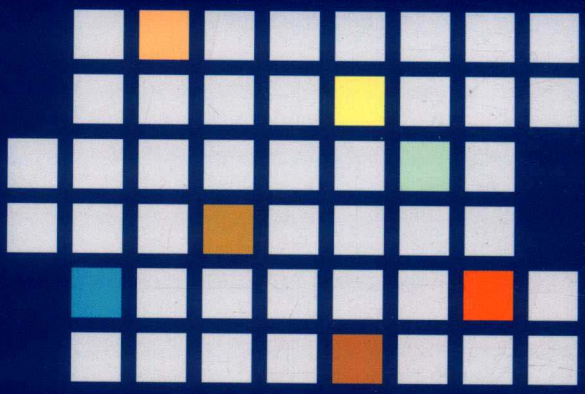
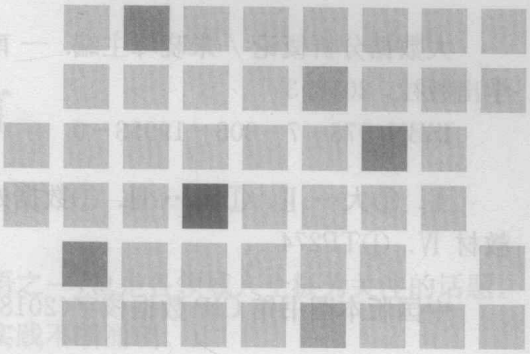


大数据分析概论

Introduction to Big Data Analysis

朱晓峰　主 编

南京大学出版社



大数据分析概论

Introduction to Big Data Analysis

主 编 朱晓峰

编 者 朱晓峰 赵柳榕 张 琳

吴海东 马小东 郑 乐

南京大学出版社

图书在版编目(CIP)数据

大数据分析概论 / 朱晓峰主编. — 南京 : 南京大学出版社，2018.3
ISBN 978-7-305-19953-0

Ⅰ. ①大… Ⅱ. ①朱… Ⅲ. ①数据处理—高等学校—教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2018)第 044959 号

出版发行　南京大学出版社
社　　址　南京市汉口路 22 号　　邮　编　210093
出 版 人　金鑫荣

书　　名　大数据分析概论
主　　编　朱晓峰
责任编辑　胥橙庭　王南雁　　编辑热线　025-83593962

照　　排　南京南琳图文制作有限公司
印　　刷　南京人民印刷厂有限责任公司
开　　本　787×1092　1/16　印张 25.25　字数 646 千
版　　次　2018 年 3 月第 1 版　2018 年 3 月第 1 次印刷
ISBN 978-7-305-19953-0
定　　价　55.00 元

网址：http://www.njupco.com
官方微博：http://weibo.com/njupco
官方微信号：njupress
销售咨询热线：(025) 83594756

前　言

“大数据分析”是当今科技行业最受欢迎的流行语之一，也是各领域人士极为关注的话题。飞速发展的中国，同样将大数据作为国家战略，企业实践不断涌现。

《大数据分析概论》是数据科学领域为数不多的理论与实践相结合的入门级教材，它通过详细剖析大数据分析基础理论和实例实训，全景展现了大数据分析各个阶段的基础知识、相关方法、关键技术和实用工具。

本书分为两个部分：第一部分，大数据分析的理论部分，包括“大数据分析概述”、“大数据分析的体系架构”、“大数据分析的关键技术”、“大数据分析的数据采集与存储”、“大数据分析的数据清洗”、“大数据分析的数据挖掘”和“数据可视化”；第二部分，大数据分析的实训部分，包括“体育行业 NBA 数据分析”、“金融行业贷款分析”、“服装行业库龄库存分析”、“公司财务数据分析”、“能源行业油井数据分析”、“政府行业财政收支分析”、“人力资源行业职位需求分析”和“互联网行业网站分析”等十二个不同行业的大数据分析。每个理论章节都以 1000 字左右的小案例引出本章内容，然后提供本章知识要点，最后提供案例思考题，并设计专门实验，方便学生认知和操作；每个行业实训，都包括背景分析、需求分析、大数据分析过程和分析结论。

本书由朱晓峰、赵柳榕讨论大纲，朱晓峰负责理论篇中的第 1—3 章；赵柳榕负责理论篇的第 4—6 章、实训篇的 9—10 章；张琳负责理论篇中的第 7 章，并和朱晓峰一起审核本书的理论部分；吴海东（福州大学）负责实训篇中的第 1—8 章；马小东（苏州国云）负责实训篇中的第 11 章，并和朱晓峰一起审核本书的实训部分；郑乐负责实训篇中的第 12 章。

本书在编写过程中，得到了南京工业大学校级教材重点项目的资助，得到了苏州国云数据科技有限公司、南京大学出版社的支持和帮助，尤其是经济与管理学院姚山季院长、苏州国云朱琼琼项目总监、南京大学出版社吴汀老师给予的项目申报、行业案例、分析工具、原始数据等方面的指导和帮助，在此表示衷心的感谢！

本书作者均为从事大数据类课程教学、实践的一线教师、实践者。因此，希望本书能够对大数据类课程的教学、学习和实践提供帮助。当然，本书难免有不妥之处，恳请广大读者对此教材提出宝贵意见，以期不断改进。

编　者

2018 年 2 月

目 录

理论篇

实训篇

理论篇

第一章 大数据分析概述

案例导读

大数据分析的光环与陷阱

2009年2月19日，Nature上面有一篇文章，“Detecting influenza epidemics using search engine query data”，论述了Google基于用户的搜索日志（其中包括搜索关键词、用户搜索频率以及用户IP地址等信息）的汇总信息，成功“预测”了流感病人的就诊人数。

那么，Google为什么要做这件事情呢？在美国，由疾控中心CDC专门负责统计美国本土各个地区的疾病就诊人数，然后汇总并公布。但是，这个公布的数据一般要延迟两周左右，也就是说当天的流感的全国就诊人数，要在两周之后才知道，Google就利用搜索引擎搭建了一个预测平台，把这个数据提前公布出来。因此，Google做的工作并不是实际意义上的预测什么时候流感来，而是将CDC已经获得但是没及时公布的数据提前给“猜”出来，然后公布出来。“越及时的数据，价值越高”，数据是有价值属性的。所以，Google的工作无论在公共管理领域还是商业领域都具有重大的意义。

Google成功“预测”流感病人的例子成为经典案例的深层次原因在于，如果在这个案例上成功了，Google就真正证明了大数据是“万能的”。因为Google在这项研究中对于数据的处理，只用了很简单的Logistic回归关系，却成功地预测了复杂的流感规模的问题。Google用了简单的方法，预测复杂的问题，充分证明了Google的大数据价值观——大就是一切！

大数据的观点之一认为，海量的数据可以弥补模型的不足，如果数据足够大，理论模型甚至根本就不需要。这种观点目前仍然处于争论中，偏重理论和实证（强调数据和统计方法）的专家们对此既惶恐又试图批判。但无论如何，Google对于流感预测的研究无疑站在了支持大数据的一方，如果Google的案例是成功的，那么或许，拥有海量数据就真的意味着可以解决任意复杂的问题，大数据解决大问题！

截至Nature发表论文的时候，Google的预测还是准确的，不过到后来就发生了很大的偏差，偏差最大甚至高出了标准值（CDC公布的结果）将近一倍①。

Google预测的失败也确实是过度地依赖于数据，导致很多被忽略了的因素对预测的结果产生了很大的影响。对客观世界进行预测需要模型，模型首先来自于理论构造，其次需要数据对模型进行训练、对模型进行优化完善。大数据观点强调模型对数据训练的依赖，而尽可能地忽略理论构造这一部分的意义，这就有可能带来隐患。

因此，Google的案例既是一个很好的大数据的应用，同时也为大数据在未来的发展道路起到了很好的指示灯作用。

① 数据大湿的博客. 从Google预测流感引发的大数据反思[EB/OL]. (2015-07-12). http://blog.sina.com.cn/s/blog_1464091000102vmeb.html.

学习目标

- 理解大数据分析产生的背景和现状
- 理解和掌握大数据的真正含义与价值
- 理解和掌握大数据分析的基本概念
- 熟悉正确的大数据分析思维
- 了解大数据分析的影响与价值
- 熟悉大数据分析的过程与对象
- 熟悉大数据分析的流程

1.1 大数据分析的背景与基础

大数据分析是数学与计算机科学相结合的产物，在20世纪早期就已确立，但直到计算机的出现才使得实际操作成为可能，并使得大数据分析得以推广。在学习大数据分析时，应该首先了解它的产生背景和基础。

1.1.1 大数据分析的背景

大数据分析的产生有其深刻的时代背景和历史的必然性，是IT技术的发展变革以及商务应用需求驱动的必然结果。

1. 数据的价值，已有时日

数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。社会对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。数据在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业存在已有时日，却因为近年来互联网和信息行业的发展而引起社会关注。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然很多企业可能并没有意识到由数据爆炸性增长带来的问题隐患，但是随着时间的推移，社会将越来越多地意识到数据对企业的重要性。

三分技术，七分数据，得数据者得天下。维克托·迈尔-舍恩伯格在《大数据时代》一书中举了诸多例证，都是为了说明一个道理：在大数据时代已经到来时要用大数据思维去发掘大数据的潜在价值。

大数据就是核心竞争力。全世界都在高呼大数据时代来临的优势：一家超市如何从一个17岁女孩的购物清单中，发现了她已怀孕的事实；或者将啤酒与尿不湿放在一起销售，神奇地提高了双方的销售额。实际上，数据已经无处不在，衣食住行、喜怒哀愁、吃喝玩乐都以数据的形式存在。通过数据来记录这个世界，再通过研究数据去发现这个世界。正如IBM所言：大数据时代——用智慧的分析洞察、构建智慧的地球。

2. 数据的数量，与日俱增

数据的数量，到底有多大？一组名为“互联网上一天”的数据显露无疑：一天之中，互联网产生的全部内容可以刻满1.68亿张DVD；发出的邮件有2 940亿封之多（相当于美国两年的纸质信件数量）；发出的社区帖子达200万个（相当于《时代》杂志770年的文字量）；卖出的手