

AUTHOR **NATHAN YAU**



DATA POINTS

VISUALIZATION THAT MEANS SOMETHING

Ex. 01



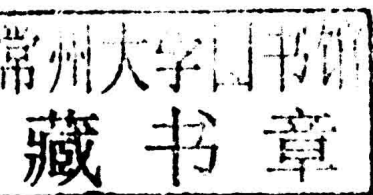
FLOWINGDATA.COM

WILEY

Data Points

Visualization That Means Something

Nathan Yau



WILEY

Data Points: Visualization That Means Something

EXECUTIVE EDITOR: Carol Long
SENIOR PROJECT EDITOR: Adaobi Obi Tulton
TECHNICAL EDITOR: Jen Lowe
PRODUCTION EDITOR: Kathleen Wisor
COPY EDITOR: Apostrophe Editing Services
EDITORIAL MANAGER: Mary Beth Wakefield
FREELANCER EDITORIAL MANAGER: Rosemarie Graham
ASSOCIATE DIRECTOR OF MARKETING: David Mayhew
MARKETING MANAGER: Ashley Zurcher
BUSINESS MANAGER: Amy Knies
PRODUCTION MANAGER: Tim Tate
VICE PRESIDENT AND EXECUTIVE GROUP PUBLISHER: Richard Swadley
VICE PRESIDENT AND EXECUTIVE PUBLISHER: Neil Edde
ASSOCIATE PUBLISHER: Jim Minatel
PROJECT COORDINATOR, COVER: Katie Crocker
COMPOSITOR: James D. Kramer, Happenstance Type-O-Rama
PROOFREADER: James Saturino, Word One
INDEXER: Johnna VanHoose Dinse
COVER DESIGNER: Ryan Sneed
COVER IMAGE: Nathan Yau
CHAPTER PHOTOGRAPHS: Serena Yau

Published by
John Wiley & Sons, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

Copyright © 2013 by John Wiley & Sons, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 978-1-118-46219-5
ISBN: 978-1-118-46222-5 (ebk)
ISBN: 978-1-118-65493-4 (ebk)
ISBN: 978-1-118-65493-4 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2012956416

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

Data Points

To Bea

About the Author

[Nathan Yau](#) has written and created graphics for FlowingData since 2007, a site on visualization, statistics, and design and believes that visualization is a medium that can be used as both a tool and a way to express data. He is the author of *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*, also published by Wiley.

Yau has a master's degree in statistics from the University of California, Los Angeles, and has a Ph.D. in statistics with a focus on visualization and personal data. He is a statistician.

About the Technical Editor

[Jen Lowe](#) is an Associate Research Scholar at the Spatial Information Design Lab at Columbia University where she experiments with new forms of data visualization and communication. She has a bachelor's degree in applied mathematics and a master's degree in information science. She has publications in applied meteorology and astronomy in journals such as the *Journal of Atmospheric and Oceanic Technology*, and has spoken internationally on the relationship between data, visualization, and identity. A sucker for a good argument, she'll nearly always side with complexity.

Acknowledgments

Thank you to everyone I spoke to and received feedback from in the process of writing this book. You provided valuable insight and gave me fresh perspectives. Thank you to those who graciously let me feature their beautiful work.

Thank you to FlowingData readers who continue to support and always give me something to be excited about. This book would not exist without you.

Thank you to Wiley Publishing for helping me write a book worth reading and to Jen Lowe for making sure my words meant something.

Thank you to Serena Yau for taking pictures.

Thank you to the statisticians, computer scientists, data scientists, designers, and everyone else involved in this hodge-podge data world for making it fun and always keeping it interesting.

Thank you to all those who have taught me.

Thank you to my wife, my parents, and all other family and friends for their support and encouragement.

Introduction

What is good visualization? It is a representation of data that helps you see what you otherwise would have been blind to if you looked only at the naked source. It enables you to see trends, patterns, and outliers that tell you about yourself and what surrounds you. The best visualization evokes that moment of bliss when seeing something for the first time, knowing that what you see has been right in front of you, just slightly hidden. Sometimes it is a simple bar graph, and other times the visualization is complex because the data requires it.

THE PROCESS

A data set is a snapshot in time that captures something that moves and changes. Collectively, data points form aggregates and statistical summaries that can tell you what to expect. These are your means, medians, and standard deviations. They describe the world, countries, and populations and enable you to compare and contrast things. When you push down on the data, you get details about the individuals and objects within the population. These are the stories that make a data set human and relatable.

Data in an abstract sense that includes information and facts is the foundation of every visualization. The more you understand about the source and the stronger base that you build, the greater the potential for a compelling data graphic. This is the part that a lot of people miss: Good visualization is a winding process that requires statistics and design knowledge. Without the former, the visualization becomes an exercise only in illustration and aesthetics, and without the latter, one of only analyses. On their own, these are fine skills, but they make for incomplete data graphics. Having skills in both provides you with the luxury—which is growing into a necessity—to jump back and forth between data exploration and storytelling.

This book is for those interested in the process of design and analysis, where each chapter represents a step toward visualization that means something. It is about visualization that is more than large printed numbers with clipart. It is about making sense of data. Visualization creation is iterative, and the cycle is always a little different for each new dataset.

The first part of *Data Points* helps you know your data and what it means to visualize it. Because of what data represents—people, places, and things—there is always important context attached to the factual numbers. Who is the data about? Where is the data from? When was it collected? We are responsible for this human part of the computer-generated output, too. On top of that, most datasets are estimated, so they are not the absolute truth; there is uncertainty and variability attached, just like in real life.

In the middle of the book, you go into exploration mode. You have the freedom to ask questions and try to answer them by digging through the data. Look for patterns, relationships, and anything that does not look right. Missing values are common, as are typos. This is a great time to play and experiment, to look around from different angles, and maybe you'll find something unexpected. Maybe that bit ends up being the most interesting part of the story. For whatever reason, the exploration stage is skipped too often, and lack of understanding shows in the final product. Take the time to get to know your data and what it represents, and the visualization improves exponentially.

When you find the underlying narratives, the next step is typically to communicate your results to a wider audience. This is the last section, when you put your design hat on. A graphic for a small audience of four people who are familiar with the subject matter and have read every significant paper on the topic will be different than a graphic for a large, general audience unfamiliar with the complex background implied by the numbers.

Again, these stages are not meant as a step-by-step guide. If you work with data already, you know that it is common to discover a need for new data as you explore what you already have. Similarly, the design process can force you to see details that you didn't notice before, which takes you back to exploration or to the beginning. If you are new to data, you must learn this process while reading this book to feel confident enough to use it in your own projects. The back and forth between data and story is the fun stuff.

Data Points is a complement to my previous book *Visualize This*. The first book serves as an introduction to the tools available and offers concrete programming examples, whereas this book describes the full process and thinking that goes into larger data projects and is software-independent. In other words, the two books feed off each other. *Visualize This* provides technical guidance for those ready to make their own graphics, and *Data Points* describes a process for data and visualization so that you can create better and more thoughtful things.

MORE THAN A TOOL

Throughout this book, visualization is referred to as a medium rather than a specific tool. When you approach visualization as an unyielding tool, it is easy to get caught thinking that almost every graphic would be better as a bar graph. This is true for a lot of charts, but it must be in the right context. In an analysis setting, yes, you often want graphs that read the quickest and most accurately; however, from another point of view, such a comment might be premature. What if emotion and curiosity are the goals? Visualization is a way to represent data, an abstraction of the real world, in the same way that the written word can be used to tell different kinds of stories. Newspaper articles aren't judged on the same criteria as novels, and data art should be critiqued differently than a business dashboard.

That said, there are rules to follow regardless of the visualization type. These aren't dictated by design or statistics. Rather they are governed by human perception, and they ensure accuracy when readers interpret encoded data. There are only a handful of these, such as to properly size by area when that aspect is the actual visual cue, and all the rest are suggestions.

You must distinguish between rules and suggestions. You should follow the former almost always, whereas suggestions are rooted in opinion and vary among individuals and situations. Many beginners make the mistake that advice is concrete, and they lose the context in which the data is presented in. For example, Edward Tufte suggests stripping charts of all junk, but the definition of junk can change. What needs to be stripped from one chart might need to stay in another. In the words of Tufte, "Most principles of design should be greeted with some skepticism."

Similarly, people often cite the work of statisticians William Cleveland and Robert McGill on perception and accuracy. They found that position along a common scale, such as with a scatterplot, was decoded most accurately, followed by length, angle, and then slope. These results are based on research trials and were reproduced in other studies, so it is easy to mistake Cleveland and McGill's findings as rules. However, Cleveland also notes that the mark of a good graph is not only how fast you can read it, but also what it shows. Does it enable you to see what you could not see before?

You must come back to the data for worthwhile visualization. Fortunately, you have plenty of data to play with, and the source keeps growing. Every week for the past few years, there is an article that describes the flood of data and the risk of drowning in it, but you see, the amount is controlled, and you can easily filter and aggregate it. Storage is cheap and practically infinite, which means more potential happy feelings for those who know how to swim. The challenge is learning to dive deeper.

Okay, I'm psyched. Let's have some fun.

Data Points



Contents

	<i>Introduction.....</i>	<i>xi</i>
1	Understanding Data	1
	What Data Represents.....	2
	Variability.....	20
	Uncertainty.....	30
	Context.....	35
	Wrapping Up.....	41
2	Visualization: The Medium	43
	Analysis and Exploration.....	45
	Information Graphics and Presentation.....	58
	Entertainment.....	69
	Data Art.....	74
	The Everyday.....	81
	Wrapping Up.....	89
3	Representing Data	91
	Visualization Components.....	93
	Putting It Together.....	115
	Wrapping Up.....	132
4	Exploring Data Visually	135
	Process.....	136
	Visualizing Categorical Data.....	143
	Visualizing Time Series Data.....	154
	Visualizing Spatial Data.....	165
	Multiple Variables.....	176
	Distributions.....	193
	Wrapping Up.....	199

5	Visualizing with Clarity	201
	Visual Hierarchy.....	202
	Readability.....	205
	Highlighting.....	221
	Annotation.....	228
	Do the Math.....	236
	Wrapping Up.....	239
6	Designing for an Audience	241
	Common Misconceptions.....	242
	Present Data to People.....	254
	Things to Consider.....	258
	Putting It Together.....	268
	Wrapping Up.....	273
7	Where to Go from Here	277
	Visualization Tools.....	278
	Programming.....	283
	Illustration.....	288
	Statistics.....	289
	Wrapping Up.....	289
	<i>Index</i>	291



Understanding Data

When you ask people what data is, most reply with a vague description of something that resembles a spreadsheet or a bucket of numbers. The more technically savvy might mention databases or warehouses. However, this is just the format that the data comes in and how it is stored, and it doesn't say anything about what data is or what any particular dataset represents. It's an easy trap to fall in because when you ask for data, you usually get a computer file, and it's hard to think of computer output as anything but just that. Look beyond the file though, and you get something more meaningful.

WHAT DATA REPRESENTS

Data is more than numbers, and to visualize it, you must know what it represents. Data represents real life. It's a snapshot of the world in the same way that a photograph captures a small moment in time.

Look at Figure 1-1. If you were to come across this photo, isolated from everything else, and I told you nothing about it, you wouldn't get much out of it. It's just another wedding photo. For me though, it's a happy moment during one of the best days of my life. That's my wife on the left, all dolled up, and me on the right, wearing something other than jeans and a T-shirt for a change. The

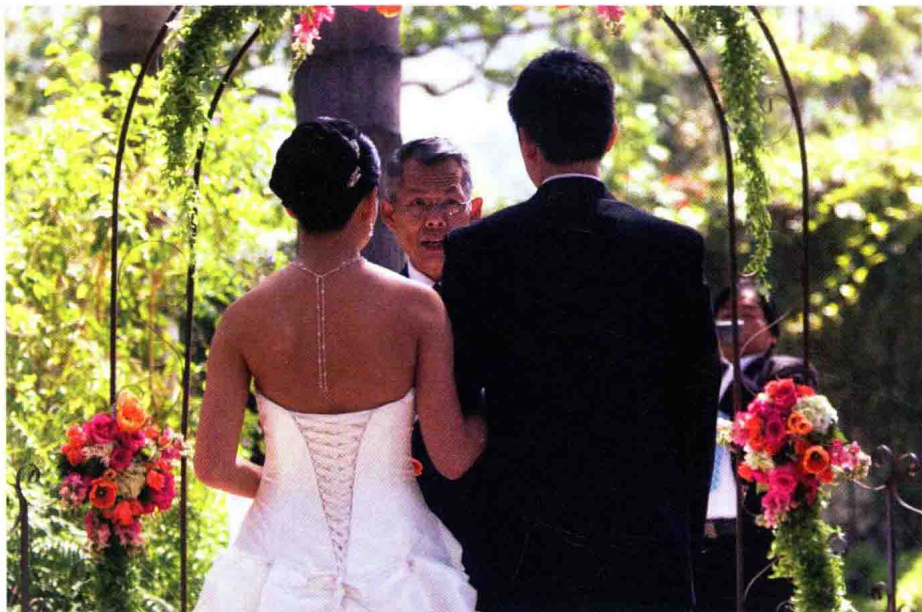


FIGURE 1-1 A single photo, a single data point

pastor who is marrying us is my wife's uncle, who added a personal touch to the ceremony, and the guy in the back is a family friend who took it upon himself to record as much as possible, even though we hired a photographer. The flowers and archway came from a local florist about an hour away from the venue, and the wedding took place during early summer in Los Angeles, California.

That's a lot of information from just one picture, and it works the same with data. (For some, me included, pictures are data, too.) A single data point can have a who, what, when, where, and why attached to it, so it's easy for a digit to become more than a toss in a bucket. Extracting information from a data point isn't as easy as looking at a photo, though. You can guess what's going on in the photo, but when you make assumptions about data, such as how accurate it is or how it relates to its surroundings, you can end up with a skewed view of what your data actually represents. You need to look at everything around, find context, and see what your dataset looks like as a whole. When you see the full picture, it's much easier to make better judgments about individual points.

Imagine that I didn't tell you those things about my wedding photo. How could you find out more? What if you could see pictures that were taken before and after?



FIGURE 1-2 Grid of photos