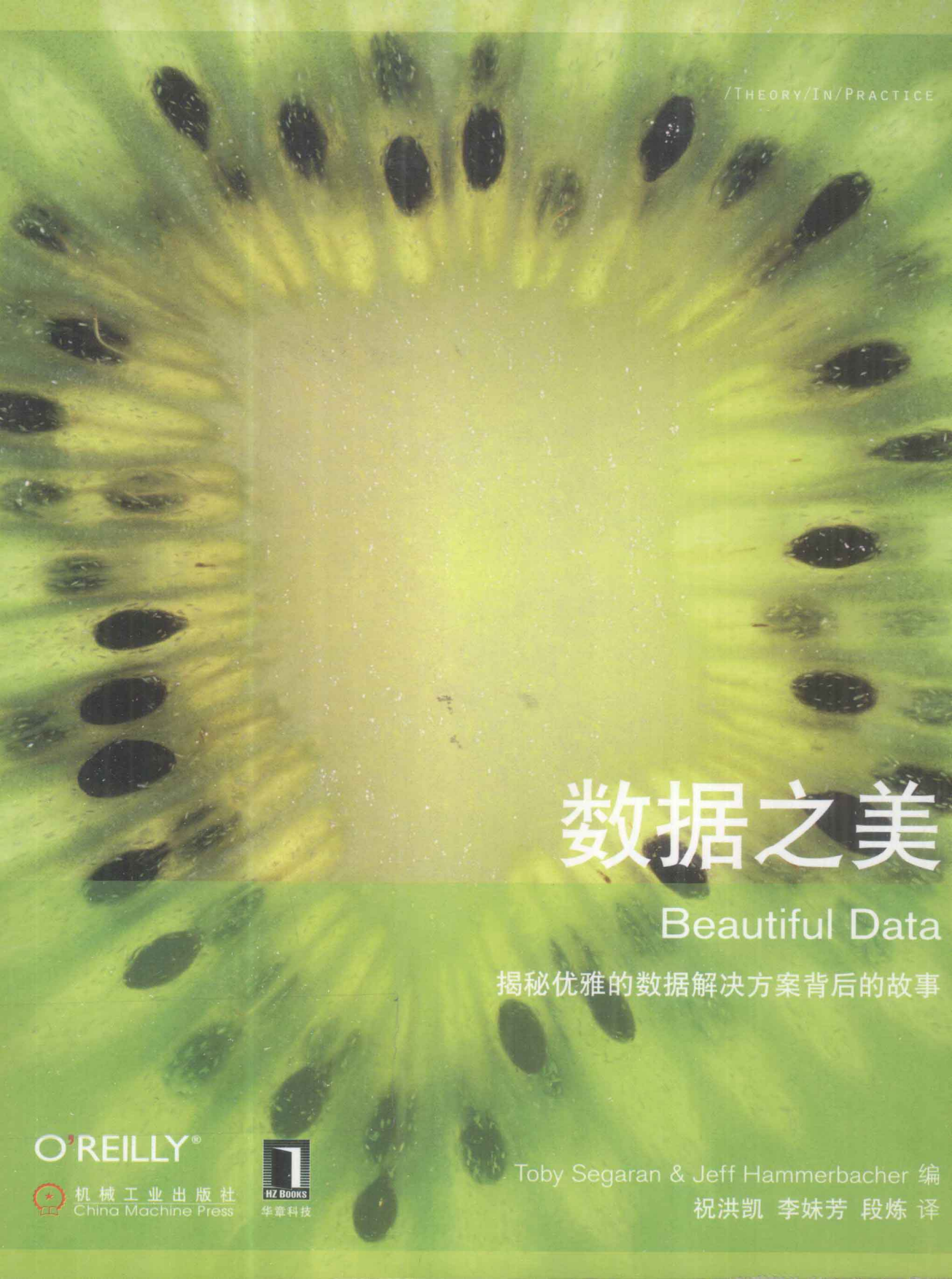


数据之美

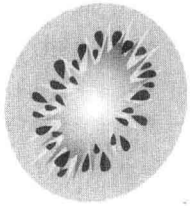
Beautiful Data

揭秘优雅的数据解决方案背后的故事

O'REILLY®
机械工业出版社
China Machine Press
HZ BOOKS
华章科技

Toby Segaran & Jeff Hammerbacher 编
祝洪凯 李妹芳 段炼 译

数据之美



Jeff Hammerbacher 编

祝洪凯 李妹芳 段炼 译

O'REILLY®

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

O'Reilly Media, Inc.授权机械工业出版社出版

机械工业出版社

图书在版编目（CIP）数据

数据之美/（美）托比（Segaran, T.）等编；祝洪凯，李妹芳，段炼译. －北京：机械工业出版社，2010.8

（O'Reilly精品图书系列）

书名原文：Beautiful Data: The Stories Behind Elegant Data Solutions

ISBN 978-7-111-31512-4

I. 数… II. ①托… ②祝… ③李… ④段… III. 数据处理 IV. TP274

中国版本图书馆CIP数据核字（2010）第153394号

北京市版权局著作权合同登记

图字：01-2009-6775号

书　　名/　数据之美

书　　号/　ISBN 978-7-111-31512-4

责任编辑/　秦健

封面设计/　Mark Paglietti，张健

出版发行/　机械工业出版社

地　　址/　北京市西城区百万庄大街22号（邮政编码100037）

印　　刷/　北京京师印务有限公司

开　　本/　178毫米×233毫米　16开本　25.25印张（含2印张彩插）

版　　次/　2010年10月第1版　2010年10月第1次印刷

定　　价/　75.00元（册）

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

购书热线：(010) 68326294; 88379649; 68995259

投稿热线：(010) 88379604

读者信箱：hzjsj@hzbook.com

O'Reilly Media, Inc.介绍

为了满足读者对网络和软件技术知识的迫切需求，世界著名计算机图书出版机构O'Reilly Media, Inc.授权机械工业出版社，翻译出版一批该公司久负盛名的英文经典技术专著。

O'Reilly Media, Inc.是世界上在 Unix、X、Internet 和其他开放系统图书领域具有领导地位的出版公司，同时也是联机出版的先锋。

从最畅销的*The Whole Internet User's Guide & Catalog*（被纽约公共图书馆评为20世纪最重要的50本书之一）到GNN（最早的Internet门户和商业网站），再到WebSite（第一个桌面PC的Web服务器软件），O'Reilly Media, Inc.一直处于Internet发展的最前沿。

许多书店的反馈表明，O'Reilly Media, Inc.是最稳定的计算机图书出版商 —— 每一本书都一版再版。与大多数计算机图书出版商相比，O'Reilly Media, Inc.具有深厚的计算机专业背景，这使得O'Reilly Media, Inc.形成了一个非常不同于其他出版商的出版方针。O'Reilly Media, Inc.所有的编辑人员以前都是程序员，或者是顶尖级的技术专家。O'Reilly Media, Inc.还有许多固定的作者群体 —— 他们本身是相关领域的技术专家、咨询专家，而现在编写著作，O'Reilly Media, Inc.依靠他们及时地推出图书。因为O'Reilly Media, Inc.紧密地与计算机业界联系着，所以O'Reilly Media, Inc.知道市场上真正需要什么图书。

译者序

我一直对数据挖掘很感兴趣，尤其是通过对海量、抽象甚至枯燥的数据进行挖掘分析后，利用数据可视化工具展现出来的那种绚丽多彩、富含意蕴的数据之美更是令我痴迷、叹为观止。本书涉及领域很广，各领域的精英们向我们娓娓道来相关领域的数据信息系统的架构的设计，包括Yahoo！的云存储架构、Deep Web数据抓取、Facebook的信息平台、自然语言处理、“凤凰号”火星探测器的图像数据处理、探索数据生命的DNA漫谈，甚至是Radiohead视频的制作、旧金山的次贷危机等。

阅读完本书之后，我自己的一个很大的收获是对于自己比较了解的领域，如云存储、Deep Web、NLP等有了进一步的理解和实践指导，而对于那些完全不熟悉的领域，如探索数据生命、火星探测器、制作Radiohead视频等则更是开阔了视野，不但对数据有了新的认识，而且激发了思考问题的一些新的思维方式。

这本书令我很感怀的另一方面是，我发现这些“数据科学家”在兢兢业业构建平台处理数据的过程中，虽然遇到了很多困难和挑战，但是却依然如此坚持、执着地探索数据之美。在翻译本书过程中，这种激情不仅激励着我完成这本书的翻译，同时也激励着我在生活、工作中要有毅力和恒心。而纵观我身边的阿里巴巴云计算的同事们 —— 这些“阿里数据科学家”们，也无一不是那种永远充满着激情致力于我们的“飞天”梦想！

这是我翻译的第一本书，很感激机械工业出版社华章公司编辑陈冀康先生慷慨地引我入门，并且对因为我前段时期项目开发非常紧张而导致翻译进度几乎停滞的宽容和理解表示深深感激。感谢所有其他为本书付出努力的人们。

由于时间和精力有限，本书的疏漏、错误之处在所难免，还望各位读者不吝批评指正。

李妹芳

2010年6月26日

目录

前言

当我们第一次接触为《代码之美》编写“续集”的想法时，这次是关于数据也就是这本书，我们觉得这个想法令人兴奋且很有挑战性。现在收集、可视化和处理数据涉及每个专业领域和日常生活的诸多方面，一个大数据集在范围上将是难以想象的广泛。因此，我们联系了一组相当多样化的群体，这些人的工作让我们钦佩。当他们中的大多数都同意撰稿时，我们感到异常兴奋。

这本书就是我们努力的结果，我们希望它能够展示数据处理工作可以多么的广泛（和美丽）。在本书中，你将了解从和政府协作到和火星登陆器一起工作的各个方面；你将了解如何使用统计程序、制作可视化应用、混合Radiohead视频；你将看到地图、DNA和一些我们真正只能称之为“数据哲学”的内容。

本书的版权收益贡献给知识共享组织（Creative Commons）和阳光基金会（the Sunlight Foundation），它们致力于通过解放数据使世界变得更美好。我们希望你将会考虑你和数据亲身“邂逅”的经历如何塑造了世界。

本书的组织方式

本书的章节贯彻一条较为松散的曲线：从数据收集到数据存储、组织、检索、可视化及最后的数据分析。

第1章：在数据中观察生活。作者Nathan Yau着眼于在新兴的个人数据收集领域的两个项目背后的动机和挑战。

第2章：美丽的人们：设计数据收集方法时牢记用户。Jonathan Follett和Matthew Holm讨论了在Web上向人们收集数据时，信任、说服和测试的重要性。

第3章：火星上的嵌入式图像数据处理。J. M. Hughes分析了设计在太空旅行下能够正常工作的数据处理系统所面临的挑战。

第4章：PNUTShell中的云存储设计。Brian F. Cooper、Raghu Ramakrishnan和Utkarsh Srivastava描述了雅虎所设计的软件系统，该系统将其全球分布式数据中心转换为支持现代Web应用的通用存储平台。

第5章：信息平台和数据科学家的兴起。Jeff Hammerbacher以Facebook的数据团队的历史演化作为特例，追溯了信息处理工具以及驱动这些工具的人们的演化。

第6章：照片档案的地理之美。Jason Dykes和Jo Wood吸引人们注意一个志愿者组织收集的彩色可视化空间数据的普及性及其力量。

第7章：数据发现数据。Jeff Jonas和Lisa Sokol阐述了思考数据的新方式，为了完全管理这些数据，很多人需要采用这种方式。

第8章：实时的可移动数据。Jud Valeski深入分析了Web上实时的分布式社会和定位数据当前存在的局限，讨论了解决该问题的一个可能方案。

第9章：探寻Deep Web。Alon Halevy和Jayant Madhavan描述了G公司开发的用于搜索当前“受困”于Web表单之后的数据的工具。

第10章：构建Radiohead的“House of Cards”。Aaron Koblin和Valdean Klump讲述了一个涉及激光、编程和“骑在巴士背上”的惊险故事，故事以一个获奖音乐视频结束。

第11章：都市数据可视化。Michal Migurski详细描述了释放和美化一些我们身边的最重要的数据的过程。

第12章：Sense.us的设计。Jeffrey Heer重塑了作为社会空间的数据可视化，并使用这种新视角来探索历时150年的美国人口普查数据。

第13章：数据所做不到的。Coco Krumme关注于证明人们在很多方面误解和误用数据的实验性工作。

第14章：自然语言语料库数据。Peter Norvig通过从Web上获取的1兆规模的自然语言词汇语料数据，带领读者走进一些令人回味的实践。

第15章：数据中的生命：DNA漫谈。Matt Wood和Ben Blackburne描述了数据之美，即DNA和创造、捕捉和处理数据需要的大量基础设施。

第16章：美化真实世界中的数据。Jean-Claude Bradley、Rajarshi Guha、Andrew Lang、Pierre Lindenbaum、Cameron Neylon、Antony Williams和Egon Willighagen展示了“众包”（crowdsourcing）和高度透明的结合如何提高了药物发现的研究。

第17章：数据浅析：探索形形色色的社会定型。Brendan O'Connor和Lukas Biewald展示了当让人们匿名对其他人的图片进行打分时所表现出来的关联和模式。

第18章：旧金山海湾之殇：次贷危机的影响。Hadley Wickham、Deborah F. Swayne和David Poole通过使用开源软件和公共数据资源，带领读者走进对近年来旧金山海湾地区的住房危机的详尽研究。

第19章：美丽的政治数据。Andrew Gelman、Jonathan P. Kastellec和Yair Ghitza展示了统计和数据可视化工具是如何帮助我们加深对社会进行组织的政治进程的理解。

第20章：连接数据。Toby Segaran探索了对Web上可获取的大量的数据集进行连接的挑战性和可能性。

本书使用的体例

本书遵循以下字体体例：

斜体（*Italic*）

表示新的术语、URL、Email地址、文件名和文件扩展名。

等宽字体（`Constant width`）

用于程序清单以及段落中的程序单元如变量或函数名称、数据库、数据类型、环境变量、声明和关键字。

等宽粗体字（**`Constant width bold`**）

显示命令或者其他由用户输入的文本。

等宽斜体字（*`Constant width italic`*）

表示必须根据用户提供的值或者由上下文决定的值进行替代的文本。

使用本书的样例代码

本书是为了帮助你完成工作。通常来说，你可以在你的程序和文档中使用本书的代码。除非你使用了本书的大量代码，否则你无需联系我们获取许可。例如，写一个程序用到本书的几段代码不需要获得许可，销售和分发O'Reilly丛书的代码需要获得许可；引用本书的样例代码来解决一个问题不需要获得许可，使用本书的大量代码到你的产品文档中需要获得许可。

我们不要求你（引用本书时）给出出处，但是如果你这么做，我们对此表示感谢。出处通常包含标题、作者、出版社和ISBN。例如：“*Beautiful Data*, edited by Toby Segaran and Jeff Hammerbacher. Copyright 2009 O'Reilly Media, Inc., 978-0-596-15711-1.”。

如果你觉得你对本书样例代码的使用超出了这里给出的许可范围，请与我们联系：*permissions@oreilly.com*。

联系方式

如果您对本书有任何意见和问题，请联系出版社：

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街2号成铭大厦C座807室（100035）
奥莱利技术咨询（北京）有限公司

O'Reilly的每一本书都有专属网站，你可以在那找到关于本书的相关信息，包括勘误列表、示例代码以及其他的信息。本书的网站地址是：

http://www.oreilly.com/catalog/9780596157111/

对于本书的评论和技术性的问题，请发送电子邮件到：

bookquestions@oreilly.com

关于本书的更多信息、会议、资料中心和网站，请访问以下网站：

http://www.oreilly.com
http://www.oreilly.com.cn

第1章

在数据中观察生活

Nathan Yau

在不远的过去，Web就是关于共享、广播和分发。但是潮流在变：Web已经走进了个人的世界。每个月，新的应用如雨后春笋般诞生，这些应用可以追踪、监测和分析人们的行为习惯，使他们更好地了解自己和周围的世界。人们可以对饮食习惯、运动、上网时间、性生活、每月的自行车旅行、睡觉、情绪和资产进行在线跟踪。如果你对自己生活中的某一部分感兴趣，很可能存在某种应用可以对它进行跟踪。

当然，个人数据收集不是什么新鲜事。在20世纪30年代，英国的社会研究小组Mass Observation，收集了关于日常生活的各个方面的数据 —— 如胡须、眉毛、驾驶员的叫喊和手势，以及人们在战争纪念碑前的行为，这些都是为了对英国有更好的理解。然而，数据收集方法从1930年开始已经有了提高；现在不再仅仅是一支笔和便条纸，或者是一个手工的计数器。数据可以通过手机和手持电脑自动收集，因此每天都可以持续地上传数据和信息流到服务器、数据库和所谓的数据仓库。

随着数据收集技术的进步，数据流也发展为比Mass Observation[译注1]参与者报告的计算计数要强大得多的技术。数据可以实时修改，因此，人们期望获取最新的信息。

但是，只是简单地为人们提供数以GB计的数据是不够的。并非所有人都是统计学家或者

译注1：Mass Observation是一个英国社会研究组织。它通过500名未经过训练的志愿者的观察来记录英国人的日常生活。请参考*http://en.wikipedia.org/wiki/Mass_Observation*。