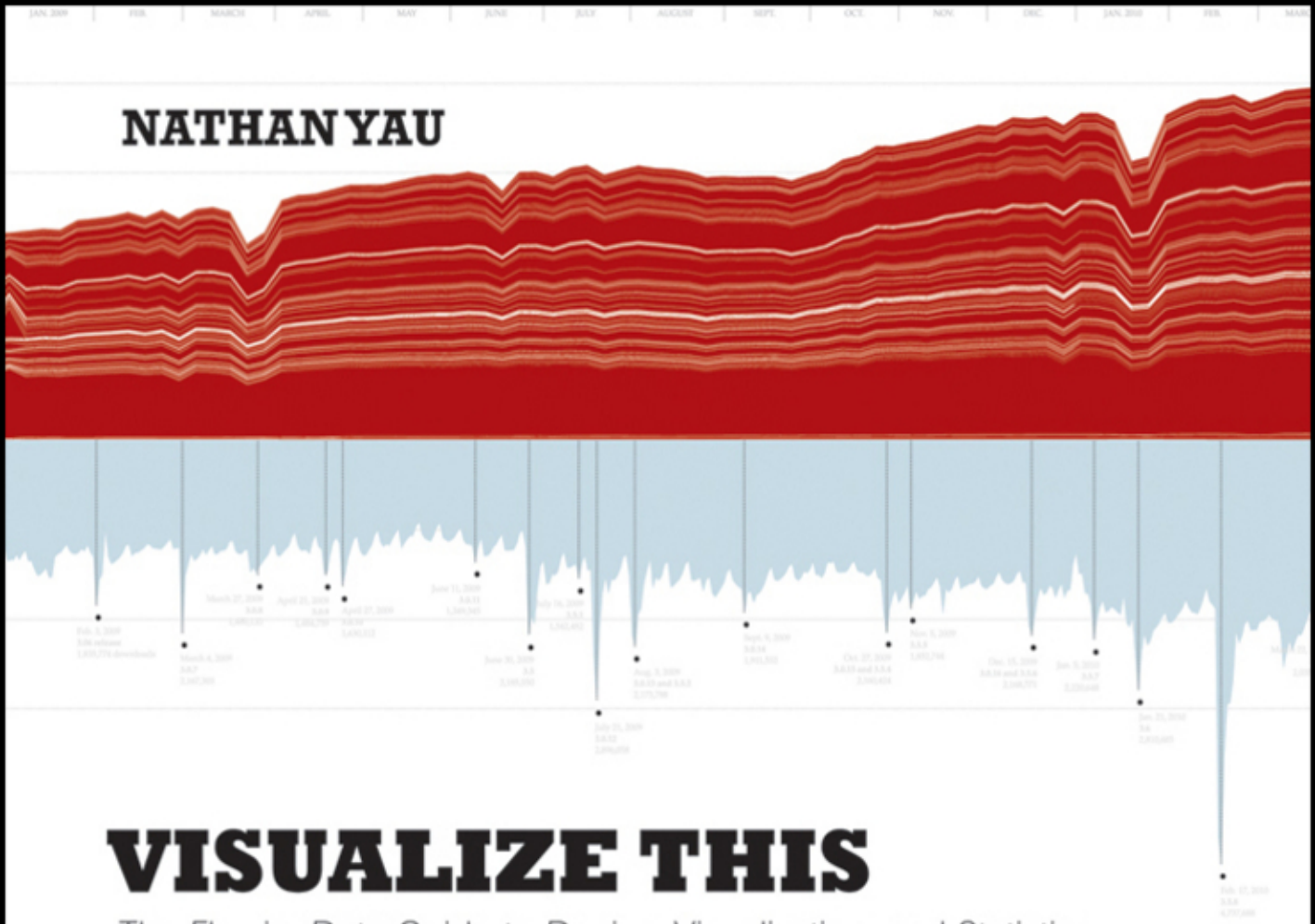
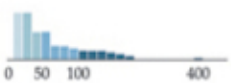
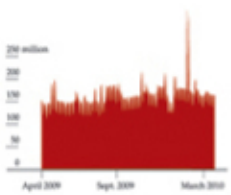
NATHAN YAU

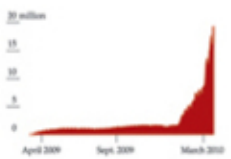


VISUALIZE THIS

The FlowingData Guide to Design, Visualization, and Statistics







Visualize This

Visualize This

The FlowingData Guide to Design, Visualization, and Statistics

Nathan Yau

Wiley Publishing, Inc.

Visualize This: The FlowingData Guide to Design, Visualization, and Statistics

Published by
Wiley Publishing, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

Published by Wiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 978-0-470-94488-2
ISBN: 978-1-118-14024-6 (ebk)
ISBN: 978-1-118-14026-0 (ebk)
ISBN: 978-1-118-14025-3 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Control Number: 2011928441

To my loving wife, Bea

About the Author

Since 2007, NATHAN YAU has written and created graphics for FlowingData, a site on visualization, statistics, and design. Working with groups such as *The New York Times*, CNN, Mozilla, and SyFy, Yau believes that data and information graphics, while great for analysis, are also perfect for telling stories with data.

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

About the Technical Editor

KIM REES is co-founder of Periscopic, a socially conscious information visualization firm. A prominent individual in the visualization community, Kim has over seventeen years of experience in the interactive industry. She has published papers in the *Parsons Journal of Information Mapping* and the InfoVIS 2010 Proceedings, and has spoken at the O'Reilly Strata Conference, WebVisions, AIGA Shift, and Portland Data Visualization. Kim received her bachelor of arts in Computer Science from New York University. Periscopic has been recognized in CommArts Insights, Adobe Success Stories, and awarded by the VAST Challenge, CommArts Web Picks, and the *Communication Arts Interactive Annual*. Recently, Periscopic's body of work was nominated for the Cooper-Hewitt National Design Awards.

Credits

Executive Editor
Carol Long

Senior Project Editor
Adaobi Obi Tulton

Technical Editor
Kim Rees

Senior Production Editor
Debra Banninger

Copy Editor
Apostrophe Editing Services

Editorial Director
Robyn B. Siesky

Editorial Manager
Mary Beth Wakefield

Freelancer Editorial Manager
Rosemarie Graham

Marketing Manager
Ashley Zurcher

Production Manager
Tim Tate

Vice President and Executive Group Publisher
Richard Swadley

Vice President and Executive Publisher
Barry Pruett

Associate Publisher
Jim Minatel

Project Coordinator, Cover
Katie Crocker

Compositor
Maureen Forys,
Happenstance Type-O-Rama

Proofreader
Nancy Carrasco

Indexer
Robert Swanson

Cover Image
Nathan Yau

Cover Designer
Ryan Sneed

Acknowledgments

THIS BOOK would not be possible without the work by the data scientists before me who developed and continue to create useful and open tools for everyone to use. The software from these generous developers makes my life much easier, and I am sure they will keep innovating.

My many thanks to FlowingData readers who helped me reach more people than I ever imagined. They are one of the main reasons why this book was written.

Thank you to Wiley Publishing, who let me write the book that I wanted to, and to Kim Rees for helping me produce something worth reading.

Finally, thank you to my wife for supporting me and to my parents who always encouraged me to find what makes me happy.

Contents

Introduction . . . . . *xv*

1 **Telling Stories with Data** 1
- More Than Numbers 2
- What to Look For 8
- Design 13
- Wrapping Up 20

2 **Handling Data** 21
- Gather Data. 22
- Formatting Data 38
- Wrapping Up 52

3 **Choosing Tools to Visualize Data.** 53
- Out-of-the-Box Visualization 54
- Programming 62
- Illustration 76
- Mapping 80
- Survey Your Options 88
- Wrapping Up 89

4 **Visualizing Patterns over Time** 91
- What to Look for over Time 92
- Discrete Points in Time. 93
- Continuous Data 118
- Wrapping Up 132

5 **Visualizing Proportions** 135
- What to Look for in Proportions. 136
- Parts of a Whole 136

Proportions over Time 161
Wrapping Up 178

6 Visualizing Relationships 179
What Relationships to Look For 180
Correlation 180
Distribution 200
Comparison 213
Wrapping Up 226

7 Spotting Differences 227
What to Look For 228
Comparing across Multiple Variables 228
Reducing Dimensions 258
Searching for Outliers 265
Wrapping Up 269

8 Visualizing Spatial Relationships 271
What to Look For 272
Specific Locations 272
Regions 285
Over Space and Time 302
Wrapping Up 325

9 Designing with a Purpose 327
Prepare Yourself 328
Prepare Your Readers 330
Visual Cues 334
Good Visualization 340
Wrapping Up 341

Index *343*

Introduction

Data is nothing new. People have been quantifying and tabulating things for centuries. However, while writing for FlowingData, my website on design, visualization, and statistics, I've seen a huge boom in just these past few years, and it keeps getting better. Improvements in technology have made it extremely easy to collect and store data, and the web lets you access it whenever you want. This wealth in data can, in the right hands, provide a wealth of information to help improve decision making, communicate ideas more clearly, and provide a more objective window looking in at how you look at the world and yourself.

A significant shift in release of government data came in mid-2009, with the United States' launch of Data.gov. It's a comprehensive catalog of data provided by federal agencies and represents transparency and accountability of groups and officials. The thought here is that you should know how the government spends tax dollars. Whereas before, the government felt more like a black box. A lot of the data on Data.gov was already available on agency sites scattered across the web, but now a lot of it is all in one place and better formatted for analysis and visualization. The United Nations has something similar with UNdata; the United Kingdom launched Data.gov.uk soon after, and cities around the world such as New York, San Francisco, and London have also taken part in big releases of data.

The collective web has also grown to be more open with thousands of Application Programming Interfaces (API) to encourage and entice developers to do something with all the available data. Applications such as Twitter and Flickr provide comprehensive APIs that enable completely different user interfaces from the actual sites. API-cataloging site ProgrammableWeb reports more than 2,000 APIs. New applications, such as Infochimps and Factual, also launched fairly recently and were specifically developed to provide structured data.

At the individual level, you can update friends on Facebook, share your location on Foursquare, or tweet what you're doing on Twitter, all with a few clicks on a mouse or taps on a keyboard. More specialized applications enable you to log what you eat, how much you

weigh, your mood, and plenty of other things. If you want to track something about yourself, there is probably an application to help you do it.

With all this data sitting around in stores, warehouses, and databases, the field is ripe for people to make sense of it. The data itself isn't all that interesting (to most people). It's the information that comes out of the data. People want to know what their data says, and if you can help them, you're going to be in high demand. There's a reason that Hal Varian, Google's chief economist, says that statistician is the sexy job of the next 10 years, and it's not just because statisticians are beautiful people. (Although we are quite nice to look at in that geek chic sort of way.)

Visualization

One of the best ways to explore and try to understand a large dataset is with visualization. Place the numbers into a visual space and let your brain or your readers' brains find the patterns. We're good at that. You can often find stories that you might never have found with just formal statistical methods.

John Tukey, my favorite statistician and the father of exploratory data analysis, was well versed in statistical methods and properties but believed that graphical techniques also had a place. He was a strong believer in discovering the unexpected through pictures. You can find out a lot about data just by visualizing it, and a lot of the time this is all you need to make an informed decision or to tell a story.

For example, in 2009, the United States experienced a significant increase in its unemployment rate. In 2007, the national average was 4.6 percent. In 2008, it had risen to 5.8 percent. By September 2009, however, it was 9.8 percent. These national averages tell only part of the story though. It's generalizing over an entire country. Were there any regions that had higher unemployment rates than others? Were there any regions that seemed to be unaffected?

The maps in Figure I-1 tell a more complete story, and you can answer the preceding questions after a glance. Darker-colored counties are areas that had relatively higher unemployment rates, whereas the lighter-colored counties had relatively lower rates. In 2009, you see a lot of regions with rates greater than 10 percent in the west and most areas in the east. Areas in the Midwest were not hit as hard (Figure I-2).

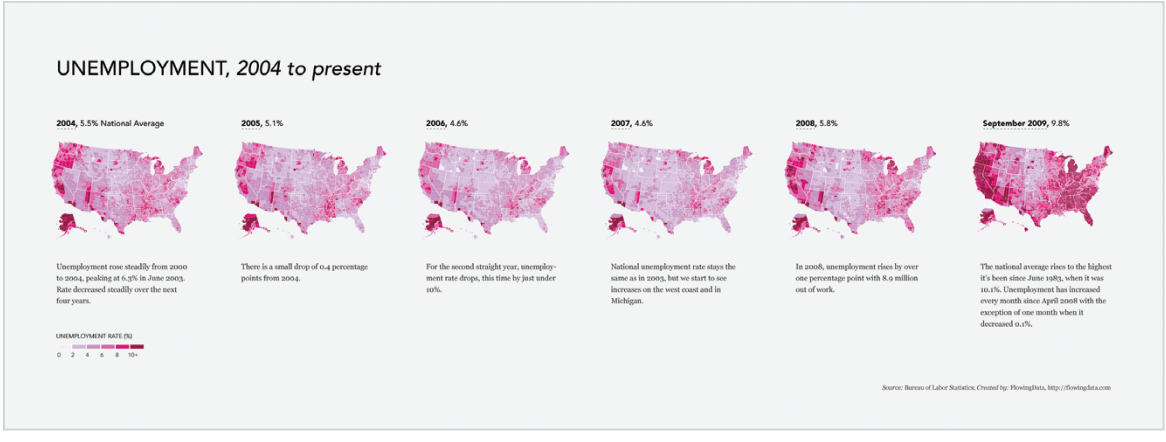


FIGURE I-1 Maps of unemployment in the United States from 2004 to 2009

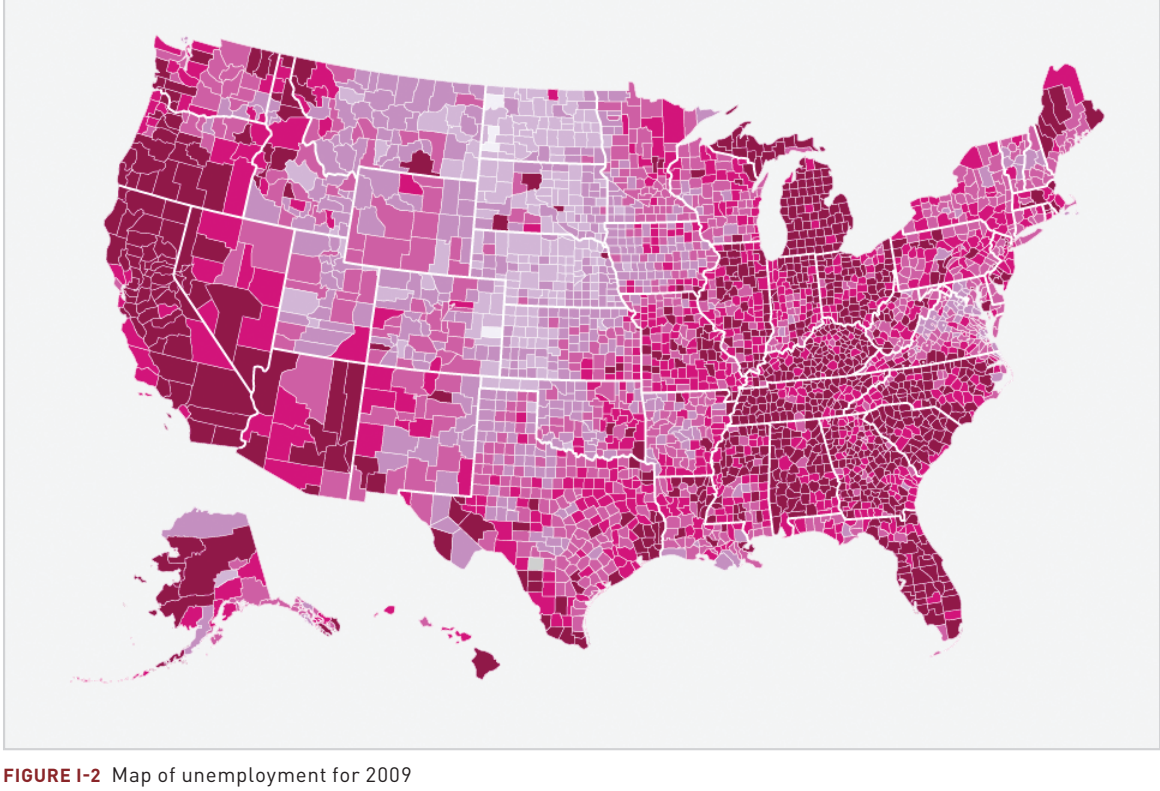


FIGURE I-2 Map of unemployment for 2009

You couldn't find these geographic and temporal patterns so quickly with just a spreadsheet, and definitely not with just the national averages. Also, although the county-level data is more complex, most people can still interpret the maps. These maps could in turn help policy makers decide where to allocate relief funds or other types of support.

The great thing about this is that the data used to produce these maps is all free and publicly available from the Bureau of Labor Statistics. Albeit the data was not incredibly easy to find from an outdated data browser, but the numbers are there at your disposal, and there is a lot sitting around waiting for some visual treatment.

The Statistical Abstract of the United States, for instance, exists as hundreds of tables of data (Figure I-3), but no graphs. That's an opportunity to provide a comprehensive picture of a country. Really interesting stuff. I graphed some of the tables a while back as a proof of concept, as shown in Figure I-4, and you get marriage and divorce rates, postal rates, electricity usage, and a few others. The former is hard to read and you don't get anything out of it other than individual values. In the graphical view, you can find trends and patterns easily and make comparisons at a glance.

News outlets, such as *The New York Times* and *The Washington Post* do a great job at making data more accessible and visual. They have probably made the best use of this available data, as related stories have come and passed. Sometimes data graphics are used to enhance a story with a different point of view, whereas other times the graphics tell the entire story.

Graphics have become even more prevalent with the shift to online media. There are now departments within news organizations that deal only with interactives or only graphics or only maps. *The New York Times*, for example, even has a news desk specifically dedicated to what it calls computer-assisted reporting. These are reporters who focus on telling the news with numbers. *The New York Times* graphics desk is also comfortable dealing with large amounts of data.

Visualization has also found its way into pop culture. Stamen Design, a visualization firm well known for its online interactives, has provided a Twitter tracker for the MTV Video Music Awards the past few years. Each year Stamen designs something different, but at its core, it shows what people are talking about on Twitter in real-time. When Kanye West had his little outburst during Taylor Swift's acceptance speech in 2009, it was obvious what people thought of him via the tracker.

Table 126. **Marriages and Divorces—Number and Rate by State: 1990 to 2007**

[**2,443.5 represents 2,443,500.** By place of occurence. See Appendix III]

State	Marriages [1]						Divorces [3]					
	Number (1,000)			Rate per 1,000 population [2]			Number (1,000)			Rate per 1,000 population [2]		
	1990	2000	2007	1990	2000	2007	1990	2000	2007	1990	2000	2007
U.S. [4]	**2,443.5**	**2,329.0**	**2,204.6**	**9.8**	**8.3**	**7.3**	**1,182.0**	**(NA)**	**(NA)**	**4.7**	**4.1**	**3.6**
Alabama	43.1	45.0	42.4	10.6	10.3	9.2	25.3	23.5	19.8	6.1	5.4	4.3
Alaska	5.7	5.6	5.8	10.2	8.9	8.4	2.9	2.7	3.0	5.5	4.4	4.3
Arizona [5]	36.8	38.7	39.5	10.0	7.9	6.2	25.1	21.6	24.5	6.9	4.4	3.9
Arkansas	36.0	41.1	33.7	15.3	16.0	11.9	16.8	17.9	16.8	6.9	6.9	5.9
California	237.1	196.9	225.8	7.9	5.9	6.2	128.0	(NA)	(NA)	4.3	(NA)	(NA)
Colorado	32.4	35.6	29.2	9.8	8.6	6.0	18.4	(NA)	21.2	5.5	(NA)	4.4
Connecticut	26.0	19.4	17.3	7.9	5.9	4.9	10.3	6.5	10.7	3.2	2.0	3.1
Delaware	5.6	5.1	4.7	8.4	6.7	5.5	3.0	3.2	3.9	4.4	4.2	4.5
District of Columbia	5.0	2.8	2.1	8.2	5.4	3.6	2.7	1.5	1.0	4.5	3.0	1.6
Florida	141.8	141.9	157.6	10.9	9.3	8.6	81.7	81.9	86.4	6.3	5.3	4.7
Georgia	66.8	56.0	64.0	10.3	7.1	6.7	35.7	30.7	(NA)	5.5	3.9	(NA)
Hawaii	18.3	25.0	27.3	16.4	21.2	21.3	5.2	4.6	(NA)	4.6	3.9	(NA)
Idaho	14.1	14.0	15.4	13.9	11.0	10.3	6.6	6.9	7.4	6.5	5.4	4.9
Illinois	100.6	85.5	75.3	8.8	7.0	5.9	44.3	39.1	32.8	3.8	3.2	2.6
Indiana	53.2	34.5	51.2	9.6	5.8	8.1	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Iowa	24.9	20.3	20.1	9.0	7.0	6.7	11.1	9.4	7.8	3.9	3.3	2.6
Kansas	22.7	22.2	18.6	9.2	8.3	6.7	12.6	10.6	9.2	5.0	4.0	3.3
Kentucky	49.8	39.7	33.6	13.5	10.0	7.9	21.8	21.6	19.7	5.8	5.4	4.6
Louisiana	40.4	40.5	32.8	9.6	9.3	7.6	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Maine	11.9	10.5	10.1	9.7	8.3	7.7	5.3	5.8	5.9	4.3	4.6	4.5
Maryland	46.3	40.0	35.5	9.7	7.7	6.3	16.1	17.0	17.4	3.4	3.3	3.1
Massachusetts	47.7	37.0	38.4	7.9	6.0	6.0	16.8	18.6	14.5	2.8	3.0	2.2
Michigan	76.1	66.4	59.1	8.2	6.7	5.9	40.2	39.4	35.5	4.3	4.0	3.5
Minnesota	33.7	33.4	29.8	7.7	6.9	5.7	15.4	14.8	(NA)	3.5	3.1	(NA)
Mississippi	24.3	19.7	15.7	9.4	7.1	5.4	14.4	14.4	14.2	5.5	5.2	4.9
Missouri	49.1	43.7	39.4	9.6	7.9	6.7	26.4	26.5	22.4	5.1	4.8	3.8
Montana	6.9	6.6	7.1	8.6	7.4	7.4	4.1	2.1	3.6	5.1	2.4	3.7
Nebraska	12.6	13.0	12.4	8.0	7.8	7.0	6.5	6.4	5.5	4.0	3.8	3.1
Nevada	120.6	144.3	126.4	99.0	76.7	49.3	13.3	18.1	16.6	11.4	9.6	6.5
New Hampshire	10.5	11.6	9.4	9.5	9.5	7.1	5.3	7.1	5.1	4.7	5.8	3.9
New Jersey	58.7	50.4	45.4	7.6	6.1	5.2	23.6	25.6	25.7	3.0	3.1	3.0
New Mexico [5]	13.3	14.5	11.2	8.8	8.3	5.7	7.7	9.2	8.4	4.9	5.3	4.3
New York [5]	154.8	162.0	130.6	8.6	8.9	6.8	57.9	62.8	55.9	3.2	3.4	2.9
North Carolina	51.9	65.6	68.1	7.8	8.5	7.5	34.0	36.9	37.4	5.1	4.8	4.1
North Dakota	4.8	4.6	4.2	7.5	7.3	6.6	2.3	2.0	1.5	3.6	3.2	2.4
Ohio	98.1	88.5	70.9	9.0	7.9	6.2	51.0	49.3	37.9	4.7	4.4	3.3
Oklahoma	33.2	15.6	26.2	10.6	4.6	7.3	24.9	12.4	18.8	7.7	3.7	5.2
Oregon	25.3	26.0	29.4	8.9	7.8	7.8	15.9	16.7	14.8	5.5	5.0	4.0
Pennsylvania	84.9	73.2	71.1	7.1	6.1	5.7	40.1	37.9	35.3	3.3	3.2	2.8
Rhode Island	8.1	8.0	6.8	8.1	8.0	6.4	3.8	3.1	3.0	3.7	3.1	2.8
South Carolina	55.8	42.7	31.4	15.9	10.9	7.1	16.1	14.4	14.4	4.5	3.7	3.3
South Dakota	7.7	7.1	6.2	11.1	9.6	7.7	2.6	2.7	2.4	3.7	3.6	3.1
Tennessee	68.0	88.2	65.6	13.9	15.9	10.6	32.3	33.8	29.9	6.5	6.1	4.9
Texas	178.6	196.4	179.9	10.5	9.6	7.5	94.0	85.2	79.5	5.5	4.2	3.3
Utah	19.4	24.1	22.6	11.2	11.1	8.6	8.8	9.7	8.9	5.1	4.5	3.4
Vermont	6.1	6.1	5.3	10.9	10.2	8.6	2.6	5.1	2.4	4.5	8.6	3.8
Virginia	71.0	62.4	58.0	11.4	9.0	7.5	27.3	30.2	29.5	4.4	4.3	3.8
Washington	46.6	40.9	41.8	9.5	7.0	6.5	28.8	27.2	28.9	5.9	4.7	4.5
West Virginia	13.0	15.7	13.0	7.2	8.7	7.2	9.7	9.3	9.0	5.3	5.2	5.0
Wisconsin	38.9	36.1	32.2	7.9	6.8	5.8	17.8	17.6	16.1	3.6	3.3	2.9
Wyoming	4.9	4.9	4.8	10.7	10.3	9.3	3.1	2.8	2.9	6.6	5.9	5.5

NA Not available. [1] Data are counts of marriages performed, except as noted. [2] Based on total population residing in area; population enumerated as of April 1 for 1990 and 2000; estimated as of July 1 for all other years. [3] Includes annulments. [4] U.S. total for the number of divorces is an estimate which includes states not reporting. Beginning 2000, divorce rates based solely on the combined counts and populations for reporting states and the District of Columbia. The collection of detailed data of marriages and divorces was suspended in January 1996. [5] Some figures for marriages are marriage licenses issued.

Source: U.S. National Center for Health Statistics, National Vital Statistics Reports (NVSR), *Births, Marriages, Divorces, and Deaths: Provisional Data for 2007*, Vol. 56, No. 21, July 14, 2008 and prior reports.

FIGURE I-3 Table from the Statistical Abstract of the United States

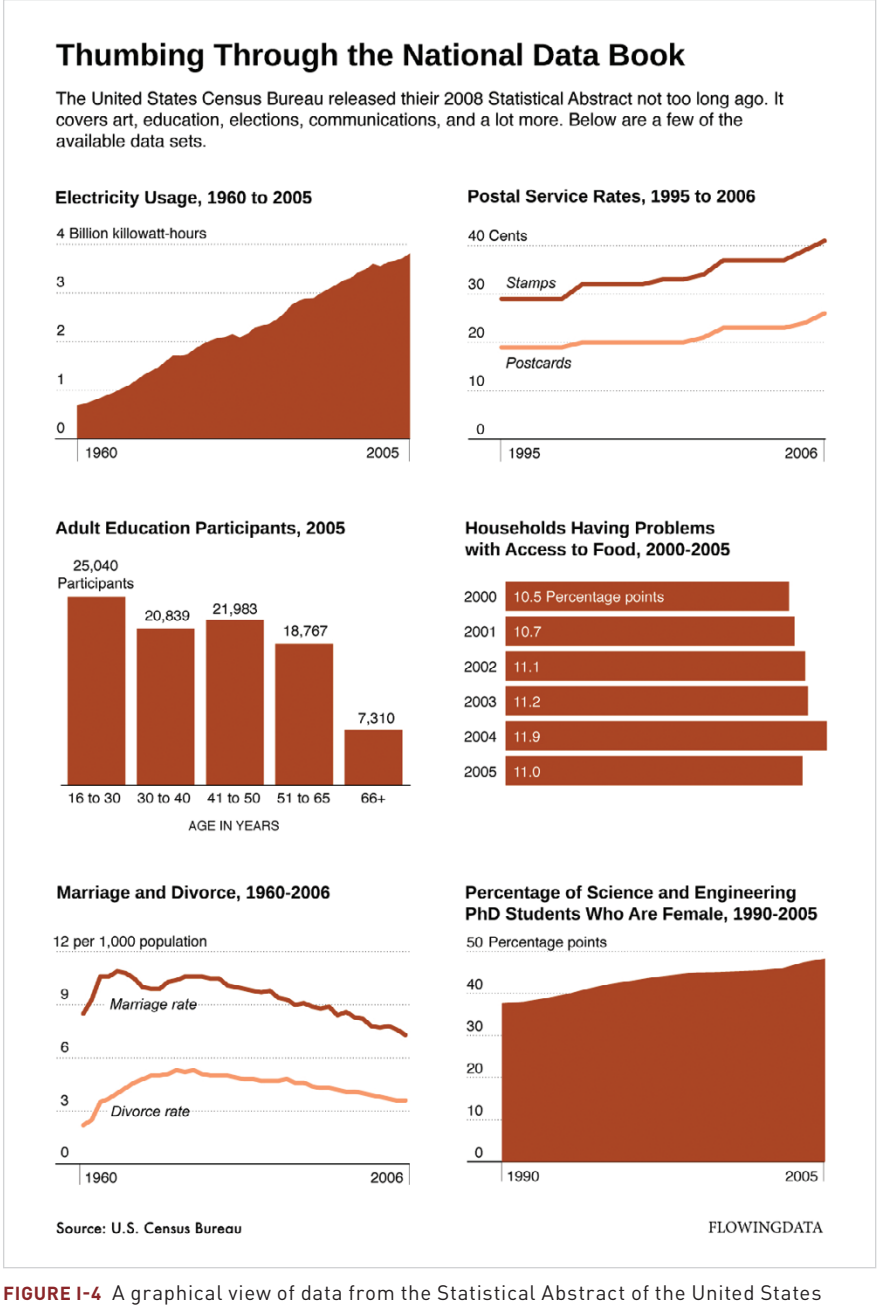


FIGURE I-4 A graphical view of data from the Statistical Abstract of the United States

At this point, you enter a realm of visualization less analytical and more about feeling. The definition of visualization starts to get kind of fuzzy. For a long time, visualization was about quantitative facts. You should recognize patterns with your tools, and they should aid your analysis in some way. Visualization isn't just about getting the cold hard facts. Like in the case of Stamen's tracker, it's almost more about the entertainment factor. It's a way for viewers to watch the awards show and interact with others in the process. Jonathan Harris' work is another great example. Harris designs his work, such as *We Feel Fine* and *Whale Hunt*, around stories rather than analytical insight, and those stories revolve around human emotion over the numbers and analytics.

Charts and graphs have also evolved into not just tools but also as vehicles to communicate ideas—and even tell jokes. Sites such as GraphJam and Indexed use Venn diagrams, pie charts, and the like to represent pop songs or show that a combination of red, black, and white equals a Communist newspaper or a panda murder. Data Underload, a data comic of sorts that I post on FlowingData, is my own take on the genre. I take everyday observations and put it in chart form. The chart in Figure I-5 shows famous movie quotes listed by the American Film Institute. It's totally ridiculous but amusing (to me, at least).

So what is visualization? Well, it depends on who you talk to. Some people say it's strictly traditional graphs and charts. Others have a more liberal view where anything that displays data is visualization, whether it is data art or a spreadsheet in Microsoft Excel. I tend to sway more toward the latter, but sometimes find myself in the former group, too. In the end, it doesn't actually matter all that much. Just make something that works for your purpose.

▶ Find more Data Underload on FlowingData at `http://datafl.ws/underload`

Whatever you decide visualization is, whether you're making charts for your presentation, analyzing a large dataset, or reporting the news with data, you're ultimately looking for truth. At some point in time, lies and statistics became almost synonymous, but it's not that the numbers lie. It's the people who use the numbers who lie. Sometimes it's on purpose to serve an agenda, but most of the time it's inadvertent. When you don't know how to create a graph properly or communicate with data in an unbiased way, false junk is likely to sprout. However, if you learn proper visualization techniques and how to work with data, you can state your points confidently and feel good about your findings.

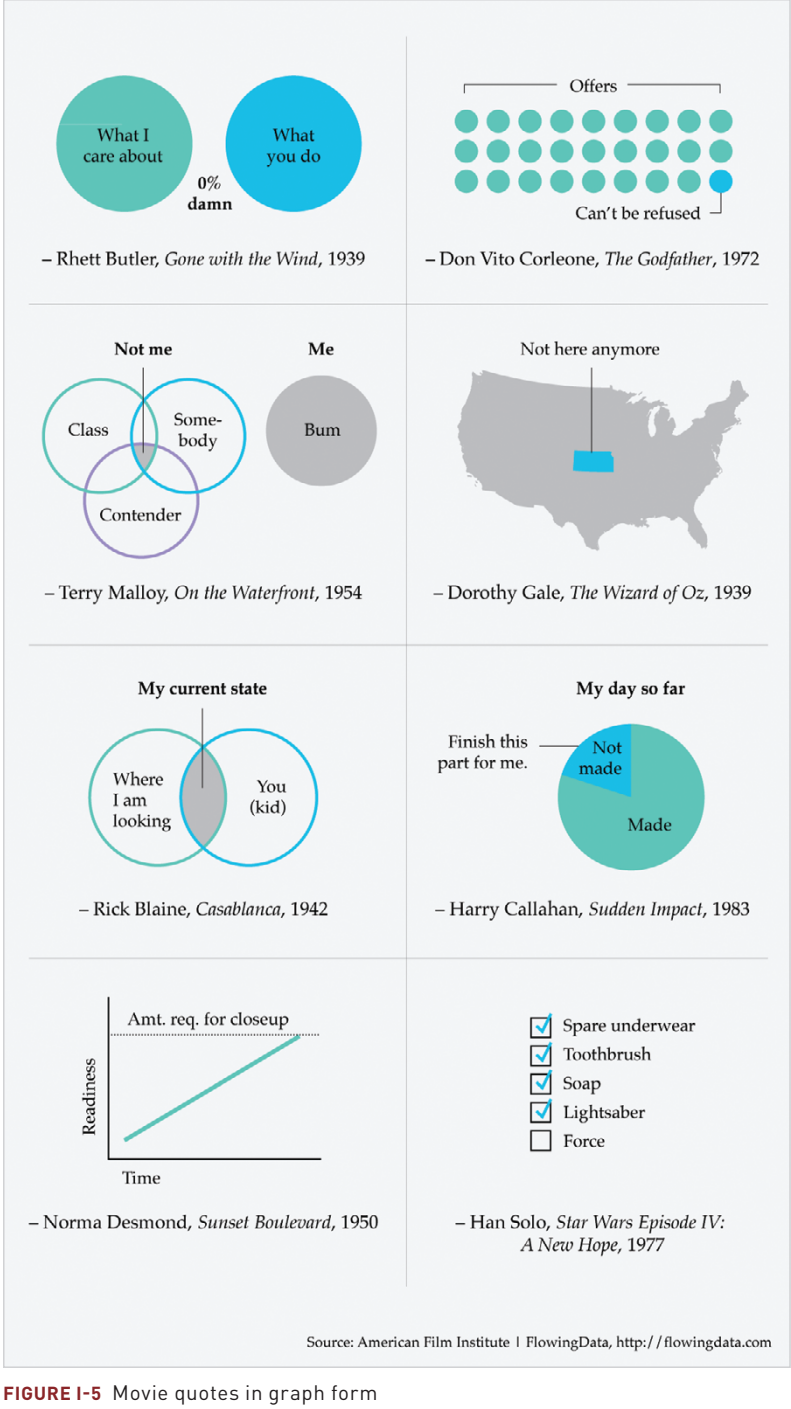


FIGURE I-5 Movie quotes in graph form

Learning Data

I got my start in statistics during my freshman year in college. It was a required introductory course toward my unrelated electrical engineering degree. Unlike some of the horror stories I've heard, my professor was extremely enthusiastic about his teaching and clearly enjoyed the topic. He quickly walked up and down the stairs of the lecture hall as he taught. He waved his hands wildly as he spoke and got students involved as he walked by. To this day, I don't think I've ever had such an excited teacher or professor, and it's undoubtedly something that drew me into the area of data and eventually what led to graduate school in statistics four years later.

Through all my undergraduate studies, statistics was data analysis, distributions, and hypothesis testing, and I enjoyed it. It was fun looking at a dataset and finding trends, patterns, and correlations. When I started graduate school though, my views changed, and things got even more interesting.

Statistics wasn't just about hypothesis testing (which turns out isn't all that useful in a lot of cases) and pattern-finding anymore. Well, no, I take that back. Statistics was still about those things, but there was a different feel to it. Statistics is about storytelling with data. You get a bunch of data, which represents the physical world, and then you analyze that data to find not just correlations, but also what's going on around you. These stories can then help you solve real-world problems, such as decreasing crime, improving healthcare, and moving traffic on the freeway, or it can simply help you stay more informed.

A lot of people don't make that connection between data and real life. I think that's why so many people tell me they "hated that course in college" when I tell them I'm in graduate school for statistics. I know you won't make that same mistake though, right? I mean, you're reading this book after all.

How do you learn the necessary skills to make use of data? You can get it through courses like I did, but you can also learn on your own through experience. That's what you do during a large portion of graduate school anyway.

It's the same way with visualization and information graphics. You don't have to be a graphic designer to make great graphics. You don't need a statistics PhD either. You just need to be eager to learn, and like almost everything in life, you have to practice to get better.

I think the first data graphics I made were in the fourth grade. They were for my science fair project. My project partner and I pondered (very deeply I am sure) what surface snails move on the fastest. We put snails on rough and smooth surfaces and timed them to see how long it took them to go a specific distance. So the data at hand was the times for different surfaces, and I made a bar graph. I can't remember if I had the insight to sort from least to greatest, but I do remember struggling with Excel. The next year though when we studied what cereal red flour beetles preferred, the graphs were a snap. After you learn the basic functionality and your way around the software, the rest is quite easy to pick up. If that isn't a great example of learning from experience, then I don't know what is. Oh, and by the way, the snails moved fastest on glass, and the red flour beetles preferred Grape Nuts, in case you were wondering.

This is basic stuff we're talking about here, but it's essentially the same process with any software or programming language you learn. If you've never written a line of code, R, many statisticians' computing environment of choice, can seem intimidating, but after you work through some examples, you start to quickly get the hang of things. This book can help you with that.

I say this because that's how I learned. I remember when I first got into more of the design aspects of visualization. It was the summer after my second year in graduate school, and I had just gotten the great news that I was going to be a graphics editor intern at *The New York Times*. Up until then, graphics had always been a tool for analysis (with the occasional science fair bar graph) to me, and aesthetics and design didn't matter so much, if at all. Data and its role in journalism didn't occur to me.

So to prepare, I read all the design books I could and went through a guide on Adobe Illustrator because I knew that's what *The New York Times* used. It wasn't until I actually started making graphics though when I truly started learning. When you learn by doing, you're forced to pick up what is necessary, and your skills evolve as you deal with more data and design more graphics.

How to Read This Book

This book is example-driven and written to give you the skills to take a graphic from start to finish. You can read it cover to cover, or you can pick your spots if you already have a dataset or visualization in mind. The chapters are organized so that the examples are self-contained. If you're new to data, the early chapters should be especially useful to you. They cover how to approach your data, what you should look for, and the tools available to you. You can see where to find data and how to format and prepare it for visualization. After that, the visualization techniques are split by data type and what type of story you're looking for. Remember, always let the data do the talking.

Whatever way you decide to read this book, I highly recommend reading the book with a computer in front of you, so that you can work through examples step-by-step and check out sources referred to in notes and references. You can also download code and data files and interact with working demos at `www.wiley.com/visualizethis` and `http://book.flowingdata.com`.

Just to make things completely clear, here's a flowchart in Figure I-6 to help you figure what spots to pick. Have fun!

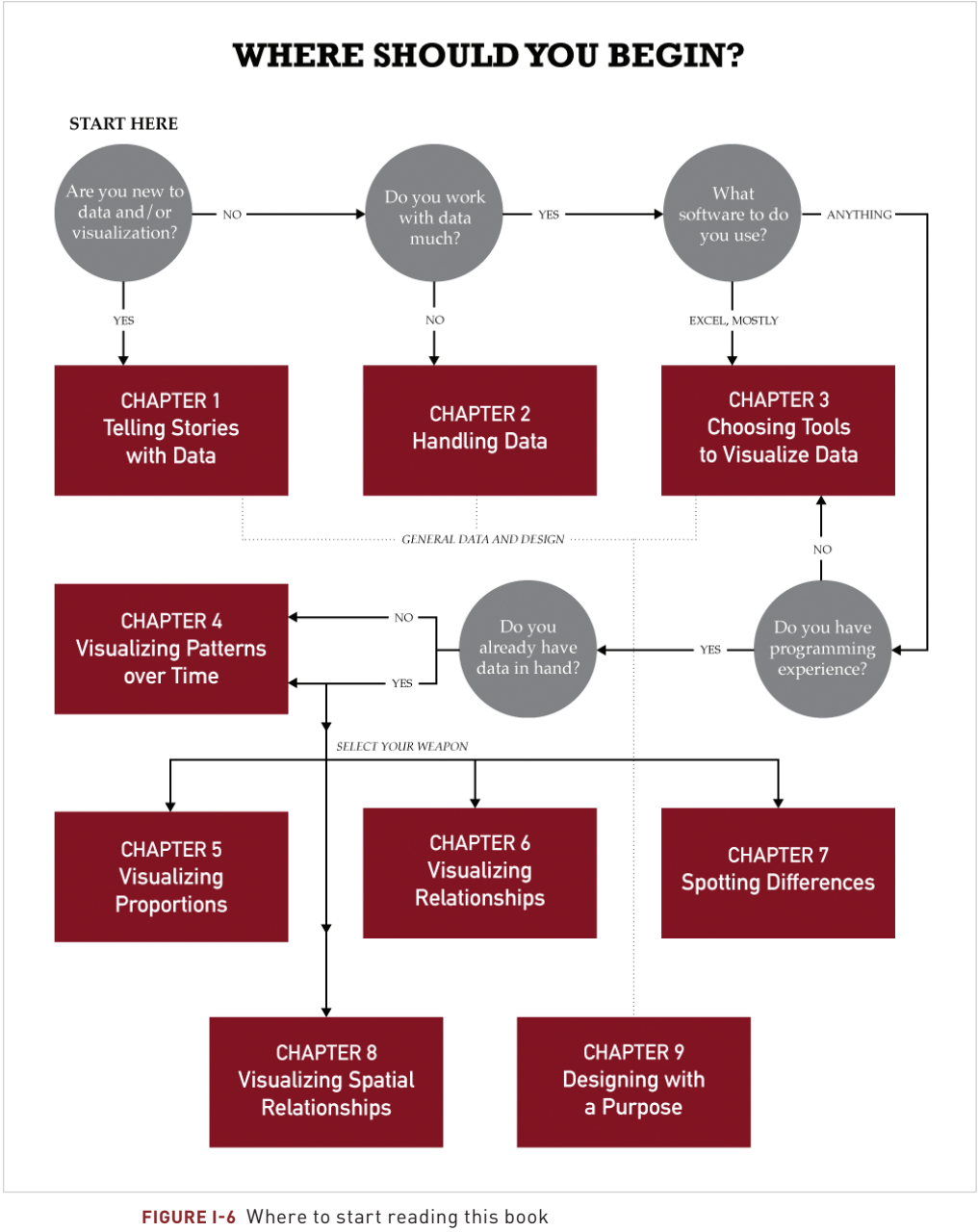


FIGURE I-6 Where to start reading this book

Telling Stories with Data

1

Think of all the popular data visualization works out there—the ones that you always hear in lectures or read about in blogs, and the ones that popped into your head as you were reading this sentence. What do they all have in common? They all tell an interesting story. Maybe the story was to convince you of something. Maybe it was to compel you to action, enlighten you with new information, or force you to question your own preconceived notions of reality. Whatever it is, the best data visualization, big or small, for art or a slide presentation, helps you see what the data have to say.

More Than Numbers

Face it. Data can be boring if you don't know what you're looking for or don't know that there's something to look for in the first place. It's just a mix of numbers and words that mean nothing other than their raw values. The great thing about statistics and visualization is that they help you look beyond that. Remember, data is a representation of real life. It's not just a bucket of numbers. There are stories in that bucket. There's meaning, truth, and beauty. And just like real life, sometimes the stories are simple and straightforward; and other times they're complex and roundabout. Some stories belong in a textbook. Others come in novel form. It's up to you, the statistician, programmer, designer, or data scientist to decide how to tell the story.

This was one of the first things I learned as a statistics graduate student. I have to admit that before entering the program, I thought of statistics as pure analysis, and I thought of data as the output of a mechanical process. This is actually the case a lot of the time. I mean, I did major in electrical engineering, so it's not all that surprising I saw data in that light.

Don't get me wrong. That's not necessarily a bad thing, but what I've learned over the years is that data, while objective, often has a human dimension to it.

For example, look at unemployment again. It's easy to spout state averages, but as you've seen, it can vary a lot within the state. It can vary a lot by neighborhood. Probably someone you know lost a job over the past few years, and as the saying goes, they're not just another statistic, right? The numbers represent individuals, so you should approach the data in that way. You don't have to tell every individual's story. However, there's a subtle yet important difference between the unemployment rate increasing by 5 percentage points and several hundred thousand people left jobless. The former reads as a number without much context, whereas the latter is more relatable.

Journalism

A graphics internship at *The New York Times* drove the point home for me. It was only for 3 months during the summer after my second year of graduate school, but it's had a lasting impact on how I approach data. I didn't just learn how to create graphics for the news. I learned how to report

data as the news, and with that came a lot of design, organization, fact checking, sleuthing, and research.

There was one day when my only goal was to verify three numbers in a dataset, because when *The New York Times* graphics desk creates a graphic, it makes sure what it reports is accurate. Only after we knew the data was reliable did we move on to the presentation. It's this attention to detail that makes its graphics so good.

Take a look at any *New York Times* graphic. It presents the data clearly, concisely, and ever so nicely. What does that mean though? When you look at a graphic, you get the chance to understand the data. Important points or areas are annotated; symbols and colors are carefully explained in a legend or with points; and the *Times* makes it easy for readers to see the story in the data. It's not just a graph. It's a graphic.

The graphic in Figure 1-1 is similar to what you will find in *The New York Times*. It shows the increasing probability that you will die within one year given your age.

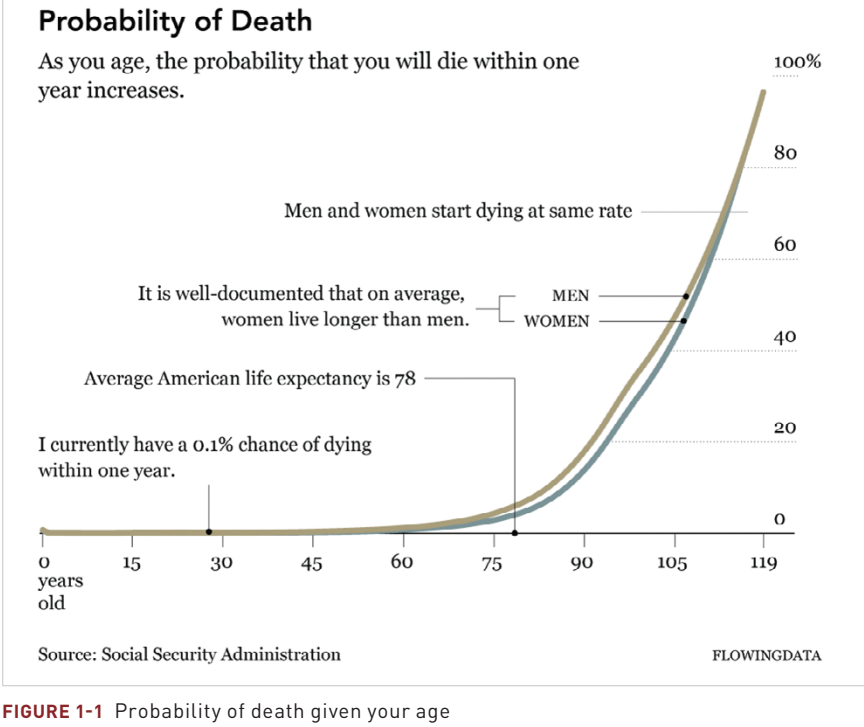


FIGURE 1-1 Probability of death given your age

▶ Check out some of the best *New York Times* graphics at `http://datafl.ws/nytimes`.

The base of the graphic is simply a line chart. However, design elements help tell the story better. Labeling and pointers provide context and help you see why the data is interesting; and line width and color direct your eyes to what's important.

Chart and graph design isn't just about making statistical visualization but also explaining what the visualization shows.

NOTE

See Geoff McGhee's video documentary "Journalism in the Age of Data" for more on how journalists use data to report current events. This includes great interviews with some of the best in the business.

Art

The New York Times is objective. It presents the data and gives you the facts. It does a great job at that. On the opposite side of the spectrum, visualization is less about analytics and more about tapping into your emotions. Jonathan Harris and Sep Kamvar did this quite literally in *We Feel Fine* (Figure 1-2).

FIGURE 1-2 We Feel Fine by Jonathan Harris and Sep Kamvar

The interactive piece scrapes sentences and phrases from personal public blogs and then visualizes them as a box of floating bubbles. Each bubble represents an emotion and is color-coded accordingly. As a whole, it is like individuals floating through space, but watch a little longer and you see bubbles start to cluster. Apply sorts and categorization through the interface to see how these seemingly random vignettes connect. Click an individual bubble to see a single story. It's poetic and revealing at the same time.

There are lots of other examples such as Golan Levin's *The Dumpster*, which explores blog entries that mention breaking up with a significant other; Kim Asendorf's *Sumedicina*, which tells a fictional story of a man running from a corrupt organization, with not words, but graphs and charts; or Andreas Nicolas Fischer's physical sculptures that show economic downturn in the United States.

The main point is that data and visualization don't always have to be just about the cold, hard facts. Sometimes you're not looking for analytical insight. Rather, sometimes you can tell the story from an emotional point of view that encourages viewers to reflect on the data. Think of it like this. Not all movies have to be documentaries, and not all visualization has to be traditional charts and graphs.

▶ Interact and explore people's emotions in Jonathan Harris and Sep Kamvar's live and online piece at `http://wefeelfine.org`.

▶ See FlowingData for many more examples of art and data at `http://datafl.ws/art`.

Entertainment

Somewhere in between journalism and art, visualization has also found its way into entertainment. If you think of data in the more abstract sense, outside of spreadsheets and comma-delimited text files, where photos and status updates also qualify, this is easy to see.

Facebook used status updates to gauge the happiest day of the year, and online dating site OkCupid used online information to estimate the lies people tell to make their digital selves look better, as shown in Figure 1-3. These analyses had little to do with improving a business, increasing revenues, or finding glitches in a system. They circulated the web like wildfire because of their entertainment value. The data revealed a little bit about ourselves and society.

Facebook found the happiest day to be Thanksgiving, and OkCupid found that people tend to exaggerate their height by about 2 inches.

▶ Check out the OkTrends blog for more revelations from online dating such as what white people really like and how not to be ugly by accident: `http://blog.okcupid.com`.

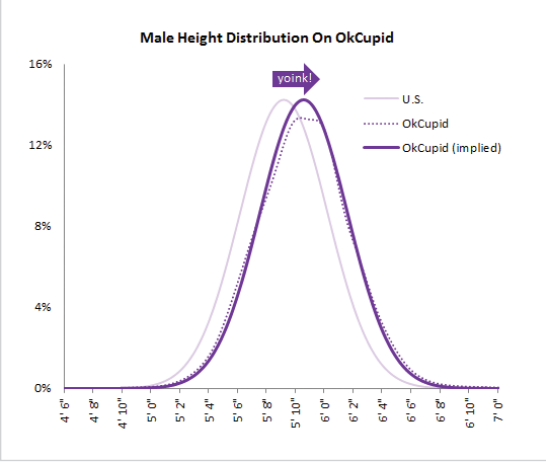


FIGURE 1-3 Male Height Distribution on OkCupid

Compelling

Of course, stories aren't always to keep people informed or entertained. Sometimes they're meant to provide urgency or compel people to action. Who can forget that point in *An Inconvenient Truth* when Al Gore stands on that scissor lift to show rising levels of carbon dioxide?

For my money though, no one has done this better than Hans Rosling, professor of International Health and director of the Gapminder Foundation. Using a tool called Trendalyzer, as shown in Figure 1-4, Rosling runs an animation that shows changes in poverty by country. He does this during a talk that first draws you in deep to the data and by the end, everyone is on their feet applauding. It's an amazing talk, so if you haven't seen it yet, I highly recommend it.

The visualization itself is fairly basic. It's a motion chart. Bubbles represent countries and move based on the corresponding country's poverty during a given year. Why is the talk so popular then? Because Rosling speaks with conviction and excitement. He tells a story. How often have you seen a presentation with charts and graphs that put everyone to sleep? Instead Rosling gets the meaning of the data and uses that to his advantage. Plus, the sword-swallowing at the end of his talk drives the point home. After I saw Rosling's talk, I wanted to get my hands on that data and take a look myself. It was a story I wanted to explore, too.

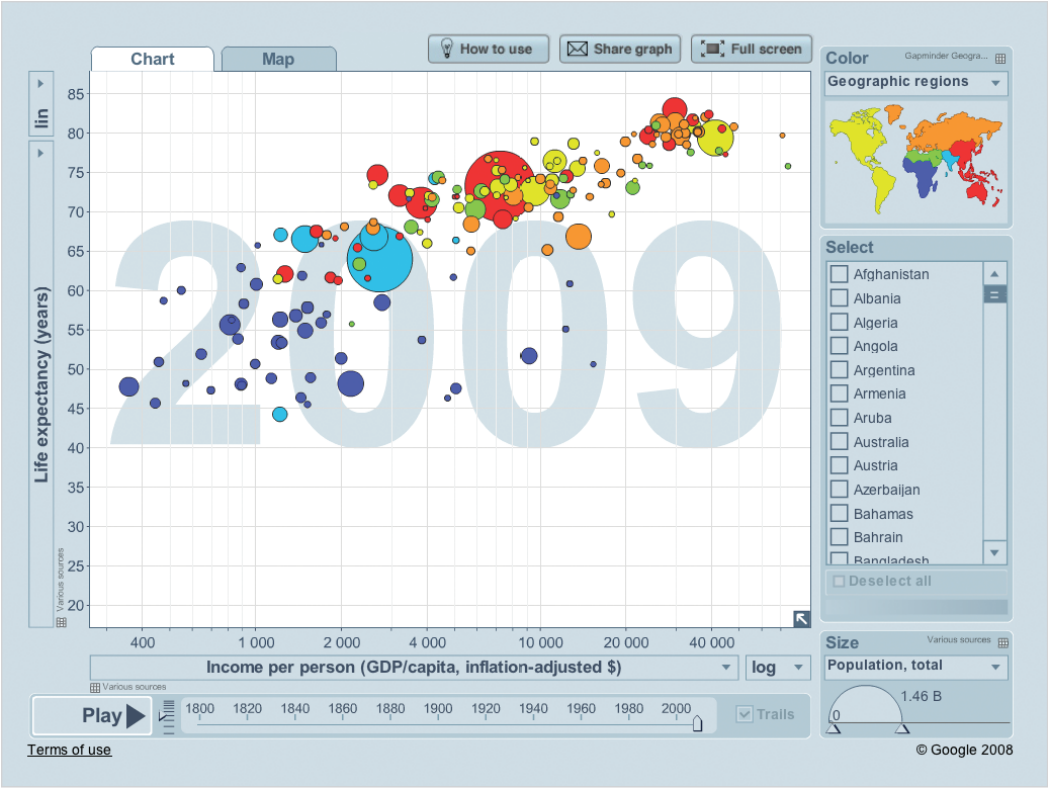


FIGURE 1-4 Trendalyzer by the Gapminder Foundation

I later saw a Gapminder talk on the same topic with the same visualizations but with a different speaker. It wasn't nearly as exciting. To be honest, it was kind of a snoozer. There wasn't any emotion. I didn't feel any conviction or excitement about the data. So it's not just about the data that makes for interesting chatter. It's how you present it and design it that can help people remember.

When it's all said and done, here's what you need to know. Approach visualization as if you were telling a story. What kind of story are you trying to tell? Is it a report, or is it a novel? Do you want to convince people that action is necessary?

Think character development. Every data point has a story behind it in the same way that every character in a book has a past, present, and future. There are interactions and relationships between those data points. It's up

▶ Watch Hans Rosling wow the audience with data and an amazing demonstration at `http://datafl.ws/hans`.

to you to find them. Of course, before expert storytellers write novels, they must first learn to construct sentences.

What to Look For

Okay, stories. Check. Now what kind of stories do you tell with data? Well, the specifics vary by dataset, but generally speaking, you should always be on the lookout for these two things whatever your graphic is for: patterns and relationships.

Patterns

Stuff changes as time goes by. You get older, your hair grays, and your sight starts to get kind of fuzzy (Figure 1-5). Prices change. Logos change. Businesses are born. Businesses die. Sometimes these changes are sudden and without warning. Other times the change happens so slowly you don't even notice.

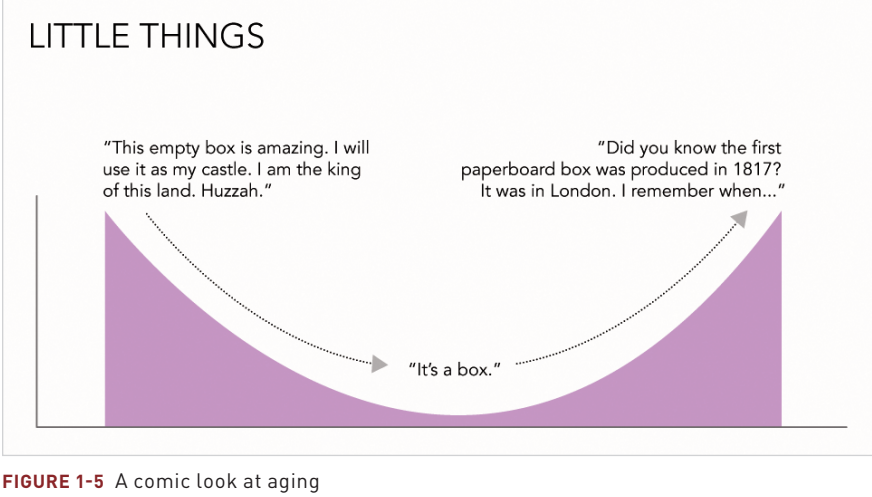


FIGURE 1-5 A comic look at aging

Whatever it is you're looking at, the change itself can be interesting as can the changing process. It is here you can explore patterns over time. For example, say you looked at stock prices over time. They of course increase

and decrease, but by how much do they change per day? Per week? Per month? Are there periods when the stock went up more than usual? If so, why did it go up? Were there any specific events that triggered the change?

As you can see, when you start with a single question as a starting point, it can lead you to additional questions. This isn't just for time series data, but with all types of data. Try to approach your data in a more exploratory fashion, and you'll most likely end up with more interesting answers.

You can split your time series data in different ways. In some cases it makes sense to show hourly or daily values. Other times, it could be better to see that data on a monthly or annual basis. When you go with the former, your time series plot could show more noise, whereas the latter is more of an aggregate view.

Those with websites and some analytics software in place can identify with this quickly. When you look at traffic to your site on a daily basis, as shown in Figure 1-6, the graph is bumpier. There are a lot more fluctuations.

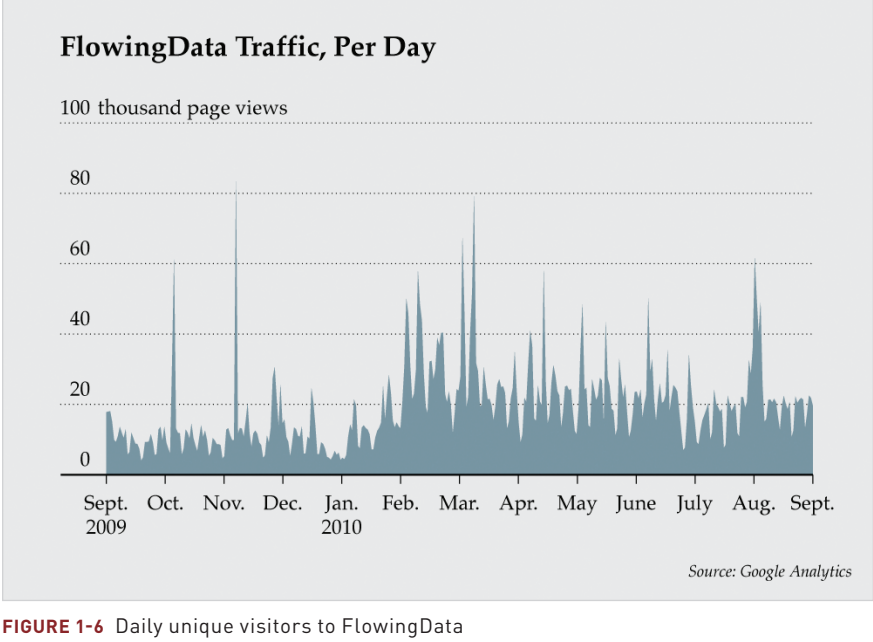


FIGURE 1-6 Daily unique visitors to FlowingData

When you look at it on a monthly basis, as shown in Figure 1-7, fewer data points are on the same graph, covering the same time span, so it looks much smoother.

I'm not saying one graph is better than the other. In fact, they can complement each other. How you split your data depends on how much detail you need (or don't need).

Of course, patterns over time are not the only ones to look for. You can also find patterns in aggregates that can help you compare groups, people, and things. What do you tend to eat or drink each week? What does the President usually talk about during the State of the Union address? What states usually vote Republican? Looking at patterns over geographic regions would be useful in this case. While the questions and data types are different, your approach is similar, as you'll see in the following chapters.

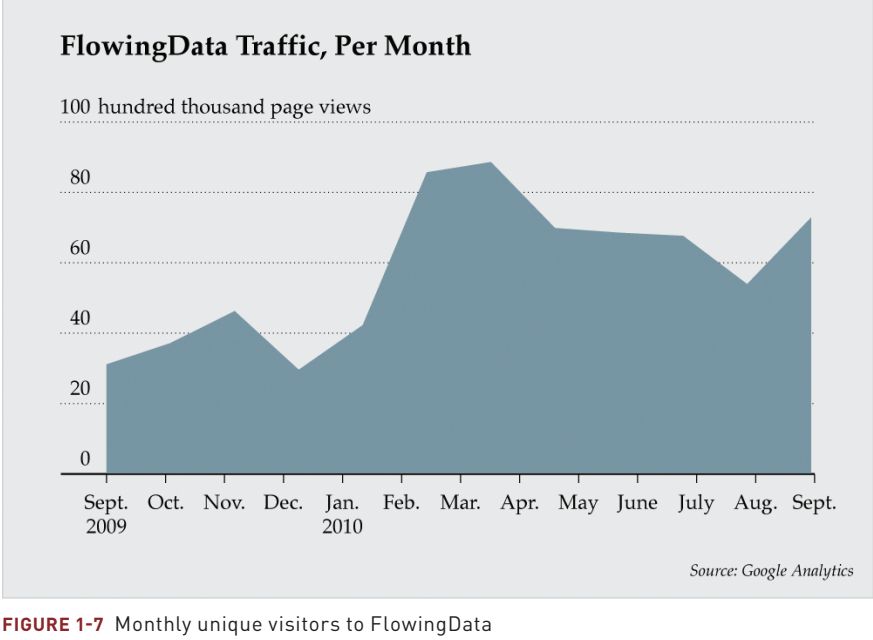


FIGURE 1-7 Monthly unique visitors to FlowingData