 课时，每周的课时数是 2 个课时，没有上机。按照一个学期 16 周计算，总课时数是 32 个课时，本书的内容则无法完全讲授，必须做出裁剪。建议选择第 1 章、第 2 章、第 3 章、第 4 章、第 14 章、第 17 章进行讲授。

五、本书的参考文献

在编写本书的过程中，我们参考了大量文献，包括论文、期刊、书籍以及网上资源。我们把这些参考文献按照各章进行分组，放在本书的最后（第 18 章）。

六、本书代码的使用

本书提供了大量可运行的 Python 代码，我们将通过网站（www. rdjg. com. cn）[①] 提供这些代码的 Notebook 文件（*.ipynb）。用户可以在 Jupyter Notebook 中打开这些 Notebook 文件，运行这些代码，加深对代码的理解，并且可以进行简单修改，然后执行它，马上看到修改后的效果。学习编程，一种行之有效的办法是照葫芦画瓢。在理解这些代码的基础上，读者可以借鉴已有代码，编写功能更强大的应用程序。

① 在中国人民大学出版社工商管理分社网站（www. rdjg. com. cn）查找本教材，转到本教材页面，即可下载相关资源。