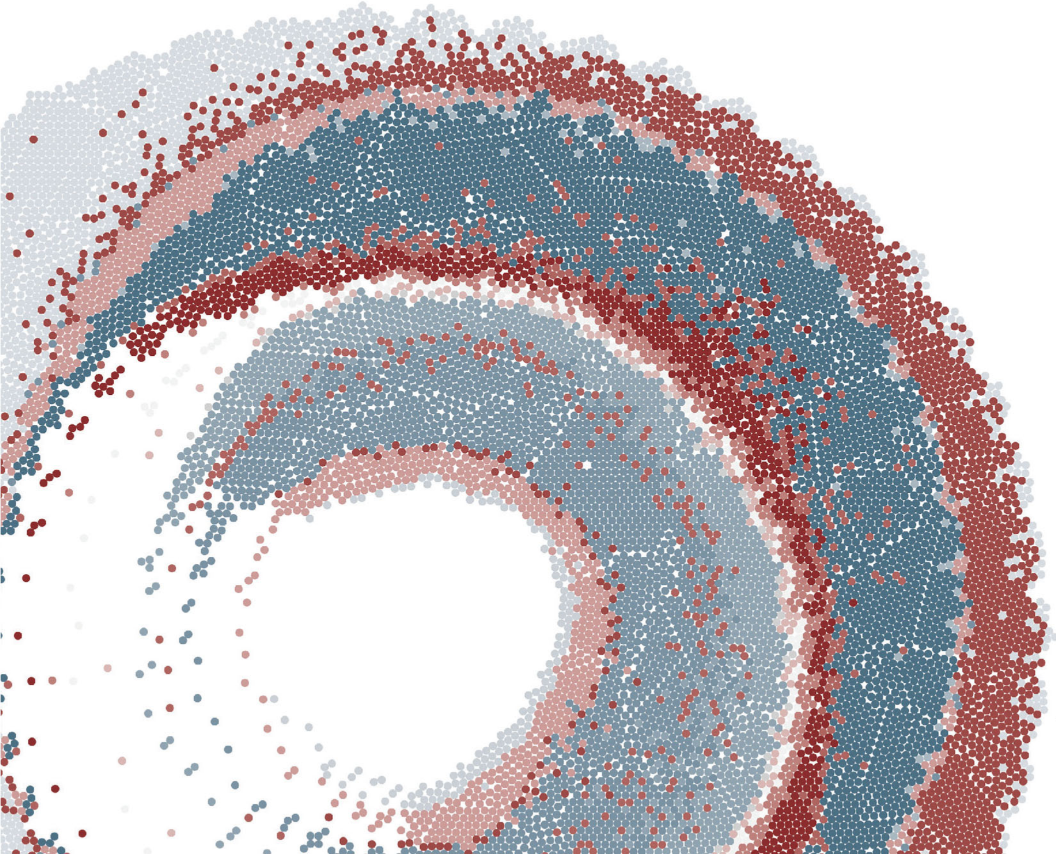
SECOND EDITION

VISUALIZE THIS

The FlowingData Guide to Design, Visualization, and Statistics

NATHAN YAU



WILEY

Visualize This

Second Edition

Visualize This

The FlowingData Guide to Design, Visualization, and Statistics

Second Edition

Nathan Yau

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada and the United Kingdom.

ISBNs: 9781394214860 (Paperback), 9781394214884 (ePDF), 9781394214877 (ePUB)

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

If you believe you've found a mistake in this book, please bring it to our attention by emailing our reader support team at wileysupport@wiley.com with the subject line "Possible Book Errata Submission."

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Control Number: 2024935194

Cover image: © Nathan Yau
Cover design: Wiley

To Bea, Caleb, and Audrey

About the Author

Nathan Yau has a PhD in statistics from the University of California at Los Angeles, with a focus on visualization for presenting and communicating data to everyone. He was the winner of a FastCompany Innovation by Design Award for Graphic Design & Data Visualization, he has won Information is Beautiful awards, and he was featured in *The Best American Infographics*. He has worked as a researcher and for mainstream publications. His work leans toward practical and has reached millions of people. Since 2007, Yau has written, analyzed, and made graphics for FlowingData, his site on visualization, statistics, and design. Yau's goal is to help people understand data, and he believes visualization—from statistical charts to information graphics to data art—is the best way to get there.

About the Technical Editor

Jan Willem Tulp (TULP interactive) is an award-winning data experience designer from The Netherlands. As an independent data visualization designer with more than 12 years of experience, he has worked for a wide variety of clients, such as World Bank, *Scientific American*, Google News Lab, European Space Agency, the Dutch Railways (NS), and Nielsen. Tulp speaks regularly at international conferences, such as Open VisConf, IEEE VIS in Practice, Visualized, Indo Data Week, and OutlierConf. His work has been published in a number of books and magazines, including *The Functional Art*, *Design for Information*, and *The Book of Circles*. The nature of his projects ranges from interactive exploratory tools to data-driven storytelling to experimental visualizations that push the boundaries. Occasionally, his work is shown in exhibitions. One of his current projects is in the permanent collection for 3 years of Ars Electronica.

Acknowledgments

This book would not have been possible without the work of statisticians, data scientists, cartographers, analysts, and designers before me, who developed and continue to create useful tools for everyone. If you are one of these people, I thank you.

Many thanks to the FlowingData readers who helped me, an introvert who likes to think about data, reach more people than I ever could have imagined. They are one of the main reasons why this book was written.

Thank you to my wife for supporting me and to my parents who always encouraged me to find what makes me happy. Thank you to my kids for their perspective, which makes work more meaningful and life fuller.

Contents

Introduction

Data is everywhere, and one of the best ways to explore a dataset is with visualization. Place the numbers into a visual space and let your brain find the patterns. We're good at that. Discover insights that you wouldn't see in a spreadsheet alone. From here, you can use visualization to communicate to others, from an audience of one to millions.

For a long while, visualization was more of a quantitative and technical exercise. Show data, get out of the way, and let the data speak. This approach works sometimes, but it assumes that data speaks a language that everyone understands and that it always speaks definitively and in absolutes. However, data is not always so straightforward, and the insights are often not so certain.

Over the past 17 years of writing for FlowingData, a site on visualization, statistics, and design, I've seen an evolution. Visualization was mostly an analysis tool when I started my studies but it has developed into a medium to tell stories with data. You can show just the facts, but you can also evoke emotion, entertain, and compel change.

In my own work, visualization is a way to understand data, share what I find, and, most importantly, make sense of what's going on around me. I follow an iterative process of answering questions with data, visualizing the answers, and then asking more questions. Repeat until there are no more questions. While the general analysis and visualization process remained about the same since the first edition of this book, the steps to carry out the process were refined and the tools shifted, varying by the year you asked me.

This is the second edition of *Visualize This*. When I wrote the first edition more than a decade ago, visualization in practice was in a different place. The tools were different (like Flash), people tended to follow stricter design guidance (like ratios between data and ink), the purpose behind visualization was narrower (such as analysis and quantitative insights only), and organizations were still figuring out what data to make public (which feels less open at times these days).

As a reflection of my own evolving process, this edition provides all new examples, explanations, and guidance, with a focus on making charts to communicate. This is how to visualize data from my point of view, and it isn't the only way to do things. For me, it's what works best. My hope is that after working through this book, you'll know the mechanics of chart-making, be able to refine your process to fit your specific needs and form your own opinions about what makes visualization great.

LEARNING DATA VISUALIZATION

I got my start in statistics during my first year in college. It was a required introductory course toward my electrical engineering degree. The professor was refreshingly enthusiastic about his teaching and clearly enjoyed the topic. He quickly walked, nearly running, 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 whizzed by. His excitement drew me into studying data and eventually led to graduate school, studying statistics four years later.

Throughout my undergraduate studies, statistics was procedural data analysis, distributions, and hypothesis testing. I enjoyed it. It was fun to look at a new dataset to find trends, patterns, and correlations. When I started graduate school, though, my field of vision widened, and things got more interesting. My appreciation for statistics grew.

Statistics became less about hypothesis testing, bell curves, and coin flips. It became more about telling stories with data. You get a bunch of data, which represents the real world, and then you analyze and interpret that data to make sense of what's going on around you. These stories can inform public policy, business, technology, health, happiness, and everyday life.

The ubiquity of data means the process of communicating data comes in handy in many places. However, a lot of people don't have the time or know how to connect data to real life. You can be the bridge between abstract numbers and insight.

How do you learn the necessary skills to visualize data usefully? These days, there are courses and degrees you can earn in data visualization, but you can also learn through practice without a dedicated degree. I've never taken a visualization course.

The first charts I made from scratch 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 crawl a specific distance. So, the data was clocked times for different surfaces, and I made a bar chart. I can't remember if I had the insight to sort from least to greatest, but I do remember struggling with Microsoft Excel. Charts were easier after that, though. Once you learn the basic functionality and your way around the software, the rest is easier to learn. (By the way, the snails moved fastest on glass, in case you were wondering.)

It's the same process with any software or programming language you learn. As my career extended beyond the fourth-grade science fair project on snails, I learned how to visualize data as I went. I learned R to analyze data in school and more so later for work. I joined a research group using Python for data collection and PHP for web applications, so I learned those languages to not be totally useless. I wanted to make interactive and animated graphics for the web, so I learned Flash, and when Flash died, I learned JavaScript. To prepare for a graphics internship, I studied all the data design books I could, but it wasn't until I struggled making graphics with Adobe Illustrator when I figured out how to make charts for a general audience.

If you've never written a line of code or used a hefty software package, the process can seem intimidating, but after you work through some examples, you start to get the hang of things. This book can help you with that.

HOW TO USE THIS BOOK

This book is example-driven, with practical steps for how to use a mix of visualization tools and understand different types of data. With each example, you start with a dataset and work through the process of asking questions, learning about data, and communicating insights to a wider audience.

Each chapter includes data, code, and files you can download. Download everything at `www.wiley.com/go/visualizethis2e` or `https://book.flowingdata.com/vt2`. The files will make it easier to work through examples step-by-step, poke at the data if you are curious, and apply what you learn to other datasets.

You can read this book cover to cover or pick your spots if you already have a dataset or visualization in mind. The chapters are organized by data type and what you want to visualize. The sections within each chapter discuss what to look for in your data and the chart types that can help you and others see relevant patterns.

By the end, you should be able to visualize your own data and design publication-ready graphics. Have fun in the process.

Ch.1

Telling Stories with Data

Think of the data visualization works that you enjoy—the ones that you see online, that appear in lectures, and that you associate with quality. Most likely the works that popped into your head 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 assumptions. Maybe it made you smile. Whatever it is, the best data visualization, big or small, for art or a slide presentation, shows patterns that you could not see otherwise.

MORE THAN NUMBERS

My interest in visualization began as a new statistics student ready to analyze all the datasets. Charts were a tool I could use to understand data better, and I would occasionally export an image to stick in a report. That was about it.

I approached chart-making from a technical point of view, without giving much thought to what type of chart worked best, who was going to look at my work, or how to design around insight and story. I just needed to figure out how to make a chart so that I could move on to the rest of my analysis.

However, the more I worked with data, the more I learned about its complexity, subjectivity, and how it related to the rest of the world. At the same time, we were interacting with data more through computers, phones, and connected devices. Data intertwined with the everyday instead of with just a spreadsheet that analysts opened at work, and I grew interested in how data would play a role in understanding ourselves better.

A couple of years into graduate school, a graphics internship at a major news publication got me thinking about visualization's role in the presentation and communication of data. How did it differ from visualization for exploratory data analysis? Then with FlowingData, I suddenly got a taste of what it was like for a visualization project to communicate data to millions of people. I felt like I was onto something, so I kept going. I was hooked. What started as a side project to keep in touch with classmates became my full-time dream job.

See the classic *Exploratory Data Analysis* by John Tukey (Pearson, 1977), which introduced a novel idea at the time to use visualization to study data.

Over the years, visualization matured beyond just an analysis tool. It became a way to communicate data to nonprofessionals. It could be fun. Visualization grew into a medium to tell stories with data, and like any good medium, it lets you tell different types of stories.

STATISTICALLY INFORMATIVE

Statistical stories probably come to mind for most people when it comes to data and visualization. In a journalistic context, the stories often follow a familiar article format with charts coupled with narrative. The charts show the data, and the narrative, in the form of text and annotations, describe what the data is about and provide context for the numbers. Think data projects by news organizations like the *New York Times*, the *Washington Post*, and *Reuters*.

You can also find statistical stories in a more analytical context, such as in reports, presentations, and analysis results. Maybe these aren't stories in the traditional sense, but the data you work with is about something, and that something is what makes visualization meaningful.

In 1874, the United States Census Bureau published a *Statistical Atlas of the United States*. It provided a graphical summary of the data collected for the 1870 decennial count with maps and charts. In the present day, it's like looking at a snapshot in time that shows what life was like.

Check out the *Statistical Atlas of the United States* from the 1870s at `https://datafl.ws/7l4`.

More recently, I wondered if I could use the visual forms of the original atlas to take a current snapshot. I used the most recently available data to make a revised atlas. For example, as shown in Figure 1.1, a breakdown of population by state and race was designed using the original 19th century aesthetic and wordage.

This idea of data snapshots that we can look at centuries from now drives most of my work. How do things look now, and how will things look 100 years from now? What do these snapshots look like for individuals using the data they collect (actively and passively) through their phones and devices?

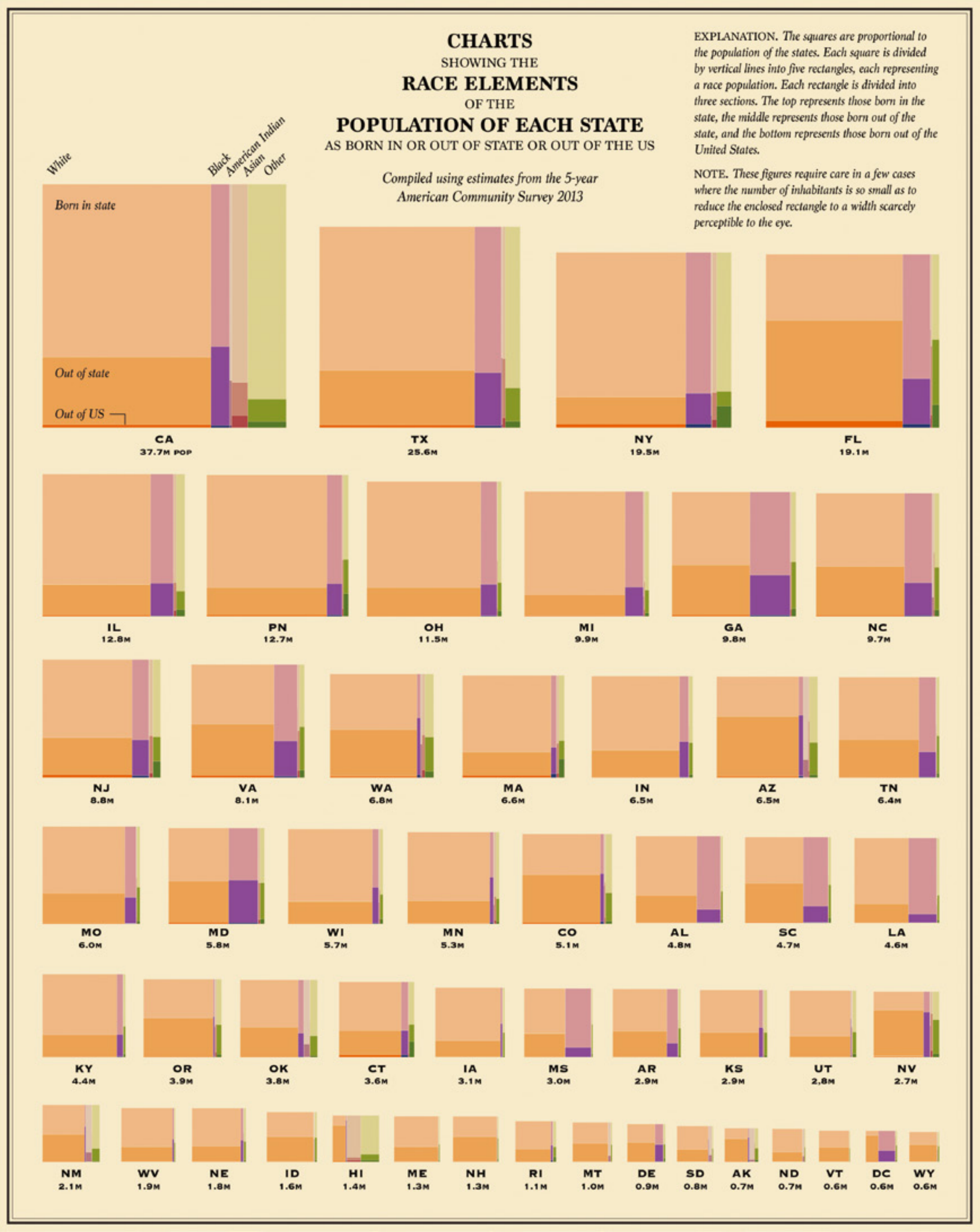
We can use data for insight, and the insight coupled with context gives us stories. This helps people make better informed decisions in both work and everyday life.

ENTERTAINING

Statistics. People would ask me what I studied, and either their eyes would glaze over in disinterest, or they would groan about the introductory statistics course they hated in college. They remembered bell curves and hypothesis tests something or other. Occasionally, someone would feign interest, and while I appreciated the effort, I knew better.

Data can be boring if you don't know how to interpret it. It might as well be gibberish. The fun thing about visualization is that people can see patterns in pictures that are more difficult to understand through equations and text.

Over the years, more people grew to appreciate data, and charts grew into a form of entertainment. People tell jokes with charts, draw comics, explore fun curiosities, and create social media-based businesses under the premise of infotainment.



CHARTS SHOWING THE RACE ELEMENTS OF THE POPULATION OF EACH STATE AS BORN IN OR OUT OF STATE OR OUT OF THE US

Compiled using estimates from the 5-year American Community Survey 2013

EXPLANATION. *The squares are proportional to the population of the states. Each square is divided by vertical lines into five rectangles, each representing a race population. Each rectangle is divided into three sections. The top represents those born in the state, the middle represents those born out of the state, and the bottom represents those born out of the United States.*

NOTE. *These figures require care in a few cases where the number of inhabitants is so small as to reduce the enclosed rectangle to a width scarcely perceptible to the eye.*

FIGURE 1.1 *"Revised Statistical Atlas of the United States,"* Nathan Yau / 2007-Present FlowingData / `https://flowingdata.com/2015/06/16/reviving-the-statistical-atlas-of-the-united-states-with-new-data` / *last accessed February 08, 2024.*

A lot of the projects I publish on FlowingData are for my own entertainment, such as the one in Figure 1.2. A question pops up, and I try to answer it with data. But if I'm interested in something, at least one other person must be curious, too. I think that's one of the foundations of the Internet.

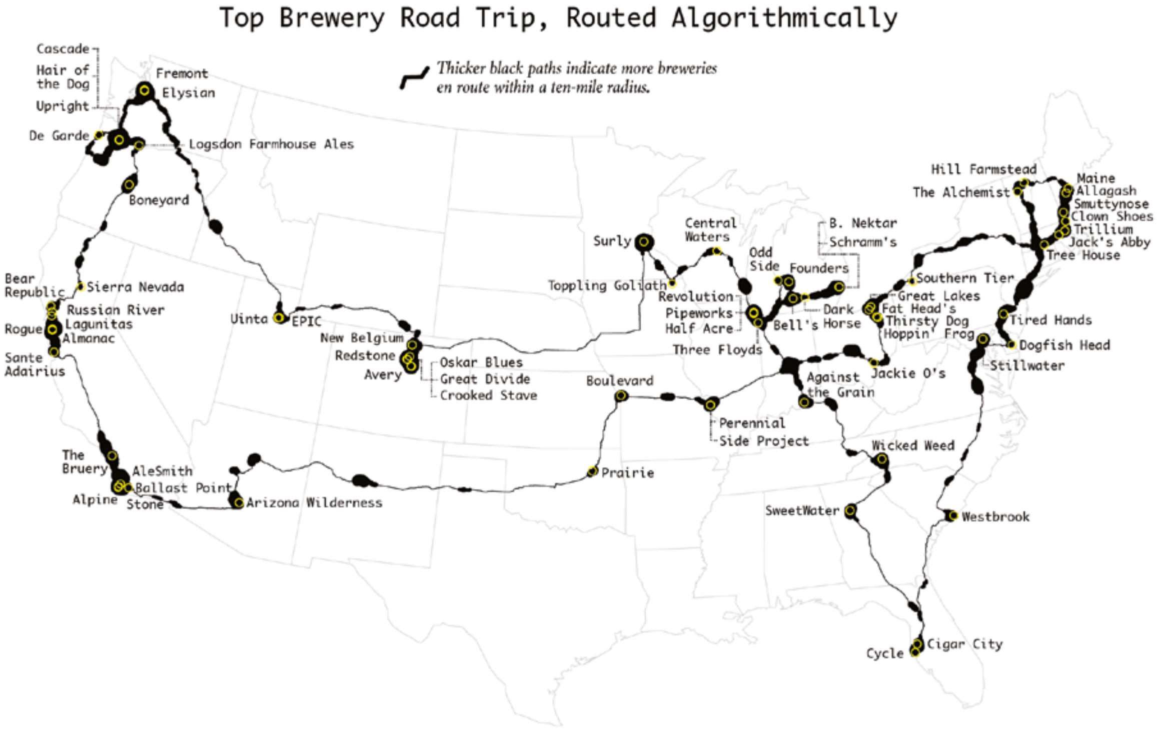


FIGURE 1.2 *"Top Brewery Road Trip, Routed Algorithmically,"* Nathan Yau / 2007-Present FlowingData / `https://flowingdata.com/2015/10/26/top-brewery-road-trip-routed-algorithmically` / *last accessed February 08, 2024.*

Every year, the beer review site RateBeer publishes the top 100 breweries based on preferences and user ratings. With a fascination for road trips and an appreciation of fine beer, I wondered what a road trip through the breweries on the list in the lower United States would look like. And, where there is one excellent brewery, there are usually others nearby, so I also wanted to know the places to stop in between the top breweries. The map shows the way to the top breweries in 2014 routed by travel times and a genetic algorithm, or a set of rules that stepped through possible solutions until it converged to an optimal route.

The visualization might not be optimized for lightning-fast decision-making, but it did seem to entertain a good number of people.

EMOTIONAL

Visualization as a field of study has a tendency toward optimized insights. This makes sense for analysis. You want to explore data quickly and efficiently so that you can evaluate from various angles.

However, if we were always after the most efficient and perceptually accurate visualization, we should just use bar charts most of the time. Or better yet, skip the visualization and just show a table for full accuracy. (I am exaggerating, but not by much.) This is not my favorite path.

Sometimes you want to visualize data in a way that reflects meaning beyond the quantitative insights. In Figure 1.3, I explored what makes people happy.

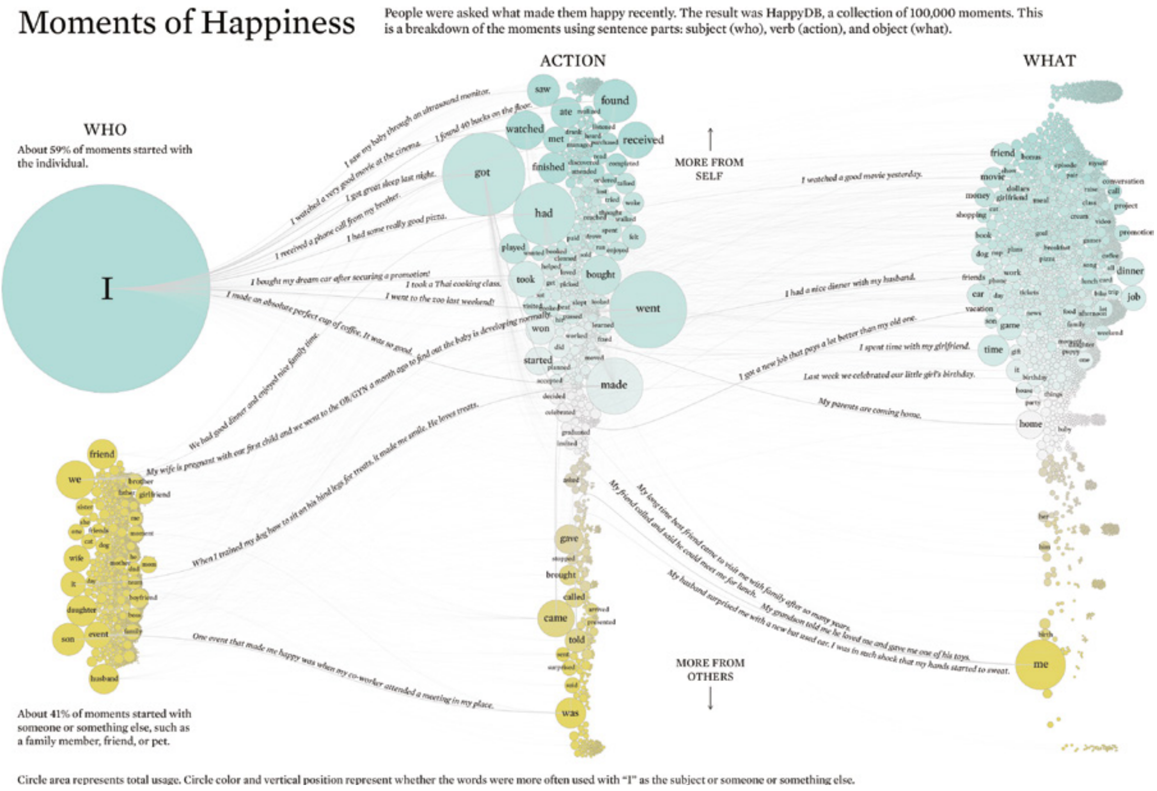


FIGURE 1.3 *"Counting Happiness and Where it Comes From,"* Nathan Yau / 2007-Present FlowingData / `https://flowingdata.com/2021/07/29/counting-happiness` / *last accessed February 08, 2024.*

Researchers asked 10,000 participants to list 10 things that recently made them happy. The result was HappyDB, a collection of 100,000 happy moments. For each moment, I parsed out the subject, verb, and object to better see what makes people happy overall. While the aggregates help you see the big picture, I was most interested in the moments and the individual words used.

See FlowingData for more examples of data art at `https://datafl.ws/art`.

How do we use visualization to feel through data? There are lots of examples such as Jonathan Harris and Sep Kamvar's *We Feel Fine* (Scribner, 2009), which examined emotions through connected vignettes; Giorgia Lupi and Stefanie Posavec's *Dear Data* (Princeton, 2016), which was a year-long data drawing project that used unique visual representations to communicate via post cards; and Stamen Design's *Atlas of Emotions* (2016), a collaboration with the Dalai Lama and Paul Ekman, which explored the range of human emotions.

While data can seem dry and concrete, it can also represent less measurable things, and visualization helps bring that aspect of data to life.

COMPELLING

It's possible for visualization to be more than one thing at a time. When a project is informative, entertaining, and emotional, it can also be compelling.

No one has done this better than the late Hans Rosling, who was a professor of international health and director of the Gapminder Foundation. Using a tool called Trendalyzer, as shown in Figure 1.4, Rosling ran an animation that showed changes in poverty by country. He did this during a talk that first draws you in to the data, and by the end, everyone is on their feet applauding. A standing ovation for data. Amazing.

The visualization itself is straightforward these days. Rosling's presentations compelled many to implement their own versions of Trendalyzer with various tools. Bubbles represent countries and move based on the corresponding country's poverty during a given year. Why is the talk so popular then? It's in Rosling's presentation style and framing. He tells a story. How often have you seen a presentation with charts and graphs that makes you drowsy? Rosling used the meaning of the data to his advantage and found a way to engage his audience. The sword-swallowing at the end of his talk tends to form a lasting impression, too.

Watch Hans Rosling wow the audience with data at `https://datafl.ws/hanstalk`.

As statistician John Tukey wrote in his 1977 book *Exploratory Data Analysis*, "The greatest value of a picture is when it forces us to notice what we never expected to see." Visualization allows you to show data with context, and framed as a story, you can help people understand concepts that are often too complex on their own.

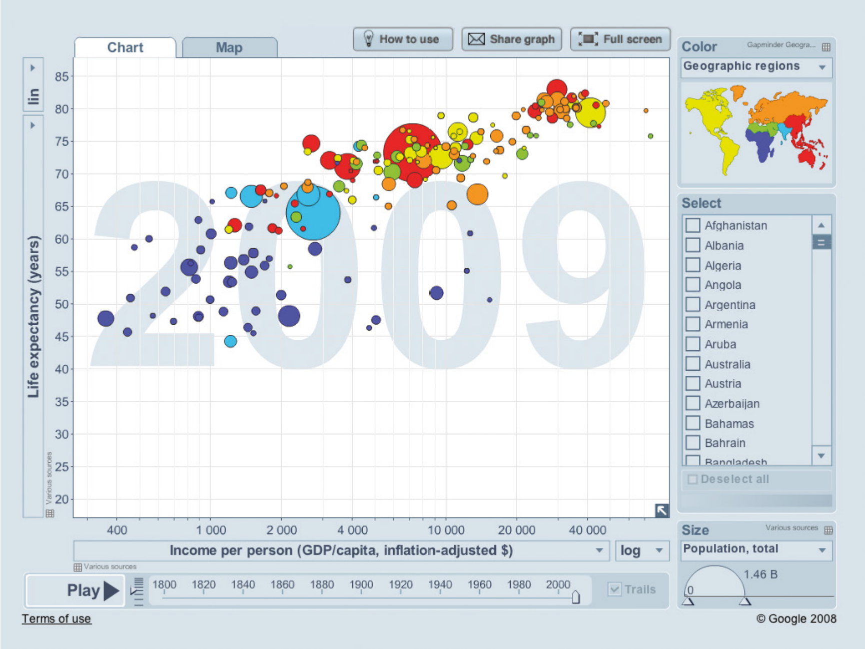


FIGURE 1.4 *Trendalyzer by the Gapminder Foundation*

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?

Every data point has a story behind it in the same way that a character in a movie has a past, present, and future. There are interactions and relationships between the data points. It's up to you to find them.

ASK QUESTIONS ABOUT THE DATA

That's the challenge. You must figure out what the data is about and what stories to tell based on what you find. Some might tell you to just let the data speak, as if you could plug a dataset into your favorite charting software and a magical, visual tale comes out. If that were the case, we could end the book here, but as of this writing, there's still more to the process.

A single dataset, even a small one with a few data points, can be visualized in many ways. Add more data variables and observations, and the possibilities for chart types, geometries, colors, formats, and dimensions multiply. To demonstrate, I visualized a single dataset, life expectancy by country over time, with 25 different charts (see Figure 1.5).

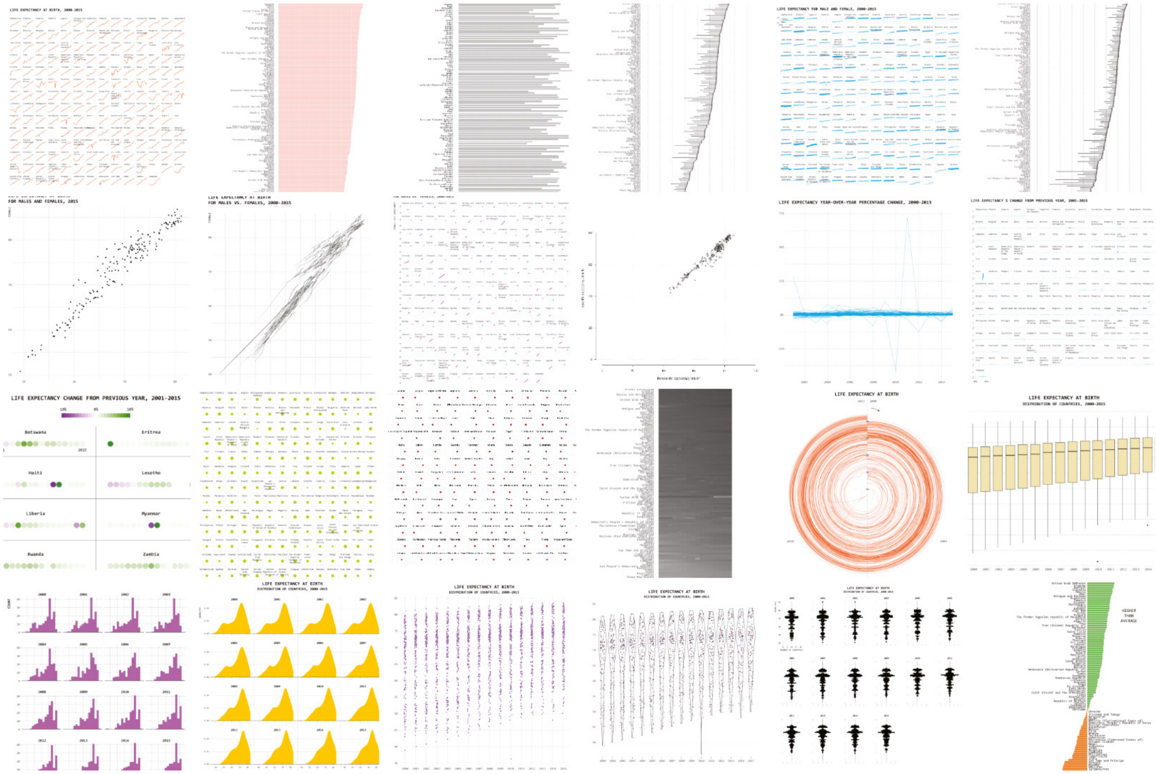


FIGURE 1.5 *"One Dataset, Visualized 25 Ways,"* FlowingData / `https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways` / *last accessed 08 February, 2024.*

I could've made a lot more charts by grouping regions, focusing on specific countries, or highlighting a range of time. I could've switched to less traditional visualization methods. I didn't even get to annotating and explaining the data.

Try it yourself. Think of all the ways to visualize two numbers, say 5 and 10. You could draw 10 circles and 5 squares, draw 10 squares and 5 circles, draw a shape that's twice the area of another shape, use a darker shade to indicate a higher number, or use a line that connects the two numbers with a defined axis.

With so much fun to be had, you need a way to filter. Find the relevant parts in the data and work through the noise. This is basically the field of statistics, which I don't have time to detail in its entirety right now, but I've found that the best way to analyze data and to provide focus to your visualizations is to ask questions about it. Use these questions and the resulting answers to verify quality, explore the meaning of the data, and communicate insight.

VERIFICATION

While you're looking for the stories in data, you should always question what you see. Remember, numbers don't always mean truth. Data is subjective in the way it's collected, who collected it, and what is collected.

In my younger days, data checking was my least favorite part of visualizing data. It seemed like a chore when I just wanted to make some charts. However, over the years, I've grown to appreciate verifying data as an important part of the visualization process. Weak data leads to mistakes and misinterpretations, whereas high confidence in your data makes it a lot easier to have high confidence in your visualization.

Basically, what you're looking for is stuff that makes no sense. Maybe there was an error at data entry and someone added an extra zero or missed one. Maybe there were connectivity issues during a data scrape and some bits got mucked up in random spots. Maybe the data was collected in haste and does not represent what you think it does.

Note: *Data scraping* is a way to automate the process of retrieving data on the Web. You'll learn how to do this in Chapter 3.

Data always has its imperfections. You need to work with them if you plan to understand the data. Here are some questions to ask early in the process:

- Does the sample represent the full population?
- Why are there so many gaps in the data, and are those gaps relevant to the existing data?
- Are the outliers errors in measurement or true standouts?
- How reliable is the data?
- Did you make an error in your calculations?
- How does the data hold up against your expectations?

This is not an exhaustive list. Some people spend their entire academic careers figuring out this stuff. But for our purposes, make sure the data is good before you waste all your time analyzing and visualizing junk.

EXPLORATION

Most of my visualization projects stem from an everyday curiosity, which I've learned to note immediately note because I have a terrible memory. They are questions like how people earn an income (`https://datafl.ws/7nz`), whether I am old or not (`https://datafl.ws/7n1`), whether it is too late for a career change (`https://datafl.ws/7o1`), or how much toilet paper I should buy at the store when I restock (`https://datafl.ws/7o0`). I try to answer these deeply profound questions with data.