资深数据分析咨询师多年经验结晶，内容全面而深入，为高效利用数据分析工具和算法进行数据分析提供翔实指导

通过大量典型数据分析案例，全面阐释分类分析、聚类分析、数据可视化及预测方面的各种技术和方法，为快速掌握并灵活运用数据分析技术提供最佳实践指南

大数据技术丛书

Practical Data Analysis

实用数据分析

（美） Hector Cuesta 著

刁晓纯 陈堰平◎译

机械工业出版社
China Machine Press

大数据技术丛书

Practical Data Analysis

实用数据分析

（美） Hector Cuesta 著

刁晓纯 陈堰平◎译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

实用数据分析 /（美）奎斯塔（Cuesta，H.）著；刁晓纯，陈堰平译 . —北京：机械工业出版社，2014.8
（大数据技术丛书）
书名原文：Practical Data Analysis

ISBN 978-7-111-47623-8

I. 实… II. ① 奎… ② 刁… ③ 陈… III. 统计数据－统计分析 IV. O212.1

中国版本图书馆 CIP 数据核字（2014）第 185713 号

本书版权登记号：图字：01-2013-9377

实用数据分析

[美] Hector Cuesta 著

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：秦 健 责任校对：董纪丽
印 刷：北京市荣盛彩色印刷有限公司 版 次：2014 年 9 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：15.5
书 号：ISBN 978-7-111-47623-8 定 价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

The Translator's Words 译者序

2013 年 7 月 20 日，我参加了“数据驱动企业 分析变革商业”2013 第三届大数据世界论坛。期间一个偶然的机会我认识了机械工业出版社的编辑王春华。当时我对她介绍说，我在中国邮政集团公司从事数据方面的工作已经有 3 年了，作为国内大型央企我们所开展的数据分析涉及非常广泛的范围，既跨越了银行保险、公共服务、电子商务及速递物流等行业，也包括了对客户、渠道、价格、实物网效率、经营业绩等多方面的分析。但是我们也遇到了很多问题，包括组织方面的，也有方法效率方面的。带着这些工作中的问题，我问她有没有相关方面的书籍可供参考。她没有直接回答我，却推荐了这本书，说感兴趣的话可以分享给国内的读者。就这样一来二去，我抱着探索和尝试的精神，利用 2013 年春节假期对本书进行了翻译。令人惊喜的是，本书所介绍的广泛案例、先进的方法以及诸多便利的工具都对数据分析工作有很多帮助和值得借鉴的地方。在翻译过程中，我的主要体会有三方面：

第一，本书包含丰富的案例。本书介绍的案例涉及垃圾邮件的分类分析、图像匹配案例、流行病暴发事件分析、社交网络的数据获取和分析、对文本型数据进行情感分析、股票价格以及黄金价格走势等。其中图像匹配的案例让我想到了很多，现实生活中我经常会看到一些漂亮的实物，但是除了一些品牌以外，的确很难找到具体的出处。如果可以上传拍摄到的图片，再通过图像匹配技术在互联网上找到最贴近的网上商家，那么这种大数据技术的应用方式可以更大范围地扩展电子商务交易的范畴。

第二，本书所涉内容包含了数据分析全流程，包括了数据准备和处理、多类型建模、数据可视化展示等。初次接触数据分析的读者可以由浅入深地了解分析的全貌。

第三，本书充分体现了大数据的特点，既介绍了对结构化数据的处理也介绍了对非结构化数据的处理，数据类型丰富。书中所涉数据包括时间序列数据、数值型数据、多维度数据和社交媒体数据、文本型数据等多种数据形式，可以帮助读者获得对数据分析的真知灼见。

本书是我和陈堰平共同翻译的成果，我负责翻译除了第6章和第7章以外的全部内容。后期机械工业出版社的编辑做了大量的文字整理工作。因为大数据是一个比较新颖的领域，一些术语在业界还没有得到统一，书中会有一些内容是按照原文的直意进行翻译，难免有不完整或者偏颇的地方，欢迎广大读者与我交流沟通，我的邮箱是 jacqueline_dut@hotmail.com，请大家批评指正。

刁晓纯

Foreword 序

“从数据到信息，再从信息到知识”这样的发展阶段，已经变成老生常谈并且不再适应当下发展的需要。随着大数据分析以及在大量分散的个人数据集中获取数据并进行分析的需求逐步呈现，数据驱动领域的实践者对于应用一系列分析方法存在着需求。无论是在数据准备和数据清洗期间，或在数据探索期间，使用计算工具的需要已经变得很迫切。然而，数据分析人员在各自领域挖掘可获取的数据潜藏的丰富内涵时，基本理论的复杂性对分析方法的应用提出了挑战。在一些领域中，文本形式的数据甚至隐藏了可能毁掉一笔成功生意的秘密。还有一些领域，对社会化网络的分析和对情感的分类分析将揭示出一些针对信息传播和政策制订的新策略。

我和我的学生们主要的研究领域是计算流行病学方面。这一学科主要是通过设计并应用工具来研究疾病在大量人群中如何进行传播。复杂的模拟建模被用于预测或者至少给出某种流行病最有可能的发展轨道。此类模型的开发依赖于数据的可获得性或者说能否提取出人口数据和疾病特征的相关变量。无论是包含人口构成信息的普查数据，还是描述个人疾病变化的医疗文字，对此类数据的探索是一项很有挑战性的工作。正如很多使用数据分析法的领域一样，计算流行病学从本质上说也是一个多学科的专业。当一些数据资源会显示出一个蚊子可能产下多少颗卵时，另外一些资源则会指出蚊子与人口间可能进行的交互所导致登革热和西尼罗病毒流行的概率。为了将信息转化为知识，计算科学家、生物学家、生物统计学家以及公共健康实践者们必须通力合作。正是对成熟的可视化工具的应用才让这些来自不同学科的科学家以及实践者们得以探索数据并分享他们的真知灼见。

2011 年秋季学期，我遇到了 Hector Cuesta，那时他以访问学者的身份加入了计算流行病研究实验室。我马上意识到 Hector 不仅仅是一名出色的编程人员，同时也是一位实践者，他愿意将计算机范式应用在不同背景的问题之中。他的专长体现在诸多计算机语言和工具方面，

包括 Python、CUDA、Hadoop、SQL 和 MPI。这些语言让他能够对来自不同领域的复杂问题提出建设性的解决方案。在本书里，Hector Cuesta 展示了如何将不同的数据分析工具应用于不同领域的问题中。采用不同类型的数据集来推动大家探索使用有效的计算方法，这些计算方法已经被轻而易举地应用在其他一些有问题的领域。因此，本书既是一本参考指南，同时对数据分析的实践者而言又是一本从数据获取信息，再从信息发现知识的实践教程。

Armin R. Mikler
计算机科学与工程专业教授
计算流行病学和反应分析中心主任
美国北德克萨斯大学

Preface 前言

本书提供了一系列现实中将数据转化为洞察力的案例。书中覆盖了广泛的数据分析工具和算法，用于进行分类分析、聚类分析、数据可视化、数据模拟以及预测。本书的目标是帮助你了解数据从而找到相应的模式、趋势、相互关系以及洞察力。

书中所包括的实用项目充分利用了 MongoDB、D3.js 和 Python 语言并采用代码片段和详细描述的方式向读者呈现本书的核心概念。

本书组织结构

第 1 章探讨数据分析的基本原理和数据分析步骤。

第 2 章解释如何清洗并准备好数据来开展分析，同时介绍了数据清洗工具 OpenRefine 的使用方式。

第 3 章展示在 JavaScript 可视化框架下应用 D3.js 语言来实现各类数据的可视化方法。

第 4 章介绍了应用朴素贝叶斯算法（Naïve Bayes）来区分垃圾文本的一种二元分类法。

第 5 章展示了一个应用动态时间规整方法来寻找图像间相似性的项目。

第 6 章解释了如何使用随机游走算法和可视化的 D3.js 动画技术来模拟股票价格的内容。

第 7 章介绍核岭回归（Kernel Ridge Regression，KRR）的原理以及如何使用此方法和时间序列数据来预测黄金价格。

第 8 章描述如何使用支持向量机的方法进行分类分析。

第 9 章介绍了对流行病进行模拟计算的基本概念并解释如何应用细胞自动机方法、D3.js 和 JavaScript 语言来实现对流行病爆发的模拟。

第 10 章解释如何应用 Gephi 从 Facebook 获取你的社会化媒体图谱并使之实现可视化。

第 11 章解释如何应用 Twitter 的应用程序编程接口（API）来获取 Twitter 的数据。读者也将看到如何改进文本分类分析方法并将其应用于情感分析。这一过程是在自然语言工具包

（Natural Language Toolkit, NLTK）中应用了朴素贝叶斯算法。

第 12 章介绍在 MongoDB 数据库中进行基本操作以及进行分组、过滤和聚合的方法。

第 13 章详细介绍如何在 MongoDB 数据库中应用 MapReduce 编程模型。

第 14 章解释了如何使用 Wakari 平台，同时介绍了 IPython 中运用 Pandas 进行数据处理和使用 PIL 图像处理库的方法。

附录提供书中所使用的软件工具的详细安装信息。

本书技能要求

使用本书的基本要求是掌握如下技术：

- ❑ Python
- ❑ OpenRefine
- ❑ D3.js
- ❑ mlpy
- ❑ Natural Language Toolkit (NLTK)
- ❑ Gephi
- ❑ MongoDB

本书读者对象

本书主要面向那些希望能够实际开展数据分析和数据可视化的软件开发人员、分析人员、计算机科学家。同时，本书也希望能够为读者提供包含时间序列数据、数值型数据、多维度数据和社会化媒体数据、文本型数据等多种数据形式的、内容完备的真实项目，以帮助读者获得对数据分析的真知灼见。读者不需要具备数据分析的经验，但仍需要对统计学和 Python 编程有基础性的了解。

本书排版约定

在本书中，你将发现很多文本格式的差别，用来区分不同类型的信息。下面给出了一些文本格式的例子以及对这些文本格式的解释。

代码段会采用如下格式：

```
beta = 0.003
gamma = 0.1
sigma = 0.1

def SIRS_model(X, t=0):
```

```
  r = scipy.array([- beta*X[0]*X[1] + sigma*X[2]
    ,  beta*X[0]*X[1] - gamma*X[1]
    ,  gamma*X[1] ] -sigma*X[2])
  return r
```

当我们希望你注意某段特别的代码时，我们会着重标注相关的文字和信息，具体如下：

```
[[215  10   0]
 [153  72   0]
 [ 54 171   0]
 [  2 223   0]
 [  0 225   0]
 [  0 178  47]
 [  0  72 153]
 [  0   6 219]
 [  0   0 225]
 [ 47   0 178]
 [153   0  72]
 [219   0   6]
 [225   0   0]]
```

任何命令行的输入和输出都将采用如下方式：

```
db.runCommand( { count: TweetWords })
```

新术语和**重要的文字**将被加粗。你在屏幕、菜单或者对话框中看到的文字示例将会采用如下方式进行显示，例如：“接下来我们能够看到如下的屏幕显示，我们将点击 Map Reduce 选项。”

下载示例代码

通过你的账户登录 http://www.packtpub.com，你能够在示例代码文件夹中找到你在 Packt 出版社网站所购买的全部书籍。如果你的书籍是在其他地方采购的，那么可以登录 http://www.packtpub.com/support 进行注册，我们将通过电子邮件的方式直接将示例代码文件发送给你。

勘误声明

尽管我们已经竭尽所能来确保我们的内容准确无误，但错误难免存在。如果你在我们的任何一本书中发现了任何错误，无论是正文还是代码，对于你所反馈的信息我们都将非常感激。这样能够帮助其他读者避免困惑，同时也能够帮助我们提升该书后续版本的质量。如果你找到任何勘误，请通过 http://www.packtpub.com/submit-errata 进行反馈，具体路径是：选择你的书籍，点击勘误提交表格的链接，然后输入勘误的详细内容。一旦你所提供的勘误被确认，你的提交将被接受，相关勘物内容将上传到我们的网站，或者增加到任何现有相关的勘误主题栏目中。任何已有的勘误结果可以在 http://www.packtpub.com/support 网址查到，并可以通过栏目标题进行选择。

评审者简介 *About the Reviewers*

Mark Kerzner 拥有法学、数学和计算机科学学位。他已经从事软件设计多年，并从2008年起开始设计基于Hadoop的系统。他现在是SHMsoft的总裁，这家企业为不同领域提供Hadoop应用，同时他也是《Hadoop Illumiated》一书（项目）的联合作者。他创作或联合创作了很多书籍和专题。

我要感谢我的同事们，特别是Sujee Maniyam对我的帮助，并且我还要感谢我多才多艺的家人对我的帮助。

Sampath Kumar博士 是Telangana大学应用统计学习的助理教师和系主任。他完成了理学学士、理学硕士和博士阶段的学习与研究。他拥有5年研究生教学经验，有超过4年的公司工作经验。他的专长是利用SPSS、SAS、R、Minitab、MATLAB等软件进行数据统计。他是SAS和MATLAB软件高级编程人员。他在不同的应用学科和纯粹统计专业，如预测建模、应用回归分析、多变量数据分析、运营管理等方面具有教学经验。他目前是博士生导师。

Rciky J. Sethi 目前是Madsci网络的研发总监，Massachusetts大学医学中心和Mass Amherst的科学研究人员。Sethi博士的研究偏重于自然科学领域中的多个交叉学科，他在计算机显示、社会化计算、科学学习领域主要采用机器学习的方法和基于物理学的建模。他在加州大学伯克利分校获得了分子和细胞生物物理学士学位，在南加州大学获得了物理和管理信息系统专业硕士学位，并在加州大学获得了计算机科学博士学位。他是30多篇论文和书籍章节的作者或联合作者，同时也是加州大学和南加州大学信息科学学院NSF计算机创新奖获得者。

Suchita Tripathi 在Allahabad大学人类学专业完成了她的理学硕士和博士的学习。她同时也熟练使用SPSS软件和其他计算机应用进行数据分析。她精通海地语、英语和日语。她在Sendai、日本等地的培训学校进行了初级和中级日语培训并获得了很多认证。她是6篇文章以

及一本专著的作者。她在 GGV 中心大学人类学和部落发展系执教 2 年。她的主要研究领域是城市人类学、人类灾害、语言学和人类考古学。

我要对我的父母和我所爱的家庭为我提供的精神支持和祝福表示感谢。

Jarrell Waggoner 博士 是 Groupon 的一位软件工程师，主要应用内部工具开展销售分析并进行市场需求预测。他在南加州大学获得了计算机科学与工程专业的博士学位，并在计算机显示和图像处理方面完成了诸多科学项目，包括一项国家人文基金会资助的文件图像处理项目、一项美国国防高级研究计划局竞赛项目：建设事件再识别系统、一项由美国空军科学研究局资助的材料科学影响识别处理项目。他是开源软件的忠实支持者，在他的研究中使用了很多开源语言、操作系统和软件开发框架。他所从事的开源项目及成果以及他的研究工作可以在 Github（https://github.com/malloc47）和他的个人主页（http://www.malloc47.com）上找到。

致谢 *Acknowledgments*

我要把这本书献给我的妻子 Yolanda 和我可爱的孩子们 Damiana 和 Issac，因为他们为我的生活带来了无比的快乐。同时把这本书献给我的父母 Elena 和 Miguel，感谢他们对我长期的支持和爱护。

我同时也要感谢 Packt 出版社的出版团队，特别要感谢 Anurag Banerjee、Erol Staveley、Edward Gordon、Anugya Khurana、Neeshma Ramakrishnan、Arwa Manasawala、Manal Pednekar、Pragnesh Bilimoria 和 Unnati Shah。

谢谢我的朋友们对本书草稿所提出的有益建议和改进，他们是 Abel Valle、Oscar Manso、Ivan Cervantes、Agustin Ramos、Rene Cruz 博士、Adrian Trueba 博士和 Sergio Ruiz。我也要特别感谢那些技术评审者，他们花费了大量时间对本书草稿提出了细致的反馈。

我还要感谢 Armin Mikler 博士对我的鼓励和对本书的推荐。最后，作为重要的灵感之源，我要感谢我的导师 Jesus Figueroa-Nazuno 博士，他也是我之前的主管。

Contents 目录