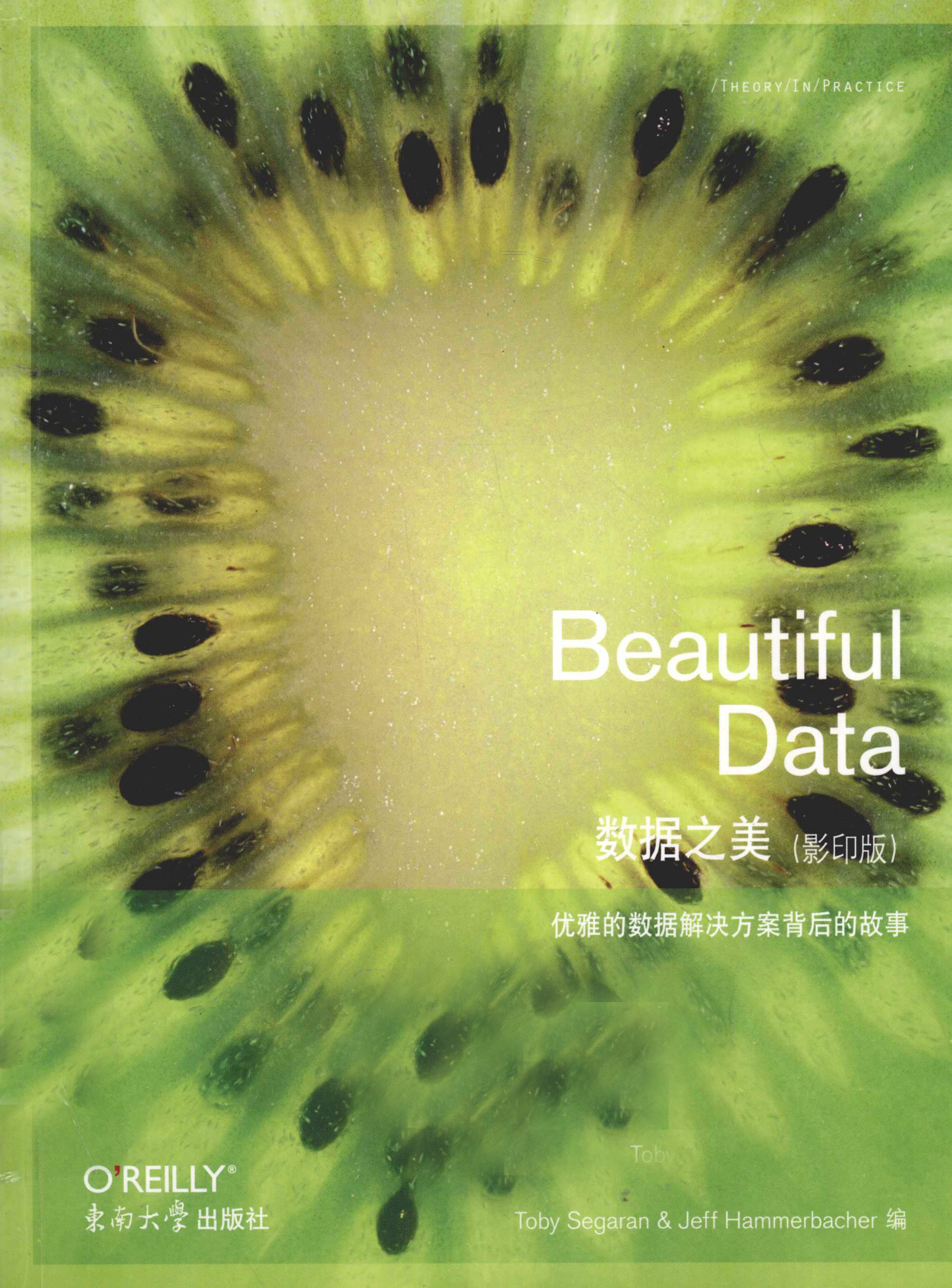
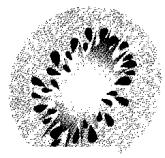

/THEORY/IN/PRACTICE
Beautiful Data
数据之美（影印版）
优雅的数据解决方案背后的故事
O'REILLY®
東南大學出版社
Toby Segaran & Jeff Hammerbacher 编

数据之美(影印版)

Beautiful Data



O'REILLY®

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

东南大学出版社

图书在版编目（CIP）数据

数据之美：英文/（美）西格兰（Segaran, T.），（美）哈梅巴赫（Hammerbacher, J.）著. —影印本. —南京：东南大学出版社，2010.6

书名原文：Beautiful Data

ISBN 978-7-5641-2272-0

Ⅰ. ①数… Ⅱ. ①西… ②哈… Ⅲ. ①数据处理－英文 Ⅳ. ① TP274

中国版本图书馆CIP数据核字（2010）第089204号

江苏省版权局著作权合同登记

图字：10-2010-152号

数据之美（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　汉
网　　址：http://press.seu.edu.cn
电子邮件：press@seu.edu.cn
印　　刷：扬中市印刷有限公司
开　　本：787毫米×980毫米　16开本
印　　张：黑白 24印张　彩色 2印张
字　　数：644千字
版　　次：2010年6月第1版
印　　次：2010年6月第1次印刷
书　　号：ISBN 978-7-5641-2272-0
印　　数：1~1800册
定　　价：62.00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

All royalties from this book will be donated to Creative Commons and the Sunlight Foundation.

Preface

When we were first approached with the idea of a follow-up to *Beautiful Code*, this time about data, we found the idea exciting and very ambitious. Collecting, visualizing, and processing data now touches every professional field and so many aspects of daily life that a great collection would have to be almost unreasonably broad in scope. So we contacted a highly diverse group of people whose work we admired, and were thrilled that so many agreed to contribute.

This book is the result, and we hope it captures just how wide-ranging (and beautiful) working with data can be. In it you'll learn about everything from fighting with governments to working with the Mars lander; you'll learn how to use statistics programs, make visualizations, and remix a Radiohead video; you'll see maps, DNA, and something we can only really call "data philosophy."

The royalties for this book are being donated to Creative Commons and the Sunlight Foundation, two organizations dedicated to making the world better by freeing data. We hope you'll consider how your own encounters with data shape the world.

How This Book Is Organized

The chapters in this book follow a loose arc from data collection through data storage, organization, retrieval, visualization, and finally, analysis.

Chapter 1, *Seeing Your Life in Data*, by Nathan Yau, looks at the motivations and challenges behind two projects in the emerging field of personal data collection.

Chapter 2, *The Beautiful People: Keeping Users in Mind When Designing Data Collection Methods*, by Jonathan Follett and Matthew Holm, discusses the importance of trust, persuasion, and testing when collecting data from humans over the Web.

Chapter 3, *Embedded Image Data Processing on Mars*, by J. M. Hughes, discusses the challenges of designing a data processing system that has to work within the constraints of space travel.

Chapter 4, *Cloud Storage Design in a PNUTShell*, by Brian F. Cooper, Raghu Ramakrishnan, and Utkarsh Srivastava, describes the software Yahoo! has designed to turn its globally distributed data centers into a universal storage platform for powering modern web applications.

Chapter 5, *Information Platforms and the Rise of the Data Scientist*, by Jeff Hammerbacher, traces the evolution of tools for information processing and the humans who power them, using specific examples from the history of Facebook's data team.

Chapter 6, *The Geographic Beauty of a Photographic Archive*, by Jason Dykes and Jo Wood, draws attention to the ubiquity and power of colorfully visualized spatial data collected by a volunteer community.

Chapter 7, *Data Finds Data*, by Jeff Jonas and Lisa Sokol, explains a new approach to thinking about data that many may need to adopt in order to manage it all.

Chapter 8, *Portable Data in Real Time*, by Jud Valeski, dives into the current limitations of distributing social and location data in real time across the Web, and discusses one potential solution to the problem.

Chapter 9, *Surfacing the Deep Web*, by Alon Halevy and Jayant Madhavan, describes the tools developed by Google to make searchable the data currently trapped behind forms on the Web.

Chapter 10, *Building Radiohead's House of Cards*, by Aaron Koblin with Valdean Klump, is an adventure story about lasers, programming, and riding on the back of a bus, and ending with an award-winning music video.

Chapter 11, *Visualizing Urban Data*, by Michal Migurski, details the process of freeing and beautifying some of the most important data about the world around us.

Chapter 12, *The Design of Sense.us*, by Jeffrey Heer, recasts data visualizations as social spaces and uses this new perspective to explore 150 years of U.S. census data.

Chapter 13, *What Data Doesn't Do*, by Coco Krumme, looks at experimental work that demonstrates the many ways people misunderstand and misuse data.

Chapter 14, *Natural Language Corpus Data*, by Peter Norvig, takes the reader through some evocative exercises with a trillion-word corpus of natural language data pulled down from across the Web.

Chapter 15, *Life in Data: The Story of DNA*, by Matt Wood and Ben Blackburne, describes the beauty of the data that is DNA and the massive infrastructure required to create, capture, and process that data.

Chapter 16, *Beautifying Data in the Real World*, by Jean-Claude Bradley, Rajarshi Guha, Andrew Lang, Pierre Lindenbaum, Cameron Neylon, Antony Williams, and Egon Willighagen, shows how crowdsourcing and extreme transparency have combined to advance the state of drug discovery research.

Chapter 17, *Superficial Data Analysis: Exploring Millions of Social Stereotypes*, by Brendan O'Connor and Lukas Biewald, shows the correlations and patterns that emerge when people are asked to anonymously rate one another's pictures.

Chapter 18, *Bay Area Blues: The Effect of the Housing Crisis*, by Hadley Wickham, Deborah F. Swayne, and David Poole, guides the reader through a detailed examination of the recent housing crisis in the Bay Area using open source software and publicly available data.

Chapter 19, *Beautiful Political Data*, by Andrew Gelman, Jonathan P. Kastellec, and Yair Ghitza, shows how the tools of statistics and data visualization can help us gain insight into the political process used to organize society.

Chapter 20, *Connecting Data*, by Toby Segaran, explores the difficulty and possibilities of joining together the vast number of data sets the Web has made available.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`
: Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

`Constant width bold`
: Shows commands or other text that should be typed literally by the user.

`Constant width italic`
: Shows text that should be replaced with user-supplied values or by values determined by context.

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 you're 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 O'Reilly 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 product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Beautiful Data*, edited by Toby Segaran and Jeff Hammerbacher. Copyright 2009 O'Reilly Media, Inc., 978-0-596-15711-1."

If you feel your use of code examples falls outside fair use or the permission given here, feel free to contact us at *permissions@oreilly.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/9780596157111

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

Safari® Books Online

When you see a Safari® Books Online icon on the cover of your favorite technology book, that means the book is available online through the O'Reilly Network Safari Bookshelf.

Safari offers a solution that's better than e-books. It's a virtual library that lets you easily search thousands of top tech books, cut and paste code samples, download chapters, and find quick answers when you need the most accurate, current information. Try it for free at *http://my.safaribooksonline.com*.

CONTENTS

CHAPTER ONE

Seeing Your Life in Data

Nathan Yau

IN THE NOT-TOO-DISTANT PAST, THE WEB WAS ABOUT SHARING, BROADCASTING, AND DISTRIBUTION. But the tide is turning: the Web is moving toward the individual. Applications spring up every month that let people track, monitor, and analyze their habits and behaviors in hopes of gaining a better understanding about themselves and their surroundings. People can track eating habits, exercise, time spent online, sexual activity, monthly cycles, sleep, mood, and finances online. If you are interested in a certain aspect of your life, chances are that an application exists to track it.

Personal data collection is of course nothing new. In the 1930s, Mass Observation, a social research group in Britain, collected data on various aspects of everyday life—such as beards and eyebrows, shouts and gestures of motorists, and behavior of people at war memorials—to gain a better understanding about the country. However, data collection methods have improved since 1930. It is no longer only a pencil and paper notepad or a manual counter. Data can be collected automatically with mobile phones and handheld computers such that constant flows of data and information upload to servers, databases, and so-called data warehouses at all times of the day.

With these advances in data collection technologies, the data streams have also developed into something much heftier than the tally counts reported by Mass Observation participants. Data can update in real-time, and as a result, people want up-to-date information.

It is not enough to simply supply people with gigabytes of data, though. Not everyone is a statistician or computer scientist, and not everyone wants to sift through large data sets. This is a challenge that we face frequently with personal data collection.

While the types of data collection and data returned might have changed over the years, individuals' needs have not. That is to say that individuals who collect data about themselves and their surroundings still do so to gain a better understanding of the information that lies within the flowing data. Most of the time we are not after the numbers themselves; we are interested in what the numbers mean. It is a subtle difference but an important one. This need calls for systems that can handle personal data streams, process them efficiently and accurately, and dispense information to nonprofessionals in a way that is understandable and useful. We want something that is more than a spreadsheet of numbers. We want the story in the data.

To construct such a system requires careful design considerations in both analysis and aesthetics. This was important when we implemented the Personal Environmental Impact Report (PEIR), a tool that allows people to see how they affect the environment and how the environment affects them on a micro-level; and your.flowingdata (YFD), an in-development project that enables users to collect data about themselves via Twitter, a microblogging service.

For PEIR, I am the frontend developer, and I mostly work on the user interface and data visualization. As for YFD, I am the only person who works on it, so my responsibilities are a bit different, but my focus is still on the visualization side of things. Although PEIR and YFD are fairly different in data type, collection, and processing, their goals are similar. PEIR and YFD are built to provide information to the individual. Neither is meant as an endpoint. Rather, they are meant to spur curiosity in how everyday decisions play a big role in how we live and to start conversations on personal data. After a brief background on PEIR and YFD, I discuss personal data collection, storage, and analysis with this idea in mind. I then go into depth on the design process behind PEIR and YFD data visualizations, which can be generalized to personal data visualization as a whole. Ultimately, we want to show individuals the beauty in their personal data.

Personal Environmental Impact Report (PEIR)

PEIR is developed by the Center for Embedded Networked Sensing at the University of California at Los Angeles, or more specifically, the Urban Sensing group. We focus on using everyday mobile technologies (e.g., cell phones) to collect data about our surroundings and ourselves so that people can gain a better understanding of how they interact with what is around them. For example, DietSense is an online service that allows people to self-monitor their food choices and further request comments from dietary specialists; Family Dynamics helps families and life coaches document key features of a family's daily interactions, such as colocation and family meals; and Walkability helps residents and pedestrian advocates make observations and voice their concerns about neighborhood

walkability and connections to public transit.* All of these projects let people get involved in their communities with just their mobile phones. We use a phone's built-in sensors, such as its camera, GPS, and accelerometer, to collect data, which we use to provide information.

PEIR applies similar principles. A person downloads a small piece of software called Campaignr onto his phone, and it runs in the background. As he goes about his daily activities—jogging around the track, driving to and from work, or making a trip to the grocery store, for example—the phone uploads GPS data to PEIR's central servers every two minutes. This includes latitude, longitude, altitude, velocity, and time. We use this data to estimate an individual's impact on and exposure to the environment. Environmental pollution sensors are not required. Instead, we use what is already available on many mobile phones—GPS—and then pass this data with context, such as weather, into established environmental models. Finally, we visualize the environmental impact and exposure data. The challenge at this stage is to communicate meaning in data that is unfamiliar to most. What does it mean to emit 1,000 kilograms of carbon in a week? Is that a lot or is that a little? We have to keep the user and purpose in mind, as they drive the system design from the visualization down to the data collection and storage.

your.flowingdata (YFD)

While PEIR uses a piece of custom software that runs in the background, YFD requires that users actively enter data via Twitter. Twitter is a microblogging service that asks a very simple question: *what are you doing right now?* People can post, or more appropriately, *tweet*, what they are doing via desktop applications, email, instant messaging, and most importantly (as far as YFD is concerned), SMS, which means people can tweet with their mobile phones.

YFD uses Twitter's ubiquity so that people can tweet personal data from anywhere they can send SMS messages. Users can currently track eating habits, weight, sleep, mood, and when they go to the bathroom by simply posting tweets in a specific format. Like PEIR, YFD shows users that it is the little things that can have a profound effect on our way of life. During the design process, again, we keep the user in mind. What will keep users motivated to manually enter data on a regular basis? How can we make data collection as painless as possible? What should we communicate to the user once the data has been logged? To this end, I start at the beginning with data collection.

Personal Data Collection

Personal data collection is somewhat different from scientific data gathering. Personal data collection is usually less formal and does not happen in a laboratory under controlled conditions. People collect data in the real world where there can be interruptions, bad network connectivity, or limited access to a computer. Users are not necessarily data experts, so when something goes wrong (as it inevitably will), they might not know how to adjust.

* CENS Urban Sensing, *http://urban.cens.ucla.edu/*