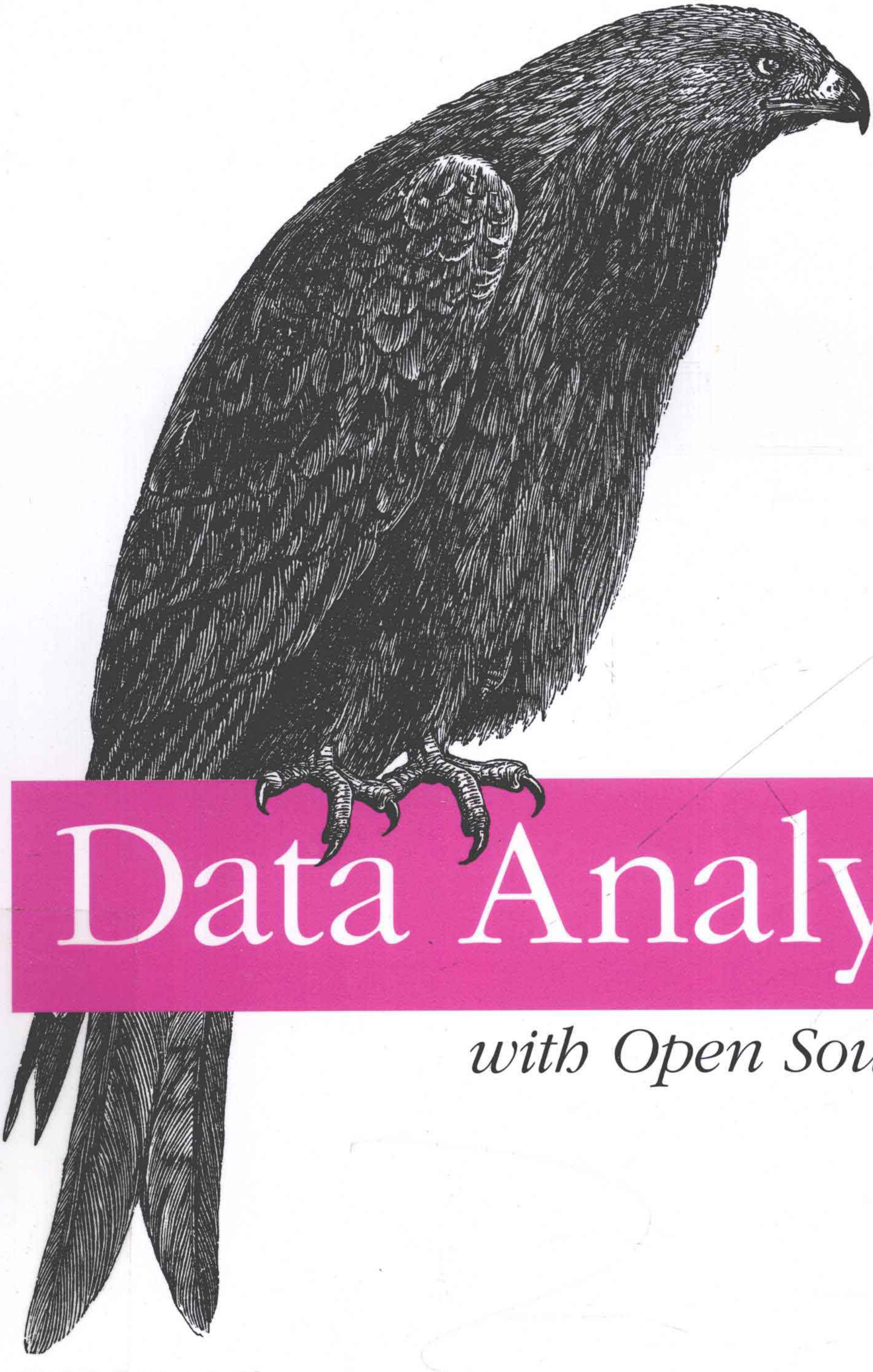
基于开源工具的数据分析（影印版）



Data Analysis

with Open Source Tools

O'REILLY®

東南大學出版社

Philipp K. Janert 著

基于开源工具的数据分析（影印版）

Data Analysis with Open Source Tools

Philipp K. Janert

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Taipei • Tokyo

O'Reilly Media, Inc. 授权东南大学出版社出版

东南大学出版社

图书在版编目（CIP）数据

基于开源工具的数据分析：英文/（美）雅奈特（Janert, P.K.）著.—影印本.—南京：东南大学出版社，2011.5

书名原文：Data Analysis with Open Source Tools
ISBN 978-7-5641-2674-2

I. ①基… II. ①雅… III. ① 软件工具－英文 IV. ① TP311.56

中国版本图书馆 CIP 数据核字（2011）第 042775 号

江苏省版权局著作权合同登记
图字：10-2010-446 号

©2010 by O'Reilly Media, Inc.

Reprint of the English Edition, jointly published by O'Reilly Media, Inc. and Southeast University Press, 2011. Authorized reprint of the original English edition, 2010 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版 2010。

英文影印版由东南大学出版社出版 2011。此影印版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc. 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

基于开源工具的数据分析（影印版）

出版发行：东南大学出版社
地　　址：南京四牌楼 2 号　　邮编：210096
出 版 人：江建中
网　　址：http://www.seupress.com
电子邮件：press@seupress.com
印　　刷：扬中市印刷有限公司
开　　本：787 毫米 × 980 毫米　16 开本
印　　张：33.25 印张
字　　数：824 千字
版　　次：2011 年 5 月第 1 版
印　　次：2011 年 5 月第 1 次印刷
书　　号：ISBN 978-7-5641-2674-2
定　　价：82.00 元（册）

本社图书若有印装质量问题，请直接与读者服务部联系。电话（传真）：025-83792328

基于开源工具的数据分析（影印版）

Data Analysis with Open Source Tools

Furious activity is no substitute for understanding.

—H. H. Williams

此为试读,需要完整PDF请访问: www.ertongbook.com

Preface

THIS BOOK GREW OUT OF MY EXPERIENCE OF WORKING WITH DATA FOR VARIOUS COMPANIES IN THE TECH industry. It is a collection of those concepts and techniques that I have found to be the most useful, including many topics that I wish I had known earlier—but didn't.

My degree is in physics, but I also worked as a software engineer for several years. The book reflects this dual heritage. On the one hand, it is written for programmers and others in the software field: I assume that you, like me, have the ability to write your own programs to manipulate data in any way you want.

On the other hand, the way I think about data has been shaped by my background and education. As a physicist, I am not content merely to describe data or to make black-box predictions: the purpose of an analysis is always to develop an understanding for the processes or mechanisms that give rise to the data that we observe.

The instrument to express such understanding is the *model*: a description of the system under study (in other words, not just a description of the data!), simplified as necessary but nevertheless capturing the relevant information. A model may be crude ("Assume a spherical cow . . . "), but if it helps us develop better insight on how the system works, it is a successful model nevertheless. (Additional precision can often be obtained at a later time, if it is really necessary.)

This emphasis on models and simplified descriptions is not universal: other authors and practitioners will make different choices. But it is essential to my approach and point of view.

This is a rather personal book. Although I have tried to be reasonably comprehensive, I have selected the topics that I consider relevant and useful in practice—whether they are part of the "canon" or not. Also included are several topics that you won't find in any other book on data analysis. Although neither new nor original, they are usually not used or discussed in this particular context—but I find them indispensable.

Throughout the book, I freely offer specific, explicit advice, opinions, and assessments. These remarks are reflections of my personal interest, experience, and understanding. I do not claim that my point of view is necessarily correct: evaluate what I say for yourself and feel free to adapt it to your needs. In my view, a specific, well-argued position is of greater use than a sterile laundry list of possible algorithms—even if you later decide to disagree with me. The value is not in the opinion but rather in the arguments leading up to it. If your arguments are better than mine, or even just more agreeable to you, then I will have achieved my purpose!

Data analysis, as I understand it, is not a fixed set of techniques. It is a way of life, and it has a name: curiosity. There is always something else to find out and something more to learn. This book is not the last word on the matter; it is merely a snapshot in time: things I knew about and found useful today.

"Works are of value only if they give rise to better ones."

(Alexander von Humboldt, *writing to* Charles Darwin, 18 September 1839)

Before We Begin

More data analysis efforts seem to go bad because of an excess of sophistication rather than a lack of it.

This may come as a surprise, but it has been my experience again and again. As a consultant, I am often called in when the initial project team has already gotten stuck. Rarely (if ever) does the problem turn out to be that the team did not have the required skills. On the contrary, I usually find that they tried to do something unnecessarily complicated and are now struggling with the consequences of their own invention!

Based on what I have seen, two particular risk areas stand out:

- The use of "statistical" concepts that are only partially understood (and given the relative obscurity of most of statistics, this includes virtually *all* statistical concepts)
- Complicated (and expensive) black-box solutions when a simple and transparent approach would have worked at least as well or better

I strongly recommend that you make it a habit to avoid all statistical language. Keep it simple and stick to what you know for sure. There is absolutely nothing wrong with speaking of the "range over which points spread," because this phrase means exactly what it says: the range over which points spread, and only that! Once we start talking about "standard deviations," this clarity is gone. Are we still talking about the *observed* width of the distribution? Or are we talking about one specific *measure* for this width? (The standard deviation is only one of several that are available.) Are we already making an implicit *assumption* about the nature of the distribution? (The standard deviation is only suitable under certain conditions, which are often not fulfilled in practice.) Or are we even confusing the *predictions* we could make if these assumptions were true with the actual data? (The moment someone talks about "95 percent anything" we know it's the latter!)

I'd also like to remind you not to discard simple methods until they have been *proven* insufficient. Simple solutions are frequently rather effective: the marginal benefit that more complicated methods can deliver is often quite small (and may be in no reasonable relation to the increased cost). More importantly, simple methods have fewer opportunities to go wrong or to obscure the obvious.

True story: a company was tracking the occurrence of defects over time. Of course, the actual number of defects varied quite a bit from one day to the next, and they were looking for a way to obtain an estimate for the typical number of expected defects. The solution proposed by their IT department involved a compute cluster running a neural network! (I am not making this up.) In fact, a one-line calculation (involving a moving average or single exponential smoothing) is all that was needed.

I think the primary reason for this tendency to make data analysis projects more complicated than they are is *discomfort*: discomfort with an unfamiliar problem space and uncertainty about how to proceed. This discomfort and uncertainty creates a desire to bring in the "big guns": fancy terminology, heavy machinery, large projects. In reality, of course, the opposite is true: the complexities of the "solution" overwhelm the original problem, and nothing gets accomplished.

Data analysis does not have to be all that hard. Although there are situations when elementary methods will no longer be sufficient, they are much less prevalent than you might expect. In the vast majority of cases, curiosity and a healthy dose of common sense will serve you well.

The attitude that I am trying to convey can be summarized in a few points:

> Simple is better than complex.
> Cheap is better than expensive.
> Explicit is better than opaque.
> Purpose is more important than process.
> Insight is more important than precision.
> Understanding is more important than technique.
> Think more, work less.

Although I do acknowledge that the items on the right are necessary at times, I will give preference to those on the left whenever possible.

It is in this spirit that I am offering the concepts and techniques that make up the rest of this book.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
  Indicates new terms, URLs, and email addresses

`Constant width`
  Used to refer to language and script elements

Using Code Examples

This book is here to help you get your job done. In general, you may use the code in this book in your programs and documentation. You do not need to contact us for permission unless youre reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from OReilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your products documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Data Analysis with Open Source Tools*, by Philipp K. Janert. Copyright 2011 Philipp K. Janert, 978-0-596-80235-6."

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at *permissions@oreilly.com.*

Safari® Books Online

Safari Books Online is an on-demand digital library that lets you easily search over 7,500 technology and creative reference books and videos to find the answers you need quickly.

With a subscription, you can read any page and watch any video from our library online. Read books on your cell phone and mobile devices. Access new titles before they are available for print, and get exclusive access to manuscripts in development and post feedback for the authors. Copy and paste code samples, organize your favorites, download chapters, bookmark key sections, create notes, print out pages, and benefit from tons of other time-saving features.

O'Reilly Media has uploaded this book to the Safari Books Online service. To have full digital access to this book and others on similar topics from OReilly and other publishers, sign up for free at *http://my.safaribooksonline.com.*

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at:

http://oreilly.com/catalog/9780596802356

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com

For more information about our books, conferences, Resource Centers, and the O'Reilly Network, see our website at:

http://oreilly.com

Acknowledgments

It was a pleasure to work with O'Reilly on this project. In particular, O'Reilly has been most accommodating with regard to the technical challenges raised by my need to include (for an O'Reilly book) an uncommonly large amount of mathematical material in the manuscript.

Mike Loukides has accompanied this project as the editor since its beginning. I have enjoyed our conversations about life, the universe, and everything, and I appreciate his comments about the manuscript—either way.

I'd like to thank several of my friends for their help in bringing this book about:

- Elizabeth Robson, for making the connection
- Austin King, for pointing out the obvious
- Scott White, for suffering my questions gladly
- Richard Kreckel, for much-needed advice

As always, special thanks go to PAUL Schrader (Bremen).

The manuscript benefited from the feedback I received from various reviewers. Michael E. Driscoll, Zachary Kessin, and Austin King read all or parts of the manuscript and provided valuable comments.

I enjoyed personal correspondence with Joseph Adler, Joe Darcy, Hilary Mason, Stephen Weston, Scott White, and Brian Zimmer. All very generously provided expert advice on specific topics.

Particular thanks go to Richard Kreckel, who provided uncommonly detailed and insightful feedback on most of the manuscript.

During the preparation of this book, the excellent collection at the University of Washington libraries was an especially valuable resource to me.

Authors usually thank their spouses for their "patience and support" or words to that effect. Unless one has lived through the actual experience, one cannot fully comprehend how true this is. Over the last three years, Angela has endured what must have seemed like a nearly continuous stream of whining, frustration, and desperation—punctuated by occasional outbursts of exhilaration and grandiosity—all of which before the background of the self-centered and self-absorbed attitude of a typical author. Her patience and support were unfailing. It's her turn now.

此为试读，需要完整PDF请访问：www.ertongbook.com

CONTENTS

PART I Graphics: Looking at Data

PART II Analytics: Modeling Data

PART III Computation: Mining Data

PART IV Applications: Using Data

CHAPTER ONE

Introduction

IMAGINE YOUR BOSS COMES TO YOU AND SAYS: "HERE ARE 50 GB OF LOGFILES—FIND A WAY TO IMPROVE OUR business!"

What would you do? Where would you start? And what would you do next?

It's this kind of situation that the present book wants to help you with!

Data Analysis

Businesses sit on data, and every second that passes, they generate some more. Surely, there *must* be a way to make use of all this stuff. But how, exactly—that's far from clear.

The task is difficult because it is so vague: there is no specific problem that needs to be solved. There is no specific question that needs to be answered. All you know is the overall *purpose*: improve the business. And all you have is "the data." Where do you start?

You start with the only thing you have: "the data." What is it? We don't know! Although 50 GB sure sounds like a lot, we have no idea what it actually contains. The first thing, therefore, is to *take a look*.

And I mean this literally: the first thing to do is to *look* at the data by plotting it in different ways and looking at graphs. Looking at data, you will notice things—the way data points are distributed, or the manner in which one quantity varies with another, or the large number of outliers, or the total absence of them.... I don't know what you will find, but there is no doubt: if you look at data, you will observe things!

These observations should lead to some reflection. "Ten percent of our customers drive ninety percent of our revenue." "Whenever our sales volume doubles, the number of

此为试读，需要完整PDF请访问：www.ertongbook.com