阅 想 时 代
Mind Times Press

DATA POINTS

Visualization That Means Something

数据之美

一本书学会可视化设计

[美] 邱南森 (Nathan Yau) ◎著　　张伸◎译

关于数据呈现的思考和方式的颠覆之作

《经济学人》年度推荐三大可视化图书之一

《大数据》作者、《经济学人》大数据主编**肯尼思·库克耶**倾情推荐

亚马逊数据和信息可视化 **TOP3 图书**

中国人民大学出版社

DATA POINTS

Visualization That Means Something

数据之美

一本书学会可视化设计

[美] 邱南森 (Nathan Yau) ◎著　　张伸◎译

中国人民大学出版社
• 北京 •

图书在版编目（CIP）数据

数据之美：一本书学会可视化设计/（美）邱南森著；张伸译.—北京：中国人民大学出版社，2013.12

ISBN 978-7-300-18612-2

Ⅰ.①数… Ⅱ.①邱… ②张… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2013）第313017号

数据之美：一本书学会可视化设计

[美] 邱南森 著

张 伸 译

Shuju zhi Mei: Yi Ben Shu Xuehui Keshihua Sheji

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010-62511242（总编室）	010-62511398（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http:// www. crup. com. cn		
	http:// www. ttrnet. com.（人大教研网）		
经　　销	新华书店		
印　　刷	北京中印联印务有限公司		
规　　格	185 mm × 230 mm　16开本	**版　　次**	2014 年2月第1版
印　　张	17.75　插页1	**印　　次**	2014 年2月第1次印刷
字　　数	308 000	**定　　价**	89.00元

Data Points：Viaualization That Means Something by Nathan Yau
ISBN: 978-1-118-46219-5

让知识沉淀于心……

The Way to Read Mind

目　录

引言　可视化是一种媒介

第1章　你真的理解数据了吗

数据表达了什么 /6

数据的可变性 /20

数据的不确定性 /29

数据所依存的背景信息 /36

第2章　数据引导可视化设计

新数据研究需要新工具 /45

信息图形和展示 /57

可视化的娱乐性 /66

走进数据艺术的世界 /72

日常生活中的可视化 /80

第3章　掌握可视化设计的原材料

各种可视化组件 /89

整合可视化组件 /108

第4章　不了解数据，一切皆是空谈

数据可视化的过程 /128

分类数据的可视化 /134

时序数据的可视化 /144

空间数据的可视化 /156

多元变量 /166

数据的分布 /179

第5章　让可视化设计更为清晰

建立视觉层次 /192

增强图表的可读性 /195

高亮显示重点内容 /211

注解可视化表达了什么 /217

从不同角度做一些计算 /223

第6章　别忘了，你是为读者进行可视化设计

可视化时常见的错误 /230

读者不同，数据展示方式不同 /242

需要注意的事项 /245

可视化步骤的整合 /255

第7章　将可视化进行到底

可视化工具 /264

编程工具 /268

插图工具 /274

数据统计 /274

结语　可视化设计，若烹小鲜

译者后记

引　言
DATA POINTS

可视化是一种媒介

什么是好的可视化设计？如果只看光秃秃的原始数据，你可能会忽视掉某些东西。好的可视化是一种表达数据的方式，能帮助你发现那些盲点。你可以通过可视化展示的趋势、模式和离群值来了解自己以及身处的世界。最好的可视化设计能让你有一见钟情的感觉，你知道眼前的东西就是你想看到的。有时候，可视化设计仅仅只是一个条形图，但大多数时候可视化会复杂得多，因为数据本来就很复杂。

可视化让数据更可信

数据集犹如即时快照，能帮助我们捕捉不断变化的事物。数据点聚集在一起就形成了数据集合以及统计汇总，可以告诉你预期的收获。这就是平均数、中位数和标准差，它们用来描述世界各地以及人口的状况，并用来比较不同的事物。你可以去了解每个数据的具体细节。这就是所谓的数据集人性化，它会使数据更加可信。

从抽象意义上说，包含信息和事实的数据是所有可视化的基础。对原始数据了解得越多，打造的基础就越坚实，也就越可能制作出令人信服的数据图表。人们往往会忽略一点：好的可视化设计是一个曲折的过程，需要具备统计学和设计方面的知识。没有前者，可视化只是插图和美术练习；而没有后者，可视化就只是分析结果。统计学和设计方面的知识都只能帮助你完成数据图形的一部分。只有同时具备了这两种技能，你才可以随心所欲地在数据研究和讲故事两者间自如转换。

这本书是为那些对设计和数据分析过程感兴趣的人而写的。我们在每一章都介绍了通

往可视化的一个步骤。在这里，可视化不只是剪贴画上大大的数字，而是向我们传递了数据的意义。可视化创作是一个迭代的过程，不同的数据集迭代周期不同。

本书第一部分主要帮助读者了解自己的数据，以及把数据可视化的意义。由于数据代表了一定的人物、地点和事物，所以除了真实的数字之外，还有重要的背景信息。数据是关于谁的？它从哪里来以及是什么时候收集的？虽然是计算机生成并输出数据，但我们也需要对这些由人处理的部分负责。除此之外，大部分数据集都是估算的，并不是绝对真实的，犹如人生一样，充满不确定性和可变性。

本书的中间部分，我们会带你进入探索模式。通过数据挖掘，你可以自由地提出问题并解答这些问题。你还可以寻找数据中的模式、关联以及所有看起来不大对劲的东西。由于拼写错误，经常会出现缺失值。你可以借此机会进行大量的实验，从不同角度观察数据。你可能会有一些意外的发现，也许最终这就是数据所能呈现的最有趣的东西。由于种种原因，人们往往会跳过探索阶段，这导致最终的成果往往让人难以理解。花一些时间去了解数据以及它们所代表的东西，能加倍提升可视化的效果。

当你找到了潜在的故事，接下来就要将其传达给更多的客户。这是本书的最后一部分，要用设计来美化一下。为 4 个熟悉该话题并且阅读过所有相关重要论文的人所做的图表，和为不熟悉这一话题的普通读者所做的图是不一样的。

这些步骤并非要按部就班地进行。如果你已经在和数据打交道了，那就会知道在研究已有数据时经常会发现需要新的数据。同样地，设计过程会迫使你看到之前没有注意到的细节，让你不得不重新回到探索阶段或者回到起点。如果你是新手，在阅读本书时就会了解到这个过程，并且你会自信能把从本书中学到的知识用到自己的项目中。在数据和故事间来回往返是很有趣的。

《数据之美》是对我的上一本书《鲜活的数据》的完美补充。《鲜活的数据》介绍了可视化设计可以使用的工具，提供了具体的编程示例；《数据之美》则描述了整个可视化的过程和思想，涉及更大的数据项目并且不涉及任何软件。换句话说，这两本书互为补充。《鲜活的数据》为准备制作图表的人提供了技术指南，而《数据之美》则描述了数据及其可视化的过程，以便帮助你创造出更好的、更有意义的东西。

可视化不只是一种工具

在本书中，我们将可视化看作是一种媒介，而非一种特定的工具。如果把可视化当成死板的工具，你很容易以为几乎所有的图形都比条形图好。对于大部分图表而言，确实如此，但前提是必须是在适合的条件下。譬如，在分析模式中，你通常会期望图表便于快速阅读且十分精确。但如果目标是激发感情和好奇心呢？可视化是一种表达数据的方式，是对现实世界的抽象表达。它像文字一样，为我们讲述各种各样的故事。报纸文章和小说不能用同一个标准来评判，同样，数据艺术也不能用商业图表的标准来衡量。

无论哪一种可视化类型都有其规则可循。这些规则并不取决于设计或统计数字，而受人类感知的支配。它们确保读者能准确解读编码数据。这样的规则很少，例如，当用面积作为视觉暗示时，要将面积按大小恰当地排序，其余的都只是建议。

你需要区分规则和建议。规则是应该时时遵循的，而建议则要具体分析，视情况而定是否采纳。很多初学者会犯这样的错误，遵循了具体的建议，结果丢失了数据的背景信息。例如，爱德华·塔夫特（Edward Tufte）建议剔除图表中所有的垃圾信息，但所谓的垃圾是相对而言的。一个图表中需要剔除的东西，在另一个图表中也许是有用的。正如塔夫特所说："大多数设计原则都应受到质疑。"

同样，统计学家威廉·克利夫兰（William Cleveland）和罗伯特·麦吉尔（Robert McGill）关于感知和精确度的研究成果也经常被人们引用。他们发现，在如散点图这样的常见图表中，位置信息是能被最精确解码的，接下来依次是长度、角度和斜率。这些结果是基于研究试验得出的，也有其他的研究支持，因此人们很容易把克利夫兰和麦吉尔的发现误当作规则。然而，克利夫兰也指出，好的图表不只是要能快速理解，还包括它显示的内容如何，以及它是否帮助你看到了之前没有看到的东西?

是时候回到值得可视化的数据上了。幸运的是，你有大量的数据可用，而且数据源始终在增长。过去几年里的每一个星期里，都会有一篇文章讲述数据洪流以及淹没其中的危险。但你知道，数据量是可控的，你可以轻松地筛选和聚集数据。数据存储费用越来越便宜，而且可以无限存储，这就意味着会"游泳"的人能得到更多的快乐。他们面临的挑战就是学习如何潜得更深。

好吧，我说得太多了，让我们来开始一段快乐的旅程吧。

第1章

你真的理解数据了吗

数据是什么？大部分人会含糊地回答说，数据是一种类似电子表格的东西，或者一大堆数字。有点儿技术背景的人会提及数据库或数据仓库。然而，这些回答只说明了获取数据的格式和数据的存储方式，并未说明数据的本质是什么，以及特定的数据集代表着什么。你很容易陷入一种误区，因为当你需要数据的时候，通常会得到一个计算机文件，你很难把计算机输出的信息看作其他任何东西。然而，透过现象看本质，就能得到更多有意义的东西。

数据表达了什么

数据不仅仅是数字。要想把数据可视化，就必须知道它表达的是什么。数据描绘了现实的世界。与照片捕捉了瞬间的情景一样，数据是现实世界的一个快照。

请看图 1—1，它和其他事物没有任何关联，我也没告诉过你关于这张照片的故事，如果你无意中看到了这张照片，那么在你看来，它只不过是一张普普通通的婚礼照片，你从中再也得不到更多的信息了。然而对我来说，它记录了我生命中最美好的时刻。照片里左边是我的妻子，穿着美丽的婚纱；右边是我，穿着和平时的 T 恤牛仔裤风格完全不同的正装。主持婚礼的牧师是我妻子的叔叔，这为婚礼增添了个性化色彩。他后面的那位家族世交承担了全程录像的重任，尽管我们也花钱雇请了一位专业摄影师。婚礼上的鲜花和拱门由距此一小时车程的一家当地供应商提供。婚礼是初夏时在加利福尼亚州洛杉矶举行的。

图 1—1　一张照片，一个数据点

仅仅一张照片就包含了如此多的信息。同样地，数据也会传递给我们大量的信息。（对一些人来说，包括我在内，照片也是数据。）一个数据点可以包含时间、地点、人物、事件、起因等因素，因此很容易让一个数字不再只是沧海一粟。可是从一个数据点中提取信息并不像看一张照片那么简单。你可以猜到照片里发生的事情，但如果对数据心存侥幸，认为它非常精确，并和周围的事物紧密相关，就会曲解真实的数据。你需要观察数据产生的来龙去脉，并把数据集作为一个整体来理解。关注全貌，比只注意到局部时更容易作出准确的判断。

想象一下，如果我没有告诉你那张照片背后的故事，你怎样才能知道更多的信息？看了婚礼前后其他的照片又会怎样呢？（见图 1—2）

在图 1—2 中，你看到了更多的照片，这些照片组成了婚礼中的一个个环节，包括新娘第一次走出来，新人宣誓，以及向双方父母和我的奶奶敬茶等。同样，这里的每一张照片背后都有故事，例如，岳父把女儿交给我时热泪盈眶；我挽着新娘走过教堂走廊时感受到了巨大的快乐和幸福。这些照片捕捉到了婚礼中从我的视角无法看到的瞬间，因此看这些照片时，我甚至感觉自己也像你一样是一个局外人。我告诉你那天的故事越多，那天的情景就变得越清晰。

图 1—2　相格

尽管如此，这些毕竟只是一组快照，你不知道这些瞬间之外还发生了什么。（当然，你可以猜测。）想要完整地知道那天的事情，要么你得在现场，要么就只能去看视频了。即便如此，你也只能从有限的几个角度看到这场婚礼，因为通常不大可能拍摄到每一个细节。例如，当我们点蜡烛时，蜡烛却总是被风吹灭，这引起了大约 5 分钟的混乱。我们划完了所有的火柴后也没能点燃蜡烛，于是婚礼策划师到处寻找可以救急的东西。幸运的是，有一位吸烟的来宾贡献了打火机。照片却错过了这一幕，还是因为它们只是提取真实事物的一个个片段。

这是我们采样的方式，你不大可能记录下一切，因为成本太高或者缺少人力，或二者皆有。你只能获取零碎的信息，然后寻找其中的模式和关联，凭经验猜测数据所表达的含义。数据是对现实世界的简化和抽象表达。当你可视化数据的时候，其实是在将对现实世界的抽象表达可视化，或至少是将它的一些细微方面可视化。可视化是对数据的一种抽象表达，所以，最后你得到的是一个抽象的抽象，真是个有趣的挑战。

无论如何，这并不是说可视化模糊了你的视角。恰恰相反，可视化能帮助你从一个个独立的数据点中解脱出来，换一个不同的角度去探索它们。可以说是见树又见林。让我们继续说说婚礼照片这个例子。图 1—3 使用了完整的婚礼数据集，图 1—1 和图 1—2 里的照片只是它的子集。每一个矩形都代表我们婚礼相册中的一张照片。它们按时间顺序排列，每一张照片都用其中的主色调来填充。

按时间顺序排列照片，你可以发现婚礼的高潮处。婚礼的高潮处摄影师拍了很多照片。你也可以看到相对平静的时候，只有很少几张照片。图中的几个高峰，毫无疑问都发生在一些重要的时刻，例如，我第一次看到新娘身穿婚纱走出来，还有婚礼刚开始的时候。之后，我们与亲朋好友合影，因此图中出现了另一个高峰。接下来是宴饮时间，略显平静，尤其是摄影师在 4 点前短暂休息的时候。然后，婚礼又开始大张旗鼓地继续进行，直到晚上 7 点左右才结束。之后，只留下我和妻子在夕阳的余晖中相依相偎。

由于照片是线性呈现，在相格（grid layout）中就看不到上述的模式。所有的事情看上去都等距发生，而实际上大部分照片是在最激动人心的时刻拍摄的。扫一眼也能大致领略到婚礼中的颜色，黑色的是西装，白色的是婚纱，花童和伴娘身着珊瑚色礼服，整个婚礼现场和签到台则被绿树环绕着。你能从中得到和看真实照片一样多的细节吗？不能，但在一开始往往没必要了解那么多细节。有时候你需要先看看总体模式，然后再放大细节。有时候只有在了解了整体以及一个独立点与整体之间的关系后，才能知道它是否值得细看。

婚礼的颜色

每一个矩形都代表婚礼中的一张照片，每张照片都用其最丰富的颜色填充。

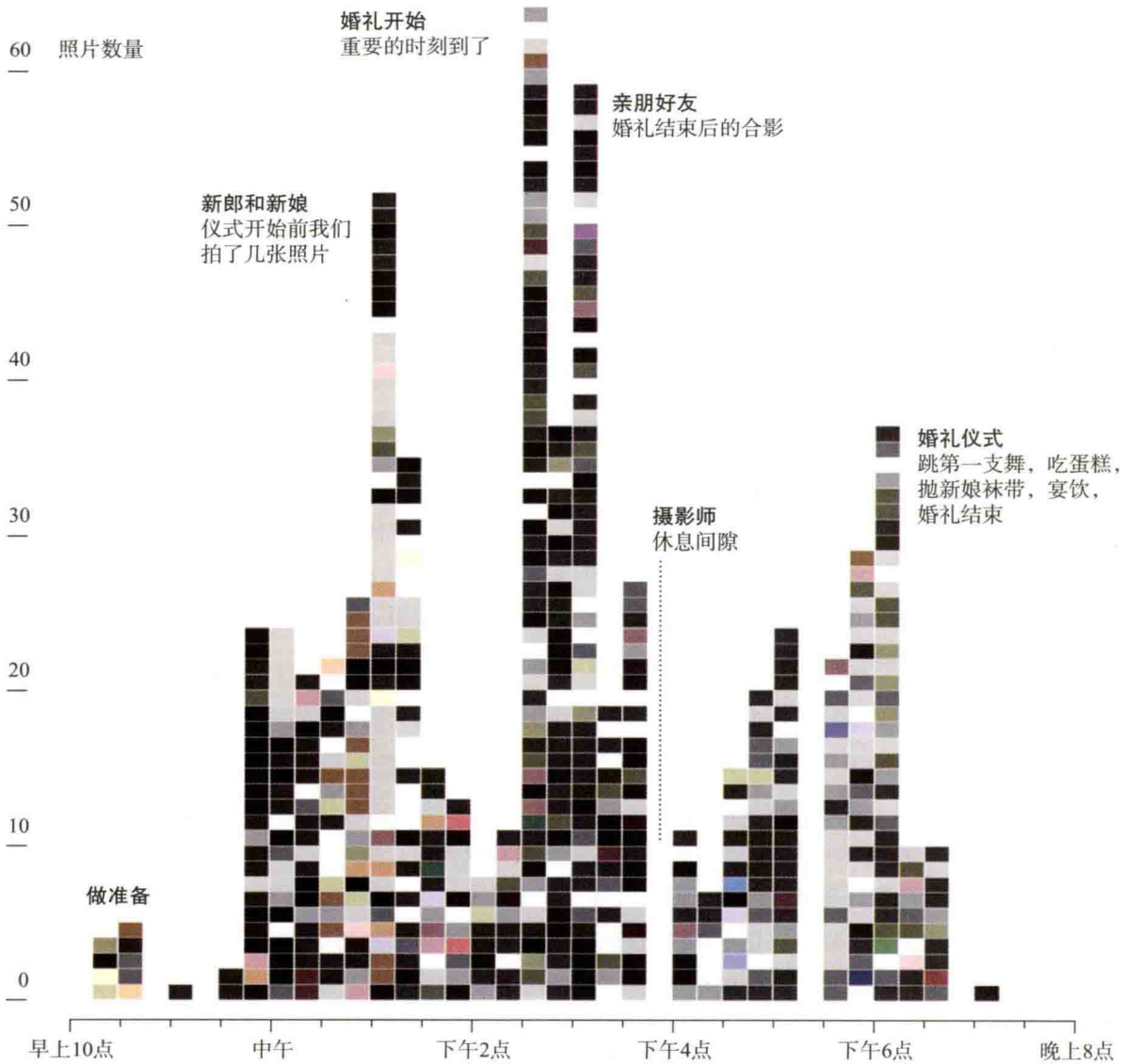

图 1—3　婚礼中的颜色

其实你可以跳出来换个角度，只去关注照片数，而暂时忽略那些颜色和一张张独立的照片。如图 1—4 所示，可能你以前看过这样的图表。它是条形图，显示了与图 1—3 一样的高峰和低谷，但是它给人的感觉不一样，提供了不同的信息。这个简单的条形图强调了每 15 分钟拍摄了多少张照片，而图 1—3 仍带有相册的感觉。

有一个重要的事实需要注意，这四个视图显示的数据相同。更确切地说，它们都描绘了我结婚那天的情景。每一张图表都用不同的方式从婚礼的各个方面展现了那一天的情景。对数据的诠释可以随着它所呈现的视觉形式而改变。对于传统的数据，通常用条形图进行考察和研究，但那并不意味着你必须失去每一张照片里所包含的感情。有时候，你需增加注解以便读者能更好地理解数据，而有时候数字传达的信息则是清晰的，可以从可视化图表中直观地获得。

数据和它所代表事物之间的关联既是把数据可视化的关键，也是全面分析数据的关键，同样还是深层次理解数据的关键。计算机可以把数字批量转换成不同的形状和颜色，但是你必须建立起数据和现实世界的联系，以便使用图表的人能够从中得到有价值的信息。

有时候很难找到这个关联，比如，当你在研究涉及成千上万陌生人这样的大规模数据时。当研究一个个体时，这种关联就明显多了。你甚至可以直接联系那个人，即使你从来没有见过他。例如，来自波特兰的软件开发者亚伦·帕拉茨基（Aaron Parecki）在 2008 年到 2012 年的三年半时间里用手机收集了 250 万个 GPS 坐标位置，每 2 ～ 6 秒就记录一个坐标点。图 1—5 是这些坐标点的地图，不同颜色代表不同的年份。

如你所期望的，这张地图显示出了帕拉茨基经常出入的那些道路和区域的颜色比其他地方要亮。他搬了几次家，你可以看到他的出行模式每年都在变。2008 年到 2010 年期间，他的出行路线（蓝色）很分散。到了 2012 年，黄色路线显示帕拉茨基看上去像是活动在几个紧挨在一起的很小的区域里。因为没有更多的背景介绍，所以你很难再说出其他信息，因为你所看到的仅仅是地理位置。但是对帕拉茨基来说，这些数据更具有个人色彩，就像那张婚礼照片对于我一样。它是三年多时间里一个人在一个城市里的足迹。因为帕拉茨基有原始的记录，有时间信息，他可以基于这些数据做出更好的决定，比如什么时候去上班最好。

然而，如果在个人的时间和地理位置信息上附加更多的信息将会怎样？如果在记录你身处何处的同时，也记下了在某些指定的时刻发生了什么，又将会怎样？这就是艺术家蒂姆·克拉克（Tim Clark）在 2010 年到 2011 年间完成的“习惯图集”（Atlas of the Habitual）项目。像帕拉