<<鲜活的数据>>

图书基本信息

书名：<<鲜活的数据>>

13位ISBN编号：9787115293817

10位ISBN编号：7115293813

出版时间：2012-10

出版时间：人民邮电出版社

作者：[美] Nathan Yau

页数：281

字数：437000

译者：向怡宁

前言

引　　言数据不是什么新鲜玩意。
早在几个世纪之前，人们就开始对数据进行量化分析并为之绘制表格了。
然而在为FlowingData（我创建的一个有关设计、可视化和统计的网站）写作时，我发觉这一领域在过去数年间出现了爆炸式的发展，而且未来还会更加蓬勃。
科技的进步使得收集和存储数据变得轻而易举，而互联网则让我们摆脱了时间和空间的束缚。
如果运用得当，这种数据的“财富”能够提供丰富的信息，帮助人们更明智地制定决策、更清楚地传达理念，而且能让我们以更为客观的角度去审视自己对世界和自身的看法。
随着2009年年中Data.gov网站的上线，美国政府数据公开化进程发生了一次重大转变。
这是一套综合的数据目录系统，由各级联邦政府机构提供，表现出各组织及官方的透明度和责任感。
比如说，国民有权利了解政府把税收收入都花在了哪里，而在此之前美国政府给人的感觉就像一个黑箱。
Data.gov上的很多数据其实在许多网站中都能找到，但现在它们都被会聚在一起，而且有着统一的格式，更加便于人们进行分析和可视化。
除了Data.gov之外，联合国也有类似的网站UNdata，英国很快也发布了Data.gov.uk，而像纽约、旧金山和伦敦等全球许多城市也都参与到了数据公开这一潮流中来。
如今的网站也变得越来越开放，有数千个API（应用编程接口）在鼓励和“怂恿”着开发人员去调用网站已有的数据做各种事情。
比如Twitter和Flickr就提供了覆盖面极广的API，开发人员可以自由定制与网站本身完全不同、五花八门的用户界面。
API编目网站ProgrammableWeb目前已收录超过2000个API。
诸如Infochimps和Factual这样的应用最近也大量涌现出来，它们存在的目的就是向人们提供结构化的数据。
在个人层面，我们可以在Facebook上结交朋友，在Foursquare上分享所在的位置，或者在Twitter上发布自己的最新动态，这所有的一切都只需要点击几次鼠标或者敲击几下键盘就能实现。
一些针对性更强的应用则方便我们记录品尝过什么美食、体重几何、情绪高低等林林总总的事情。
几乎可以这样说，只要你想对自己的某个方面进行追踪，就会有这样一款应用来帮助你实现愿望。
数据就静静地待在我们生活的每一个角落。
园子里已经果实累累，正等待着我们去采摘。
对大多数人来说，真正有意思的并不是数据本身，而是数据背后蕴涵的信息。
人们都希望知道他们的数据有何意义，而如果你能帮助他们，那么你就会大受欢迎。
难怪Google首席经济学家Hal Varian会说统计学家将是未来十年内最迷人的职业，而这绝不仅仅是因为统计学家长得好看（尽管以极客们的别样眼光来看，我们确实长得让人赏心悦目）。
可视化要想探索和理解那些大型的数据集，可视化是最有效的途径之一。
把数字置于视觉空间中，我们的大脑或者读者的大脑就会更容易发现其中潜藏的模式。
人类对图形的理解能力确实很强，往往能够从中发现一些通过常规统计方法很难挖掘到的信息。
John Tukey是我最喜爱的统计学家，也是探索性数据分析理论（Exploratory Data Analysis）的缔造者。
他精通各种统计方法和工具，而且深信图形技术在其中亦占有一席之地。
他坚信，图形的呈现方式会让人们得到许多出乎意料的结果。
只需对数据进行可视化，我们就能从中发现大量信息，而且很多情况下这也是我们制定明智决策或描述复杂事件所需要的唯一手段。
比如说，在2009年美国的失业率遭遇了一次大幅增长。
2007年的全美平均失业率是4.6%，2008年上涨到了5.8%。
而到了2009年9月，突然就攀升至9.8%。
但是这些全国平均数字只揭示了事件的一部分，它们只是概括了整个国家的总体状况。
有哪些地区的失业率高于其他地区？

又有哪些地区并未受到很大波及？
我们无法从中获得答案。
图0-1用一系列美国地图更为完整地说明了情况，而且我们只需略扫一眼就能回答上面的问题。
颜色较深的县失业率相对较高，而颜色较浅的县失业率较低。
在2009年的地图上（图0-2），我们可以看到美国西部和东部大多数地区的失业率都超过了10%，而中西部地区则未受到太大影响。
图0-1　2004—2009年美国失业率分布图如果手上只有单纯的电子表格，要想找到其中蕴涵的地区性或周期性的模式就会很花时间，而只靠前面那些全国平均数字则完全不可能。
而用地图呈现之后，虽然增加了许多县的数据，但读者的理解程度反而提高了。
这些地图有可能帮助当局决定往哪些地区划拨救济金或提供其他形式的援助。
图0-2　2009年失业率分布图这个例子的绝妙之处在于，用于产生地图的数据都是免费的，由美国劳工统计局直接面向公众开放。
尽管找到这些数据并不是那么轻而易举，但它们确实就在某个地方听候我们的差遣，而且还有更多格式化的数据正等待着我们作更好的视觉处理。
比如说，《美国统计摘要》（The Statistical Abstract of the United States）就含有数百个数据表格（见图0-3），但没有任何图表。
这简直是天赐的良机，我们可以在此基础上进行加工，展现整个国家的概貌。
这个过程将会非常有趣。
不久前我用图形描绘了其中的部分表格（见图0-4），很快就得到了美国近年来结婚率及离婚率、邮政资费、用电量等信息的直观变化情况。
单纯的表格形式很难阅读，读者只能得到一些零散的数值，而在图表化视图中，人们能够轻易地发现变化的趋势和模式，而且一眼就能作出比较。
图0-3　美国统计摘要网站中的表格图0-4　美国统计摘要网站数据的图表化视图类似《纽约时报》、《华盛顿邮报》这样的新闻机构很擅长让数据变得栩栩如生、易于理解。
它们对已有数据的利用也许是最充分的，因为经常会有相关主题的新闻故事见诸报端。
有时故事中还会插入数据图表以强调不同的观点，而有时只需要图表就能讲述整个故事。
在传统媒体向网络媒体转型的过程中，图形的应用变得更加普及。
如今的新闻机构中都已设立了专门处理交互、图表或地图数据的各种部门，比如《纽约时报》就专门为“计算机辅助报道”成立了一个新闻编辑部，旗下的记者都专注于用数据来报道新闻。
而《纽约时报》的图形编辑部处理起大量数据来也同样得心应手。
即使是在流行文化领域，可视化也占据了自己的一席之地。
Stamen Design是一家以在线交互闻名的可视化公司，他们在过去数年中一直都在对每年的MTV音乐录影带大奖颁奖时期的Twitter状态进行追踪。
Stamen Design每一次的设计都与之前有所不同，但其核心一直保持不变：实时展现人们在Twitter上的热门话题。
2009年Kanye West在Taylor Swift发表获奖感言时突然暴走，我们通过Stamen Design的追踪可以很容易地了解人们对他这种行径的看法。
现在看来，我们发现这个领域中也有偏重情绪而非分析的一面，对可视化的定义开始变得模糊起来。
在很长一段时间内人们都认为，可视化就是关于量化后的事实：我们把它们作为工具来识别事物发展的模式，转而为分析研究提供帮助。
但可视化并不仅仅与冰冷的事实有关。
就如同Stamen Design的追踪设计一样，它有着很强的娱乐因素，为观众提供了另一种方式去关注颁奖典礼，并在过程中与其他粉丝进行互动。
Jonathan Harris的设计也是一个很好的例子。
在他的We Feel Fine（我们感觉良好）和Whale Hunt（捕鲸）等作品中，Harris并不是出于分析角度，而是围绕着故事本身来进行设计，而且这些故事以人类情感为中心，超越了单纯的数字和分析行为。
图表和图形逐渐也超出了工具的范畴，发展为传达理念的载体。

GraphJam和Indexed之类的网站 就喜欢运用文氏图 、饼图等形式来戏谑流行歌曲及文化，用红白黑等颜色组合来讥讽政客，或者谴责虐待动物的行为。
我自己也在这个方向上作了一些尝试，在FlowingData上发表了系列漫画Data Underload（数据低负荷）。
在图0-5中，我用图形表现了美国电影协会评选出的一些经典电影台词——非常无厘头，但很有趣（至少对我来说如此）。
图0-5　图表形式的电影台词那么，到底什么是可视化呢？
每个人都有自己的答案。
有些人认为只有严格意义上的传统图形图表才是可视化。
而另一些人的观点则更加开放，他们认为只要是在表现数据，不管是数据艺术品还是微软Excel表格，都可以算是可视化。
我个人较为倾向于后者，但有时也发现自己站在前一阵营。
毕竟，这一问题上孰是孰非并不是那么重要，只要能达成我们的目的就行了。
不管可视化是什么，我们绘制演示用的图例也好、进行数据分析也好、用数据来报道新闻也罢，最终其实都是在寻求真相。
在某些时候，统计也会产生错误的假象，但造成错觉的并不是数字本身，而是运用数字的人。
有时候这是有意为之，但更多情况下是疏忽大意所致。
如果我们不知道如何创建合适的图形，或者不知道如何客观地看待数据，那么就会产生谬误。
但只要我们掌握了适当的可视化技巧和处理方式，就能更加自信地陈述观点，并且对自己的发现感觉良好。
学习数据我在大学一年级时开始接触统计学，当时它是一门必修的基础课，但与我的专业电气工程并没有太大关系。
讲课的教授热情极高，而且对这一领域乐此不疲。
他上课时喜欢在教室的台阶上来回走动，身体语言极为丰富，而且不时鼓励身边的学生参与讨论。
我从未遇到过如此兴奋的老师，而且毫无疑问，正是这种精神吸引我进入了数据领域，最终在四年后考上了统计学的研究生。
在本科四年中，统计学就是数据分析、频率分布和假设检验，而我一直乐在其中。
我觉得观察数据集，探索其中的趋势、模式和关联性很有意思。
但开始研究生学业之后，我的观点发生了改变，事情变得更加有趣了。
统计学不再是假设检验（结果表明，在许多情况下它并无太大作用）以及寻找模式了。
哦，不，我收回这句话。
统计学仍然与这些有关，但我对它产生了不一样的感受。
统计学其实是在用数据讲故事。
我们手头的大堆数据反映了真实的世界，然后我们对它们进行分析，得到的不只是数据的关联性，我们还能了解到身边正在发生什么。
这些故事反过来可以帮助我们解决真实世界中存在的问题，例如降低犯罪率、提高卫生意识、改善高速上的交通状况，或者只是增长我们的见识。
很多人都未能找到数据与真实生活之间的联系。
我想这也是为什么当我告诉人们我读研是为了学统计学时，大多数人都说那是他们“上学时最痛恨的一门课”。
我相信读者们不会犯同样的错误，否则你就不会选择读这本书了，不是吗？
运用数据需要一些技能，如何才能掌握呢？
你可以像我一样去学校选择正规的课程训练，但你也可以通过大量的实践经验，自学成才。
其实大多数研究生课程和自学也没有多大区别。
在可视化和信息图（infographics）方面也是如此。
并不是只有专业图形设计师才能创建优秀的图表，同样，你也不需要拿到统计学的博士学位。
你所需要的只是保持对学习的渴望，而且和生活中的所有事情一样，你需要不断练习才能变得更在行

。
我制作的第一张数据图大概是在小学四年级，那是为了应付一次课外科学研究。
我和搭档一直很想知道蜗牛在什么样的平面上会爬得更快，于是把它们放在各种粗糙或光滑的物体表面上，并计时观察它们爬过一段特定距离各需要多久。
最后我拿到了蜗牛在不同表面上爬行的时间数据，并据此制作了一张柱形图。
至于当时是否知道应该将它们按长短进行排序，我已经记不太清了，但是和Excel软件的辛苦纠缠倒是一直刻骨铭心。
不过第二年当我们研究赤拟谷盗 最喜欢吃哪种谷制品时，作图就是小菜一碟了。
当你理顺某款软件的基本功能和操作方式之后，剩下的几乎都轻而易举。
这个例子完美地说明了什么叫做从经验中学习。
噢，顺便提一句，如果你还在琢磨前面的问题，答案是蜗牛在玻璃上爬得最快，而赤拟谷盗最喜欢吃葡萄果仁麦片（Grape Nut）。
从本质上来说，学习任何软件或编程语言的过程几乎都是一样的。
如果你一行代码都没写过，那么R（许多统计学家都采用的一种计算环境）必然会让你望而生畏，而一旦你跟着完成了几个范例之后，就会很快找到窍门。
这本书能够帮助你做到这些。
之所以这样说，是因为我本人就是这样学习的。
我还记得自己第一次深入接触可视化的设计层面时的情形。
那还是我读研究生的第二年，好消息从天而降，我得知自己获得了《纽约时报》图形编辑的实习机会。
在那一刻之前，图表对我而言只是一种分析工具而已（比如小学课外活动时作的柱形图），就算其中含有一些美学和设计因素，比重也少得可怜。
而将数据用于新闻报道，这对我来说更是无从入手。
所以为了作准备，我阅读了手边能找到的所有设计书籍，以及一本Adobe Illustrator的使用指南，因为我知道《纽约时报》图形编辑部用的就是这款软件。
不过还没等我真正上手就已经开始绘制工作了。
当你被迫边学边干的时候，就不得不尽快掌握那些必需的知识，而当你开始处理更多数据、设计更多图表时，你的技能也会随之突飞猛进。
如何阅读本书本书以实例讲解为主，目的是让大家熟悉制图所需的每一个步骤，掌握每一项技能。
你可以从头开始完整地读一遍，不过如果你已经有想法在酝酿了，也可以只挑选最感兴趣的几章来读。
所有的章节都经过了精心的组织，案例是相互独立的。
如果读者对数据领域还比较陌生，那么阅读最开始的几章应该会很有帮助。
它们介绍了处理数据的方法、需要关注的重点以及各种可用的工具，便于读者了解如何获得数据，如何规范格式并为可视化作准备。
之后的几章会根据不同的数据类型和侧重面分别介绍各种可视化技巧。
请记住，永远都要让数据说话。
不管你选择何种阅读方式，我都强烈建议你在阅读时打开电脑，和我一起逐步完成每一个范例，并且浏览在注释和参考中提到的各种资源。
你也可以在网站上下载到所有的代码、数据文件和可交互演示。
为了表述得更清楚一些，图0-6给出了一张流程图，便于读者找到需要的章节。
祝大家阅读开心！

<<鲜活的数据>>

内容概要

在生活中，数据几乎无处不在，任我们取用。
然而，同样的数据给人的感觉可能会千差万别：或冰冷枯燥，让人望而生畏、百思不解其意；或生动有趣，让人一目了然、豁然开朗。
为了达到后一种效果，我们需要采用一种特别的方式来展示数据，来解释、分析和应用它。
这就是数据可视化技术。

Nathan Yau是这一创新领域的先锋。
在《鲜活的数据：数据可视化指南》中，他根据数据可视化的工作流程，先后介绍了如何获取数据，将数据格式化，用可视化工具（如R）生成图表，以及在图形编辑软件（如Illustrator）中修改以使图表达到最佳效果。
本书介绍了数十种方法（如柱形图、饼图、折线图和散点图等），以创造性的视觉方式生动讲述了有关数据的故事。
翻开本书，思维之门会豁然大开，你会发现有那么多样的手段去赋予数据全新的意义！

《鲜活的数据：数据可视化指南》主要内容包括：
学习如何用视觉化表示方式来呈现数据，让读者看到不一样的信息；
发现数据背后的故事；
探索不同的数据来源，确定有效的展示格式；
试验并对比不同的可视化工具；
寻找数据中的趋势和模式，并以适当的图表来展现它们；
设定明确的目标，并用其指引你的可视化过程。

<<鲜活的数据>>

作者简介

Nathan Yau，加州大学洛杉矶分校统计学专业在读博士、超级数据迷，专注于数据可视化与个人数据收集。他曾在《纽约时报》、CNN、Mozilla和SyFy工作过，认为数据和信息图不仅适用于分析，用来讲述与数据有关的故事也非常合适。
Yau的目标是让非专业人士读懂并用好数据。
你可以从中欣赏到他最新的数据可视化实验作品。

向怡宁，交互和视觉设计师、摇滚乐手，同时还热衷于翻译和写作。
著有《Flash组件、游戏、SWF加解密》及《就这么简单：Web开发中的可用性和用户体验》，译有《奇思妙想：15位计算机天才及其重大发现》、《瞬间之美：Web界面设计如何让用户心动》、《网站设计解构：有效的交互设计框架和模式》、《网站搜索设计：兼顾SEO及可用性的网站设计心得》等书。
他认为“一个不会弹吉他的设计师不是个好译者”。

<<鲜活的数据>>

书籍目录

<<鲜活的数据>>

章节摘录

版权页： 插图： 1.带分隔符的文本 很多人都很熟悉带分隔符的文本。
我们在前面一节的例子中就创建过以逗号分隔的文本文件。
如果把数据集看成是按行和列来分布，那么分隔符文本就是用分隔符来分开每一列。
分隔符一般用的是英文逗号（半角字符），也可以是制表符tab，或者是空格、英文分号、冒号、斜杠等任何你喜欢的字符。
不过逗号和tab是最常见的。
分隔符文本应用广泛，可以被大多数电子表格程序阅读，例如Excel或者Google Documents。
我们也可以把电子表格输出成分隔符文本。
如果你要使用多个工作表格，通常就会有多个分隔符文件，除非特殊指定。
这种格式也便于与其他人共享，因为它无需依赖于任何特定程序。
2.JavaScript对象表示法（JSON） 很多网页API都适用于这种格式。
它既能够让计算机理解，又便于人类阅读。
不过如果你眼前的数据过多，盯太久可能会头晕目眩。
该格式基于JavaScript表示法，但并不依赖于这种语言。
JSON中有许多规格说明，但只用掌握一些基础就能满足大部分需要。
JSON利用关键字和值，并且把数据条目作为对象来处理。
如果我们把JSON数据转化成逗号分隔数据（Comma-Separated Value,CSV），那么每个对象都会占一行。
大家将会在后文中看到，有很多应用、语言和函数库都支持JSON输入。
如果你打算设计便于互联网传播的数据图形，就得了解一下这种格式。
访问http:／／json.org阅读JSON的完整说明。
你不必了解这一格式的所有细节，但当你需要使用某个JSON数据源时，它还是很管用的。
3.XML XML（可扩展标记语言）是另一种互联网上的流行格式，常被用于在API间传递数据。
XML分为很多类型，规格说明也不少，但从最基本的层面来看，它就是一个文本文件，其中的值都封闭在各种标签之内。

媒体关注与评论

本书就像是一封写给Python、R、地图和数据的情书。
——FlowingData读者评论我是Nathan Yau的博客FlowingDate的忠实粉丝，本书还没出来我就预定了。果然，它完全符合我的预期：各种各样的分析、数据资源和绝对精美的图表。
——亚马逊读者评论本书写的很好，思路清晰，实例丰富，如果你经常与数据打交道，选择本书错不了。
——亚马逊读者评论

<<鲜活的数据>>

编辑推荐

数据可视化经典著作 讲解清晰、示例丰富、实用性强 创作信息图的最佳参考指南

名人推荐

“本书就像是一封写给Python、R、地图和数据的情书。
”--FlowingData读者评论“我是Nathan Yan的博客FlowingData的忠实粉丝，本书还没出来我就预订了。
果然，它完全符合我的预期：各种各样的分析、数据资源和绝对精美的图表。
”--亚马逊读者评论“本书写得很好，思路清晰，实例丰富，如果你经常与数据打交道，选择本书错不了。
”--亚马逊读者评论

<<鲜活的数据>>

版权说明