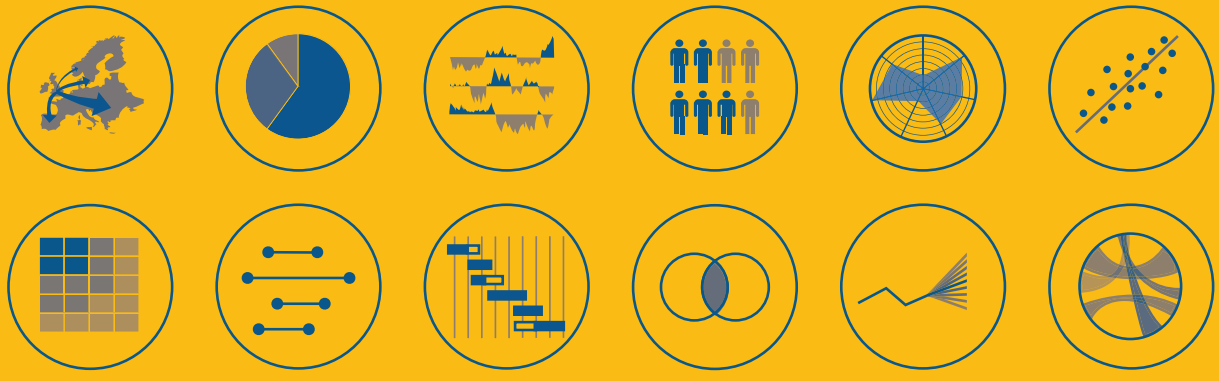
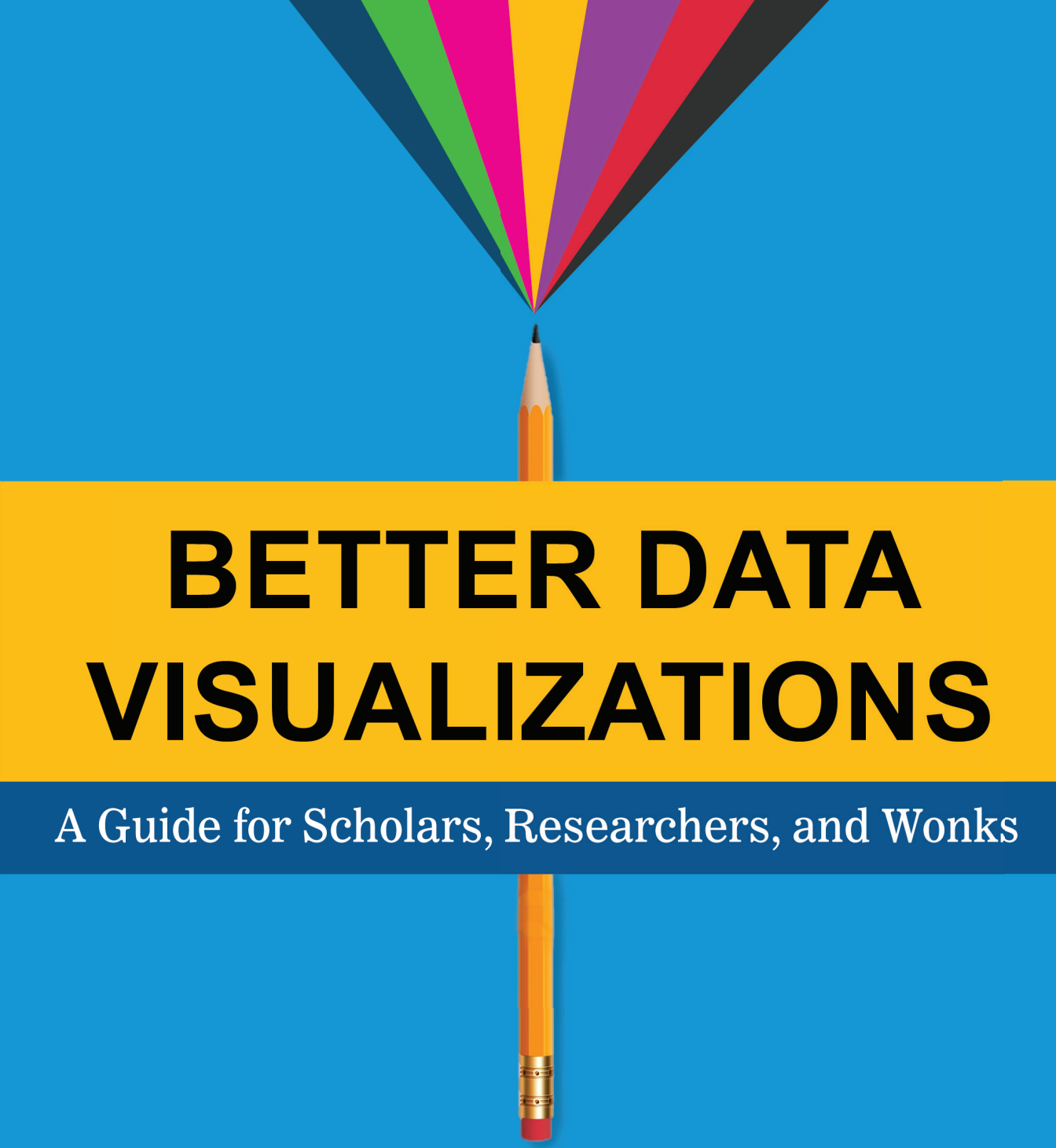
BETTER DATA VISUALIZATIONS

A Guide for Scholars, Researchers, and Wonks



Jonathan Schwabish

BETTER DATA VISUALIZATIONS



BETTER DATA VISUALIZATIONS

A Guide for Scholars, Researchers, and Wonks

Jonathan Schwabish

COLUMBIA UNIVERSITY PRESS ▸ NEW YORK

Columbia University Press
Publishers Since 1893
New York Chichester, West Sussex
cup.columbia.edu

Chapter 11, "Tables," based on Jonathan A. Schwabish, "Ten Guidelines for Better Tables," *Journal of Benefit-Cost Analysis* 11, no. 2 (2020): 151–178. Reprinted with permission.

Library of Congress Cataloging-in-Publication Data
Names: Schwabish, Jonathan A., author.
Title: Better data visualizations : a guide for scholars, researchers, and wonks / Jonathan Schwabish.
Description: New York : Columbia University Press, [2021] | Includes bibliographical references and index.
Identifiers: LCCN 2020017814 (print) | LCCN 2020017815 (ebook) | ISBN 9780231193108 (hardback) | ISBN 9780231193115 (trade paperback) | ISBN 9780231550154 (ebook)
Subjects: LCSH: Information visualization. | Visual analytics.
Classification: LCC QA76.9.I52 S393 2021 (print) | LCC QA76.9.I52 (ebook) | DDC 001.4/226—dc23
LC record available at https://lccn.loc.gov/2020017814
LC ebook record available at https://lccn.loc.gov/2020017815

Columbia University Press books are printed on permanent and durable acid-free paper.

Printed in the United States of America

For Aunt Vivi. Our Mendales. With love and Diet Coke.

CONTENTS

PART TWO: CHART TYPES

BETTER DATA VISUALIZATION

INTRODUCTION

Raise your hand if your approach to creating a graph goes something like this: You analyze some data. Write up the results. Make a graph and drop it into the report, surrounded by text. Label it something benign like "Figure 1. Average Earnings, 1990–2020." Save it as a PDF. Post it to the world.

It might have taken you months or even years to compile and analyze the data and write the report. For many, it takes far less time to design the graphs that showcase that data. You might open a program like Microsoft Excel, paste in the data, click through the drop-down menu, select one you've used dozens or hundreds of times, accept the default formatting, and paste it into the report.

But at any point in this sequence did you pause to consider what's most important about communicating the work? It's the audience. *People* will read your report. *People* will listen to you discuss your work. And yet many of us spend far too little time thinking about how we can best present our findings. Instead we use whatever default approach is quickest and easiest.

Why is this? Maybe you don't believe you have the technical skills or design know-how to create complex, attractive graphs. Or you worry it's not worth the effort, because your managers or tenure committee or whoever else won't see it as time well spent. Many people simply think that their reader will just "get it," as if everyone has seen the content a hundred times before. But many readers, especially those who can make change or implement policy, may have never seen this content before. In these cases—which are probably most of them—thinking carefully about *how* data is presented is just as important as the data itself.

This book is about how to create better, more effective visualizations of your data. It aims to expand your graphic literacy and put more graphs in your toolbox. The next time you open

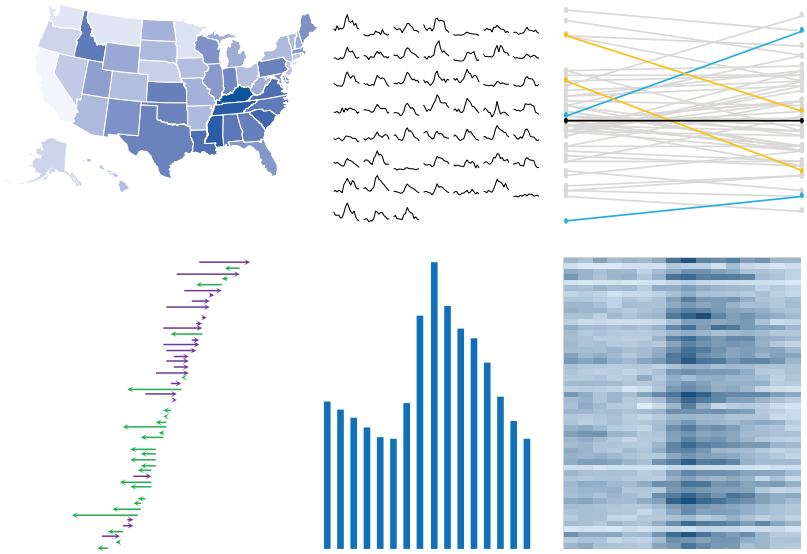
Excel, Tableau, R, or whatever your software tool of choice, you won't be bound by the graphs in the dropdown menus or the tutorial manual. This book will guide you to choose the graph that is the best fit for your data and will most effectively communicate your message.

People often tell me they can't create some of these different, nonstandard graphs because their colleague or manager or audience won't understand them. We are not born knowing instinctively how to read a bar chart or line chart or pie chart. As Scott Klein, deputy managing editor at *ProPublica* once wrote, "There is no such thing as an innately intuitive graphic. None of us are born literate in reading visualizations."

As data visualization creators, we must understand our audience and know when a different graph can engage readers—and help them expand their own graphic literacy.

▸ ▸ ▸ ▸ ▸

This book has three main parts. Part 1 covers general guidelines to creating effective visualizations. We'll learn the importance of our audience and how to consider what category of graph will best meet their needs. No data visualization book will contain every lesson to create effective graphs, but there are some best practices that can guide your work. As you go



Each of these six charts visualizes the same data: The share of people earning minimum wage or less in each state.

forward creating more visuals and seeing their effect on your audience, you'll develop your own aesthetic and learn when to bend or break these guidelines.

Part 2 is the meat of the book. We will define and discuss more than eighty graphs, categorized into eight broad categories: Comparisons, Time, Distribution, Geospatial, Relationship, Part-to-Whole, Qualitative, and Tables. We will see how each graph works and the advantages and disadvantages of each.

Graphs overlap between these categories—a bar chart, for example, can be used to show changes over time or comparisons between groups. The categorizations here are based on a graph's primary purpose. But even that's not an objective truth, and your perspective and situation may differ. I do not discuss *every single possible graph*—there are many specialized graphs in fields like architecture, biology, and engineering that are excluded here. Instead, these chapters cover the most common and flexible graphs that can showcase the sorts of data most people will need to display.

I tie these chapters together in part 3 with a chapter on building a data visualization style guide and a chapter on how to pull the different lessons together in a series of graph redesigns. If you've ever written a research paper, or even a book report, you are probably aware of the array of writing style guides, from the *Chicago Manual of Style* to the *Modern Language Association*. These guides break down writing into component parts and prescribe their proper use. A data visualization style guide does the same for graphs—defines their parts and how to style and use them. In the final chapter, we apply the lessons to redesign a series of graphs to improve how they communicate data.

This book will guide you as you explore your data and how it might be visualized. Now more than ever, content must be visual if it is to travel far. Your clients and colleagues, and your audiences of policymakers, decisionmakers, and interested readers are inundated with a flow of information. Visuals cut through that.

Anyone can improve the way they visualize and communicate their data—and you don't need a graduate degree in marketing or design or advertising. Take it from me, I started my career as an economist in the federal government.

HOW I LEARNED TO VISUALIZE MY DATA

Once I settled on declaring my economics major at the University of Wisconsin at Madison (there was an ill-fated attempt to also be a math major, but I hit a wall at Markov chains), I knew I wanted to end up in Washington, DC. I wanted to be near the center of public policy and politics. I wanted to explore the real problems of the day and help craft solutions.

I moved to DC in 2005 to join the Congressional Budget Office (CBO). My job was to help work on the long-term microsimulation model that is used to examine the Social Security system and forecast the long-term finances of the federal budget. The spring of 2005 was an exciting time to work on Social Security—President George W. Bush had made Social Security a central component of his second term. In his 2005 State of the Union address, he said, "We must pass reforms that solve the financial problems of Social Security once and for all." Reform would stall later that year, but in the course of my first few months on the job, my group at CBO estimated and analyzed the effects of dozens of policy proposals.

Five years later, I had expanded my work to include issues around policies that affected disabled workers, immigration, and food stamps (now called the Supplemental Nutrition Assistance Program or SNAP). In 2010, three of my colleagues were drafting a special report on policy options for Social Security. In it, they would show the impact of thirty different options for reform. One of the central figures in the report would show changes in taxes received by the system, benefits paid out from the system, the balance between the two, and other measures of fiscal solvency for these thirty options. It looked something like this:

	Option Name		Revenues, Outlays, and Balances as a Percentage of GDP, Year 2020	2040	2060	2080	75 Year Present Value as a Percentage of GDP	75 Year Present Value as a Percentage of Taxable Payroll	Trust Fund Exhaustion Year
	Baseline[a]	Revenues[b]	4.9	4.9	4.9	5.0	5.2	14.4	20XX
		Outlays[c]	5.2	6.2	6.0	6.3	5.8	16.0	
		Balance[d]	-0.3	-1.3	-1.1	-1.3	-0.6	-1.6	
	Option Name		Changes in Revenues, Outlays, and Balances as a Percentage of GDP, Year 2020	2040	2060	2080	Change in 75 Year Present Value as a Percentage of GDP	Change in 75 Year Present Value as a Percentage of Taxable Payroll	Change in Trust Fund Exhaustion Year
1	Increase the Payroll Tax Rate by 1 Percentage Point in 2012	Revenues	0.4	0.4	0.3	0.3	0.3	1.0	XX
		Outlays	*	*	*	*	*	*	
		Balance	0.4	0.4	0.4	0.4	0.3	1.0	
2	Increase the Payroll Tax Rate by 2 Percentage Points over 20 Years	Revenues	0.3	0.7	0.7	0.7	0.5	1.6	YY
		Outlays	*	*	*	*	*	*	
		Balance	0.3	0.7	0.7	0.8	0.6	1.6	
3	Increase the Payroll Tax Rate by 3 Percentage Points over 60 Years	Revenues	0.2	0.5	0.8	1.0	0.5	1.5	ZZ
		Outlays	*	*	*	*	*	*	
		Balance	0.2	0.5	0.9	1.1	0.5	1.4	
4	Eliminate the Taxable Maximum	Revenues	0.8	0.9	0.9	0.9	0.9	n.a.	AA
		Outlays	*	0.3	0.5	0.5	0.3	n.a.	
		Balance	0.8	0.6	0.4	0.4	0.6	n.a.	
5	Raise the Taxable Maximum to Cover 90% of Earnings	Revenues	0.3	0.4	0.4	0.4	0.4	n.a.	BB
		Outlays	*	0.1	0.2	0.2	0.1	n.a.	
		Balance	0.3	0.3	0.2	0.2	0.2	n.a.	

Author's rendering of early draft of exhibit from the Congressional Budget Office.

You don't need to be a government economist to know that members of Congress are unlikely to read something that looks like a spreadsheet. There are too many rows, too many columns, too many numbers—too much information. It was right then that I first started thinking about better ways to present this information.

This was the result. We replaced some numbers with small area charts, which give the reader an immediate visual impression of each option—which ones increased the solvency of the program and which ones did not.

Table 2.

Changes to Social Security's Finances Under Various Options with Scheduled Benefits

(Percentage of GDP)

		2020	2040	2060	2080	Annual Finances	75-Year Present Value as a Percentage of GDP	75-Year Present Value as a Percentage of Taxable Payroll
						Current Law[a]		
						Revenues and Outlays[b]		
	Revenues	4.9	4.9	4.9	5.0		5.2	14.4
	Outlays	5.2	6.2	6.0	6.3		5.8	16.0
	Balance	-0.3	-1.3	-1.1	-1.3		-0.6	-1.6
						Percentage-Point Change from Current Law[a]		
Change the Taxation of Earnings						Change in Annual Balance[c]		
1	Revenues	0.4	0.4	0.3	0.3		0.3	1.0
Increase the Payroll Tax Rate by 1 Percentage Point in 2012	Outlays[d]	*	*	*	*		*	*
	Balance	0.4	0.4	0.4	0.4		0.3	1.0
2	Revenues	0.3	0.7	0.7	0.7		0.5	1.6
Increase the Payroll Tax Rate by 2 Percentage Points Over 20 Years	Outlays[d]	*	*	*	*		*	*
	Balance	0.3	0.7	0.7	0.8		0.6	1.6
3	Revenues	0.2	0.5	0.8	1.0		0.5	1.5
Increase the Payroll Tax Rate by 3 Percentage Points Over 60 years	Outlays[d]	*	*	*	*		*	*
	Balance	0.2	0.5	0.9	1.1		0.5	1.4
4	Revenues	0.8	0.9	0.9	0.9		0.9	n.a.
Eliminate the Taxable Maximum[e]	Outlays	*	0.3	0.5	0.5		0.3	n.a.
	Balance	0.8	0.6	0.4	0.4		0.6	n.a.
5	Revenues	0.3	0.4	0.4	0.4		0.4	n.a.
Raise the Taxable Maximum to Cover 90% of Earnings[e]	Outlays	*	0.1	0.2	0.2		0.1	n.a.
	Balance	0.3	0.3	0.2	0.2		0.2	n.a.

Continued

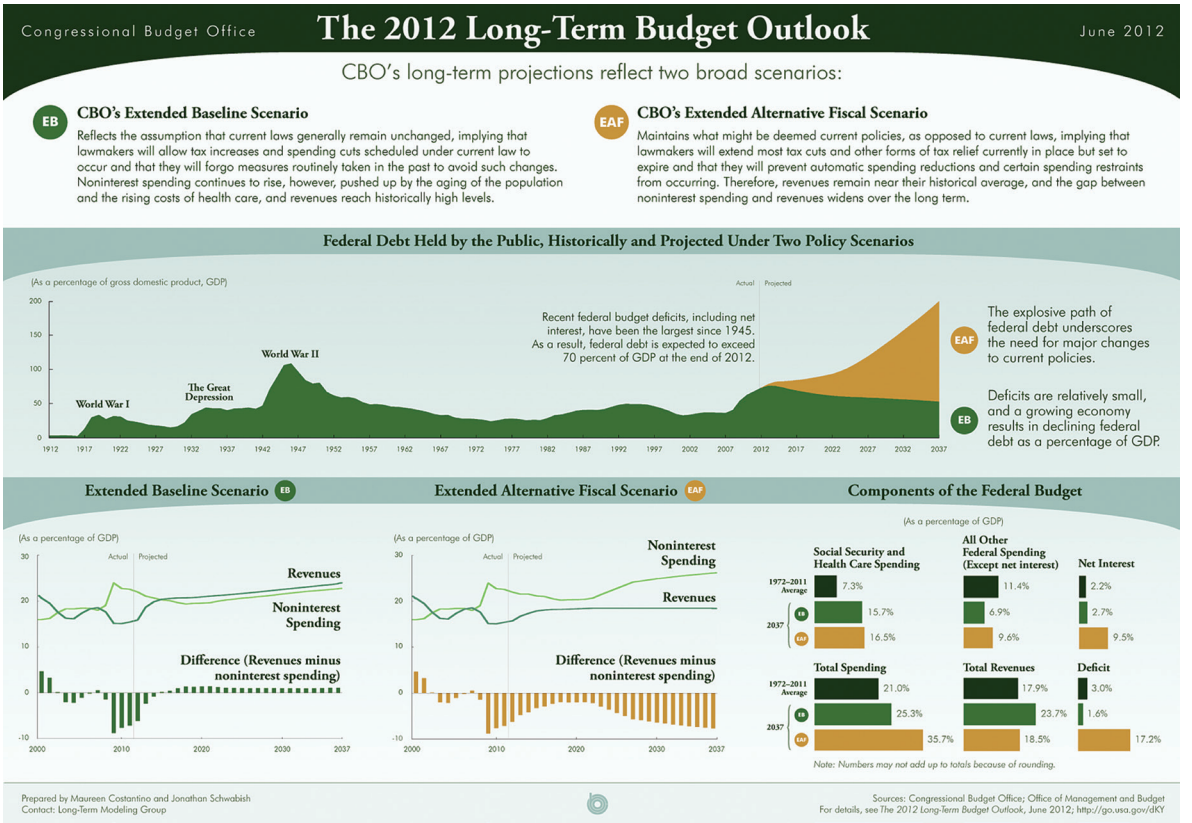
Final version of that main exhibit in the Congressional Budget Office report on Social Security. Notice that there is less data and more graphs.

Source: Congressional Budget Office.

The report worked. We received good feedback from colleagues at CBO and other agencies, as well as readers on Capitol Hill and elsewhere, noting how easy it was to read and digest the graphs. It was maybe the first time I (and perhaps the agency) thought carefully

and strategically about our data visuals. From there, I started reading books on data visualization, design, color theory, and typography.

Working with our editorial department and designers, we began to improve the graphs in our basic reports and started creating new report and graph types. We made infographics—what was then a buzzword referring (sometimes derisively) to longer graphics that combine data, text, images, and more into a single visual. In 2012, we created this infographic to accompany and summarize *The Long-Term Budget Outlook*, a 109-page report.



One-page infographic about the 2012 Long-Term Budget Outlook from the Congressional Budget Office.

Source: Congressional Budget Office.

That June, CBO's director sat in front the U.S. House Budget Committee to relay the results of our analysis. As the hearing played on a TV out in the hallway, I suddenly heard yells of, "Jon! Jon! Come out! Your infographic is on TV!"

And, sure enough, Congressman Chris Van Hollen was holding up the infographic on C-SPAN, covered with scribbles and notes. The visualization had captured and engaged the attention of one of the busiest people in America, and someone who could do something about the pressures facing the federal budget. That was the moment I knew that *how* we presented our data could matter as much as the data itself.

In 2014, I moved to the Urban Institute, a nonprofit research institution in Washington, DC, to spend half of my time conducting research and half of my time in the Communications department, helping colleagues present and visualize their data.

Since that time, I have conducted hundreds of workshops, delivered lectures around the globe, and published two books on data communication. The world, it seemed, had seen what I saw—better visual content and better presentations were the currency of research and

Maryland Congressman Chris Van Hollen holding up that Long-Term Budget Outlook infographic in a House Budget Committee hearing.

Source: C-SPAN2.

policy adoption. The advance of computing power, social media platforms, and the expanding media landscape made visual content more important, perhaps even necessary.

Today, I work with people in nonprofits, government agencies, private sector companies, and everything in between to improve how they create their graphs and communicate their content. I've worked with junior economists and analysts dealing with enormous data sets; health care workers trying to communicate results to patients, families, and hospital administrators; human resource representatives working with databases of job-seekers; advertisers and marketing executives selling products to clients; and many more.

I've seen hundreds of different kinds of data visualization challenges. The skills to meet them, unfortunately, are not yet regularly taught in schools or professional development programs. But these skills *can* be learned. We can learn how to read chart types we've never seen before, even if they are complex. And we can learn how to communicate our work in better and more effective ways.

Eventually, I discovered that one of the most important things I can show people is the incredibly wide array of graphs available to them. And that is precisely the content of this book, a survey of more than eighty types of data visualizations, from the familiar to the nonstandard.

But before we get to the library of graph types, we'll consider some of the science behind how we process visual information and some best practices and approaches to visualizing data.

PART ONE

PRINCIPLES OF DATA VISUALIZATION

VISUAL PROCESSING AND PERCEPTUAL RANKINGS

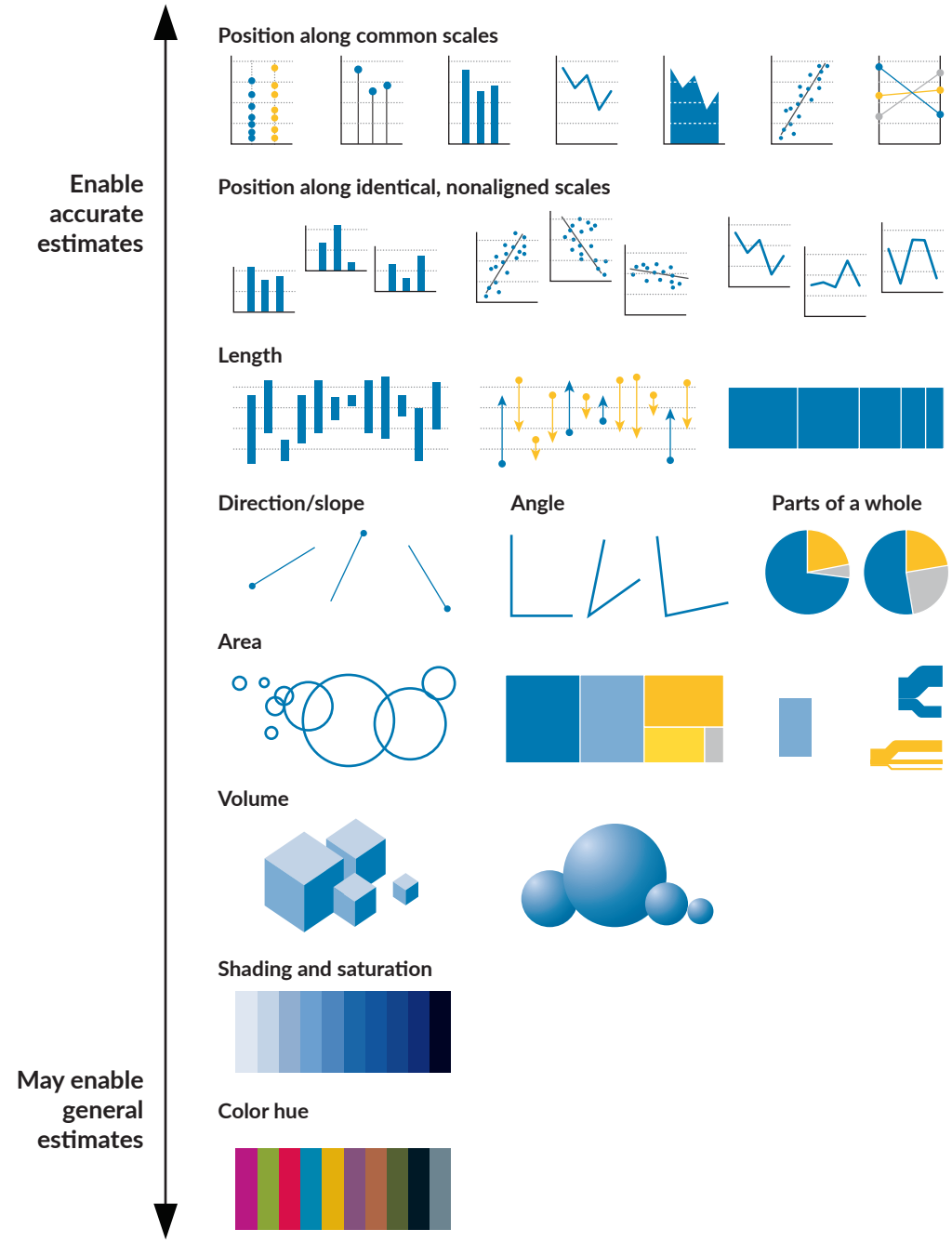
Before we start creating our charts and graphs, we need to cover some basic theory of how the brain perceives visual stimuli. This will guide you as you decide what chart type is most appropriate to visualize your data.

When we consider how to visualize our data, we must ask ourselves how accurately the reader can perceive the data values. Are some graphs better equipped to guide the reader to the specific difference between, say, 2 percent and 2.3 percent? If so, how should we think about those differences as we create our visualizations?

There's a thread of research in the data visualization field that explores this very question. Based on original research over the past thirty years or so, the image on the next page shows a spectrum of graphs—or more generally, types of data *encodings* like dots, lines, and bars—arrayed by how easily readers can estimate their value. The encodings that readers can most accurately estimate are arranged at the top, and those that enable more general estimates are at the bottom.

The rankings are unsurprising. It is easier to compare the data in line charts, bar charts, and area charts that have the same axis or baseline. Graphs on which the data are positioned on unaligned axes—think of a pair of bars that are offset from one another on different axes—are slightly harder for us to accurately discern the values.

Farther down the vertical axis are encodings based on angle, area, volume, and color. You intuitively know this: it's much easier to discern the exact data values and differences between values when reading a bar chart than when reading a map where countries are shaded with different colors.



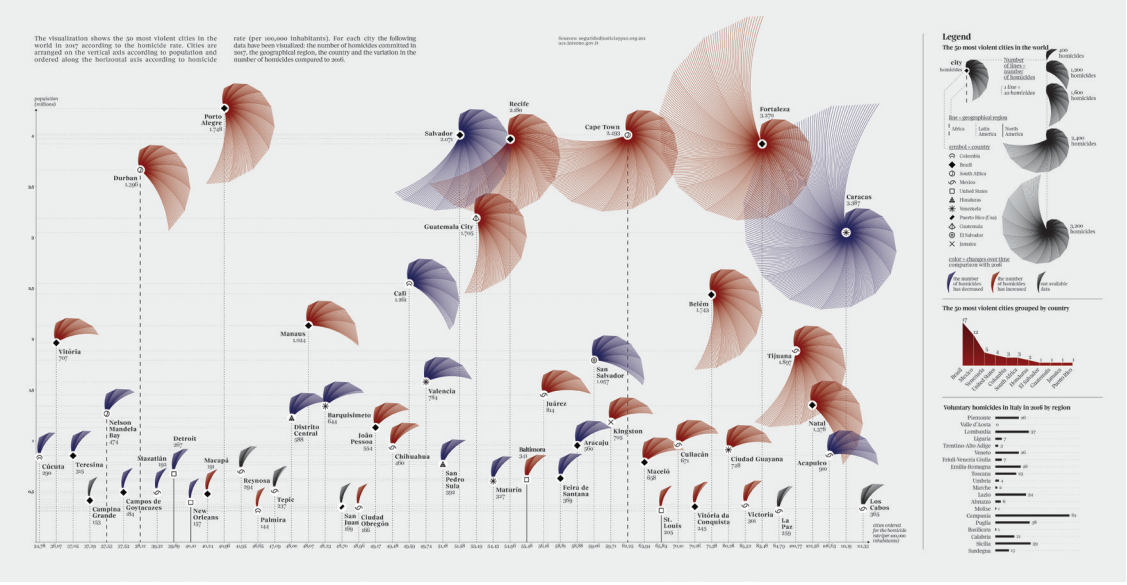
Perceptual ranking diagram. What kind of data visualization you choose to create will depend on your goals and your audience's needs, experiences, and expertise. This image is based on Alberto Cairo (2016) from research by Cleveland and McGill (1984), Heer, Bostock, and Ogievetsky (2010), and others.

Standard graphs, like bar and line charts, are so common because they are perceptually more accurate, familiar to people, and easy to create. Nonstandard graphs—those that use circles or curves, for instance—may not allow the reader to most accurately perceive the exact data values.

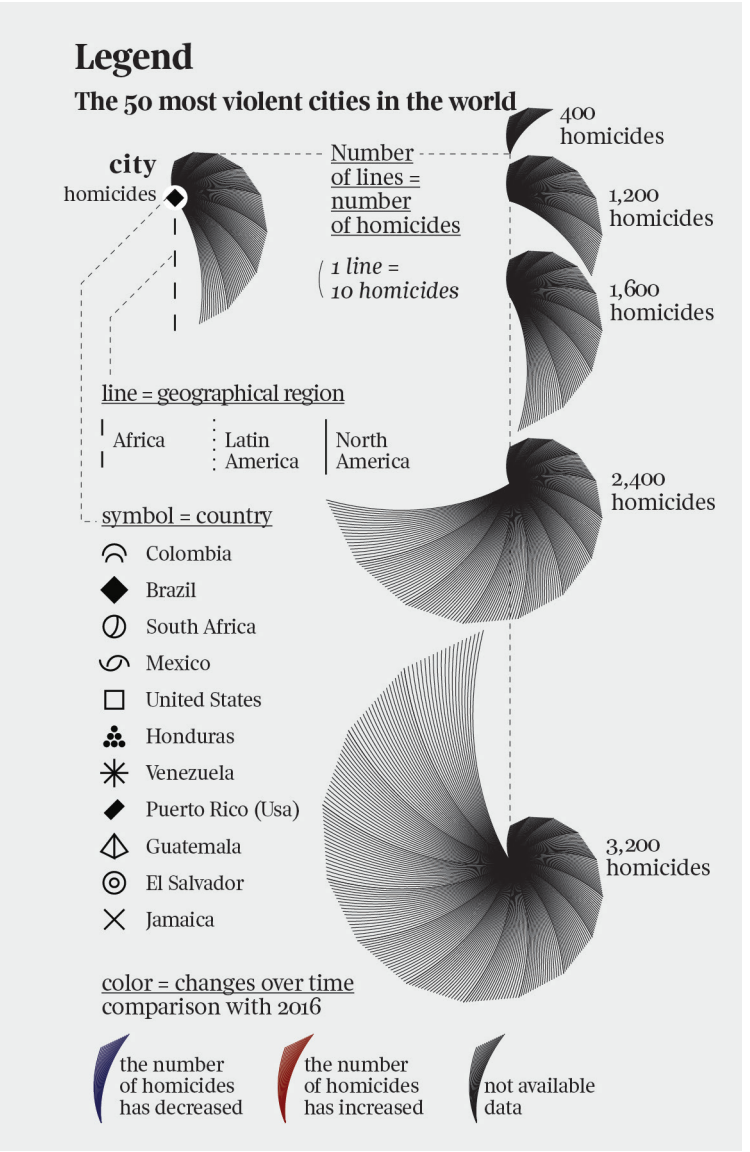
But perceptual accuracy is not always the goal. And sometimes it's not a goal at all.

Spurring readers to engage with a graph is sometimes just as important. Sometimes, it's more important. And nonstandard chart types may do just that. In some cases, nonstandard graphs may help show underlying patterns and trends in better ways that standard graphs. In other cases, the fact that these nonstandard graphs are different may make them more engaging, which we may sometimes need to first attract attention to the visualization.

This graphic from information designer Federica Fragapane shows the fifty most violent cities in the world in 2017. The vertical axis measures the population of each city and the horizontal axis captures the homicide rate per 100,000 people. The number of lines in each icon represents the number of homicides, and additional colors, shapes, and markers capture metrics like country of origin (the symbol in the middle of each), region (vertical dashed line), and change since 2016 (blue for decreases, red for increases). It could be a bar



Graphic from Frederica Fragapane for La Lettura—Correier della Serra that shows the fifty most violent cities in the world. See the next page for a closer look at the legend.



A zoom-in of the graphic from Frederica Fragapane. Notice all of the details and data elements included in each icon. It could be a bar chart or line chart, but would you then be inclined to zoom in and read it closely?

chart or a line chart or some other chart type. But if it were, would you be inclined to zoom in, read it closely, and examine it?

Data visualization is a mix of science and art. Sometimes we want to be closer to the science side of the spectrum—in other words, use visualizations that allow readers to more accurately perceive the absolute values of data and make comparisons. Other times we may want to be closer to the art side of the spectrum and create visuals that engage and excite the reader, even if they do not permit the most accurate comparisons.

Sometimes you must make your visuals interesting and engaging, even at the cost of absolute perceptual accuracy. Readers may not be as interested in the topic as we hope or may not have enough expertise to immediately grasp the content. As content creators, however, our job is to encourage people to read and use the graph, even if we "violate" perceptual rules that we know will hamper someone's ability to make the most accurate conclusions. Thinking about different audience types is not just about considering among decision makers, scholars, policymakers, and the general public—it also means thinking about different levels of interest or engagement with the visual itself. As historian Cecelia Watson writes in her book about the history and use of the semicolon, "What if we thought less about rules and more about communication, and considered it our obligation to one another to try to figure out what is really being communicated?"

We should not operate from the assumption that readers will pay attention to everything in our visual, even if we use a common, familiar chart type. Let's be honest: People see bar charts and line charts and pie charts all the time, and those charts are often boring. Boring graphs are forgettable. Different shapes and uncommon forms that move beyond the borders of our typical data visualization experience can draw readers in. Reading a graph is not like the spontaneous comprehension of seeing a photograph. Instead, reading a graph has more of the complex cognitive processes as reading a paragraph.

This isn't to say we should not concern ourselves with visual perception or allowing our readers to make the most accurate comparisons, but the goal of *engagement* can be worth a lot in its own right. Elijah Meeks, a data visualization engineer, wrote that, "Charts, like any other communication, need to be compelling to be convincing, and if your bar chart, as optimal as it may be, has been reduced to background noise by the constant hum of bar charts crossing a stakeholder's screen, then it's your responsibility to make it more compelling, even if it's not any more precise or accurate than a more simple form."

Introducing a new or different graph type can also introduce a hurdle to your reader. These can be big hurdles, like a completely new graph type or an exceptionally unusual

This graphic from an interactive visualization from the Organisation of Economic Co-Operation and Development (OECD) enables users to explore the different metrics and definitions of what it means to have a "better life." A more standard chart type, like a bar chart, might enable easier comparisons, but would it be as much fun?

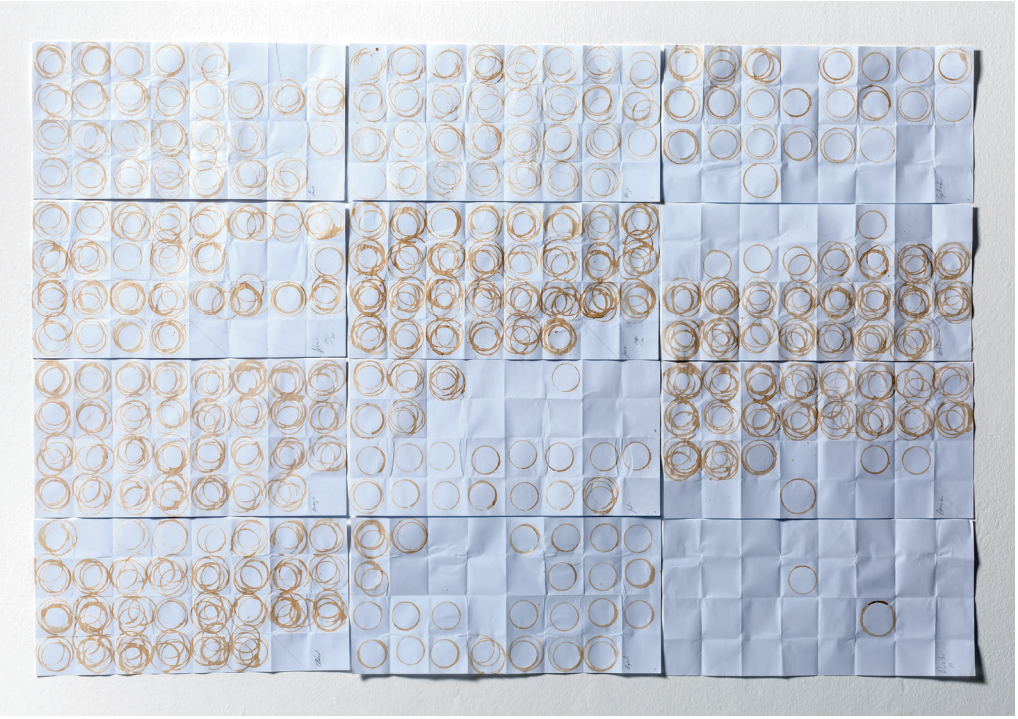
Source: Organisation for Economic Co-Operation and Development

representation of the data. Or they can be small hurdles, graphs that rank lower on the perceptual-accuracy scale or graphs that people may have only seen a few times before. To overcome these hurdles, you may need to explain how to read the graph. But that might be worth it because sometimes different charts attract reader's attention and pique their curiosity.

When should you use a nonstandard graph? Likely not for many scholarly purposes, because they do not enable the most accurate perceptions of the data. For scholarly writing, accuracy is paramount. We want our reader to clearly and efficiently compare the values we're presenting. But in other cases—headline-style or standalone graphics, blog posts, shorter briefs or reports, or graphs for social media—creating something *different* may draw people in and hold their attention just long enough to convey your argument, data, or content.

This visualization from artist and journalist Jaime Serra Palou is a lovely example of this kind of nonstandard and creative data visualization. He plots his coffee consumption every day over the course of a year by using the stains from his coffee cups. You can immediately see those parts of the year when he needed an extra burst of caffeine. Yes, a line chart might convey the same data, but would you pause to spend an extra moment reading it?

Sometimes you can do both—a nonstandard, attention-grabbing graphic accompanied by a more familiar graph next to it. What you present and how you present it depends on your audience. The Serra piece might work as the lead graphic on a book or report about coffee consumption, but more detailed charts inside might take the form of standard charts and tables. Some academic research has shown that creating novel graphs, such as



Artist and journalist Jaime Serra Palou plotted his coffee consumption every day for a year by using stains from his coffee cup.

those that enable the user to personalize the content (by inputting their own information) or are simply more aesthetically appealing, encourages readers to actively process the content.

ANSCOMBE'S QUARTET

The value of visualizing data is best illustrated by Anscombe's Quartet, published in 1973 by statistician Francis Anscombe. The Quartet demonstrates the power of graphs and how they, together with statistical calculations, can better communicate our data.

Examine the table below, which shows four pairs of data, an *X* and a *Y*.

We can make some basic observations about these data. We can see that the first three series of X's are all the same; the values of X's in the last series are all 8 except for the one 19; and the X's are all whole numbers while the Y's are not. We might even notice that the 12.7 value in the third column of Y is larger than the rest. In my experience, most people don't comment about the *relationship* between the different series, which, at the end of the day, is what we want to understand. It turns out that each of the four pairs yield the same standard information: the same average values of the X series and the Y series; the same variance for each; the same correlation between X and Y; and the same estimated regression equation.

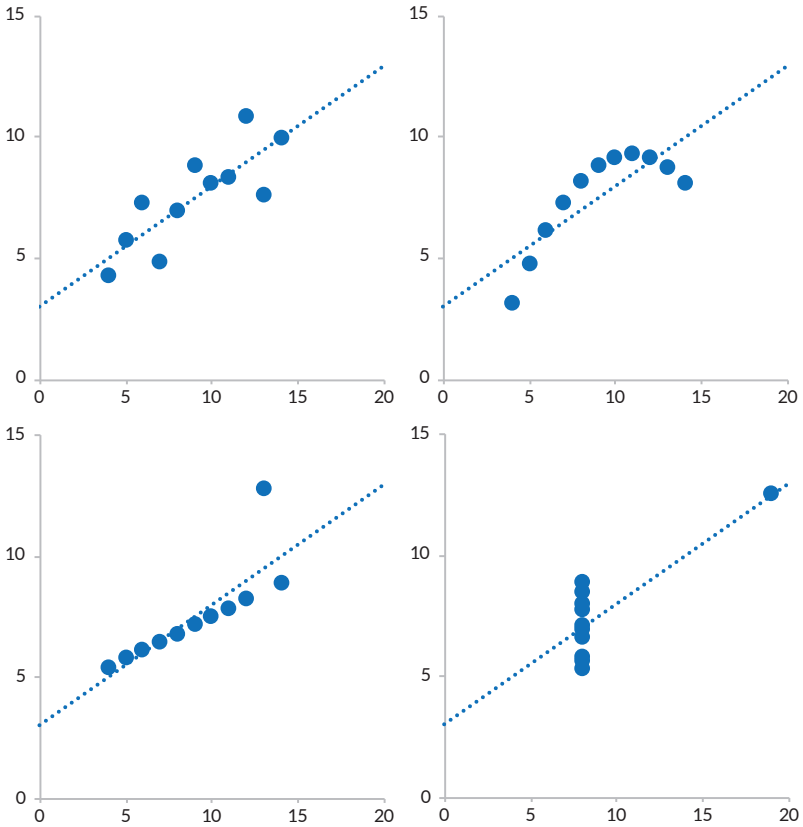
Data set		1	1	2	2	3	3	4	4
Variable		x	y	x	y	x	y	x	y
Obs. No.	1 :	10	8.0	10	9.1	10	7.5	8	6.6
	2 :	8	7.0	8	8.1	8	6.8	8	5.8
	3 :	13	7.6	13	8.7	13	12.7	8	7.7
	4 :	9	8.8	9	8.8	9	7.1	8	8.8
	5 :	11	8.3	11	9.3	11	7.8	8	8.5
	6 :	14	10.0	14	8.1	14	8.8	8	7.0
	7 :	6	7.2	6	6.1	6	6.1	8	5.3
	8 :	4	4.3	4	3.1	4	5.4	19	12.5
	9 :	12	10.8	12	9.1	12	8.2	8	5.6
	10 :	7	4.8	7	7.3	7	6.4	8	7.9
	11 :	5	5.7	5	4.7	5	5.7	8	6.9
Mean		9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance		11.0	4.1	11.0	4.1	11.0	4.1	11.0	4.1
Correlation		0.816		0.816		0.816		0.817	
Regression line		y = 3 + 0.5x		y = 3 + 0.5x		y = 3 + 0.5x		y = 3 + 0.5x	

Source: Francis Anscombe

Known as Anscombe's Quartet, this example demonstrates how difficult it is for us to pull out basic patterns and summary statistics.

When we see the same data presented in four graphs, however, we can immediately see these relationships, for example, the positive correlation in all four pairs, the curvature in the second pair that you couldn't see in the table, and the outliers 12.7 and 19.0.

We are much more likely to remember these four small graphs than we are the original table. In his bestselling book, *Brain Rules*, molecular biologist John Medina writes, "The more visual the input becomes, the more likely it is to be recognized and recalled." The more we can make our data and content visual, the more we can expect our readers to remember it and, hopefully, use it.



The data visualization representation of Anscombe's Quartet. Notice how much easier it is to see the positive relationship between the two variables, the curvature in the pattern in the top-right graph, and the outliers in the bottom two graphs.

Source: Francis Anscombe (1973).

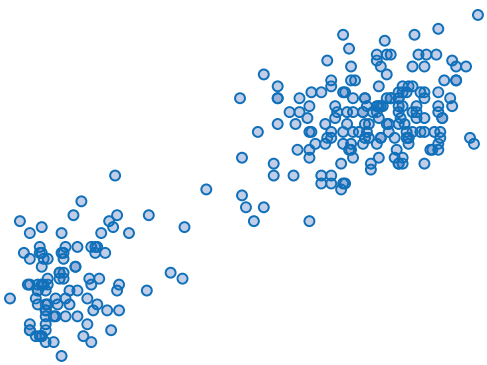
GESTALT PRINCIPLES OF VISUAL PERCEPTION

How do we perceive information? And how, as chart creators, can we use these perceptual rules to more effectively communicate our data? "Gestalt theory" is one such way we can think about how our readers will look at our graphs. Gestalt theory was developed in the early part of the twentieth century by German psychologists and refers to how we tend to organize visual elements into groups. Further developments in the field were interrupted by the rise of the Nazi regime in Germany and then by World War II, and after the war it was criticized for not having rigorous methodological methods. But the ideas persist in many disciplines, including information theory, vision science, and cognitive neuroscience.

These six organizational principles from Gestalt theory are especially useful for creating graphs and visuals that tap into our reader's visual processing network.

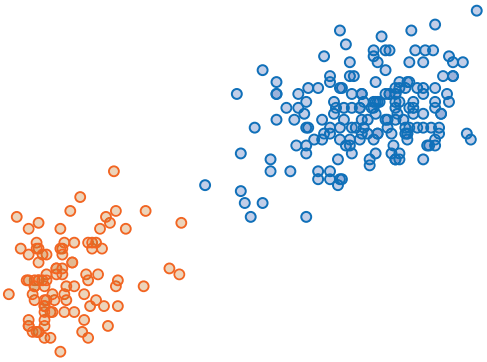
PROXIMITY

We perceive objects that are close to one another as belonging to a group. There are lots of graphical elements that we can group together: labels with points, bars with each other, or, like this graph, clusters of points in a scatterplot in which we can see two groups or clusters, one in the top-right and the other closer to the bottom-left.



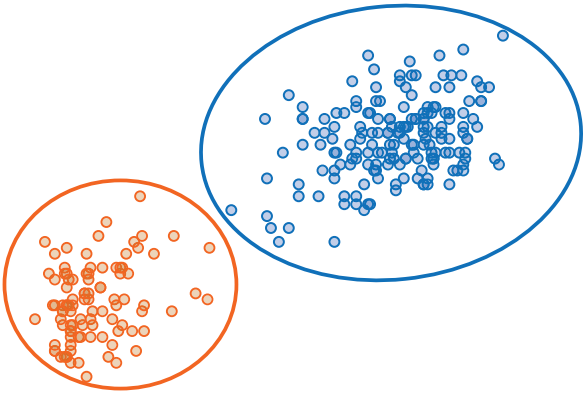
SIMILARITY

Our brains group objects that share the same color, shape, or direction. Adding color to the above scatterplot reinforces the two groups.



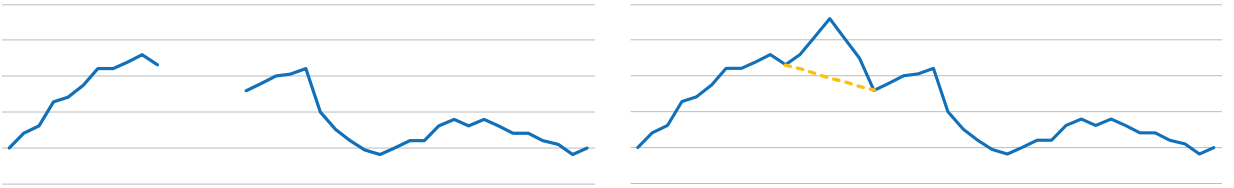
ENCLOSURE

Bounded objects are perceived as a group. Here, in addition to using color, we can enclose the two groups with circles or other shapes.



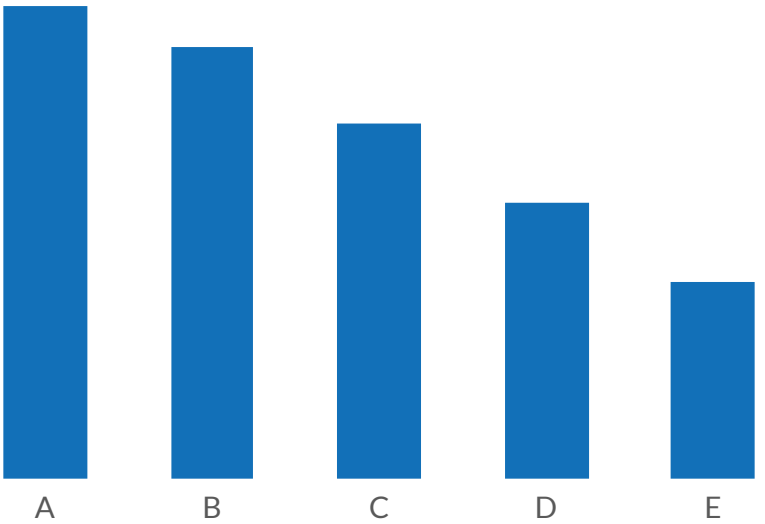
CLOSURE

Our brains tend to ignore gaps and complete structures with open areas. In its basic form, we don't have a problem viewing a simple graph that has a horizontal axis and a vertical axis as a single object because the two lines are enough for us to define the closed space. In a line chart with missing data, for example, we tend to mentally close the gap in the most direct way possible, even if there might be something different going on in that missing area. For example, in the line graph on the left, we mentally close the gap between the two segments with a straight line even though the missing data might yield a pattern that moves up and then down.



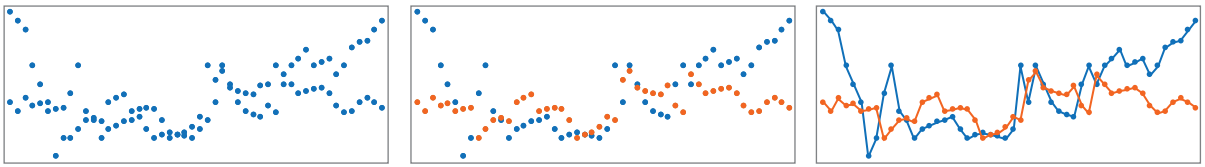
CONTINUITY

Here, objects that are aligned together or continue one another are perceived as a group. Hence, our eyes seek a smooth path when following a sequence of shapes. You don't need the horizontal axis line in this bar chart, for example, because the bars are aligned along a consistent path between the labels and the bottoms of the bars.



CONNECTION

According to this principle, we perceive connected objects as members of the same group. Take this series of dots: At first, we perceive it as a single series, a mass of blue dots. Adding color makes it clear there are two different series. Connecting the dots makes it clear how the two initially track each other but then diverge.



PREATTENTIVE PROCESSING

The concept of "preattentive processing" is a subset of Gestalt theory, and it is the visual process I consider most when creating my data visualizations. As we just saw, because our eyes can detect a limited set of visual characteristics, we combine various features of an object and unconsciously perceive them as a single image. In other words, preattentive attributes draw our attention to a specific part of an image or, in our case, a graph.

For example, try to find the four largest numbers in this table.

Table 1. Our sales grew to $600 million this year

	Q1	Q2	Q3	Q4
Bob	26	35	72	84
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	95	13	64
Jon	83	62	46	48
Karen	30	65	98	82
Ken	38	28	45	71
Lauren	98	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

Hard to do, right? Now try it with these versions that use color (on the left) and intensity (on the right) to highlight those four numbers.

Table 1. Our sales grew to $600 million this year

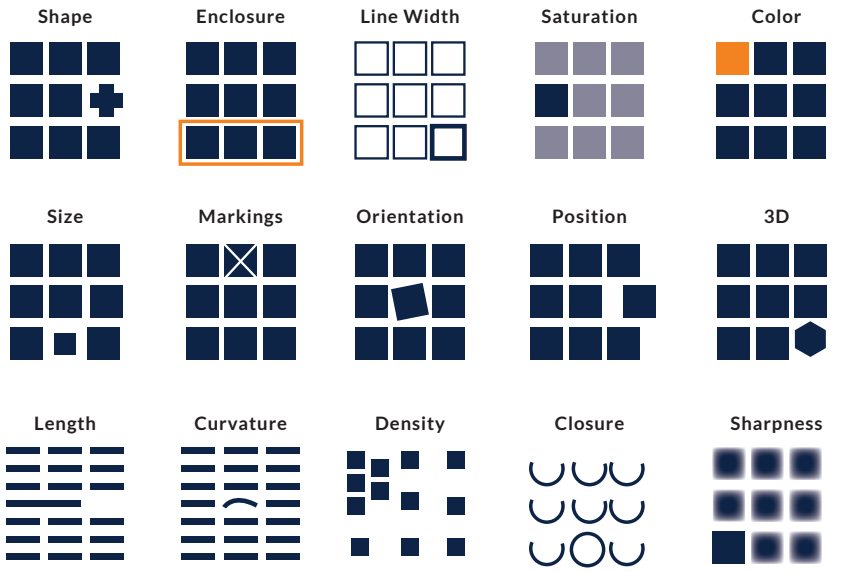
	Q1	Q2	Q3	Q4
Bob	26	35	72	84
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	95	13	64
Jon	83	62	46	48
Karen	30	65	98	82
Ken	38	28	45	71
Lauren	98	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

Table 1. Our sales grew to $600 million this year

	Q1	Q2	Q3	Q4
Bob	26	35	72	**84**
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	**95**	13	64
Jon	83	62	46	48
Karen	30	65	**98**	82
Ken	38	28	45	71
Lauren	**98**	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

Preattentive attributes here direct our attention to the large numbers immediately.

It's easier to find the numbers in these two tables than the first because the numbers are encoded using *preattentive attributes*: color and weight. Each distinction helps us effortlessly identify the key number.



Examples of preattentive attributes that we can use in our visualizations to direct our reader's attention.

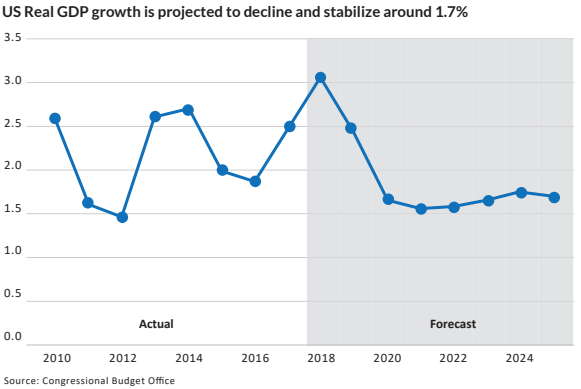
Preattentive attributes are effects that seem to pop out from their surroundings. There are many we can use to tap into our reader's visual processing network to draw their attention: shape, line width, color, position, length, and more.

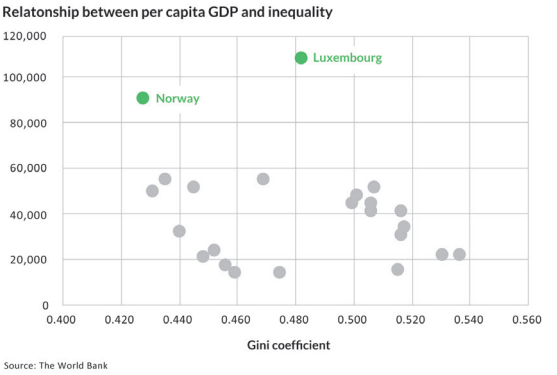
Preattentive processing works in photographs too. Consider these images of fruits and vegetables. In the photo on the left, the eye is drawn to the upper-right corner. The group of tomatoes is slightly larger than the rest and positioned away from the group. In the photograph on the right, however, the eye is not drawn to any specific position. This photograph is more evenly balanced, so no one object stands apart from the rest.

Notice how your eye gravitates toward the four tomatoes in the top-right part of the image on the left. The image on the right is balanced, so your eye doesn't immediately focus on any particular area. Photos by NordWood Themes (left) and Tim Gouw (right) on Unsplash.

We can apply these attributes to data visualization. A line chart uses the *position* of the points to indicate the data, while a bar chart uses *length*. You can use preattentive attributes to draw your audience's attention to aspects of your graphs, guiding their focus.

For example, on the next page, I can *enclose* the 'Forecast' area of the line chart on the left with the gray box—notice how it immediately draws your eye to the right side of the graph. Similarly, I can use the *color* attribute to highlight a few points in the scatterplot on the right (and keep the other dots gray).





Applying simple preattentive attributes to these graphs directs your eye to the "Forecast" area of the graph on the left and to the two highlighted countries in the graph on the right.

WRAPPING UP

With these basic rules of perception, we are now better equipped to recognize and interpret the visual features we can use to encode and highlight our data. Before we start adding more graphs to our data visualization toolbox, let's lay out some basic guidelines of more effective data visualizations—things you should keep in mind no matter what kind of graph you are creating.

FIVE GUIDELINES FOR BETTER DATA VISUALIZATIONS

Whenever I create a data visualization, whether it's static, interactive, or part of a report or blog post or even a tweet, I follow five primary guidelines.

1. Show the data
2. Reduce the clutter
3. Integrate the graphics and text
4. Avoid the spaghetti chart
5. Start with gray

Showing the data and reducing the clutter means reducing extraneous gridlines, markers, and shades that obscure the actual data. Active titles, better labels, and helpful annotations will integrate your chart with the text around it. When charts are dense with many data series, you can use color strategically to highlight series of interest or break one dense chart into multiple smaller versions.

Taken together, these five guidelines remind me of the needs of my audience and how my visuals can tell them a story.

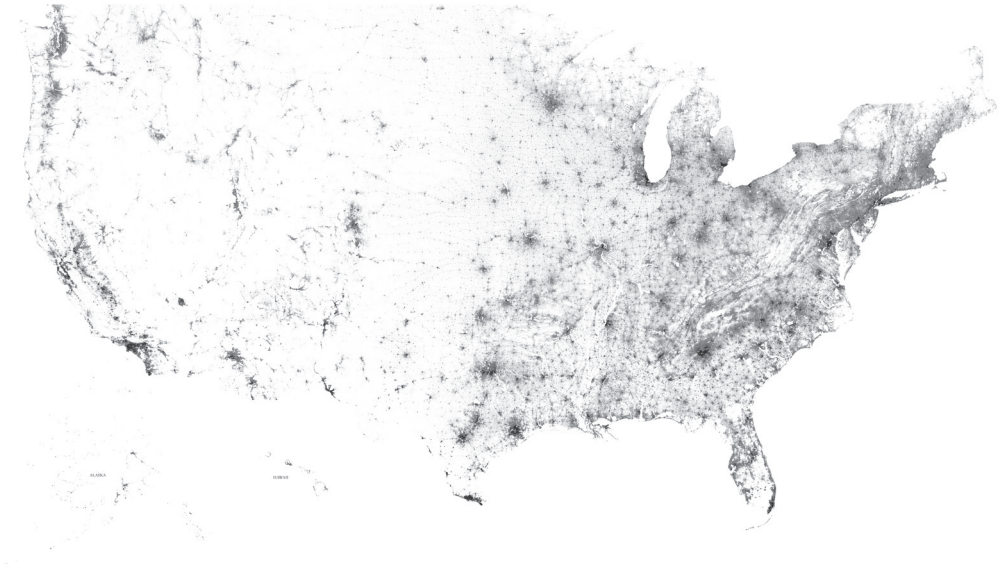
GUIDELINE 1: SHOW THE DATA

Your reader can only grasp your point, argument, or story if they see the data. This doesn't mean that *all* the data must be shown, but it does mean that you should highlight the values

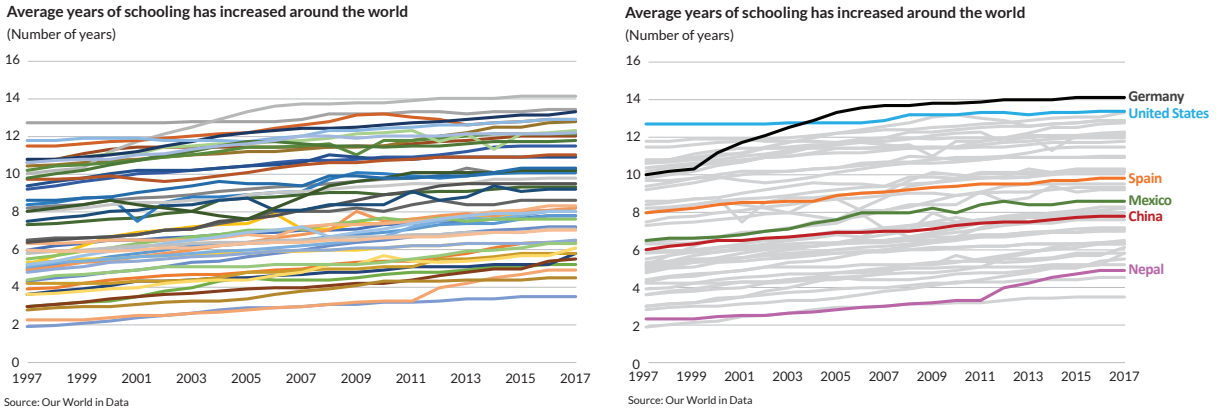
that are important to your argument. As chart creators, our challenge is deciding how much data to show and the best way to show it.

Consider this dot density map of the United States (see page 244 for more on this kind of map). It uses data from the 2010 U.S. decennial census and places a dot for each of the country's 308 million residents in their census blocks (a census block roughly corresponds to a city block). Notice how there is nothing in the image *except for the data*. There are no state borders, roads, city markers, or labels for lakes and rivers. We still recognize it as the United States because people tend to live along borders and coasts, which helps give shape to the country.

This doesn't mean we must show *all of the data all the time*. Sometimes charts show too much data, making it hard to see which data points matter most. On the next page are two line charts that both show the average number of years of schooling for fifty countries around the world. In the graph on the left, each country is assigned its own color. This makes



The Gestalt principle of *similarity* helps us see the clusters of people around the country.
Source: Image Copyright, 2013, Weldon Cooper Center for Public Service, Rector and Visitors of the University of Virginia (Dustin A. Cable, creator).



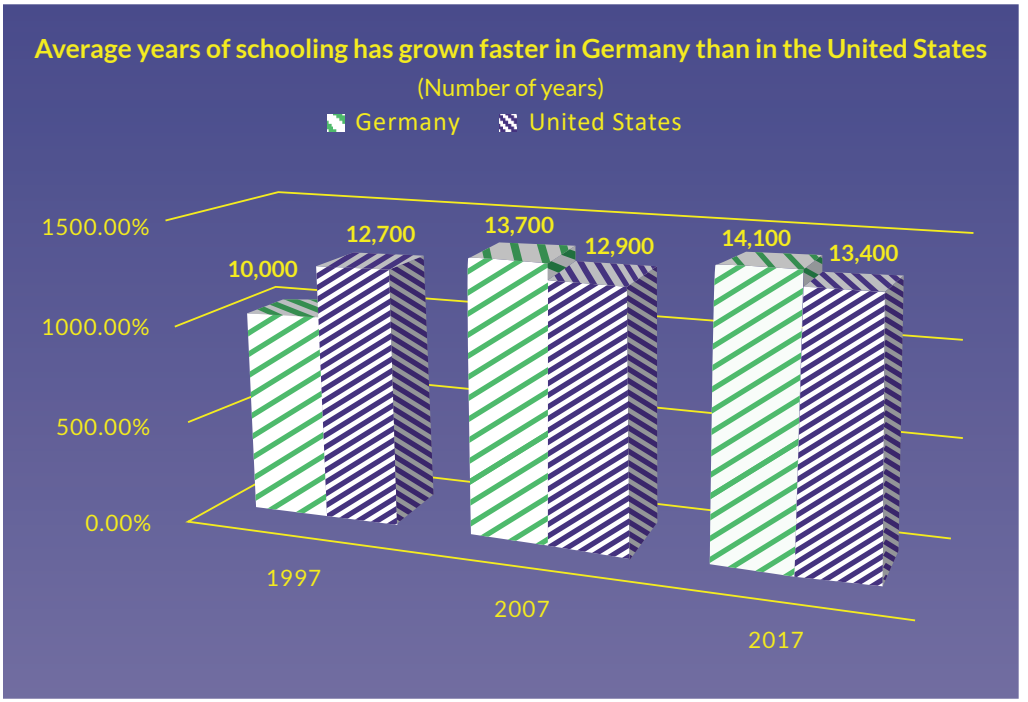
Highlighting just a few countries in the chart on the right makes it easier to read.

it busy and confusing, impossible to pick out a trend for any single country. In the graph on the right, just six countries of interest are highlighted while the remaining are set in gray, blending them into a neutral background. This gives the reader a clear view of the countries we want to highlight. It's not about showing the least amount of data, it's about showing the data that matter most.

GUIDELINE 2: REDUCE THE CLUTTER

The use of unnecessary visual elements distracts your reader from the central data and clutters the page. There are lots of different types of chart clutter we might want to avoid. There are basic elements like heavy tick marks and gridlines, which we should remove in almost every case. Some graphs use data markers like squares, circles, and triangles to distinguish between series, but when the markers overlap they jumble the patterns. Some use textured or filled gradients when simple, solid shades of color work just as well. Some use unnecessary dimensions that distort the data. And others contain too much text and too many labels, cluttering the space and crowding out the data.

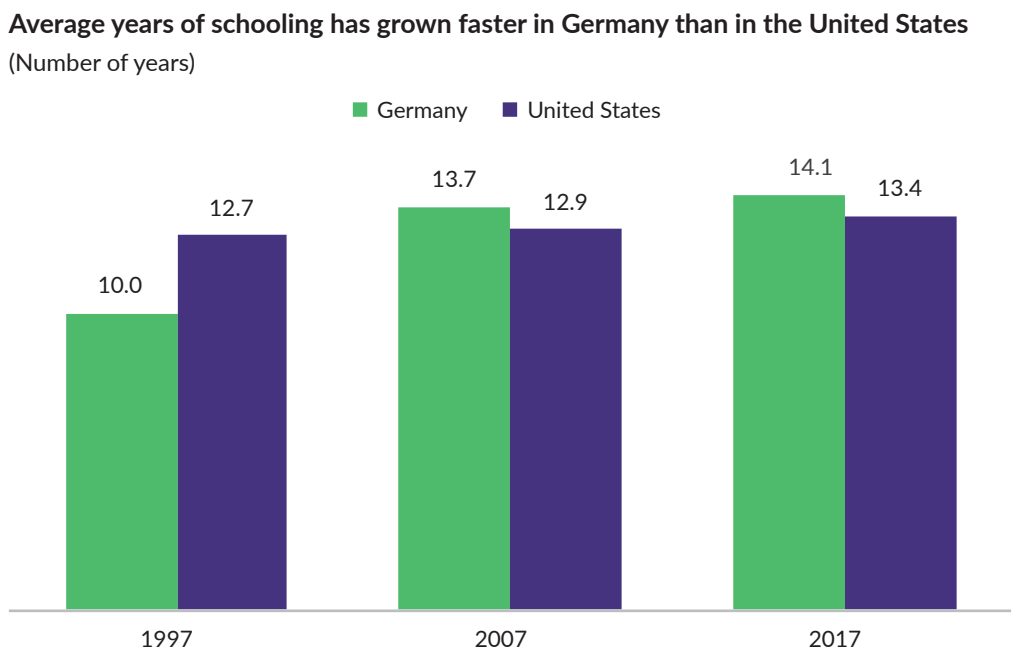
Take this three-dimensional column chart of average schooling for the United States and Germany for a few select years.



You've seen these kinds of 3D charts before—they are distracting, hard to read, and distort the data.

If you think that this looks so outlandish that no one would ever style a chart this way, you'd be wrong. I've copied the exact style from another chart, even down to the gradient styling. The three-dimensional bars and shimmering stripes, mismatched data and axis labels, the abundance of decimals that suggest a level of data precision that's not actually there—all these combine to create a graph that is difficult to read and, quite honestly, uncomfortable to look at. Also notice how the three-dimensional view distorts the data. The first bar never touches the gridline even though it should match it exactly. This distortion occurs because the unnecessary third dimension requires adding perspective to the graph. Simplifying the graph by discarding these extraneous, distracting elements and showing the data makes your argument clear and comprehensible.

While much of our understanding of perception and how our eyes and brains work is rooted in scientific research, our decisions of which graph to use, where we place labels and annotation, which colors and fonts to use, and how we lay out our visualizations is mostly subjective. There are cases where certain graphs are wrong, but many other cases call for



A basic bar chart eliminates the clutter and the distortion caused by the 3D effect, so the graph is easier to read and understand.

Data Source: Our World in Data.

nothing more than your best judgment. As you create more visualizations and read more graphs, you'll develop your own eye and aesthetic—and your own balance of art and science.

GUIDELINE 3: INTEGRATE THE GRAPHICS AND TEXT

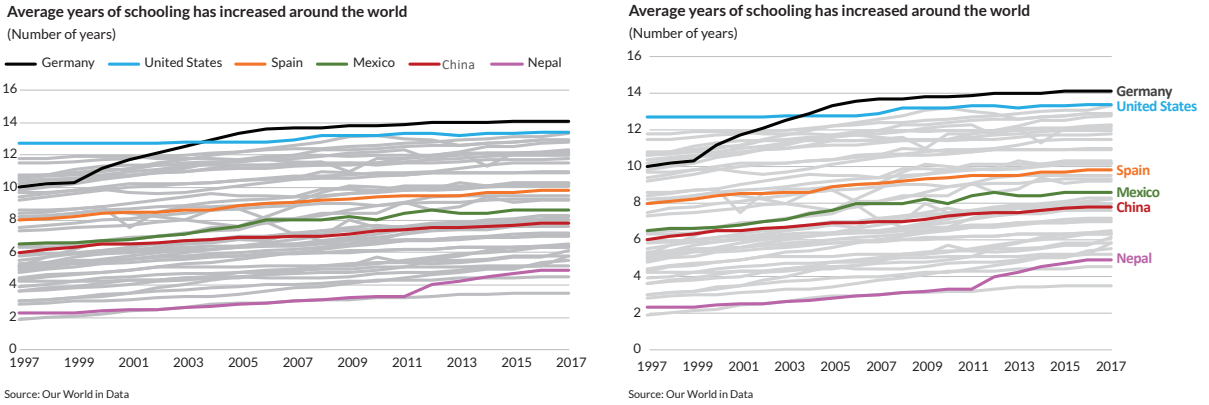
Although our primary focus on creating a visualization is the graphic elements—bars, points, or lines—the text we include in and around our graphs is just as important. Far too often, we treat the text and annotations as an afterthought, but these elements can be used to explain how to read the content in the graph as well as how to read the graph itself. Amanda Cox, the Data Editor at the *New York Times*, once said that "The annotation layer is the most important thing we do . . . otherwise it's a case of 'here it is, you go figure it out.'"

Adding the right annotations to a graph can be vitally important to your reader's comprehension. There are three ways we can integrate our graphs and our visuals: removing legends, creating active titles, and adding detail.

1. REMOVE LEGENDS WHEN POSSIBLE AND LABEL DATA DIRECTLY

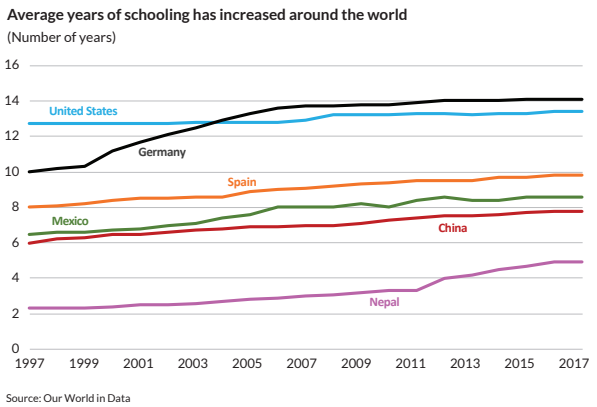
Let's start with the easiest type of annotation: Removing legends and directly labeling your data. Many software tool defaults create a data legend and place it around the chart, disconnected from the data. This forces more work upon your reader to connect each line or bar to its label. A better approach is to directly label your data series.

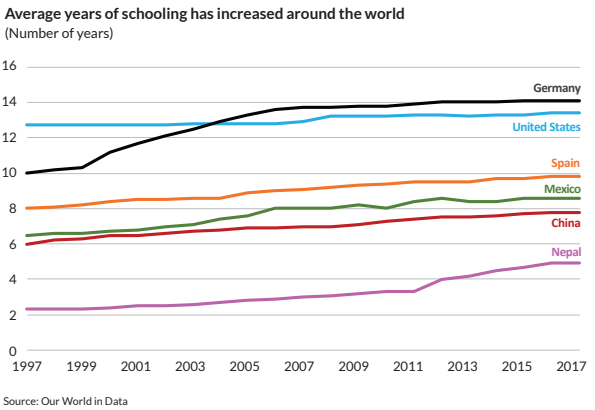
Take the line chart of average schooling for fifty countries from earlier. Rather than the default approach of putting a legend somewhere around the chart, as in the graph on the left, in the version on the right, I directly label the lines at the right end of the graph.



Help your reader more easily find the labels for the data values by placing the labels directly on the chart.

In graphs that have fewer lines, we might also be able to place the labels directly on the graph. In these cases, I try to align the labels instead of placing them in random positions. Notice how in the graph on the left of the next page your eye needs to jump around to find each label. And because we might start reading the graph with the title, the proximity of the label for the United States could give that series greater emphasis. In the version on the right, the labels are aligned along a single vertical line, making it easier to read the entire visual.

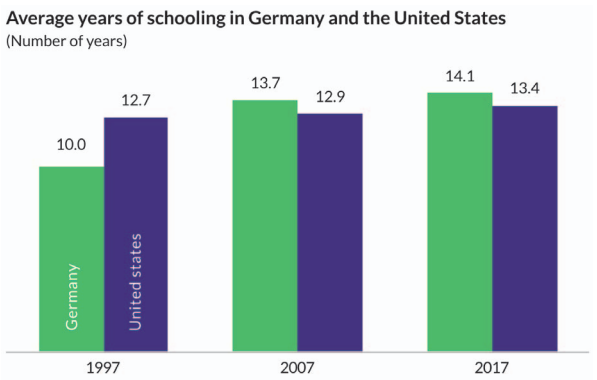


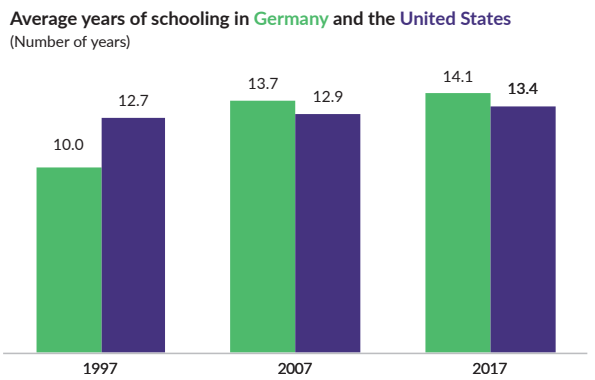


Align the labels and match the colors with the data as in the graph on the right rather than placing them in random positions.

We can take a similar approach to labeling this bar chart of average schooling in Germany and the United States. With just two countries, instead of a legend disconnected from the data, what if we added the labels inside the bars or used color in the title of the graph itself to link the title to the content of the graph?

By integrating the text and the data, we're doing a better job of considering the reader's needs. Do they need to see *every* line equally, or will including all the lines clutter the graph? Is it important to label *every* point in the scatterplot, or will highlighting just a few points suffice? How can we integrate labels and chart elements to help the reader understand the content quickly and easily?



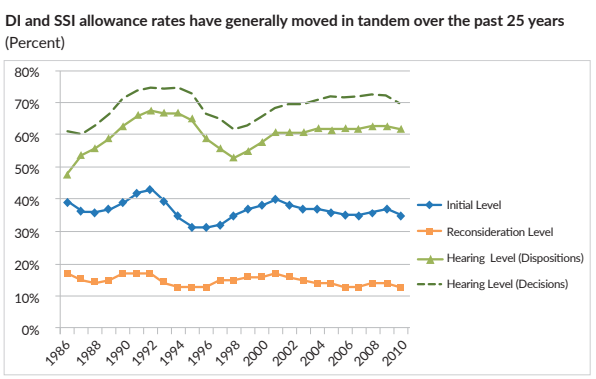


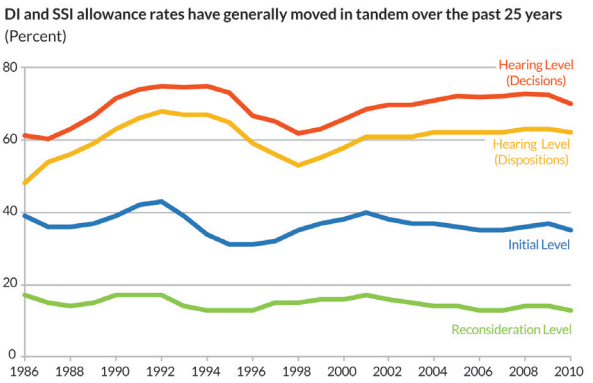
These are just two examples of how to integrate labels into the graph.

Data Source: Our World in Data.

Removing the legend isn't always possible. A bar chart with several categories or a map with different colors requires a legend, because directly labeling the chart will add too much clutter to the visual. In these cases, at least keep the order of the legend consistent with the order of the data. Notice the inconsistency between the order of the lines and the legend in this line chart from the Social Security Advisory Board. Not only do we need to jump back and forth between the lines and labels, there is an extra task of figuring out the order of the two. A redesigned version removes much of those unnecessary data markers and extra gridlines, and integrates the legend onto the chart by adding labels directly next to the lines.

We won't be able to remove the legend on every single graph we create, but we should strive to link the data and the labels as best we can, and that starts with labeling the data series on our charts.





Notice the inconsistency between the order of the lines and the legend in the chart on the left. The redesigned version removes unnecessary clutter and directly labels the lines.

Source: Social Security Advisory Board, February, 2012.

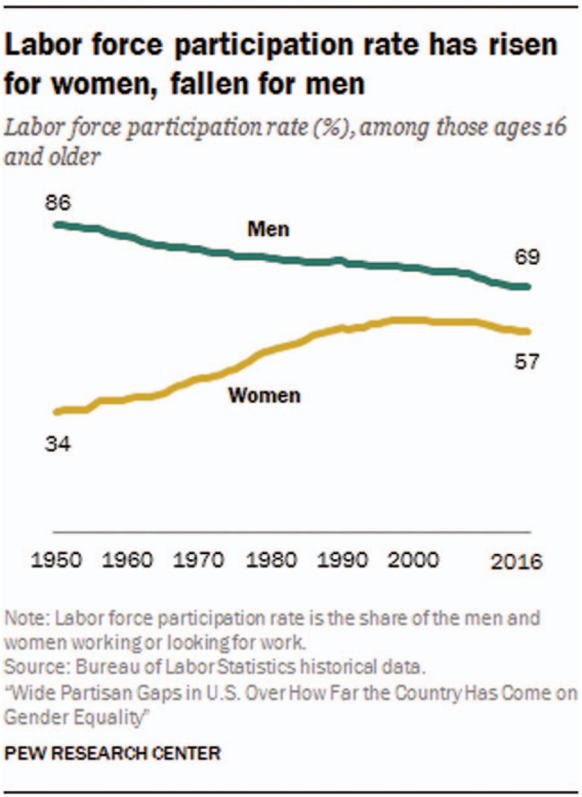
2. WRITE THE TITLE LIKE A NEWSPAPER HEADLINE

Most titles are neutral descriptions of the data, as in "Figure 1. Labor Force Participation Rate, Men and Women, 1950–2016." But better titles capture the takeaway of the chart, telling the reader what conclusions can be drawn from the data. I call these "active titles" or "headline titles." In my book on presentations, I follow the advice of author Carmine Gallo and urge presenters to use "Twitter-like headlines" in their slides. These are concise, active phrases that make it easy to understand what the slide—or chart—is aiming toward.

Too often, we attach a title to the chart that describes the data instead of the point or argument we want to make.

While "Labor Force Participation, Men and Women, 1950–2016" is certainly a correct and accurate description of the data in this graph from the Pew Research Center, it does not describe *what the reader should learn* about the labor participation rate among men and women between 1950 and 2016. The more active title that Pew uses instead—"Labor force participation rate has risen for women, fallen for men"—tells the reader exactly what they should take away from the graph.

Do people even read titles? A 2015 study from researchers at Harvard University showed that they do: "Titles and text attract people's attention, are dwelled upon during encoding,



The active title in this chart from the Pew Research Center tells you exactly what you are supposed to learn from it.

and . . . contribute to recognition and recall." If it's indeed the case that people read titles (and text more generally), then we should treat a chart's title as carefully as the chart itself.

But can so-called "active" titles make us seem biased or partisan? If we use active titles only to faithfully represent the results and showcase the message of the graph, then no. I've worked with many people who have debated my inclination for active titles by arguing that such titles will make their work appear biased. In most of those cases, I can look at the text around their chart and see a single argument for what's being shown in the graph and how to interpret the data. Their argument is right next to the graphic, but, like the legend we saw earlier, it's disconnected from it.

Active titles don't make us biased, but descriptive titles do waste an opportunity to make a clear, compelling case. Of course, short, active titles aren't always possible—you may be making more than one point or your sole goal is to simply describe the data. Generally speaking, however, integrating your graphs as part of your argument creates a more cohesive approach to making your argument and telling your story.

In the case above, Pew doesn't leave it to the reader to search for a point in the graph, but neither are they biasing the results by adding commentary in the title. They are simply foregrounding the takeaway of the visual.

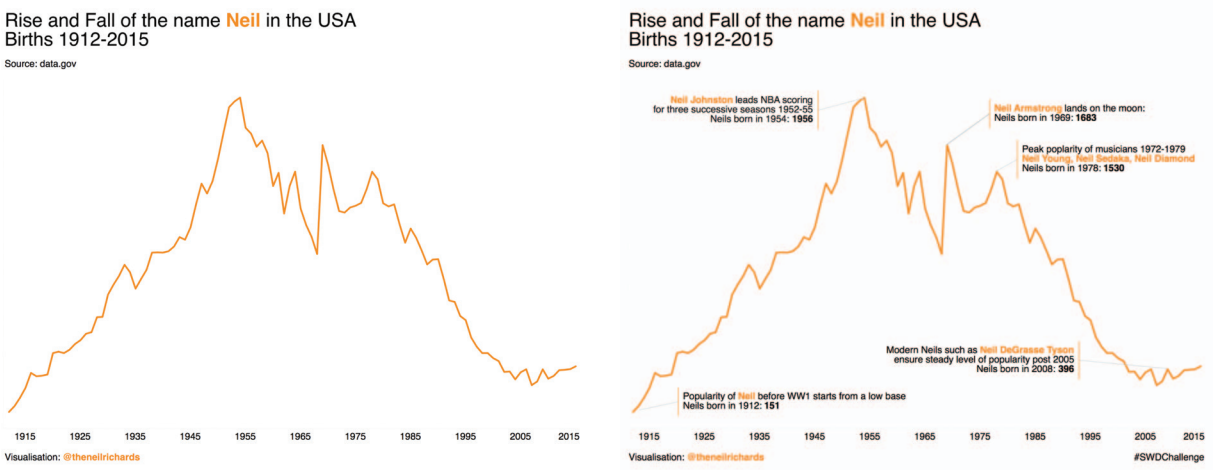
If you are having trouble coming up with a concise, active title, that may be a sign that your chart doesn't have a concise takeaway or—and maybe this is more common—you haven't thought through what you want the graph to communicate.

3. ADD EXPLAINERS

Once the chart is made and the title is settled, ask yourself, Would this chart benefit from more text?

Sometimes data sets have peaks or valleys, outliers or variations that bear explanation. Adding detail in graphs can push your argument, highlight points of interest, or (in cases of nonstandard graphs) even explain how to read it.

Take this line chart of the popularity of the name "Neil" in the United States created by, yes, Neil Richards, a data visualization consultant in the United Kingdom. Anyone could make the simple line chart on the left—it's only one data series—but with only a quick glance the reader might immediately ask some obvious questions: Why did the decline stop in the late-1960s? Why did the line spike upwards a few years later in the 1970s? And what halted the decline in the early 2000s?



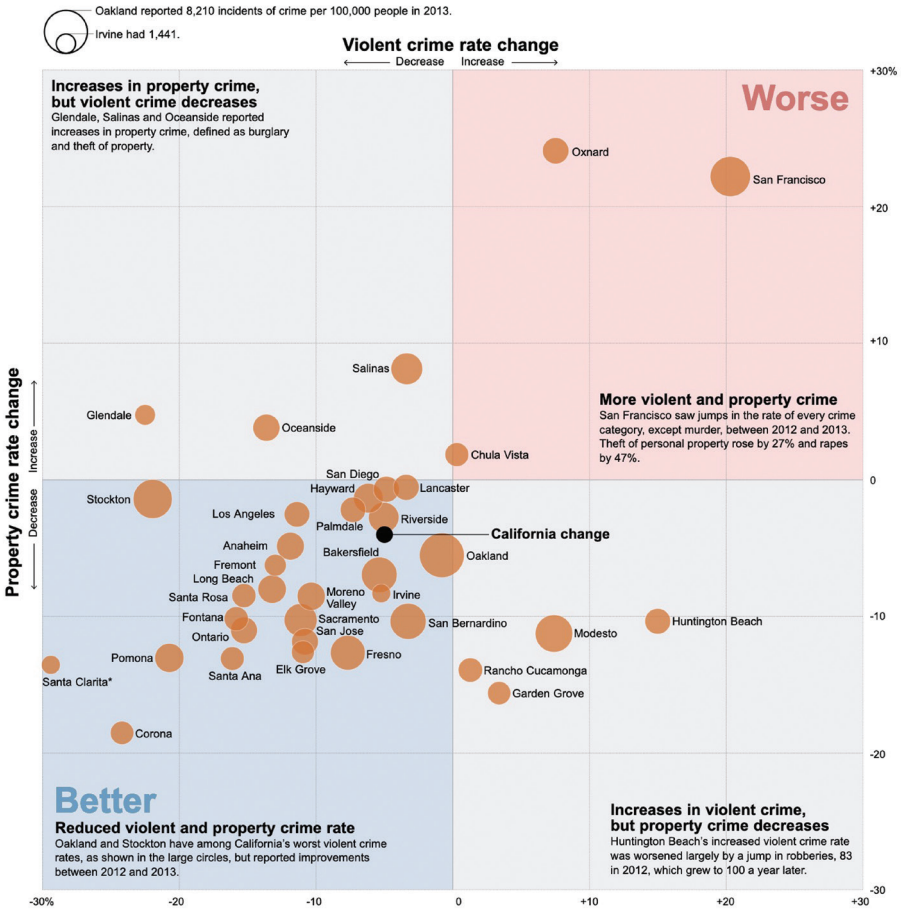
Short explainers in this graph on the right from Neil Richards explain some of the basic features of the data.

Now look at this second version of the chart with short explainers. The late-1960s spike might be attributed to Neil Armstrong landing on the moon, followed by the popularity of musicians like Neil Young, Neil Sedaka, and Neil Diamond in the 1970s. The flattening of the trend in the mid-2000s could be attributed to "modern Neils" like Neil DeGrasse Tyson. These annotations are not complicated and don't require complex programming or design techniques—they are often just interesting points in the data thoughtfully added with short sections of text.

Annotation allows readers—especially those who may have less experience with data visualization—to grasp and understand the content quickly. The bubble chart from the *Los Angeles Times* on the next page is a great illustration of how to do so. The change in the violent crime rate is plotted along the horizontal axis and the change in the property crime rate is plotted along the vertical axis for about thirty-five cities in California. The average *LA Times* reader is probably not a bubble-plot expert, so the authors have added annotations to help readers navigate the format of the graph and its content.

Notice the use of color and annotation to help the reader understand this graph. The top-right quadrant is shaded red with the word "Worse" in large, red letters. The bottom-left quadrant is shaded blue with the word "Better" in large, blue letters. Immediately, you

understand that the cities in the top-right have worsened and the cities in the bottom-left have improved. Short, bold headlines ("Reduced violent and property crime rate" in the bottom-left quadrant) explain the substance of the changes. Then, a short sentence highlights a city or two and what has transpired over the past year. This graph does an expert job of explaining *how to read* it and *how to understand the content* in it.

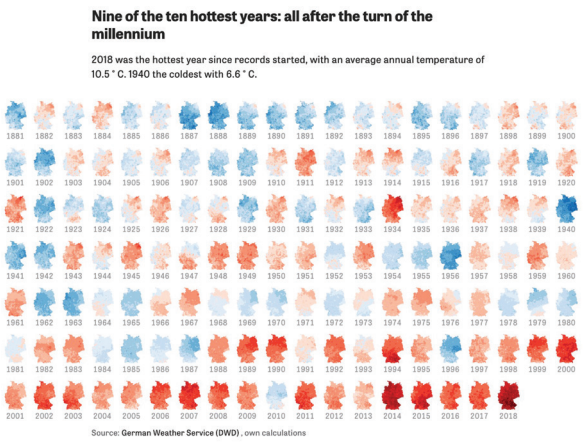


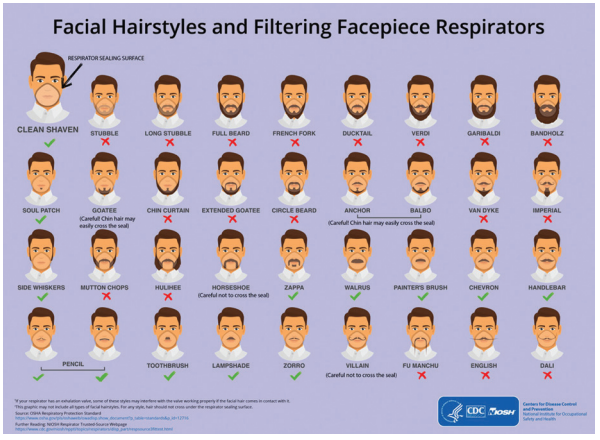
This graph from the *Los Angeles Times* is one of my favorite examples of how to use annotation to explain how to read a graph and its content.

In a 2016 interview, John Burn-Murdoch, an interactive data journalist at the *Financial Times* said, "The annotation layer is where the 'journalism' really comes into 'visual journalism.' Making a graphic is the equivalent to interviewing your source. But it's then your job to actually pick out . . . the bits the reader should know about." Not everyone is a journalist, but everyone can find ways to help our readers clearly see what's important and what we want them to learn.

GUIDELINE 4. AVOID THE SPAGHETTI CHART

It's obvious when a certain graph contains too much information—line charts that look like spaghetti, maps with dozens of colors and icons, or bar after bar after bar littering a chart. Sometimes we face the challenge of including lots of data in a single graph but we don't need to try to pack everything into a single graph.





Two examples of the small multiples approach. The graph on the left, from Zeit Online, shows the average temperature in Germany over the past 140 years. The graph on the right, from the Centers of Disease Control and Prevention, shows how facial hair can affect the fit of respirators. The Gestalt principle of *connection* helps us track the changes from one image to the next in both graphics.

One way to address the packed, single chart is to break it into smaller parts. Known as grid charts or panel charts (also called *facets, trellis charts*, or, most commonly, *small multiples*) these are smaller charts that use the same scale, axes, and scope but spread the data across multiple visuals. In other words, instead of putting all of the data on one graph, create multiple, smaller versions with variations on the basic data.

The small multiples approach isn't a new or revolutionary approach to communicating data. In 1878, photographer Eadweard Muybridge (see page 44) was tasked with determining whether a horse becomes fully airborne when it gallops. Muybridge developed a technique to take a sequence of fast-action photographs (what we now call stop-motion) of a horse at gallop. His photos proved that horses do indeed leave the ground entirely. The sequence of images, which also gives a sense of motion and animation, is an early example of small multiples.

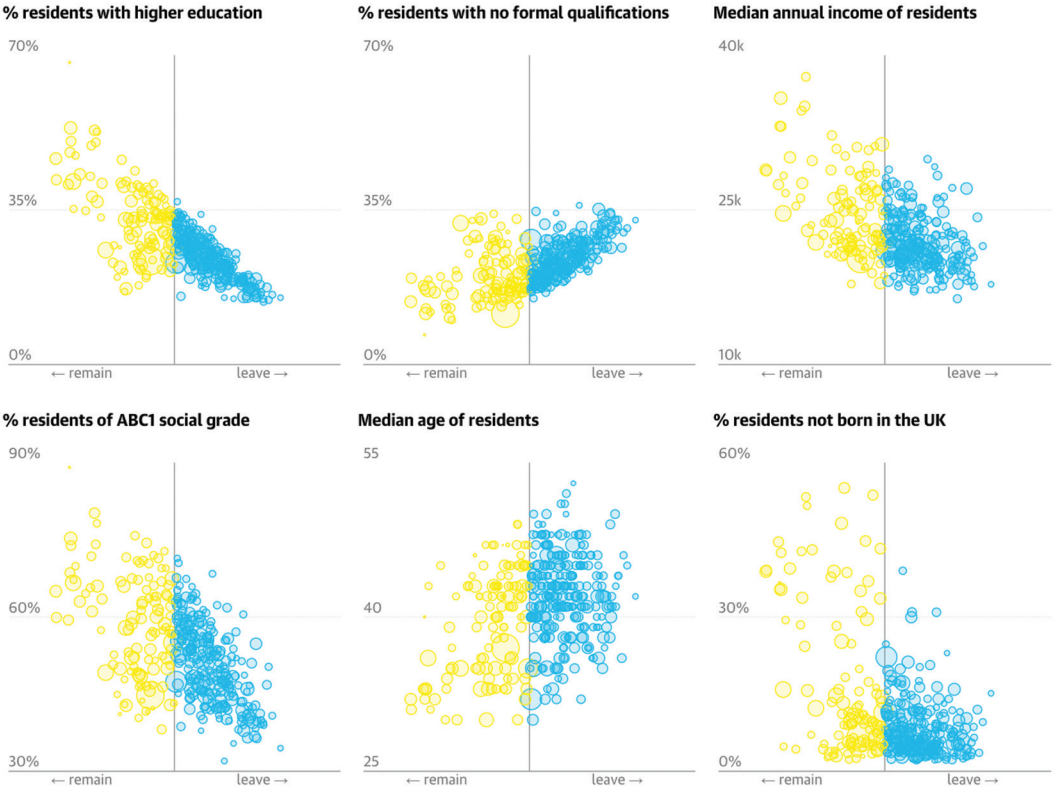
The small multiples approach has at least three advantages. First, once the reader understands how to read one chart, they know how to read all the charts. Second, you can display lots of information without confusing your reader. Third, small multiples let readers make comparisons across multiple variables. This example from the *Guardian* shows voting results in 2016 for the Brexit resolution in the United Kingdom across six different demographic variables. The horizontal axes stay unchanged and it's easy to see the direction of the relationships for each demographic measure.

But there are pitfalls to the small multiples approach that will muddy the visual if not avoided. First, the charts should be arranged in a logical order. Don't make your reader navigate around the page—use an intuitive arrangement based on something like geography or alphabetical order.

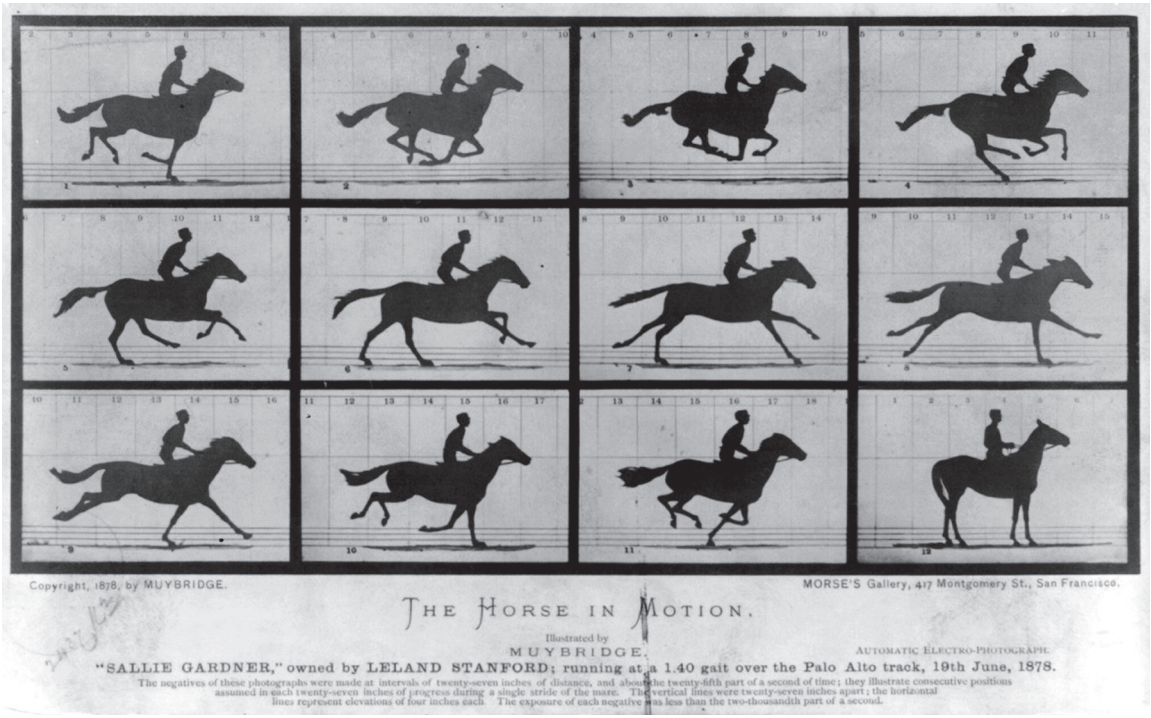
Second, the graphs should share the same layout, size, font, and color. Remember, we are breaking up one chart into many, so it should look like one chart replicated multiple times. The vertical and horizontal axes may change, but you wouldn't want to have one chart where blue dots represent "no" and another where they represent "yes." Third, small multiples should be relatively easy to read. You are not necessarily asking your reader to zoom in and uncover all the specific details in all of the graphs—the purpose is to give them a view of the overall patterns. The graphs are intended to be small, so including annotations and labels or repeating long axis labels and data markers features can overwhelm the reader.

Every area by key demographics

Comparing the results to key demographic characteristics of the local authority areas, some patterns emerge more clearly than others. The best predictor of a vote for remain is the proportion of residents who have a degree. In many cases where there are outliers to a trend, the exceptions are in Scotland.



Small multiple scatterplots from the *Guardian* shows the relationship between voting choice and six demographic variables. Notice how the Gestalt principle of *similarity* lets us easily see the two clusters of circles within each scatterplot.

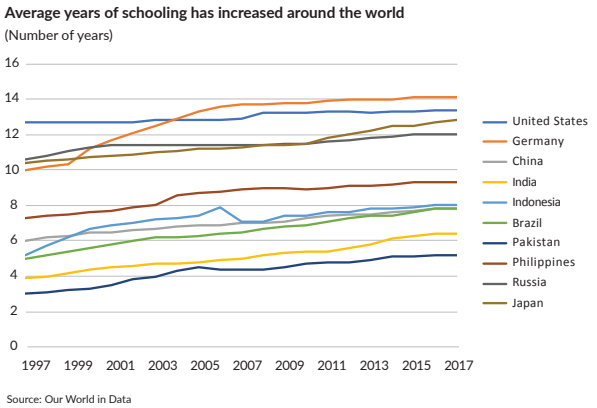


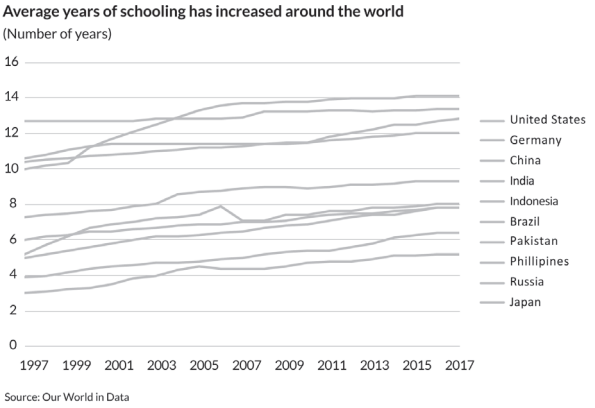
Photographer Eadweard Muybridge used the small-multiples approach back in 1878 to determine whether a horse becomes fully airborne when it gallops.

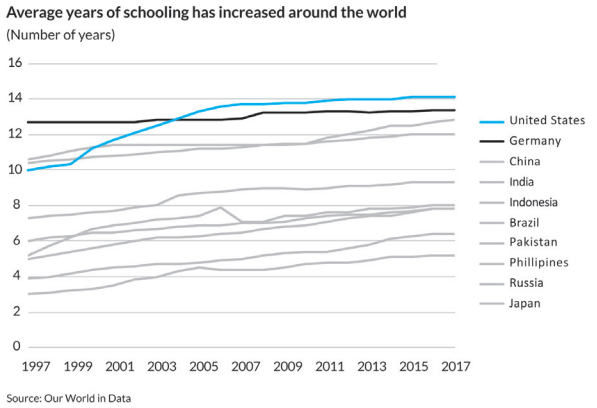
GUIDELINE 5. START WITH GRAY

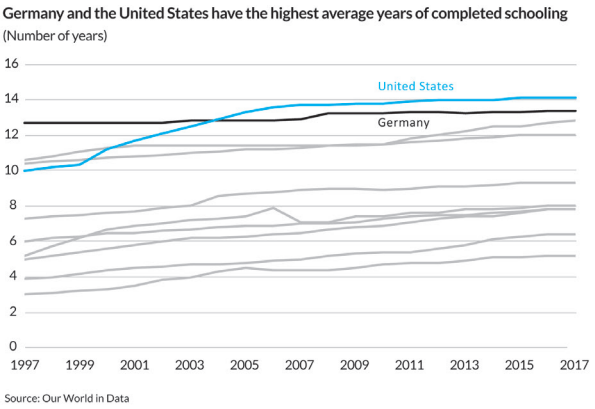
I end this section with a practical technique that I think can be an easy step to creating clear, comprehensible visualizations: *Start with gray*. Whenever you make a graph, start with all-gray data elements. By doing so, you force yourself to be purposeful and strategic in your use of color, labels, and other elements.

Consider a simpler version of the average schooling chart from earlier, this time with only ten countries as shown on the next page. With color and labels (top-left), I could put this graph in my report or handout, and with a little work (and a more active title), my reader could figure out which labels correspond to which lines. But if I make all the lines gray (top-right), the reader can't accomplish that same task because it's impossible to figure out which country is represented by which line.









Starting your graphs with all-gray data elements forces you to make purposeful, strategic decisions about where you want to direct your reader's attention.

Now I can be purposeful about what I want to do with this graph. I could add color and even vary the thickness of the lines to better highlight only, say, the two countries I want to emphasize. (Leaving all the labels in the version on the bottom-left is less useful than labeling the lines directly as in the version on the right.) Starting with gray forces us to deliberately choose what elements to put into the foreground.

DATA TYPES

The bedrock of any data visualization is the data. Without data and a good understanding of what our data is, how it was collected, and what it tells us, we are just painting pictures. This book is not the venue for a thorough review of data types and statistical methods, but a short primer can help us organize our data types just as we organize our graph types.

There are two major groupings of data types: quantitative and qualitative. Quantitative data can be measured with *numbers*, for example, distance, dollars, speed, and time. Qualitative data is *non-numerical* information, usually descriptive text like "yes or no," "satisfied or dissatisfied," or longer quotes or passages from interviews and surveys.

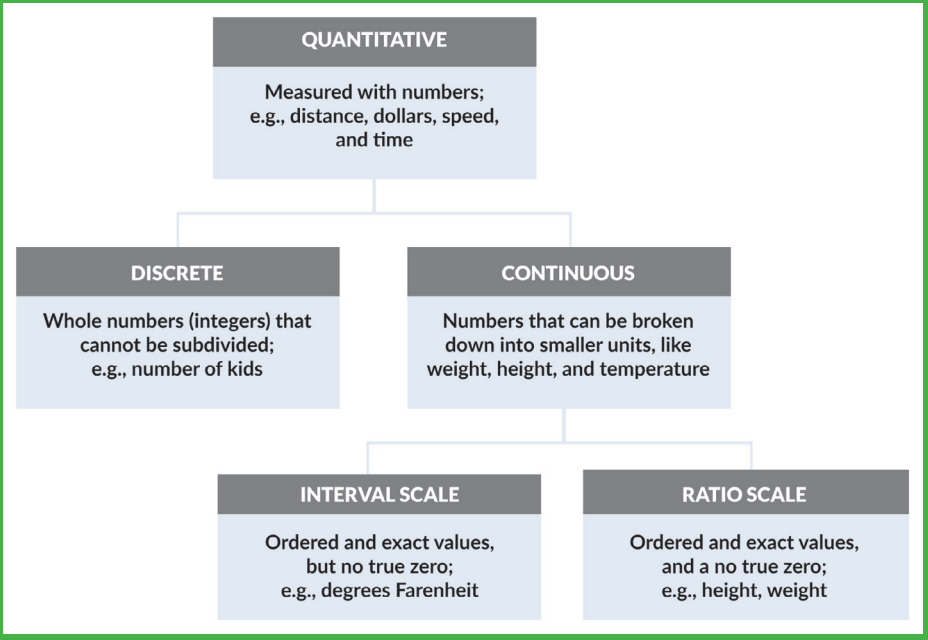
We can further break down each major data type into subcategories. On the qualitative side, we have *nominal* and *ordinal* scales. *Nominal* scales are used to label variables and don't have an order or quantitative value. In a data set of animal types, the order of lion, tiger, and bear has no meaning (aside from the song, of course). In *ordinal* scales, order does matter, but the exact size in comparison between values is unknown. Consider a survey that asks people to select between *1. Strongly Agree; 2. Agree; 3. Disagree;* and *4. Strongly Disagree*. These choices can be ordered, but the difference between 1 and 2 is not necessarily the same as the difference between 3 and 4.

On the quantitative side, data can be either *discrete* or *continuous*. Discrete data are whole numbers (integers) that cannot be subdivided. Despite national averages, no one has exactly 2.3 children. Continuous data are numbers that *can* be broken down into smaller units, like weight, height, and temperature.

Continuous data can be further broken down into two major scales: *interval* and *ratio*. The difference is what we can and cannot calculate. With *interval* scales, we know both the order and the exact differences, but they do not have a true zero value. This means we can add and subtract data measured in interval scales, but we can't multiply or divide. A classic example is temperature in degrees Farenheit: The difference between 10 and 20 degrees is the same as between 70 and 80 degrees, but we can't say that 20 degrees is twice as hot as 10 degrees, because 0 degrees is an actual value, not absolute zero.

Ratio scales have all of the characteristics of all the other scales *plus* they have an absolute zero, which means we can do all of our mathematical calculations.

Weight is a good example of a ratio scale—a person who weighs 200 pounds is twice as heavy as someone who weighs 100 pounds, and 0 pounds is the absence of weight.



DATA EQUALITY & RESPONSIBILITY

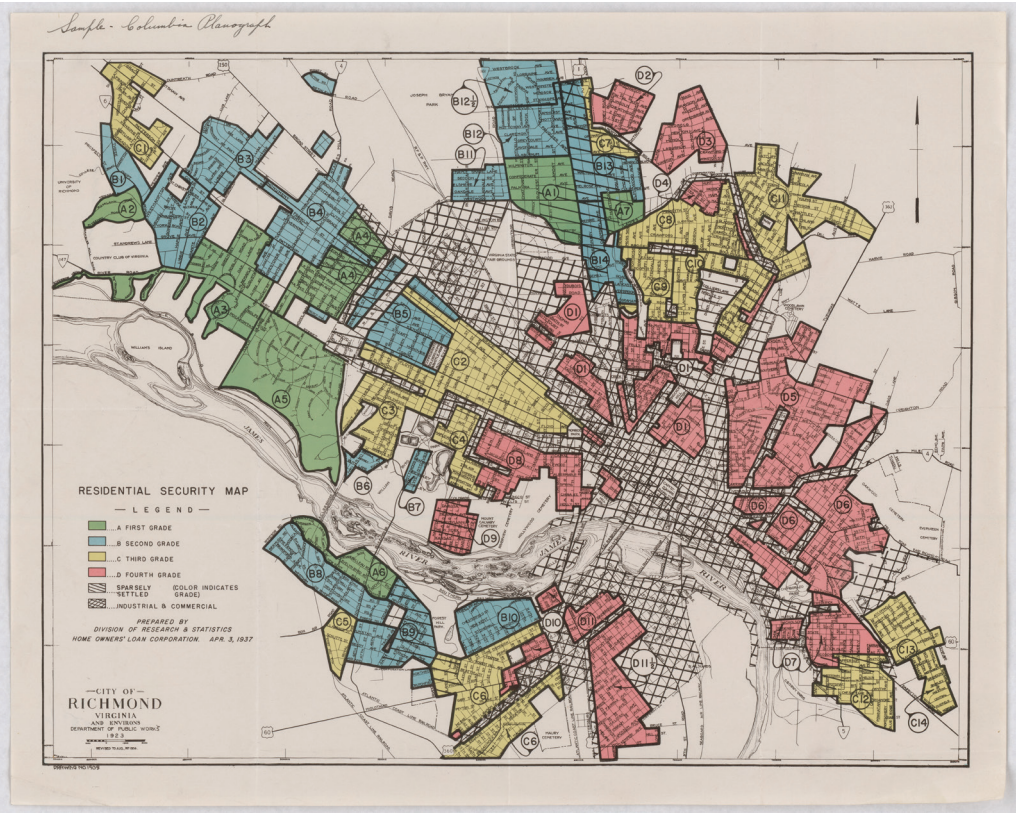
These guidelines lay out the basic approaches to effectively visualizing our data. While this is not a book about data *analysis*—how and where to get data, how to analyze underlying statistical properties, and develop statistical models—whenever we work with data it is important to recognize that visual content can have a large influence on how people use data and make decisions. As data communicators, it is therefore our responsibility to treat our work and our data as carefully and objectively as possible. It is also our responsibility to recognize where our data may suffer from underlying bias or error, or even implicit bias that data creators may themselves not even be aware of.

There are many ways in which the data we use may be biased or not representative. In their book, *Data Feminism*, Catherine D'Ignazio and Lauren Klein describe how standard practices in data science reinforce existing power inequalities. They explore how data has been used for both good and evil—to expose injustice and improve health and policy outcomes, for example, but also to surveil and discriminate. By asking who is producing the data and for whom it is being produced, we can be better stewards of our own data and our own visualizations.

Many fields are squarely built on a model of the world in which men are the only—or maybe just the most important—participants. In *Invisible Women*, author Caroline Criado Perez reveals the hidden places where inequality in even basic data resides. There are straightforward examples, like how the average smartphone is 5.5 inches long—too big for most women's hands and pants pockets. Or how the average temperature in many office buildings is five degrees too cold for women because the formula to determine the ideal temperature was developed in the 1960s based on the metabolic resting rate of a forty-year-old, 150-pound man. There are more insipient examples as well, like how women in Britain are 50 percent more likely to be misdiagnosed following a heart attack, or how car crash test dummies are based on the male body, so even though men are more likely to get into car accidents, women involved in collisions are almost 50 percent more likely to be seriously injured.

In a similar vein, the era of big data, machine learning, and artificial intelligence use more and more unseen algorithms and statistical techniques. We often know little about the data that feed these algorithms and how the models themselves may perpetuate inequality. Mathematician Cathy O'Neil explores this in her book *Weapons of Math Destruction*, from teacher quality, creditworthiness, and recidivism risk, algorithms can develop and reinforce discriminatory models of public policy.

When it comes to data visualization specifically, we must be mindful of the underlying biases and inequality in how we present our results. As just one example of how data and visualizations have been used to discriminate, consider this map of Richmond, Virginia, produced in 1937 by the Home Owners' Loan Corporation (HOLC), a federal agency tasked to appraise home values and neighborhoods across the United States. As Richard Rothstein writes in his book, *The Color of Law*, "The HOLC created color-coded maps of every metropolitan area in the nation, with the safest neighborhoods colored green and the riskiest colored red. A neighborhood earned a red color if African Americans lived in it, even if it was a solid middle-class neighborhood of



This redlining map of Richmond, Virginia, demonstrates how data and data visualization can be wielded to further systematic discrimination. Information Studies scholar Safiya Umoja Nobel argues that modern internet search engines and other algorithms are enacting new ways of discrimination and racial profiling, creating a modern form of "technological redlining."

Source: National Archives.

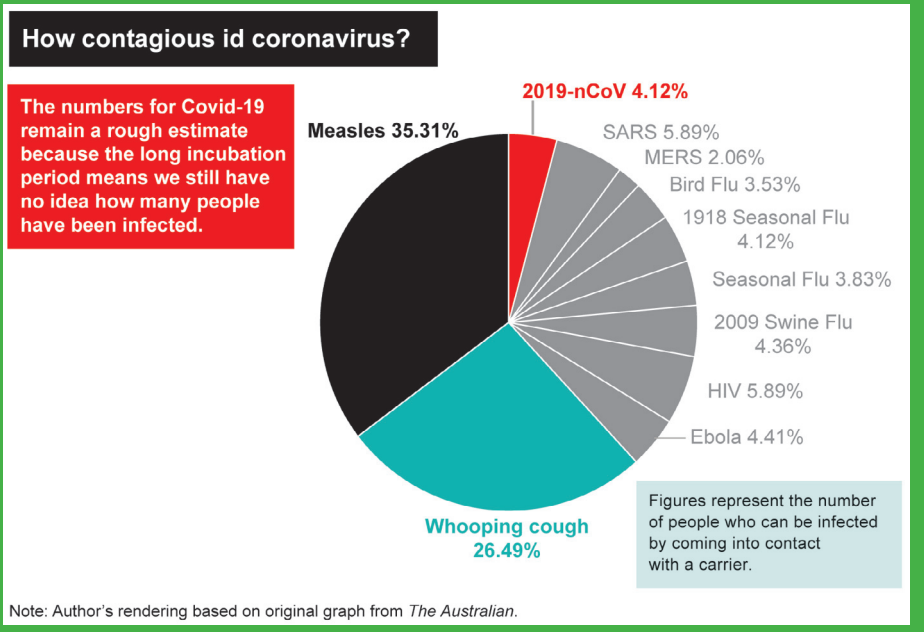
single-family homes." Systematic discrimination is and can be generated by how we use and misuse our data.

Finally, in addition to cultural differences that might arise from, say, using certain colors in different cultures, we should also be mindful of the language, shapes, and images in our visuals. Are we using language and images that are inclusive? When do we need to provide historical and social context for problems people are facing? As with developments in accessibility, diversity, and inclusion (see Chapter 12), these are all challenges with which the data visualization field is always wrestling.

DATA VISUALIZATION LESSONS FROM THE CORONAVIRUS PANDEMIC

The manuscript for this book was delivered to the publisher in March 2020, just as the coronavirus (COVID-19) pandemic was traveling around the world. As it spread, it induced massive political, economic, and societal changes, and it brought new terms into our lexicon, like "flattening the curve." In late February 2020 *The Economist* published a version of the graphic by data journalist Rosamund Pearce, based on original work by the Centers for Disease Control and Prevention. Graphics like this built awareness and facilitated action, most notably on the relatively new concept of "social distancing".

But for every graph that informed and educated, there were many others that misrepresented data or spread misinformation. One pie chart, for example, inexplicably summed contagion rates among eleven separate diseases to 100 percent and added a separate note that the "numbers for COVID-19 remain a rough estimate because



Note: Author's rendering based on original graph from *The Australian*.

the long incubation period means we still have no idea how many people have been infected." This is not responsible data visualization.

The unprecedented spread of the coronavirus gave us an opportunity to use real-time data that could be used to better understand the virus and its spread. But one reason why many graphs and charts around COVID-19 are problematic is because too many of us assume we have adequate knowledge in a particular subject area. Public health professionals, epidemiologists, and physicians have the training, insight, and experience with the health care system and modeling disease transmission to provide useful data and information. For the rest of us, without expertise in these areas, our visualization work—even as well-intentioned as it might be—can make things worse.

We often create—or are asked to create—visualizations in subject areas in which we are not experts. Sometimes this is an opportunity to explore different visualization forms and functions and try new tools. Other times, though, we may be out of our depth. We may not fully understand our data. Even if we have read the data dictionary or considered the data collection methods, we may not know enough about how the data were modeled or simulated or the reliability of their collection methods.

Under ordinary circumstances, visualization exercises might consider issues such as unemployment rates or housing options or the distribution of wealth and not life-threatening events like a viral pandemic. In these cases, we must be especially aware of how our work might be misunderstood and how it may change the thinking or behavior of our readers.

The converse of the above is also true. An epidemiologist may know a lot about modeling disease spread, but he or she may not understand how best to visualize that modeling, explain jargon, and annotate important data points. Here, it is incumbent upon the scientist to reach out to data visualization experts and graphic designers to ensure their visualization work is accessible to readers.

There is a better way forward. Instead of thinking our limited knowledge is sufficient to weigh in on every topic and every dataset, we should strive to collaborate. In the case of COVID-19, not knowing enough may lead to deadly outcomes. If we think of ourselves as journalists and seek out domain specific experts, we can work to build teams, groups, and organizations that can deliver better data, better visualizations, and better decisions.

NEXT STEPS

Now armed with basic guidelines and rules of perception, you are almost ready to start adding more graphs to your data visualization toolbox. But there's one more thing you should consider before you start encoding your data with bars, lines, and dots: the purpose of your graph.

In what format do you need to present your data to your reader or user? Do they need a static graph where you present your argument or will an interactive visualization help them explore the data and come to more and deeper conclusions? In the next chapter, we discuss the different forms and functions for visualizations and then turn to the many ways we can visualize our data.

FORM AND FUNCTION

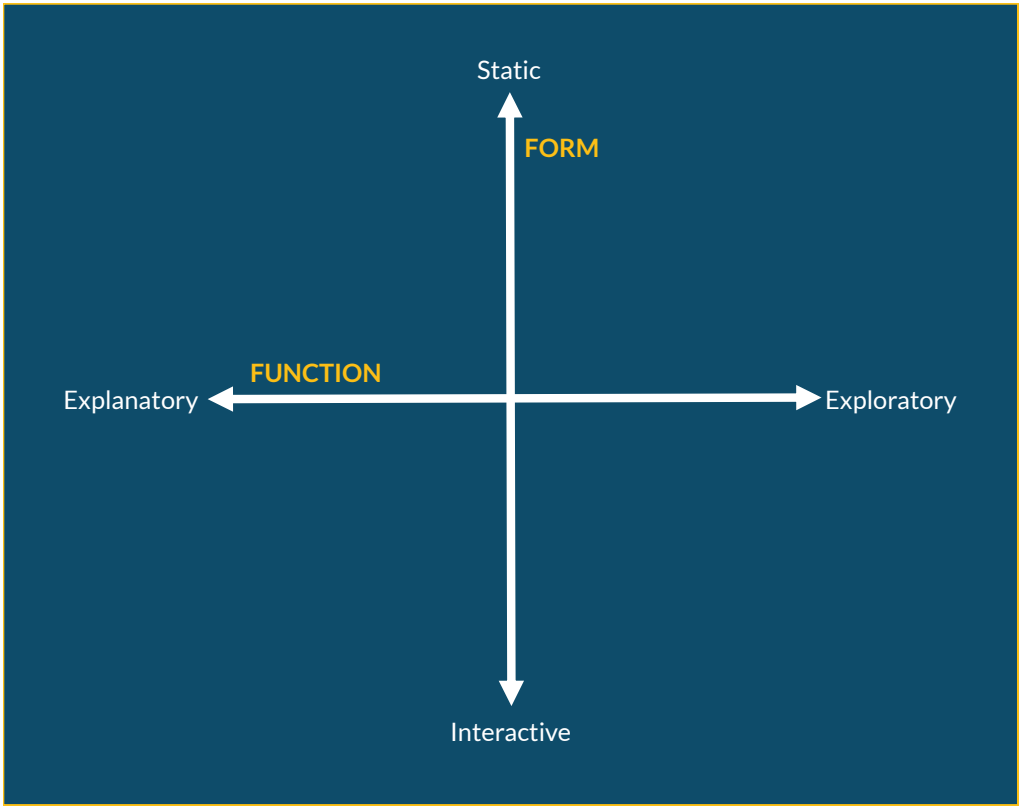
LET YOUR AUDIENCE'S NEEDS DRIVE YOUR DATA VISUALIZATION CHOICES

The primary focus of this book is on static data visualizations. That is, those visualizations that are not interactive and do not move. But interactive and animated visualizations are becoming more common even in academic fields, and the tools to make interactive visuals are, year by year, becoming easier to use, cheaper (even free), and more powerful.

Aside from the standard design decisions about color, layout, and font that apply to static visualizations, there is another set for interactive visualizations. Where will the button go? What will it look like? How will the user be guided to different parts of the visualization? Will the user need to scroll? If so, in which direction? How might the interactivity change between a desktop computer and a tablet or phone?

Discussing all of these options for storytelling, interactivity, and animation is beyond the scope of this book, but it is worth pausing to consider which graph—static or interactive—best meets the needs of our audience. Some audiences want a long, in-depth report. Others want a short brief or a blog post. And some just want to get their hands on the data to do their own work.

The broad schematic of different data visualization types shown on the next page can help you think through what kind of visualization is best for your audience. This space has two perpendicular lines. The vertical line is a spectrum of the general **FORM** of the visualization and runs from *static* to *interactive*.



All visualizations fall somewhere on the spectrums of explanatory or exploratory and static or interactive.

Static visualizations provide all the information at once and do not move. These are, for example, your basic line, bar, and pie charts.

Interactive visualizations allow a transfer of information between the user and the interface. The user clicks, swipes, or hovers, and something happens or additional information is made visible.

Sitting somewhere between the two are *animated visualizations*. These—for example, animated GIFs (Graphics Interchange Format), movies, online slideshows—do not necessarily permit the user to manipulate data points to create alternative results, but they might let the user control the pace or transitions of the visuals.

On the horizontal spectrum is the **FUNCTION** of the visualization. Here, visuals run from *explanatory* to *exploratory* visualizations.

Explanatory visualizations bring the results to the forefront; they convey the author's hypothesis or argument to the reader.

Exploratory visualizations encourage the user to explore the data or subject matter to uncover findings for themselves.

The intersection results in four quadrants.

STATIC AND EXPLANATORY

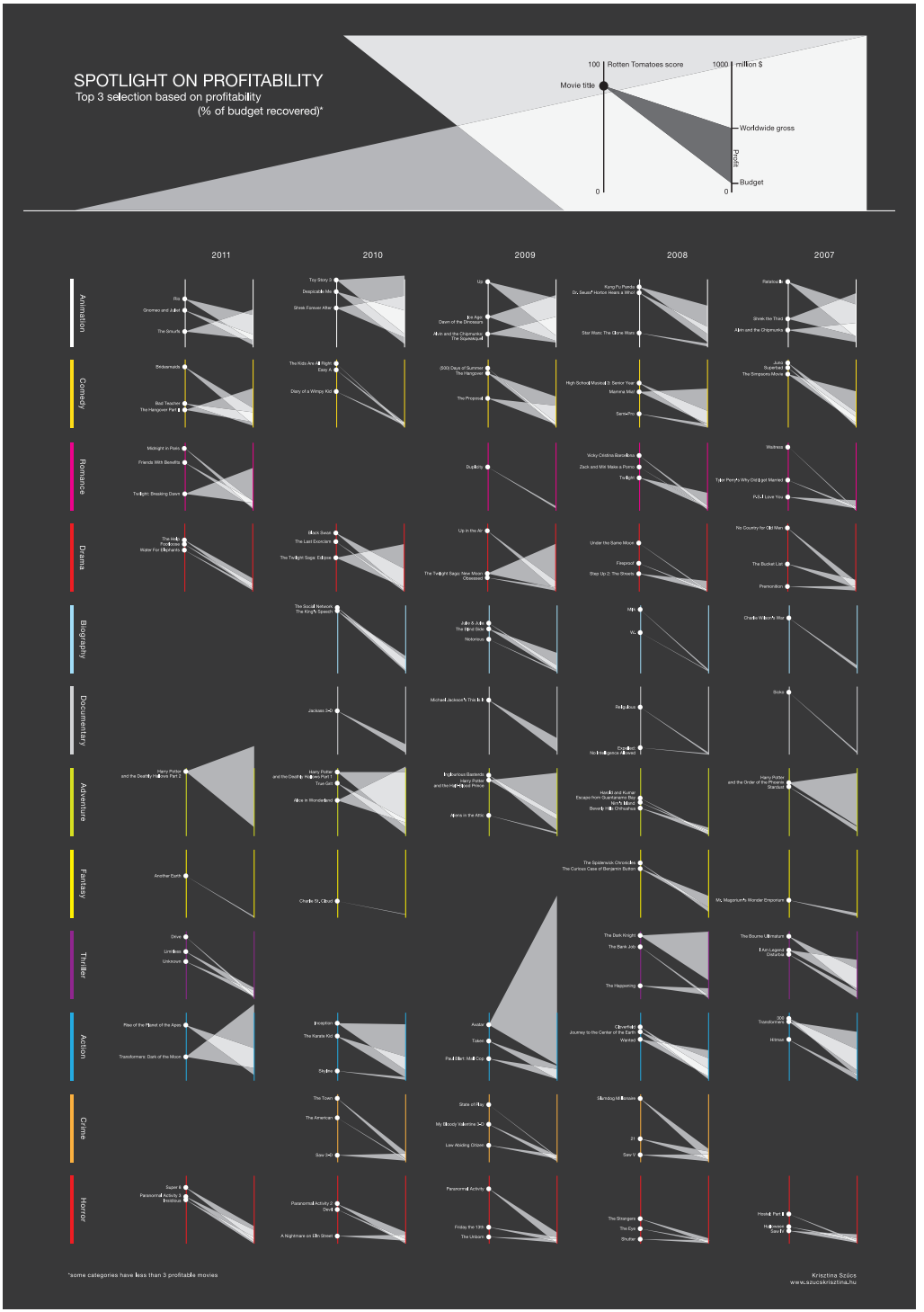
Graphs that are not interactive or do not move are typically used to illustrate a point or reinforce an argument made in the text or presentation, like a line or bar chart. These are also the visualizations we might make on our own or within our organization as we explore our data and develop our findings. This book is primarily concerned with these visuals, and we'll explore a wide array in depth in Chapters 4 through 11.

STATIC AND EXPLORATORY

These visualizations let readers interpret the data and find their own results in a static display of information. The infographic on the next page from designer Krisztina Szűcs, for example, encourages us to explore the data on our own. In each graph, the left vertical axis shows a movie's Rotten Tomatoes score (a critical score for movie quality). The right vertical axis shows the profitability of each movie, the gap between its gross revenues and budget. She doesn't make a specific argument or point out specific details. Instead, she lets you explore the data to draw your own conclusions.

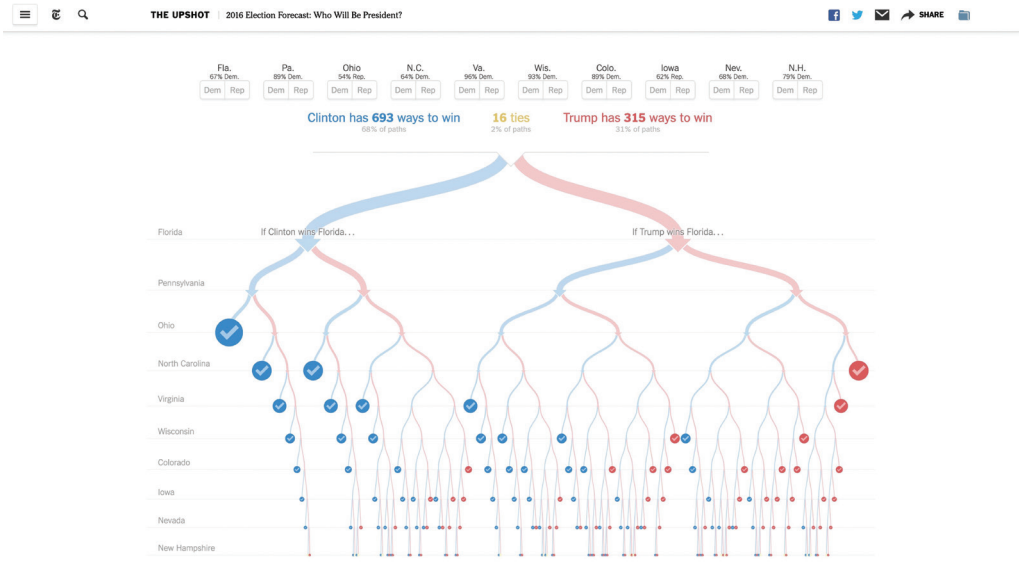
INTERACTIVE AND EXPLANATORY

Perhaps the easiest explanatory-interactive graph to consider is a static graph that has an interactive hover or rollover layered on top of it. While supremely popular just a few years ago, this approach seems to be falling out of favor as readers move toward smartphones and the need to click and hover is less important. Interactivity can engage a reader to experience a story or explain a process. To take just one example, consider the *New York Times*' "Paths



This static infographic from designer Krisztina Szűcs encourages us to explore the data on our own.

to the White House" graphic, which lets the user pick the paths by which either presidential candidate could win the election.

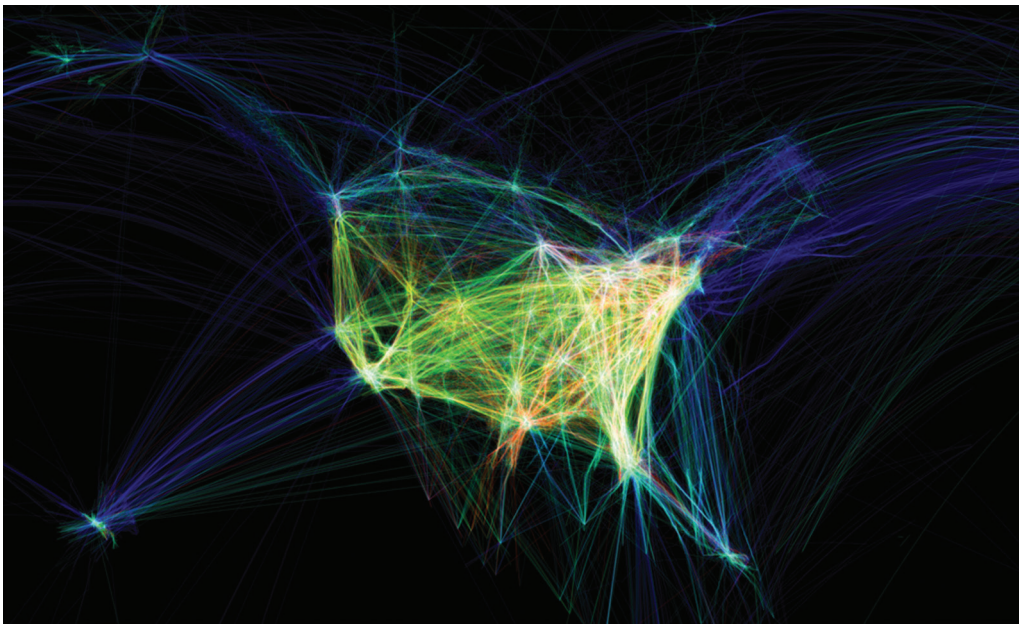


This "Paths to the White House" graphic lets the user pick the paths by which either presidential candidate could win the election.
Source: *New York Times*.

INTERACTIVE AND EXPLORATORY

These visualizations graphically present a complete data set and ask users to find interesting patterns or stories. As one example, Aaron Koblin's *Flight Patterns* project shows all flights in U.S. airspace in a twenty-four-hour period. Videos and snapshots (including his 2011 TED talk) enables users to zoom in, pause, rewind, and otherwise explore the visualization. Other times, the purpose of these graphics is to give people the data to play with or create visualizations on their own or to enable them to use the data in their own work.

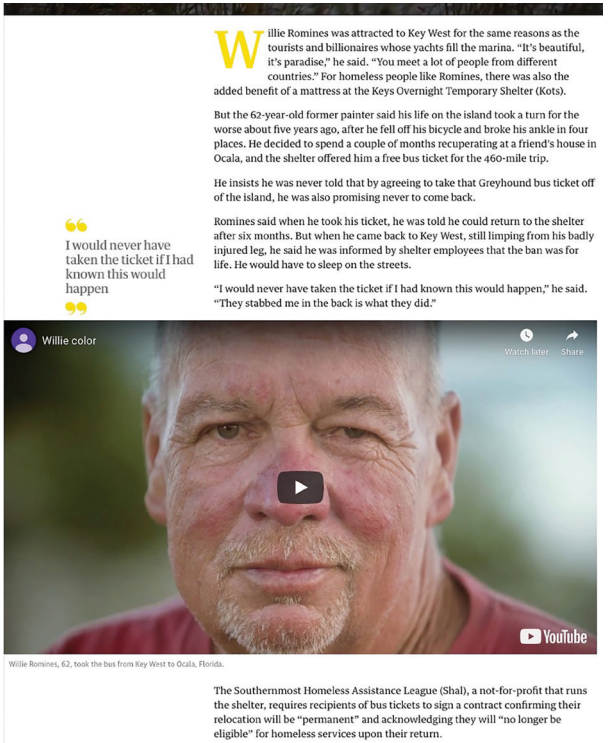
▸ ▸ ▸ ▸ ▸



This map by Aaron Koblin shows all flight paths in US airspace over twenty-four hours.

Remember these axes are *spectrums*, so a visualization can land somewhere in the middle of this space, or bring different aspects of these quadrants together. You've certainly seen examples of this from the *Guardian, New York Times*, *Washington Post*, and elsewhere—a combination of interactive and static visualizations paired with text, perhaps audio clips from interviews and pictures, all sewn together to creative a visual experience that gives you a deeper level of experience and understanding than the text alone might do. Through text, video interviews, static, and animated visualizations, *Bussed Out* from the *Guardian* in 2018 tells the story of how U.S. city governments bus homeless people out of their cities.

Another way to think about how these quadrants blend together is with animated data visualizations. Animations can take different forms and fulfill different purposes. In her book *Visualization Analysis and Design*, Tamara Munzner distinguishes between three kinds of animation: narrative storytelling (i.e., movies); transitions from one state to another; and video playback, in which the user can control the sequence by playing, pausing, and rewinding. "Animation is extremely powerful when used for transitions between two dataset configurations because it helps the user maintain context," she writes. Such transitions, therefore, help the reader (or perhaps better put, the user) see a data point or points move from one position to another.



Willie Romines was attracted to Key West for the same reasons as the tourists and billionaires whose yachts fill the marina. "It's beautiful, it's paradise," he said. "You meet a lot of people from different countries." For homeless people like Romines, there was also the added benefit of a mattress at the Keys Overnight Temporary Shelter (Kots).

But the 62-year-old former painter said his life on the island took a turn for the worse about five years ago, after he fell off his bicycle and broke his ankle in four places. He decided to spend a couple of months recuperating at a friend's house in Ocala, and the shelter offered him a free bus ticket for the 460-mile trip.

He insists he was never told that by agreeing to take that Greyhound bus ticket off of the island, he was also promising never to come back.

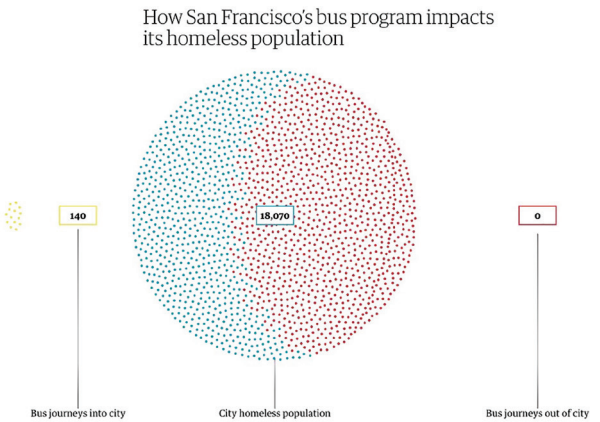
> I would never have taken the ticket if I had known this would happen

Romines said when he took his ticket, he was told he could return to the shelter after six months. But when he came back to Key West, still limping from his badly injured leg, he said he was informed by shelter employees that the ban was for life. He would have to sleep on the streets.

"I would never have taken the ticket if I had known this would happen," he said. "They stabbed me in the back is what they did."

Willie Romines, 62, took the bus from Key West to Ocala, Florida.

The Southernmost Homeless Assistance League (Shal), a not-for-profit that runs the shelter, requires recipients of bus tickets to sign a contract confirming their relocation will be "permanent" and acknowledging they will "no longer be eligible" for homeless services upon their return.



This data is partial; it does not include, for example, housed San Franciscans who become homeless while living in the city, the many homeless people who travel to and from San Francisco independently of relocation programs, or those homeless people who might have by now found a home. But it does give a rough illustration in response to Lazar's question.

If these relocation programs did not exist, and the people San Francisco has bussed out of the city had stayed put, there could be as many as 18,000 homeless people currently in the city, more than twice the current population.

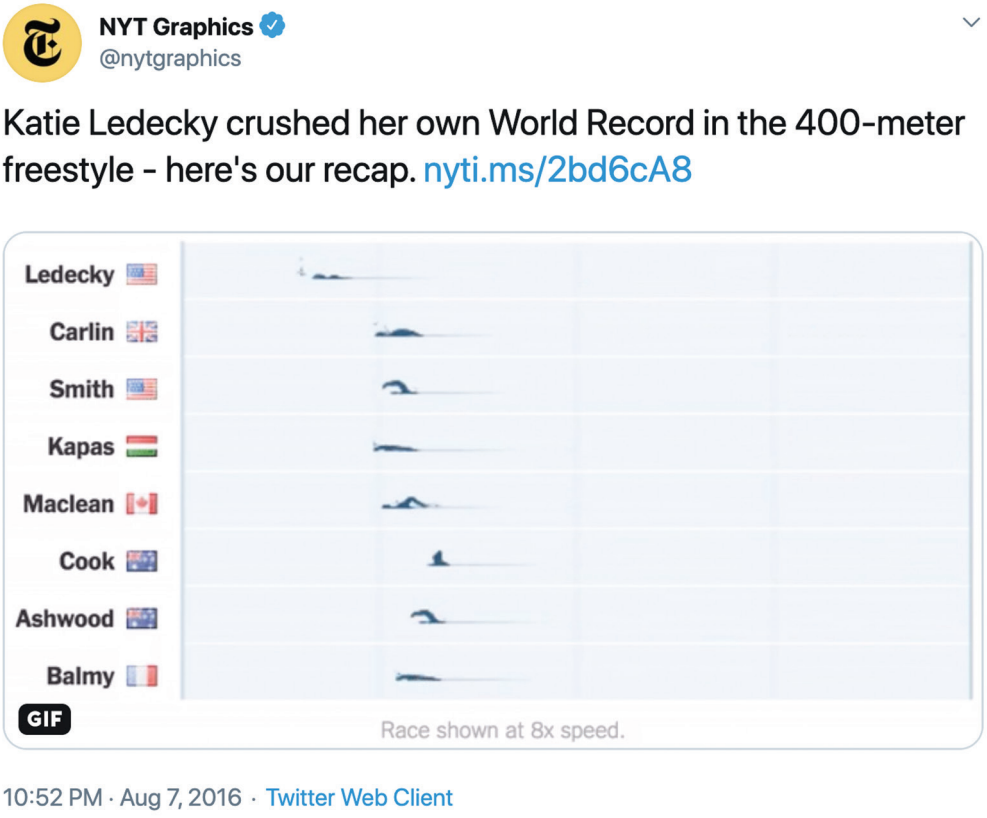
The US government mandates cities and other municipalities count their homeless street populations every two years; mayors are always keen for the tallies to show numbers are not on the rise. In 2009, New York mayor Michael Bloomberg lauded his own city's bussing scheme because it "saves the taxpayers of New York City an enormous amount of money".

Officials currently involved in running programs in Denver, Jacksonville, and Salt

Through text, video interviews, static, and animated visualizations, "Bussed Out" from the *Guardian* in 2018 tells the story of how US city governments bus homeless people out of their cities.

A simple way to think about animated visualizations is the "animated GIF." A GIF (Graphics Interchange Format) is a number of images or frames in a single file. From a data visualization standpoint, animated GIFs are powerful because they enable us to stitch together different graphs into an animated whole. They are just like cartoons: A graph with one line, followed by the same graph with another and another and another, builds up to a line graph with four lines.

Animated visualizations are especially powerful on social media because they break the stream of endless content and images. A set of Tweets from the *New York Times* during the 2016 Summer Olympics, for example, showed (sped up) swimming events so users could



A series of tweets with animated visualizations from the *New York Times* showed how US swimmer Katie Ledecky beat her opponents by wide margins.

quickly and easily "watch" an illustrated version of the event. Different than a static bar chart or table of finishing times, these animations engaged viewers and held their attention, especially on a fast-paced platform like Twitter.

There is no "right" or "wrong" quadrant here—what you choose depends on what reaction you want from your audience, where you publish your content, and what tools your audience needs to understand the data.

CHANGING HOW WE INTERACT WITH DATA

In 1997, Ben Shneiderman, then a professor of computer science at the University of Maryland at College Park, wrote what would become a mantra of online interactive data visualizations:

> "Overview first, zoom and filter, then details-on-demand."

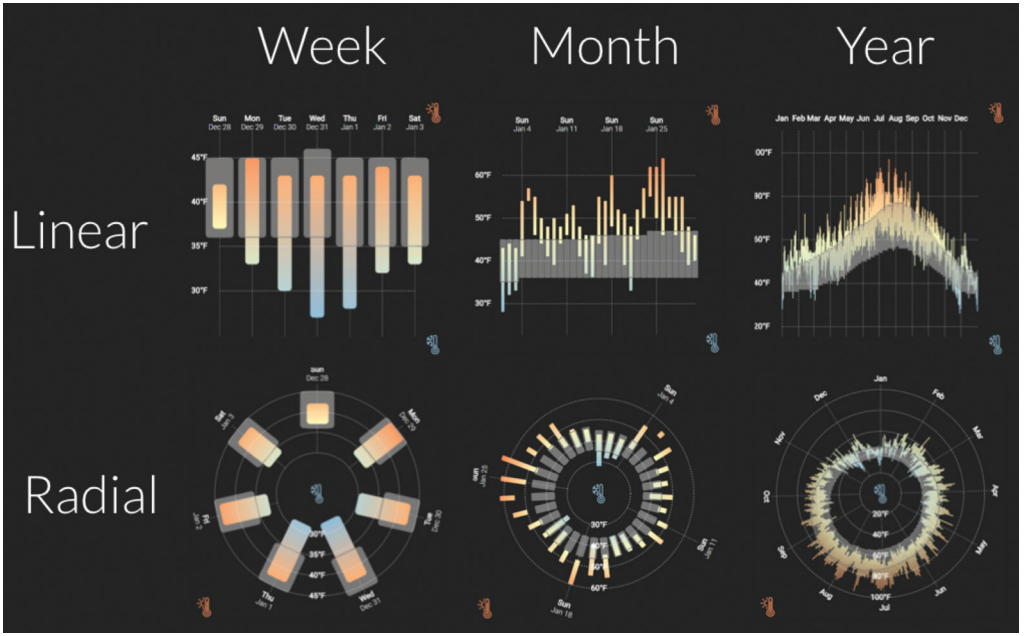
The theory was that you give users an overview of the visualization, then let them zoom in or out and filter through it, and then allow them the opportunity to reveal the details as they need (for example, through tooltips or downloads).

With the growing importance of mobile technology, however, Shneiderman's mantra is not as applicable as it once was. About a decade later, Archie Tse, the Deputy Graphic Director at the *New York Times*, argued that the increase in mobile platforms means that people just want to scroll. If the reader needs to click or do anything *besides* scroll, "something spectacular has to happen."

This argument makes perfect sense. When I think about scrolling through my newsfeed on my phone during my morning commute on the train, I don't want to sort and filter data on my device. I want the author to tell me or show me the most important features of the data. This is even more important when I'm reading technical or scientific work.

A 2018 project from Microsoft Research (see next page) tested how quickly and accurately people could read different graphs on their mobile phones. They asked a hundred people to read data from a weather app and a sleep app that used different graph types: a linear bar chart layout and a radial layout, like a circular column chart. The researchers found that people were slower to comprehend radial layouts but were no less accurate in their inferences.

Mobile use is only growing. According to the Pew Research Center, in October 2017 roughly six in ten adults (58 percent) got their news from their mobile device, compared



Microsoft researchers Brehmer, Lee, Isenberg, and Choe (2018) tested how quickly and accurately people could read different graphs on their mobile phones.

with 39 percent who got their news on a personal computer. As you consider where your audience resides in this **FORM-FUNCTION** space, keep in mind that technology, behavior, and news habits are always changing.

LET'S GET STARTED

The chapters of the rest of this book will explore families of charts and graphs based on their primary functions. Each chapter employs a different style guide borrowed from news organizations, nonprofits, and other organizations to showcase all the variety of subjective design decisions available. In Chapter 12, we'll learn more about specific styles and how to assemble the components of your own data visualization style guide.

Remember to ask yourself about your audience. You may even take this a step further and *talk* directly to your audience. Ask them what they want and what they need to better

understand your data, content, and analysis. Ground your work in the lived experience of communities and people you care about and want to reach. Some audiences want a thirty-five-page PDF report. Others want a two-page brief. Others want an eight-hundred-word blog post. Still others want a more immersive, narrative experience like you might find on a major newspaper website. And some just want the data. An academic researcher, manager, practitioner, policymaker, and a reporter may need very different things. Your visualization should match the needs of your audience.

As you consider your audience's needs, you may need to balance the *accuracy* of your graph with how it *engages* your audience. One way to think of how to do this effectively is to be empathetic to your audience's needs. Take it from Alan Alda, in his book *If I Understood You, Would I Have This Look on My Face?*: "Developing empathy and learning to recognize what the other person is thinking are both essential to good communication."

PART TWO

CHART TYPES

COMPARING CATEGORIES

The graphs in this chapter are intended to help our readers compare values across categories. Bars, lines, and dots can all let our readers compare within and between groups. In some cases, we want our reader to see both levels *and* change, or some other variable combination; in other cases, we want to focus their attention on one comparison or another.

The challenge when comparing categorical data is deciding what we want the chart to convey. Is there a primary argument or story? Is there something you can identify as the most important comparison you want the reader to make? As chart creators, we need to prioritize what we want our charts to do. By putting *every* bar or dot in the graph, we can obscure the point we wish to convey.

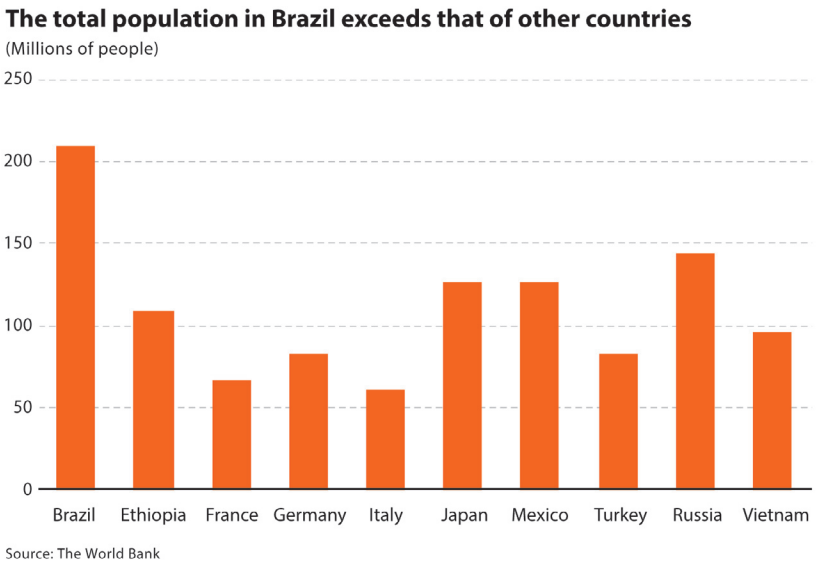
This chapter starts with the bar chart. Like the line chart that will kick off the next chapter, the bar chart is familiar to most readers, which makes it a convenient choice to guide readers as they compare categories or view changes over time. It also sits at the top of the perceptual ranking diagram. It's not necessarily the case that we must *always* give our readers the exact values, but when we do, the bar chart is an excellent choice.

Graphs in this chapter are styled roughly following the guidelines published by Eurostat, the statistical office of the European Union. Eurostat's seventy-six-page style guide covers everything from color, typography, logos, tables, layout, and more elements of a comprehensive style guide that we will discuss in Chapter 12.

BAR CHARTS

One of the most familiar data visualizations, the length or height of the rectangular bars in bar and column charts depict the value of your data. The rectangles can be arranged along the vertical axis so that the bars lie horizontally (often called a bar chart) or vertically on the horizontal axis (often called a column chart). For the sake of brevity, and the fact that whichever way you align them they are still bars, I call these bar charts throughout the book.

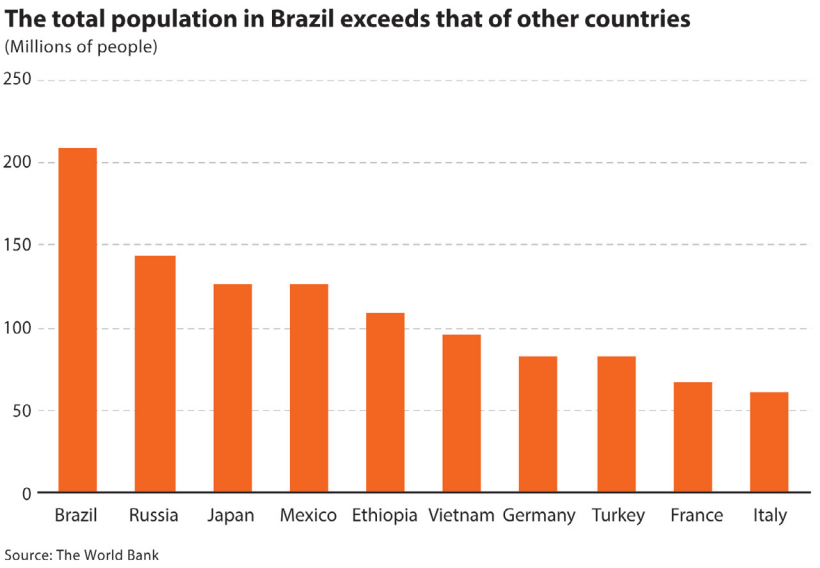
Bar charts sit at the top of the perceptual rankings list. With rectangles sitting on the same straight axis, it's easy to compare the values quickly and accurately. Bar charts are also easy to make, even with pen and paper. This one shows the total population in ten countries from around the world. It's easy to find the least (Italy) and most (Brazil) populous countries in the group, even when they are not labeled with the exact values.



The bar chart is a familiar chart that's easy to read and make. It sits at the top of the perceptual ranking matrix.

Data Source: The World Bank.

It's even easier to see the highest and lowest values when the data are sorted according to their data values. This strategy doesn't always work, however. If, for example, I was showing population levels for sixty countries, I might sort the values alphabetically, so that readers



When possible, sort the data in your bar charts. This makes it easier for your reader to find the highest and lowest values.

Data Source: The World Bank.

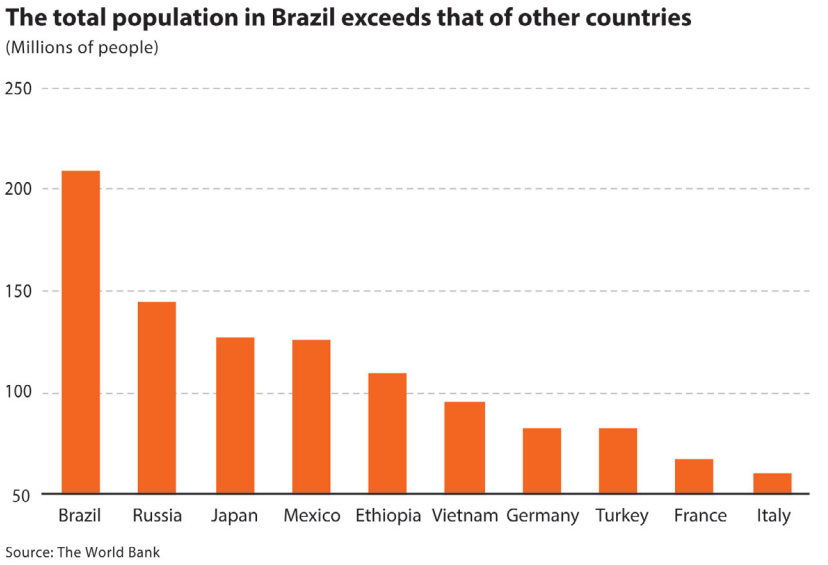
could more easily find the bar for a specific country. But if I was making an argument about the population level in a specific country or set of countries, I might sort the data so that the country or countries of interest are at one end of the graph. Alternatively, I could simply use a different color to highlight whichever bar or bars I want to set apart from the rest.

There are a few strategies to creating bar charts, many of which will apply to other charts in this chapter as well.

START THE AXIS AT ZERO

Starting the axis of bar charts at zero is a rule of thumb upon which many data visualization experts and authors agree. Because we perceive the values in the bar chart from the length of the bars, starting the axis at something other than zero may overemphasize the difference between the bars and skew our perception.

Take the bar chart of population. Because none are lower than fifty million, we might be tempted to start the axis at fifty million. After all, this would emphasize the difference between the values.

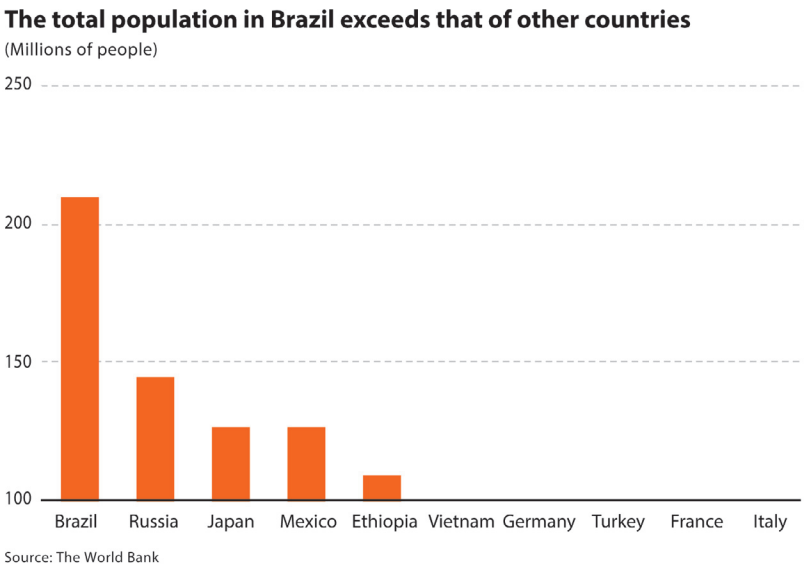


Starting the vertical axis at 50 million overmphasizes the differences in values and skews our perception of the data.

But notice what happens when we do that. The differences in values are emphasized—in fact, they are *over*emphasized. Here, it looks as though Brazil is orders of magnitude larger than Italy, when, in fact, it is only about three-and-a-half times greater. This isn't a matter of moving from accurate perception to general perception—it's a matter of moving from accurate to inaccurate.

If you want to take a more extreme view of this, imagine starting the graph at a hundred million—and why not? If starting at fifty is OK, then we can pick any arbitrary number. Now at a glance it looks like nobody lives in half of these countries!

There is emerging research in this area that suggests that perhaps starting bar charts at something other than zero does not bias our perception of the data. In one recent study, participants were better able to assess the sensitivity of the results (e.g., no effect, small effect, medium effect, or big effect) and more accurate (e.g., the size of the effect) when the vertical axis was set at a range more consistent with the variation of the data. Until more research is conducted, however, my preference is to start the axis in bar charts at zero to avoid any confusion or possibility of visual bias.



If starting the y-axis at fifty is OK, then why not one hundred?
Data Source: The World Bank.

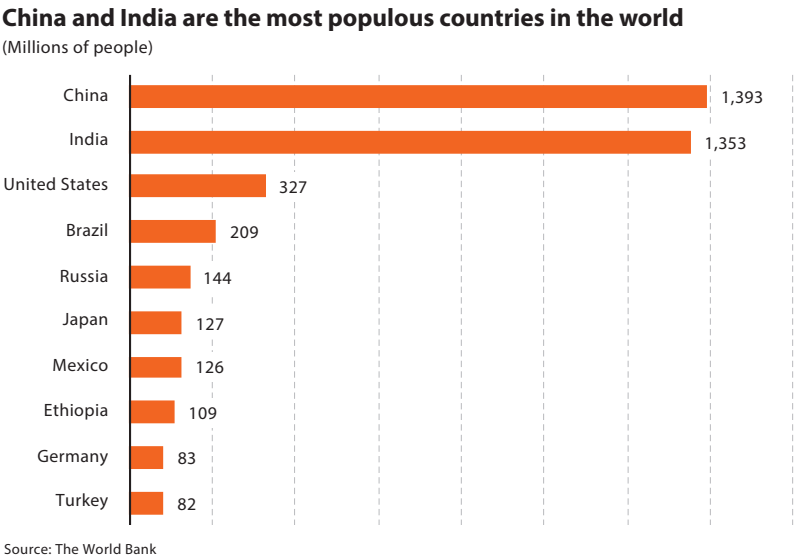
DON'T BREAK THE BAR

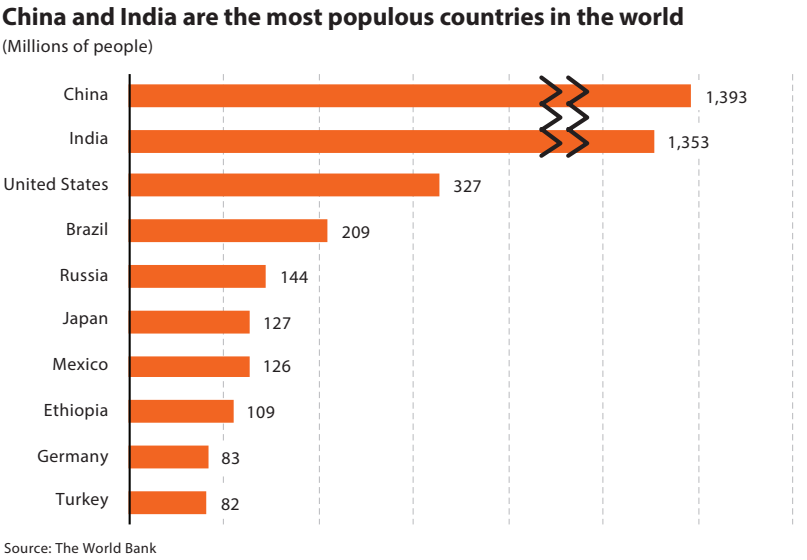
Another cardinal sin of data visualization is what is called "breaking the bar"—that is, using a squiggly line or shape to show that you've cropped one or more of the bars. It's tempting to do this when you have an outlier (see Box on page 74), but it distorts the relative values between the bars.

Let's create a bar chart of population in the ten most populous countries of the world. In 2018, China and India were the most populous countries on the planet with 1.39 billion and 1.34 billion people, respectively, followed by the United States with 327 million people. We can see how dramatically larger China and India are relative to the rest of these countries in the top graph on the next page. If we wanted to make the differences between the less populous countries larger, we could break the bars, but this makes China and India look much less populous than they are. Chopping the lengths of the bars is completely arbitrary—I can place those squiggly lines wherever I like to zoom in on the other differences. But that's not being honest with the data.

If you run into a case where you have outliers but want to show the detailed differences between the smaller values, try using more graphs. You might think of this as

a "zoom in" and "zoom out" approach—show all of your data so your reader can see the magnitude of the largest values, and then zoom in for a detailed look that omits the outliers. On the next page, I've highlighted the less populous countries to show the



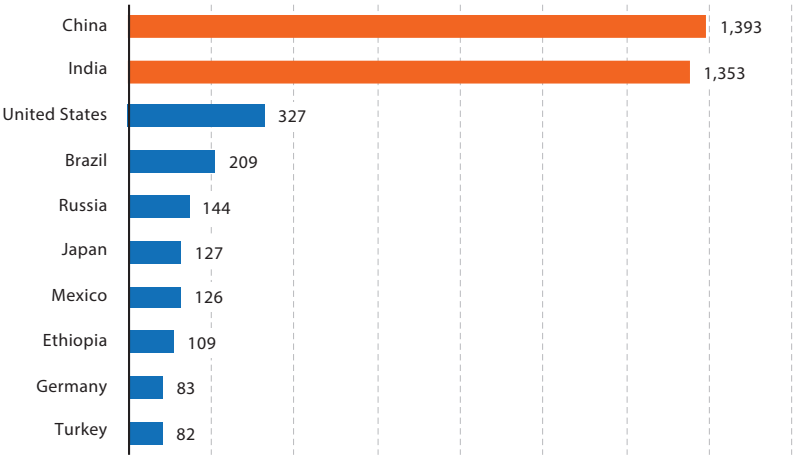


Don't break the bar in your bar charts. The break can be arbitrarily set anywhere and distort our perception of the data.

differences between them, which we can't quite see in the main graph. Adding labels and an active title is another good way to communicate the differences between smaller values to the reader.

China and India are the most populous countries in the world

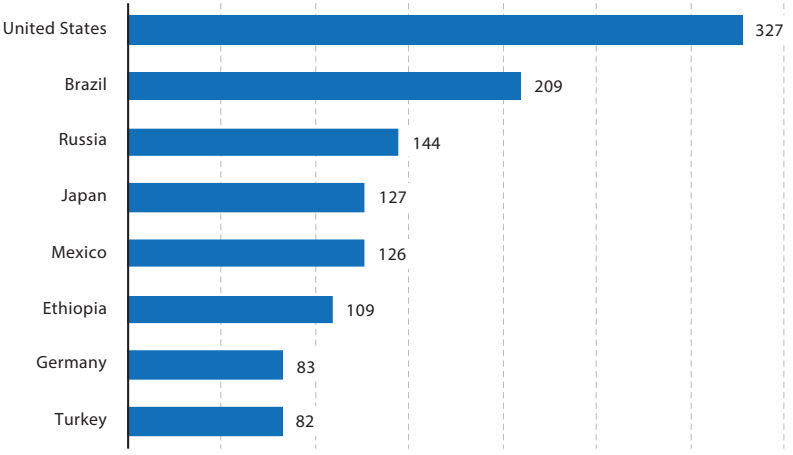
(Millions of people)



Source: The World Bank

Total population in these countries ranges from 82 million to 327 million

(Millions of people)



Source: The World Bank

In cases where you have large values or outliers but want to show the detailed differences between the smaller values, try using more graphs.

EXTREME VALUES OR OUTLIERS

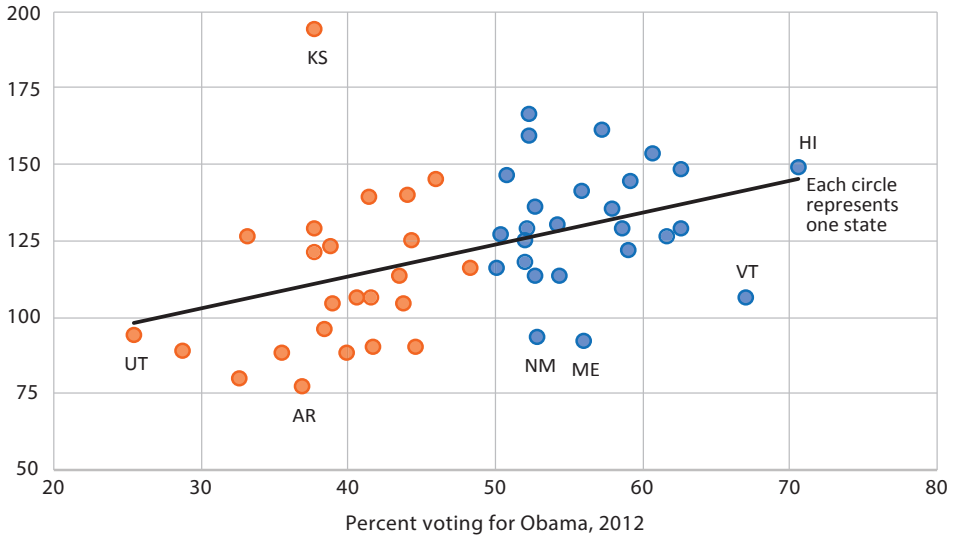
An outlier is a data point that is far away from other observations in your data. It may be due to random variability in the data, measurement error, or an actual anomaly. Outliers are both an opportunity and a warning. They potentially give you something very interesting to talk about, or they may signal that something is wrong in the data.

In 2014, Buzzfeed teamed up with the website Pornhub to look at pornography viewing by state. Using geolocation data of people accessing their site, Pornhub calculated the number of page views per person in each state. People in Kansas, they found, watched far, far more pornography than any other state in the country: 194 page views per person. Nevada was second with 166 page views.

The data went into the scatterplot below, comparing blue-state and red-state porn consumption. You can clearly see Kansas as an outlier in page views. Do people in Kansas really watch that much more porn?

Presidential politics and porn per capita

(Pornhub pageviews per capita)



Source: Pornhub views from Buzzfeed; Voting percentages from *The Guardian* and NBC News.
Scatterplot originally created by Christopher Ingraham.

Turns out the answer is no. Apparently, Pornhub's methodology assigned missing geolocation data to the geographic center of the United States, which, as it turns out, is Kansas.

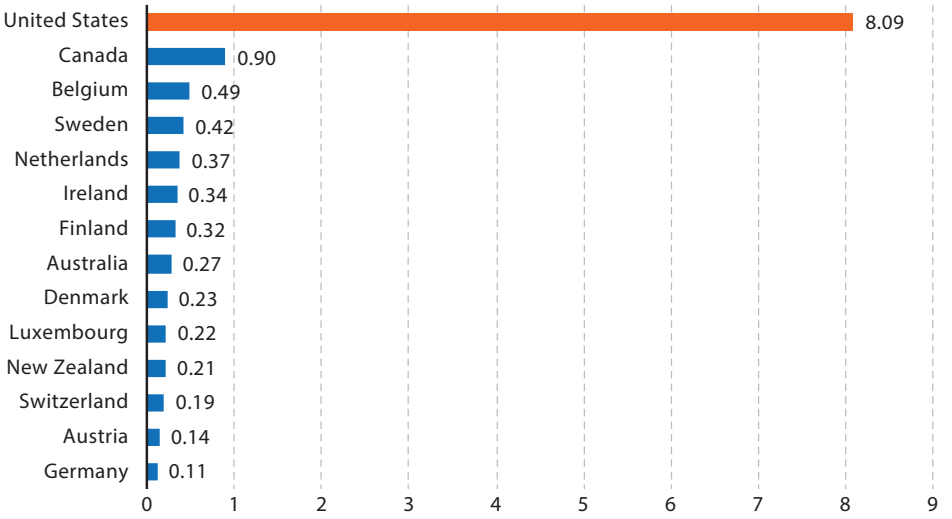
Not all outliers are mistakes, however. As just one example, we can look at the rate of physical violence from a firearm across advanced countries. In 2017, more than 8 per 100,000 people were victims of firearm violence in the United States, compared with 0.90 per 100,000 people in Canada and 0.49 per 100,000 in Belgium. In some cases, outliers are truly outliers.

There are lots of ways to test for outliers in your data, some more complex than others. One way is to simply *look* at your data. Exploring your data does not need to start with complex math and statistics—you should always visually inspect your data.

But that approach is hardly mathematical. A standard method is to compare data values to 1.5 times the interquartile range (IQR). The IQR is a simple summary of your data and is the difference between the third and first quartiles (see the Box in Chapter 6 on percentiles).

The United States has the highest rate of physical violence by firearm among advanced countries

(Rate per 100,000 people, 2017)



Source: Institute for Health Metrics and Evaluation

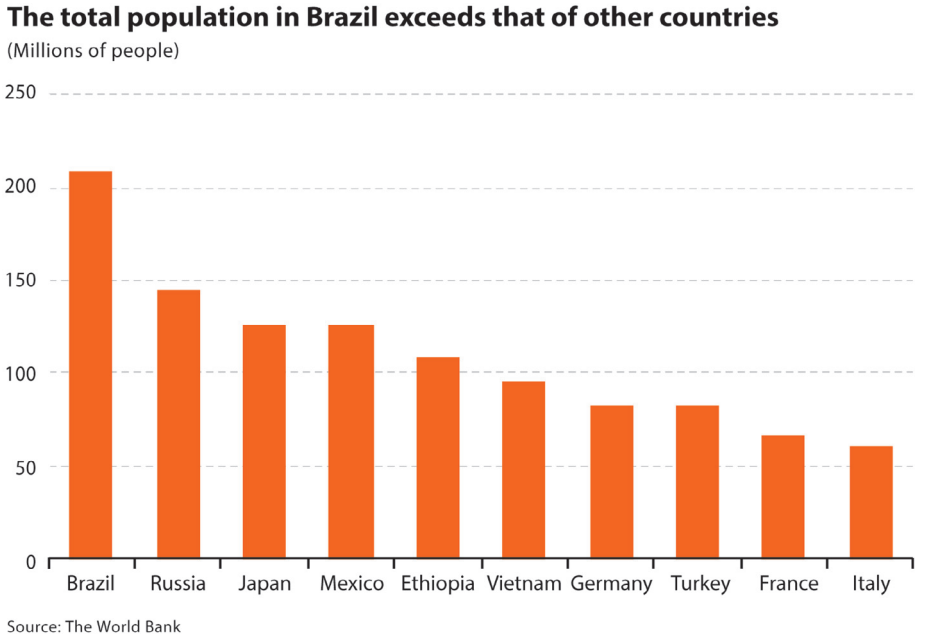
USE TICK MARKS AND GRIDLINES JUDICIOUSLY

Bar charts don't need tick marks between the bars. White space is an effective separator and deleting the tick marks reduces clutter.

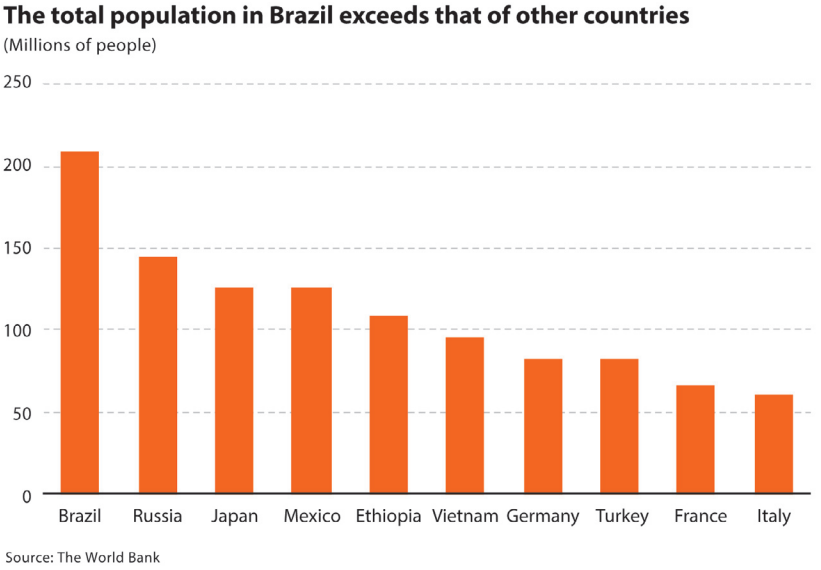
One exception is if you have a "major" category label that spans multiple bars. In such cases, larger tick marks can be helpful to group the labels (see the bottom chart on the next page).

Gridlines help the reader see the specific values for each bar and are especially useful for the bars farthest from the axis label. Because they serve as a visual guide, they can be rendered in a lighter color so the reader's eye stays on the data.

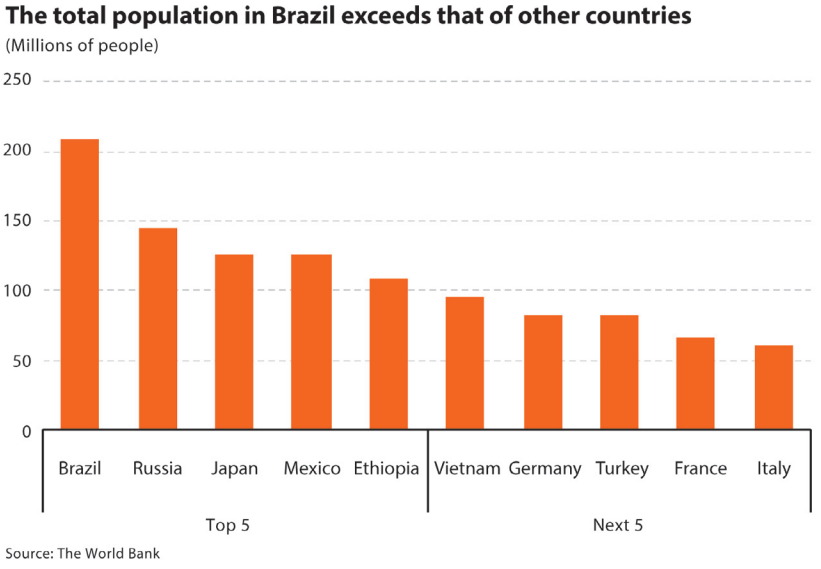
When it's important for the reader to know the *exact* values, you can add data labels to the chart. My preference is to forgo the gridlines and axis lines altogether in these cases.



In bar charts, tick marks are not necessary. The white space does the job of separating the bars.

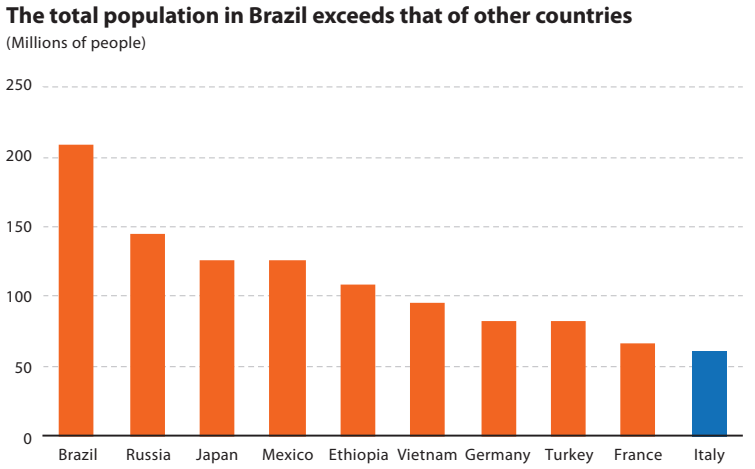


Omitting tick marks is part of removing as many non-data elements as possible.



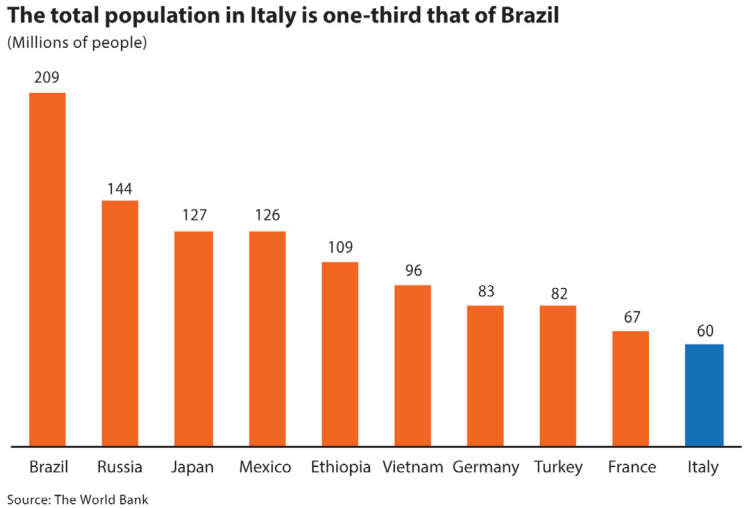
Tick marks may be necessary when you have a “major” category.

Consider Italy in the next two graphs (highlighted in blue). Without the labels, the gridline helps us see that there are more than fifty million people living in the country; with the label, it is clear that it's sixty million people and thus the gridlines are probably not necessary.



These horizontal gridlines help the reader see that, for example, there are more than fifty million people living in Italy.

Data Source: The World Bank.



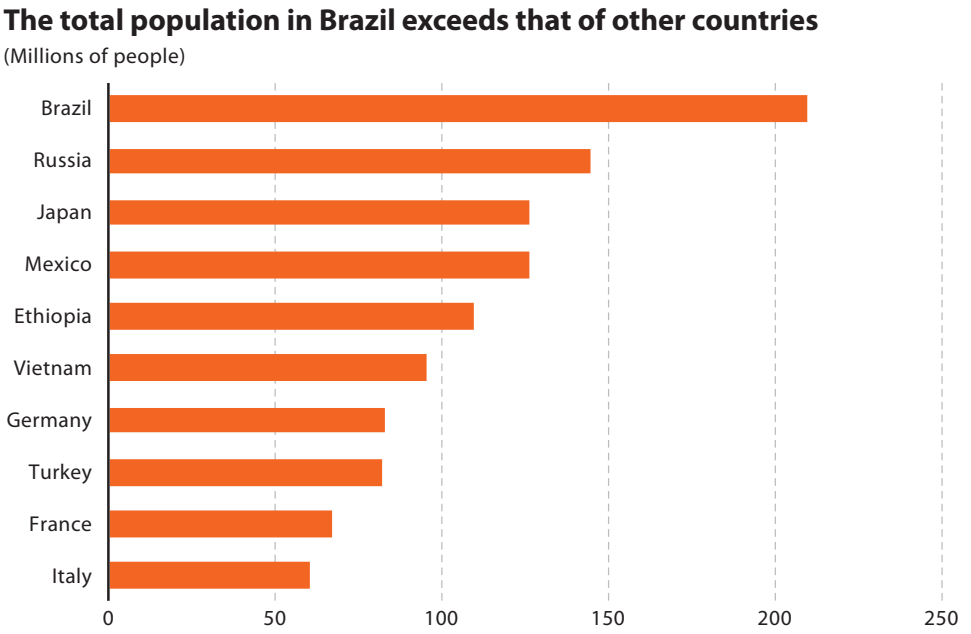
Data labels make gridlines redundant and, by extension, the vertical axis.

The graph may look too cluttered with labels if I had fifty or maybe even twenty countries, so I might include a separate table or an appendix. Deleting the gridlines when including data labels is primarily an aesthetic choice and as you continue to work with data and make your own graphs, you will develop your own style for these graphic elements.

ROTATE LONG AXIS LABELS

The default solution for long horizontal axis labels is to run the text vertically, as on the spine of a book. But this approach forces your reader to turn their head to the side. One solution is to rotate them 45 degrees, but the reader still has to turn their head. Another approach is to shrink the font size so they are aligned horizontally—though this usually makes them too small.

The most elegant solution is to simply rotate the entire graph. This still uses the same pre-attentive attribute—the length of the bars—but the axis labels are now aligned horizontally; they are easy to read with no effect on data comprehension.



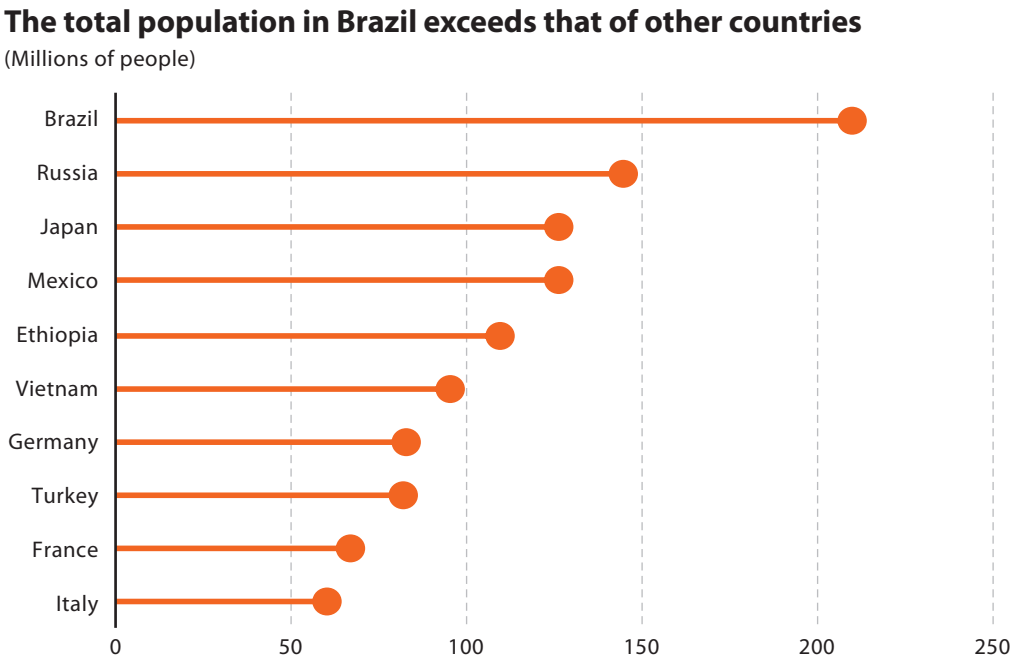
With long axis labels, consider rotating the chart to make the labels horizontal and easier to read.

Data Source: The World Bank.

VARIATIONS ON THE BAR CHART

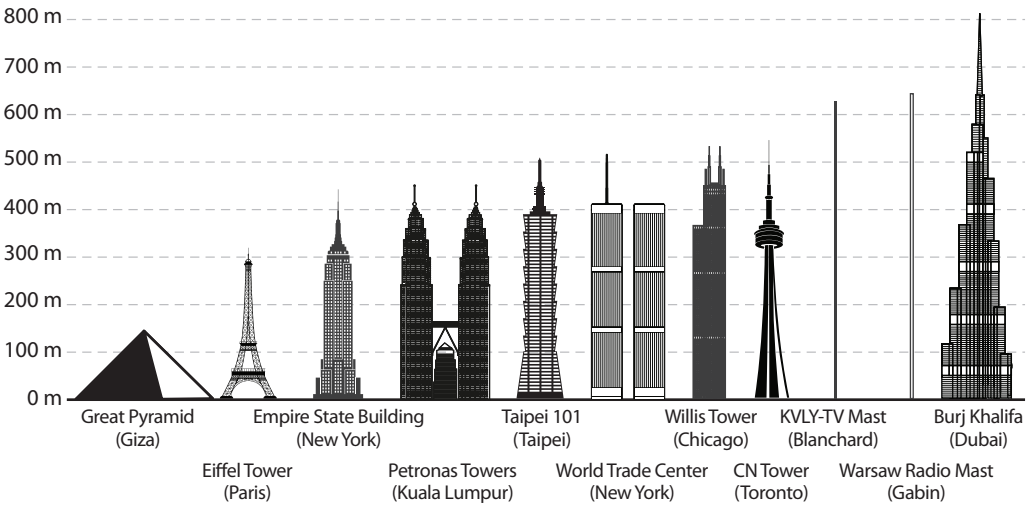
There are countless ways to modify the standard bar chart. One simple variation is to use other shapes in lieu of bars. The lollipop chart, for example, replaces the bar with a line and a dot at the end. This version lives a hair below the bar chart on our perceptual rankings, because it's not exactly clear which part of the circle encodes the value. But it removes a lot of ink from the page and gives you more white space to add labels or other annotation.

This is just one example of an alternative shape. Triangles, squares, and arrows are other options, as are bar-shaped images that reinforce your data. A chart showing data on urban growth may use building-shaped bars, and a chart on climate change may use trees for bars. Be careful with this approach, however, as readers may confuse the total *area* of the icons as a value indicator rather than just the height.



The lollipop chart replaces bars with a shape (usually a dot) and a line.

Data Source: The World Bank.

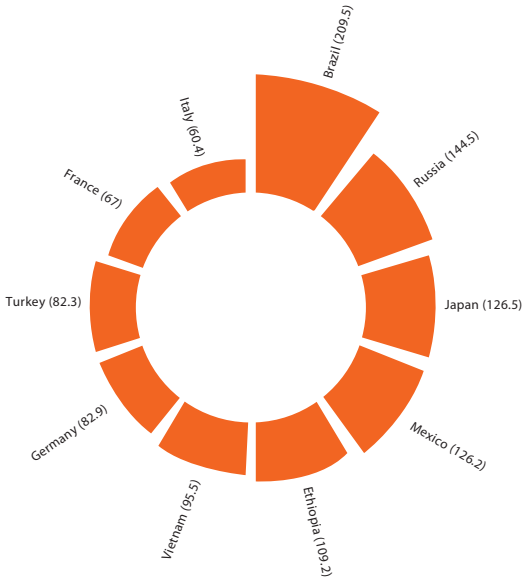


Alternative shapes, like buildings or people, can be used in lieu of the basic bar shape.

Source: Based on Wikimedia user BurjKhalifaHeight Petronas Towers

The total population in Brazil exceeds that of other countries

(Millions of people)



Change in Brazil's population from 2008 to 2018

(Millions of people)

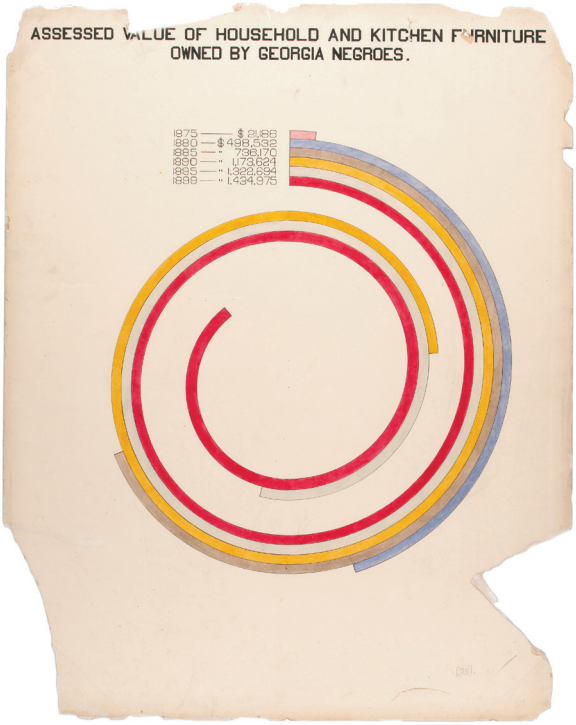
A radial bar chart wraps the standard bar chart around a circle. This chart type moves down the perceptual ranking list because it is harder to compare the heights of the bars.

Data Source: The World Bank.

Another approach to the basic bar chart is to abandon the usual grid and instead place the bars in a circle, called a radial layout. There are two common ways to do this: the radial bar chart and the circular bar chart.

The radial bar chart, also called the polar bar chart, arranges the bars to radiate outward from the center of a circle. This graph lies lowers on the perceptual ranking list because it is harder to compare the heights of the bars arranged around a circle than when they are arranged along a single flat axis. But this layout does allow you to fit more values in a compact space, and makes the radial bar chart well-suited for showing more data, frequent changes (such as monthly or daily), or changes over a long period of time.

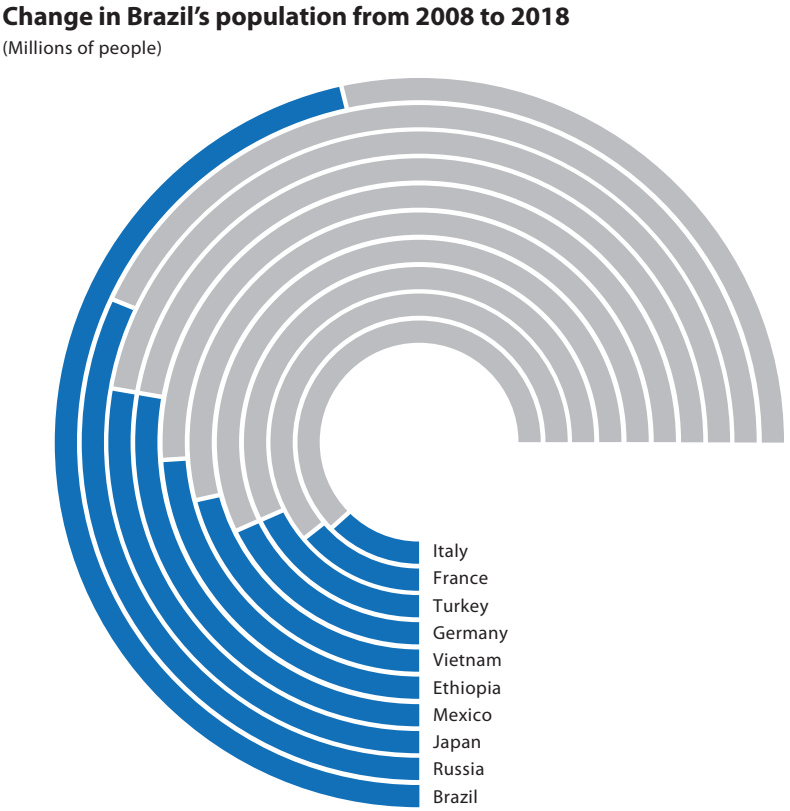
W. E. B. Du Bois used a circular bar chart in his famous *Exposition des Negres d'Ameriique* at the 1900 Paris Exposition. He included this radial bar chart in his set of infographics for



Source: W. E. B. Du Bois, Assessed Value of Household and Kitchen Furniture Owned by Georgia Negroes (1900) via Library of Congress Prints and Photographs Division.

The Georgia Negro: A Social Study, which shows the dollar value of household and kitchen furniture held by African Americans in Georgia in six years (1875, 1880, 1885, 1890, 1895, and 1899). "The end result," wrote Whitney Battle-Baptiste and Britt Rusert in their book about Du Bois's graphics, "is simultaneously easy to read and hypnotic."

Perceptually speaking, the circular bar chart is problematic because it distorts our perception of the data—in this case, the lengths of the bars don't correspond to their actual value. Consider the case where the values of two bars are the same—the ends of the bars will line up in the same position, but the lengths of the bars are not actually the same because they lie along the circumference of two different circles. Author and data visualization expert Andy



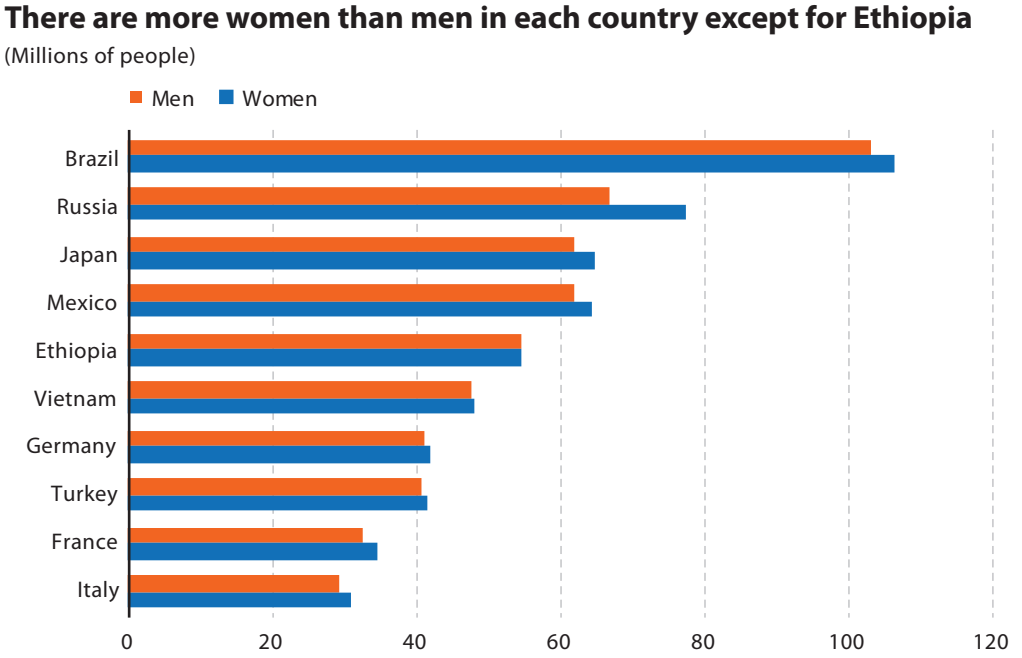
Perceptually speaking, the circular bar graph is problematic because it distorts our perception of the data. In this case, the lengths of the bars don't correspond to their actual value.

Data Source: The World Bank.

Kirk uses an Olympic footrace as a metaphor. Runners start at staggered positions on the track, but they all end up running the same distance because the runner on the outside lane has more distance to cover than the runner on the inside lane. Here, the visualization doesn't move down the perceptual ranking, but off of it altogether because it distorts the data and for that reason, I recommend avoiding them altogether.

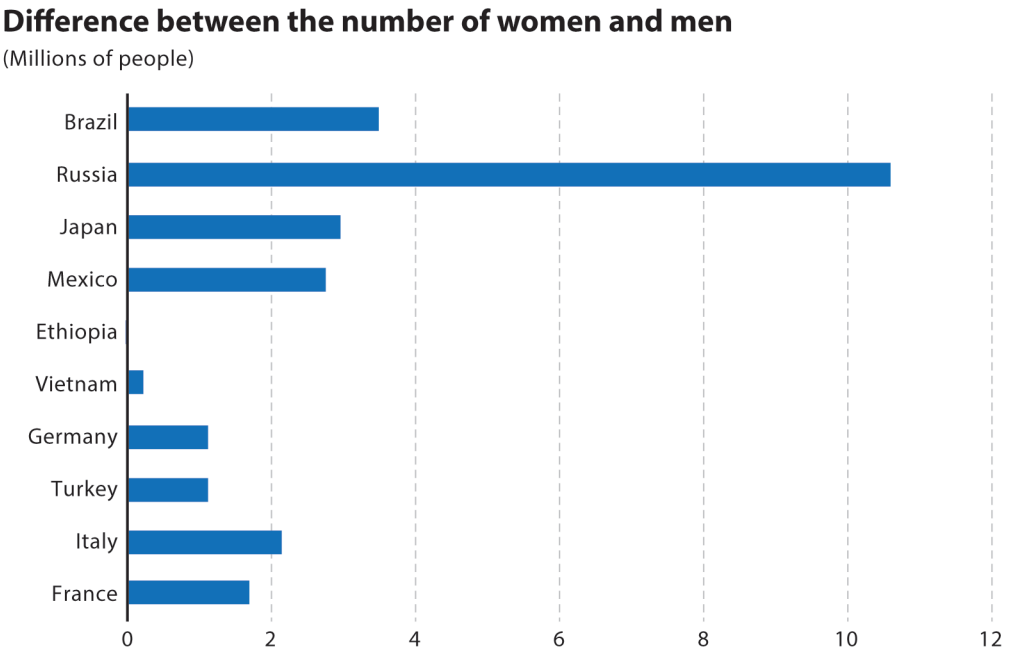
PAIRED BAR

A simple bar chart is perfect for making comparisons across categories, like comparing populations across countries. If I want to show comparisons not just across but also *within* countries, the paired bar chart is a good option. The paired bar chart will be familiar to most readers and is easy to read, and the shared baseline makes it easy to make comparisons.



A simple paired bar chart is familiar to most readers and easy to read.

Data Source: The World Bank.



Instead of showing both data values, we could show the difference between them.

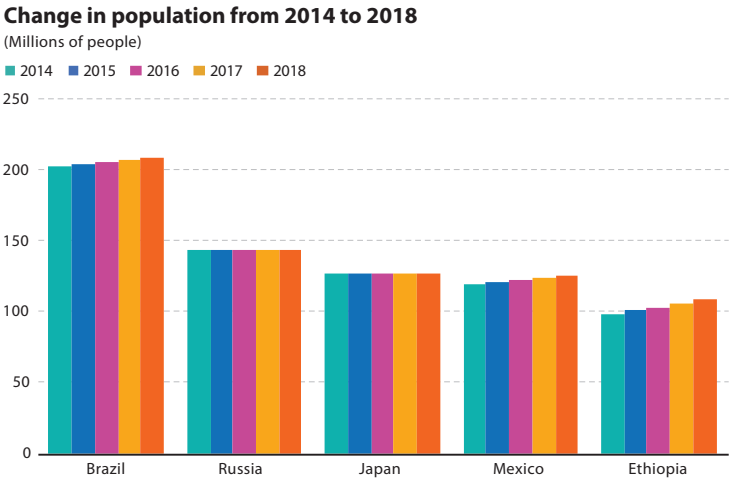
Data Source: The World Bank.

Say we want to show the number of men and women in each country in our sample. A paired bar chart allows us to do so.

Note that the paired bar chart directs the reader's attention not just to the levels, but also to the *difference*. If it's important that readers see both, this is a good option.

But if our goal is for the reader to focus only on the *difference* between the two values within each category, this isn't the most direct way to do so, because we are asking them to compare the difference in lengths. Instead, we could just show the difference between the two values in a single bar, like the one above.

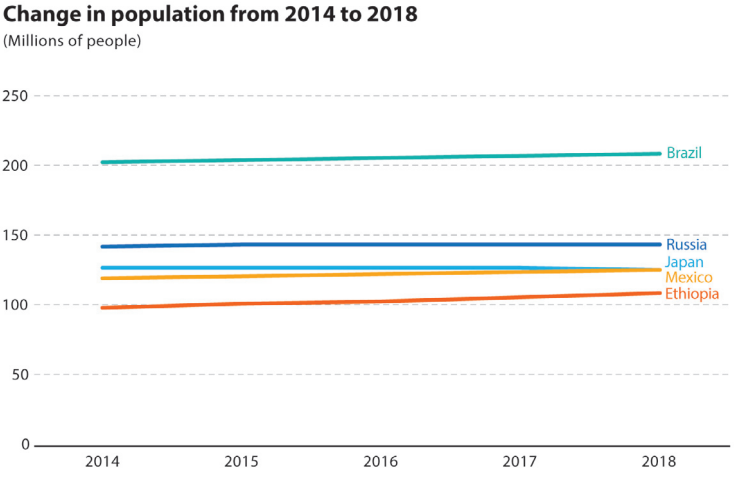
In cases where you want the reader to see the level *and* the difference, you may need a different chart entirely. I prefer the parallel coordinates plot (see page 263), the slope chart (for data that vary over time; see page 150), and the dot plot (see page 97). Remember to ask yourself, What is the goal of this graph? That question will guide you to the best way to visualize your data.



The paired bar chart can be used to show changes over time and can be used to examine changes within and between countries.

Data Source: The World Bank.

Another use for the paired bar chart is to show changes over time. And although I include the word *pair* in the title, these charts can have more than two values. The chart below, for example, shows the population in five of our countries from 2014 to 2018. This



The line chart is a more familiar way to show changes over time.

Data Source: The World Bank.

allows the reader to examine the population change *within* countries and the differences *across* countries.

The patterns in your data can also drive your chart type selection. If the values decline evenly across the different years for all categories, a paired bar chart may look fine. But if the values move around over time, a line chart (as shown earlier) or cycle chart (see Chapter 5) may make for better comparisons over time within and across each group.

There are two instances when I prefer to use bar charts rather than line charts to show changes over time. First, when there are few data points—for example, only five years—the extra ink in the five bars gives the graph more visual weight. Second, when I have discrete time intervals (and few observations), such as the first quarter of the year.

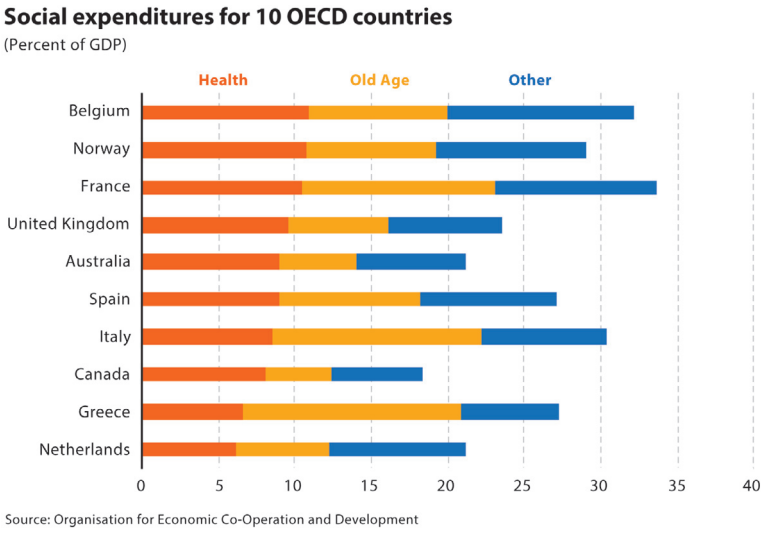
Clutter is the main issue to keep in mind when assessing whether a paired bar chart is the right approach. With too many bars, and especially when there are more than two bars for each category, it can be difficult for the reader to see the patterns and determine whether the most important comparison is between or within the different categories.

When it comes to whether a paired bar chart is too cluttered, trust your eyes and your instincts. Put yourself in your readers' shoes—try to imagine where their eyes will go when they look at the graph for the first time. If there's too much going on, you may need to break up your data, use a different chart type, or try a small multiples approach.

STACKED BAR

Another variation on the bar chart is the stacked bar chart. While the paired bar chart shows two or more data values for each category, this chart subdivides the data within each category. The categories could sum to the same total, say, 100 percent, so that the total length of the bar is the same for every group. Or the totals may differ across the groups, in which case the total length of each bar may differ. Above, I've plotted the share of gross domestic product (GDP) each of ten countries spends on support for health care, old age, and other programs. The entire length of the bars shows how much each country spends on these programs as a share of GDP.

As with the bar charts we've looked at thus far, the stacked bar chart is familiar, easy to read, and easy to create. The biggest challenge, however, is that it can be difficult to compare

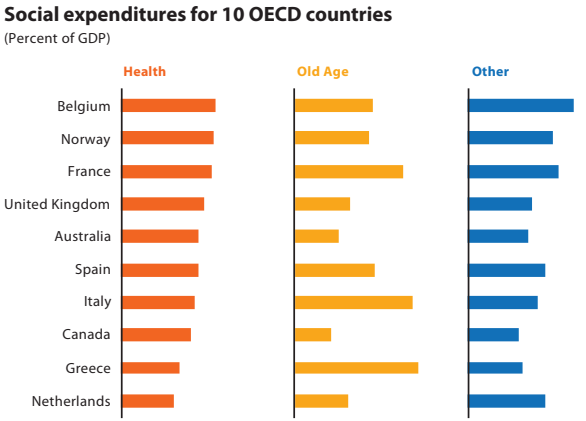


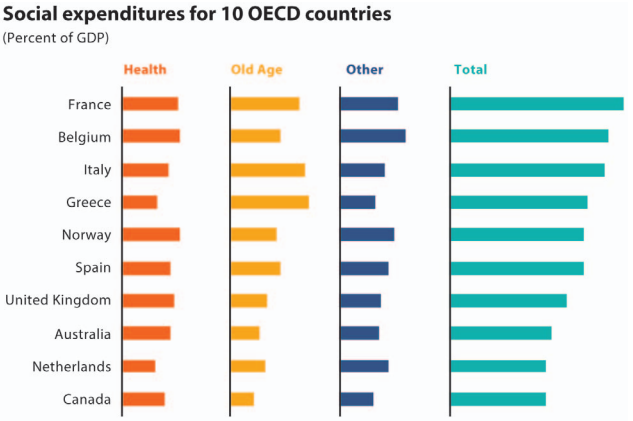
The stacked bar charts shows how different categories sum to a total. The interior series in the chart, however, are harder to compare with one another because they do not sit on the same baseline.

Data Source: Organisation for Economic Co-Operation and Development.

the different values of the segments *within* the chart. In the example above, it's easy to compare the values across the countries for the Health category, because the bar segments share the same vertical baseline. But that's harder to do with the two other series because they do not share a baseline. Which country spends more on old-age programs, Italy or Greece? You can quickly see that Italy spends more on health programs than Greece, because those segments are left-aligned on the vertical axis, but it's much harder to determine with the segments for the other categories.

One way to address the changing baseline is to break the graph apart so that each series sits on its own vertical baseline. This is a small multiples graph, arranged side by side. It's now easier to see that Greece spends more on Old Age programs than Italy. The tradeoff is that it is harder (if not impossible) to see the *total* values. But that too can be overcome: You can still break up the stacked graph and add a final segment that represents the total amount (this is not an issue when all of the series sum to 100 percent because the summed segments will all have the same length).

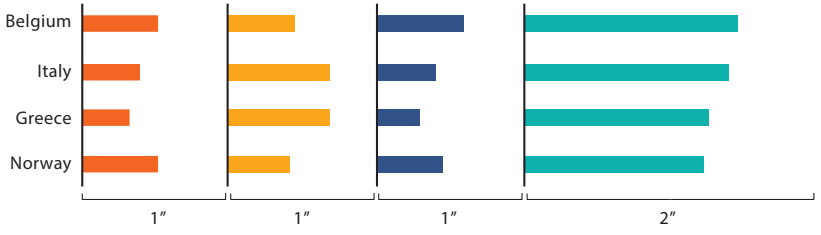




Instead of stacking all the data series together, we can break them up (either with or without the totals) to create a sort of small-multiples approach. Here, we move up to the top of the perceptual ranking list because each series sits on its own baseline.

Data Source: Organisation for Economic Co-Operation and Development.

In both versions, the horizontal spacing for each segment should be the same width, otherwise it might appear that a segment takes up a larger proportion of the space than it really does. In cases where you add the total, the width does not need to be the same as the other groups, but the increments along the axis should be the same. In other words, if the width of each segment above in which the data range from 0 percent to 50 percent is one inch wide, the total category that spans 0 percent to 100 percent should be two inches wide.

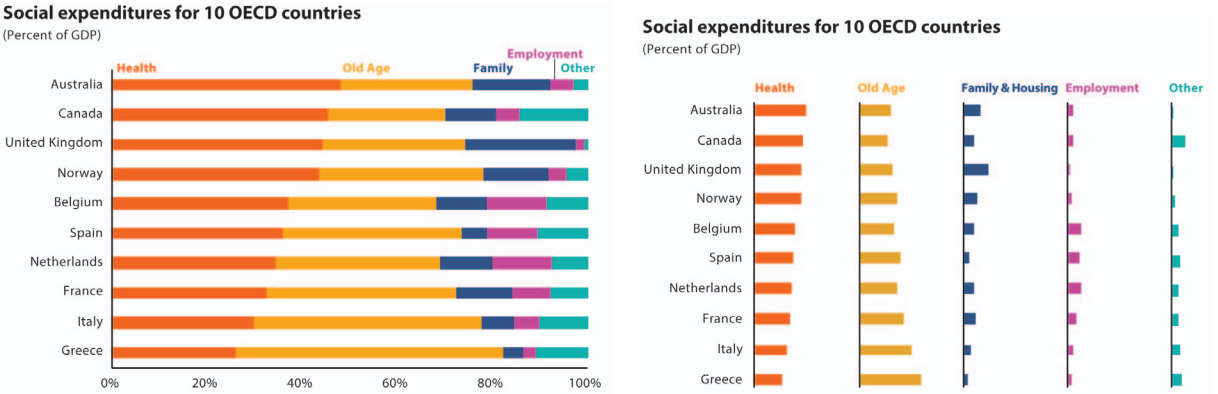


When creating these sorts of small multiples bar charts, be sure that each segment has the same width.

Even though different baselines in the standard stacked bar chart can make it more difficult to compare values, there are cases when the stacked bar chart is preferable. In this stacked bar chart, I've included more spending categories and divided them into shares of the total so the graph highlights the *distribution*. In this view, it becomes clear that around three-quarters of total government spending in these countries goes to programs for health care and old age programs. That observation is harder to see in the version on the right, where each category is placed on its own vertical baseline. Even though it is easier to compare differences in each category across countries, you don't see large differences between them.

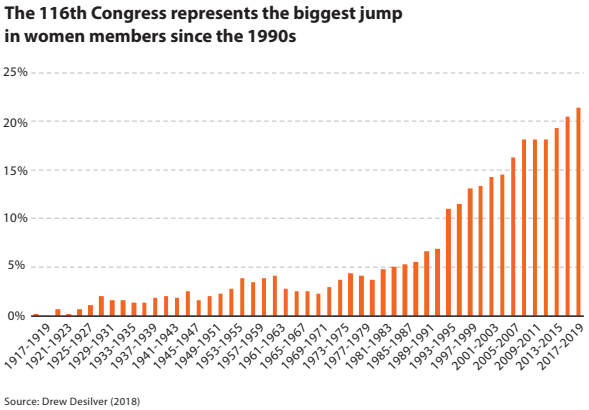
As always, identify what you want to show and where you'd like to focus your reader's attention. In these examples, the *Health* category is emphasized because the data are sorted according to those values (shown as a percent of total spending on these specific programs) and it is situated along the vertical baseline. In this layout, the other segments become secondary in comparison to health spending.

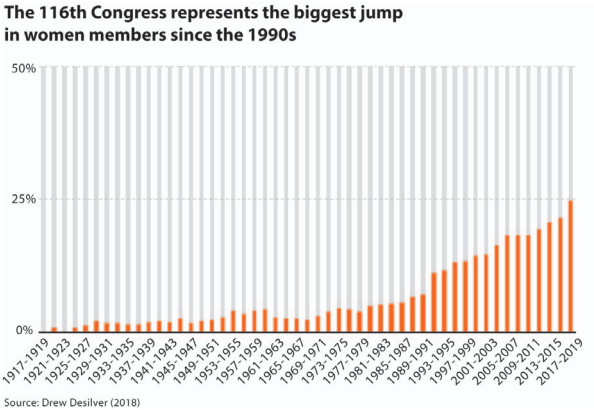
There is one other stacked bar chart that you may have come across that shows a single set of data values and the gap between them and another value (often the total). The graph on the next page uses this approach to show the share of women elected to the U.S. House of Representatives from 1917 to 2018. The version on the left shows the raw percentages; the vertical axis ranges from 0 to 25 percent. Here, you see a dramatic increase in the share of



In these examples, we can see how our ability to compare different values within and across countries varies between these two views.

Data Source: Organisation for Economic Co-Operation and Development.





A somewhat rare case where a stacked bar chart is used to focus attention on one series. This technique may be particularly valuable where the relative proportion is as important as the change.

women in Congress. The version on the right shows the same data, but stacks a gray series on top of the data values to 50 percent. In this version, we can emphasize that the share of women is still small even though that share is rising. It is in these cases—where the relative proportion is as important as the change—that this technique may be particularly valuable.

PERCENT CHANGE VS. PERCENTAGE POINT CHANGE

There is an important distinction between *percent change* and *percentage point change*, and it's a mistake that many often make.

Percent change compares an initial value *OLD* to a final value *NEW* according to this simple formula:

((NEW-OLD)/OLD) × 100.

Positive percent changes (that is, *NEW>OLD*) mean there is a percent (or percentage) increase. Negative changes (*NEW<OLD*) mean there is a decline. You can calculate differences over time or between groups; all that really matters is that you follow the formula and know that you are comparing the change relative to the initial value of *OLD*.

Now, *percentage point change* is specific to looking at raw differences in percentages. The *percentage point change* is a simpler formula:

NEW-OLD

where both are already percentages.

These are very different things. Let's take a simple example. According to the U.S. Census Bureau, there were 40.6 million people in poverty in 2016 and 39.6 million people in poverty in 2017. The poverty rate (the number of people in poverty as a percent of the total population) was 12.7 percent in 2016 and 12.3 percent in 2017.

The number of people in poverty fell by 2.3 percent. The *percent change* was

$$[(39{,}698{,}000 - 40{,}616{,}000)/40{,}616{,}000]\times 100 = [-0.023]\times 100 = -2.3\%$$

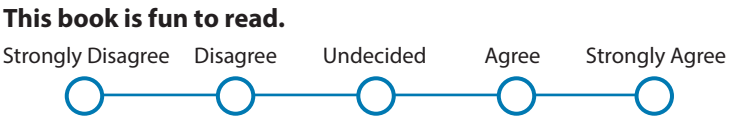
But the poverty *rate* fell by 0.4 *percentage points* over the two years:

$$12.3\% - 12.7\% = -0.4 \text{ percentage points}$$

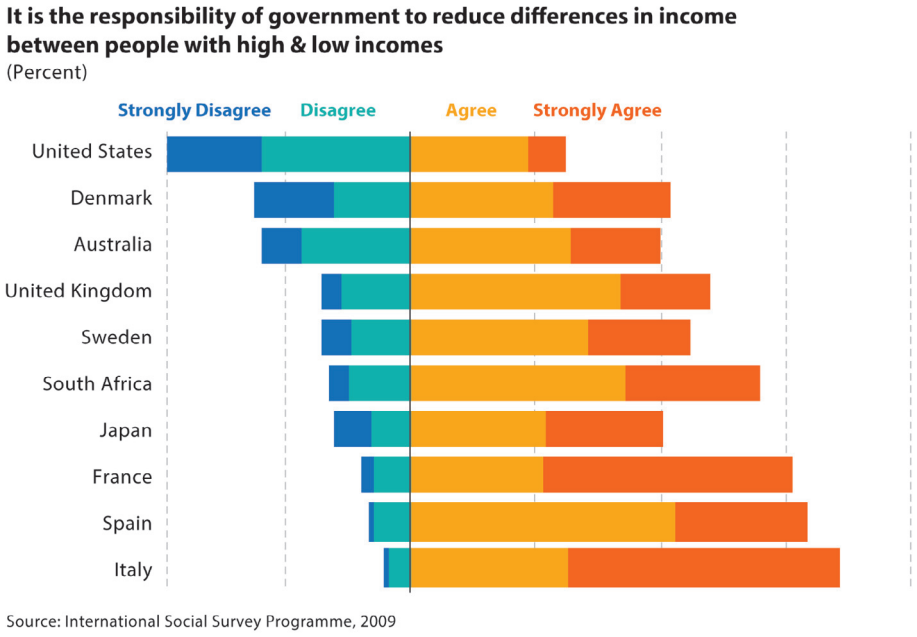
Obviously, those are two very different numbers, but people confuse them all the time. Clearly representing your data starts with clearly understanding your data, how they were collected, and how to calculate basic descriptive statistics.

DIVERGING BAR

A variation on the stacked bar chart is one in which the stacks diverge from a central baseline in opposite directions. These are often found in surveys where the responses are arrayed in ranges from, for instance, *strongly disagree* to *strongly agree*. These are often called "Likert Scales," named after the psychologist Rensis Likert, who invented the scales in the early 1930s.



In this example, drawing on data from the International Social Survey Programme, survey respondents were asked whether they believe it is the government's responsibility to reduce income inequality. By grouping the "disagrees" and the "agrees" together on either side of a central baseline, we can compare the *total* sentiment across the different countries.



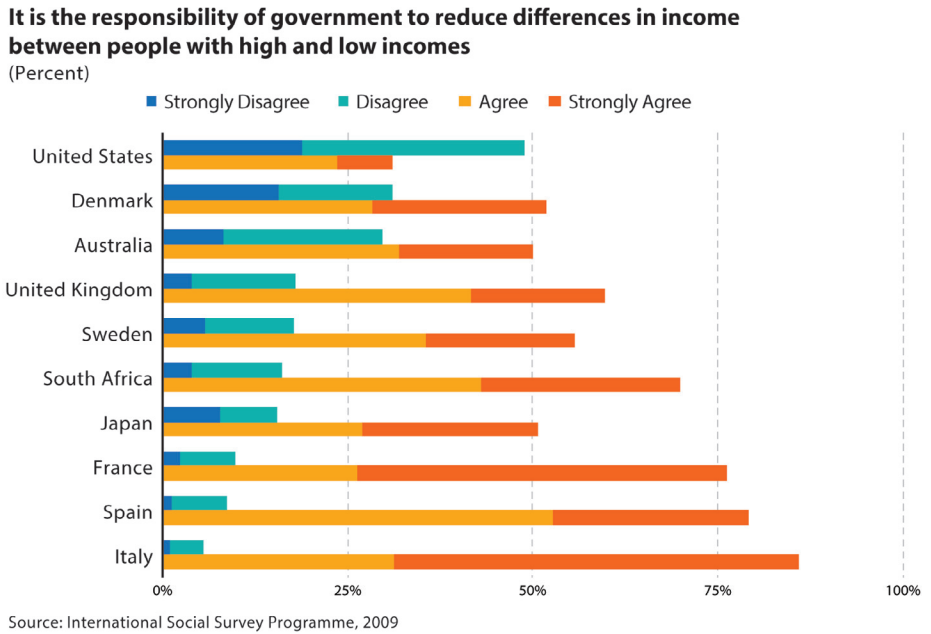
The diverging bar chart can show differences in opposing sentiments or groups, such as "agree/disagree" or "true/false."

One advantage of this chart is that the sentiments are clearly presented—the Disagrees jut out to the left (in what we might typically think of as a negative direction) and the Agrees out to the right. This works well if your audience is most interested in the *total* sentiment of each side and not necessarily comparisons between each individual component. If the individual comparisons are the primary focal point, then a paired bar chart could do the job just as well.

Why do we perceive these values to the left as negative? Throughout western history, the concept of left—and even left-handed people—has been plagued with negative connotations.

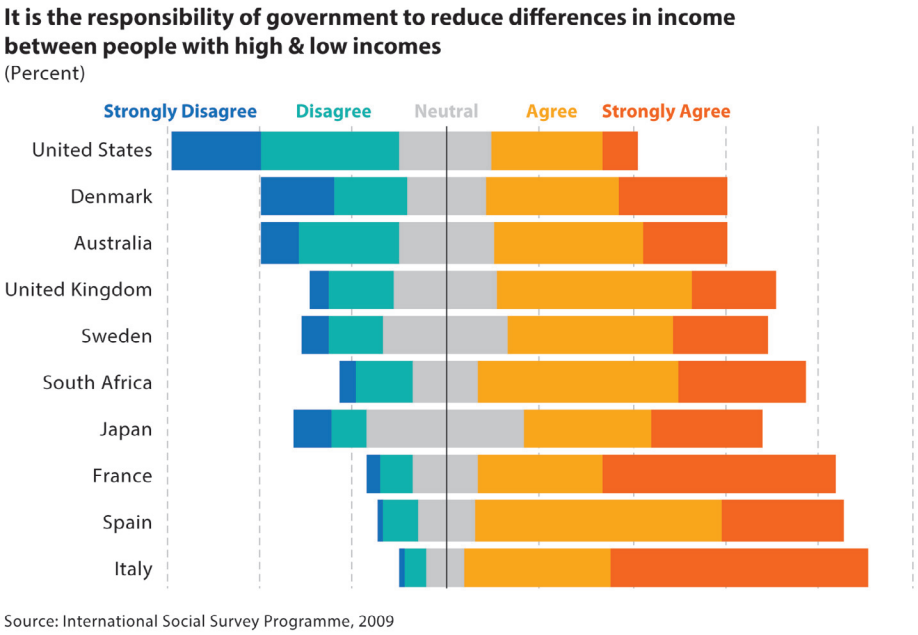
Consider the etymology of the word: *left* is derived from the Old English word *lyft*, which means "weak." In Latin the word *sinister* means the left or left-hand direction. The word *right* comes from the Old English *riht*, whose original meaning was "straight" and thus not bent or crooked. And this is why we have phrases such as "standing upright" or "do the right thing" or "the right answer," all of which connote goodness and correctness. You can also see this in other languages: In Spanish, for example, the word *derecha* means "right" and the closely-derived *derecho* means "straight."

As with the stacked bar chart, the challenge with visualizing these kinds of data is that we are comparing within *and* across the categories. Arranging the bars in opposite directions makes it difficult to compare the totals of the two groups. In other words, it's difficult to compare the *total share* of people who disagree with the *total share* of people who agree. That task is slightly easier in the paired bar chart, but then you lose the positive-negative connotation of the diverging chart. Depending on the patterns in your data and the number of categories and groups, you might find this chart looks cluttered and busy.



Taking the opposite sides of the diverging bar chart and placing them in a more standard paired bar chart approach can also work and allows us to more accurately compare the totals.

You must be especially careful using a diverging bar chart when you have a "neutral" category. By definition, the neutral survey response is neither agree nor disagree, and should therefore be grouped with neither category.

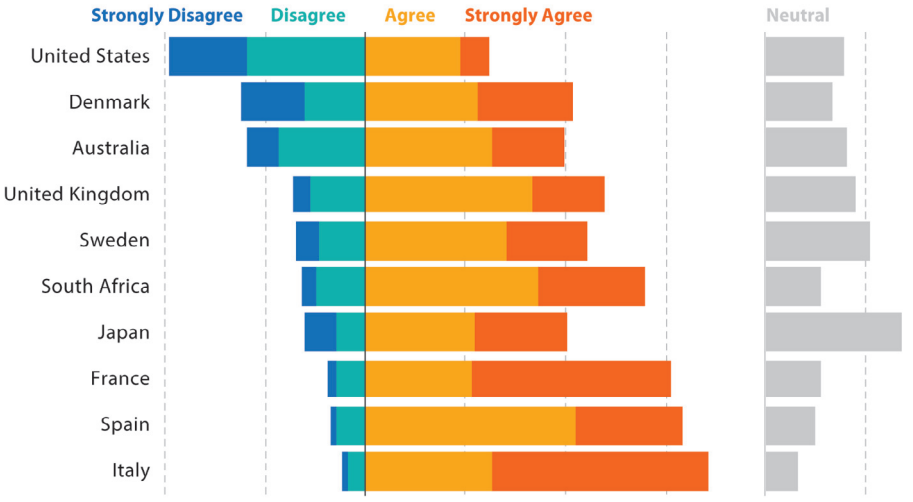


Placing the *Neutral* category of a diverging bar chart in the middle wrongly implies that the neutral responses are split between the two sentiments.

Placing the neutral category in the middle of the chart along the vertical baseline creates a misalignment between the two groups and implies the neutral responses are split between the two sentiments. It also means that none of the segments sit on a vertical baseline. Placing it to the side of the chart is a better strategy because the disagree, agree, and neutral categories now all sit on their own vertical axes, even though the neutral category is somewhat emphasized as it sits to the side (see next page).

Another alternative—regardless of whether you have a neutral category—is the stacked bar chart as shown on the next page. In this view, the different categories sum to 100 percent, and one can more easily compare the totals between the countries. A good strategy is to mark specific aggregate values to guide the reader. Here, for example, I have marked the 50-percent position to make it clear for which countries the total "agree" and "disagree" sentiments are at least half of the total.

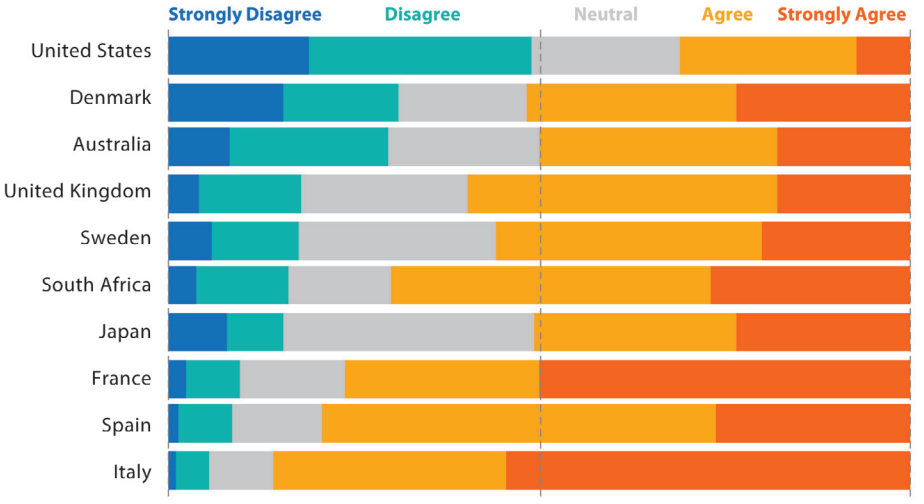
It is the responsibility of government to reduce differences in income between people with high & low incomes
(Percent)



Source: International Social Survey Programme, 2009

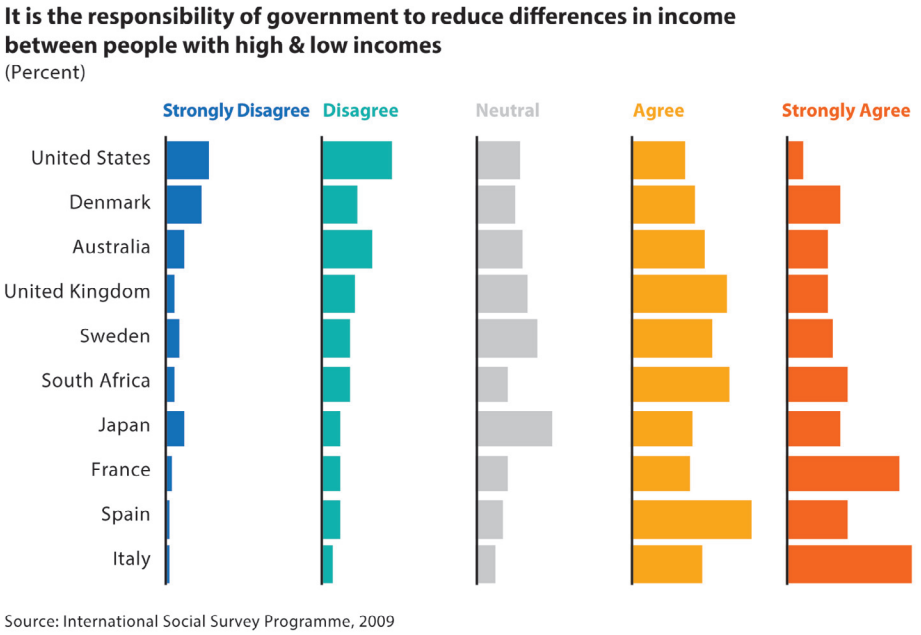
A better approach is to place the *Neutral* category off to the side of the graph.

It is the responsibility of government to reduce differences in income between people with high & low incomes
(Percent)



Source: International Social Survey Programme, 2009

A stacked bar chart can be used to show these kinds of Likert scales.



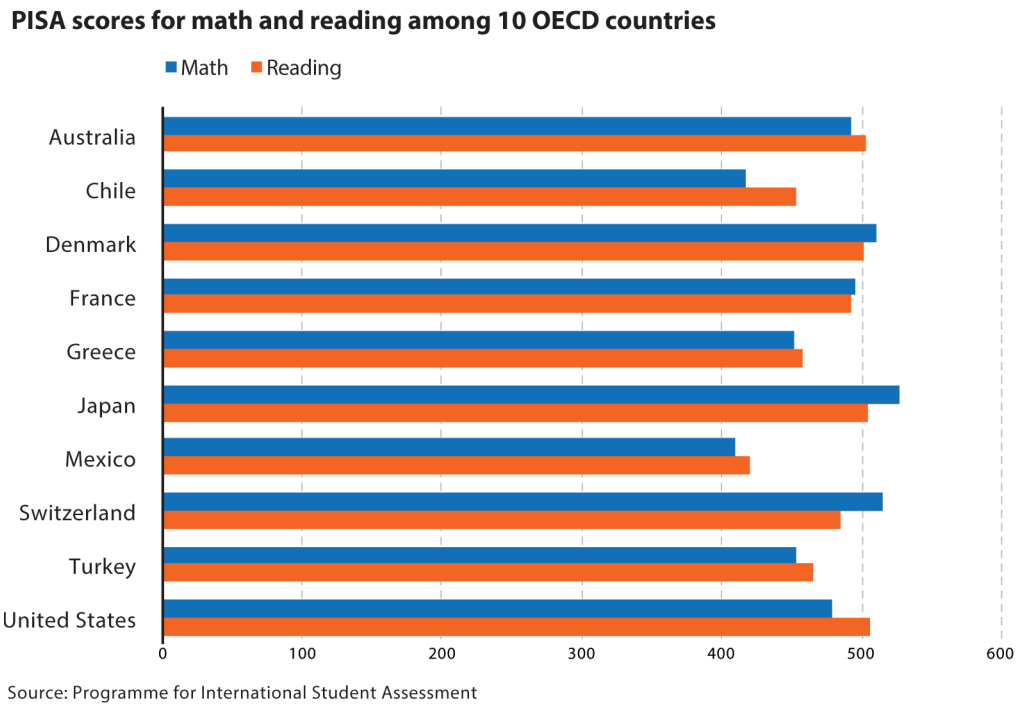
The small multiples bar chart is yet another way to visualize these kinds of data.

You could take this further and break the chart into its components, as we discussed in the previous section. In general, as with many graphs, which variation you choose will depend on your goals.

DOT PLOT

The dot plot (sometimes called a dumbbell chart, barbell chart, or gap chart) is one of my favorite alternatives to a paired or stacked bar chart. Developed by William Cleveland, one of the early pioneers in data visualization research, the dot plot uses a symbol—often but not always a circle—corresponding to the data value, connected by a line or arrow. The data values correspond to one axis and the groups to the other, which do not necessarily need to be ordered in a specific way, though sorting can help.

The dot plot is an easy way to compare categories—especially many categories—when bars might add too much ink and clutter to the page. For this example, let's look at scholastic test



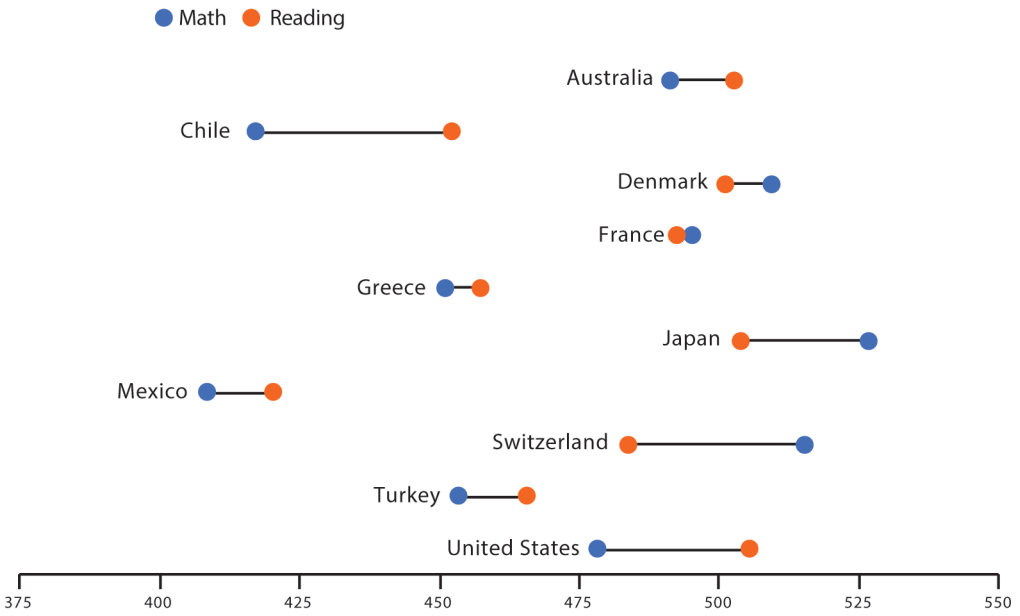
This simple bar chart shows math and reading scores across multiple countries. Bar charts are often the default visual for these kind of data, but it looks heavy and dense.

scores around the world from the Programme for International Student Assessment (PISA), an international set of achievement tests taken by fifteen-year-old students in reading, mathematics, and science. We can easily plot the mathematics and reading scores for a set of countries using a simple bar chart, but the twenty bars makes the chart heavy and dense.

By contrast, the dot plot shows the same data with a dot at each data value connected by a line to show the range or difference. The circles use less ink than the bars, which lightens the visual with more empty space. The country labels are placed close to the leftmost dot, though they could also be set off to the left along the vertical axis. If necessary, data values can be placed next to, above, or within each circle.

Dot plots are not restricted to two dots and a connecting line, nor are they restricted to simply comparing different categories. You can use dot plots to show a change between two years, for example. You could use different shapes or icons or arrows instead of lines to denote direction. You can also use more than two objects. For example, we could add science test scores to this plot, but we need to be sure to add sufficient labeling so our reader

PISA scores for math and reading among 10 OECD countries



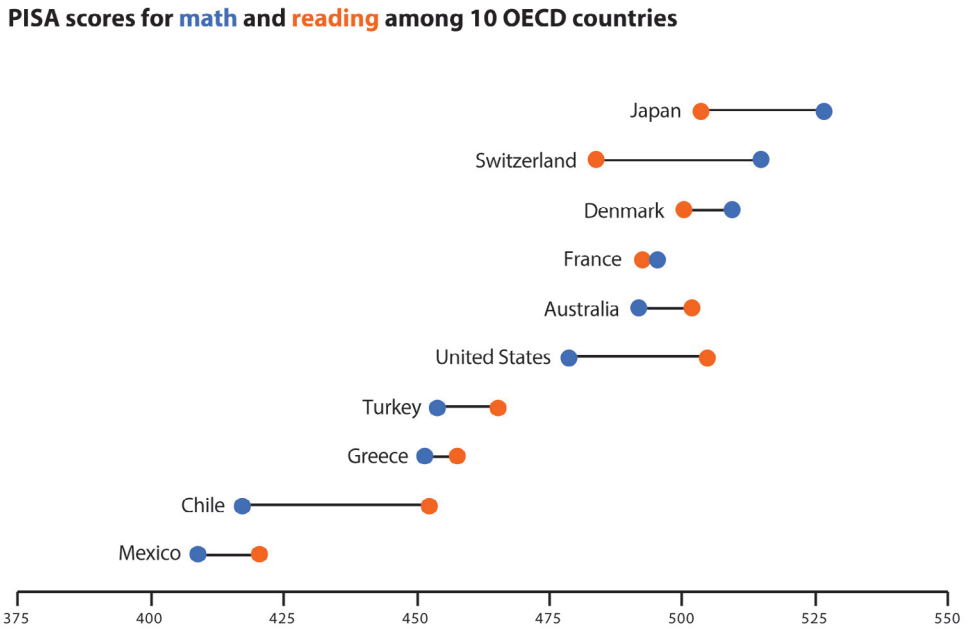
Source: Programme for International Student Assessment

The basic dot plot places a dot for each data point and connects them with a line. Notice how more white space lightens the visualization.

knows what each object on the graph represents. Axes and gridlines can be included or not, depending on how important it is for the reader to determine the exact values.

A few points of caution about the dot pot. First, it's not entirely obvious when the direction of the values change, as in the last chart. Did you notice that math scores were higher than reading scores in four of the countries in the dot plot above? That difference is not immediately evident unless the reader carefully examines the points and their coloring. In this and other cases, we should consider how sufficient annotation, clear labeling, and highlighting colors can help clarify different directions. The data are sorted by math scores in the dot plot on the next page, which helps organize the countries, but it is still not immediately clear that in only the first four countries are math scores higher than reading scores.

One approach is to split the graph into two groups, one for countries in which math scores are higher than reading scores and another for countries where the opposite occurred. In these versions (page 100), the groups are split and then sorted with larger, bold headers to distinguish them. We can also add data values—I will sometimes put them right inside the

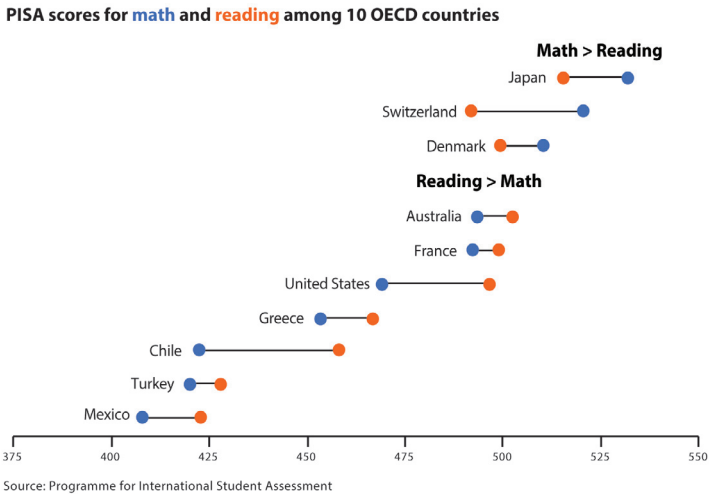


As with the basic bar chart, sorting the data in a dot plot helps organize the space for the reader.

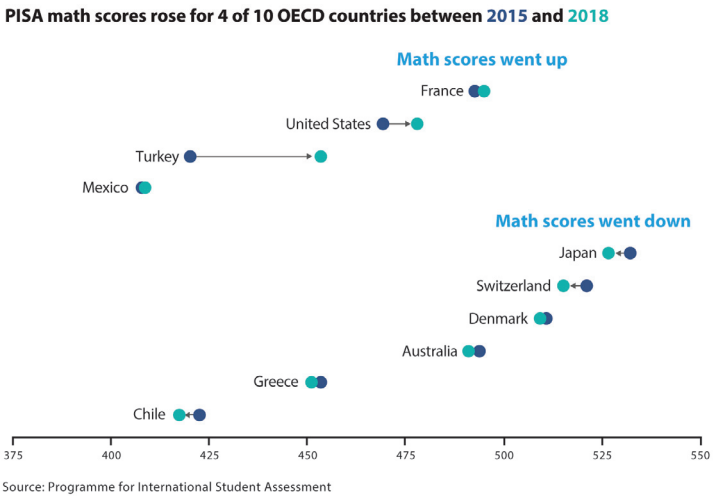
circle—but be careful because the labels can clutter the chart. An alternative is to include vertical gridlines, depending on how precisely we want to communicate the data to the reader.

When using a dot plot to show change over time, I prefer to change the linking lines to arrows, which helps make the direction clear.

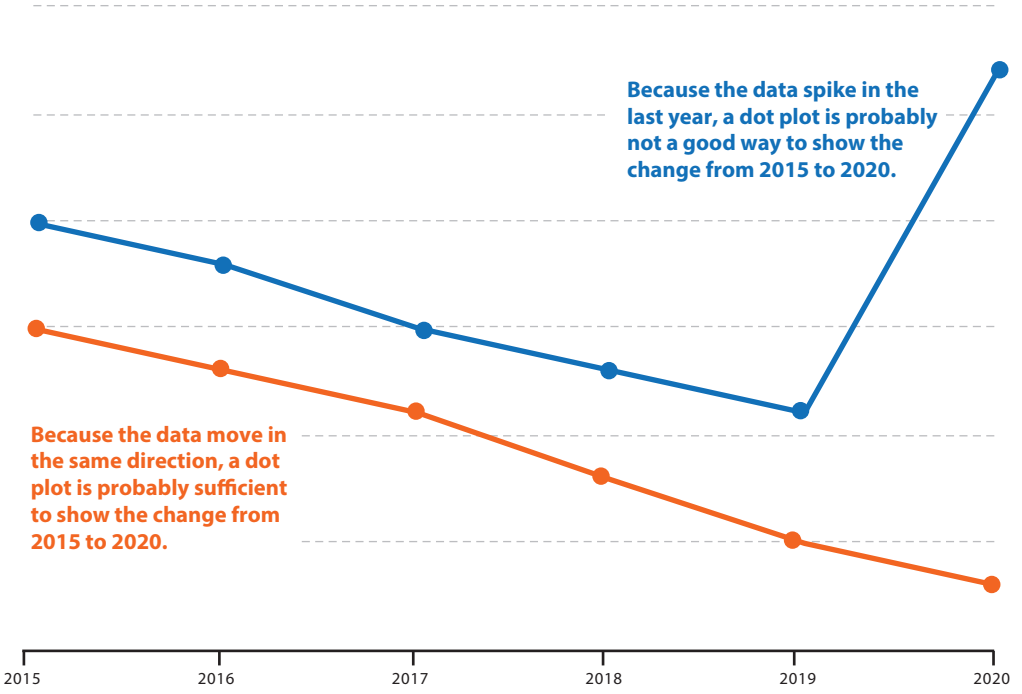
Another word of caution for dot plots that show changes over time. The dot plot is, by definition, a summary chart. It does not show all of the data in the intervening years. If the data between the two dots generally move in the same direction, a dot plot is sufficient. But if the data contain sharp variations year by year, a dot plot will obscure that pattern (as it also does for bar charts). For example, if test scores decreased between 2015 and 2019 and then increased sharply between 2019 and 2020, the dot plot would only show an overall increase, masking the change in the intervening years. In some cases, you may not have a choice—if you are using data from the decennial U.S. Census, by definition you will only have data for every ten years. That's something you can't help, but if you are familiar enough with your content, you'll know whether showing only those points is enough to clearly and accurately make your point.



Labels and annotation can clarify differences in the relationships between the values. Gridlines are not always necessary.



Dot plots can show changes over time. In these cases, I will often make the linking line an arrow to suggest the change over time.



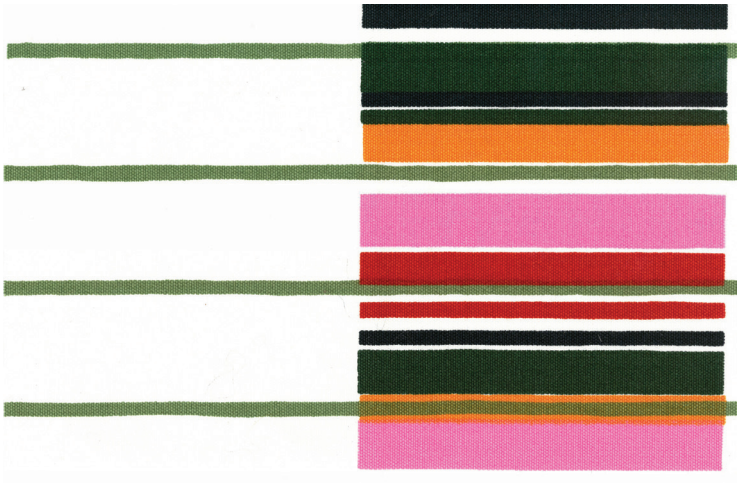
Because dot plots are essentially a summary plot, be wary of using highly variable data with dot plots.

MARIMEKKO AND MOSAIC CHARTS

Marimekko charts may look odd at first, but they are just an extension of the bar chart. This type of chart is useful when you want to make comparisons between two variables: one comparing categories and one showing how they sum to a total. The name of the chart comes from the Finnish design firm Marimekko, founded in 1951 by Armi Ratia and her husband, Viljo. Early Marimekko style featured straight, oversized, geometric patterns and bright colors.

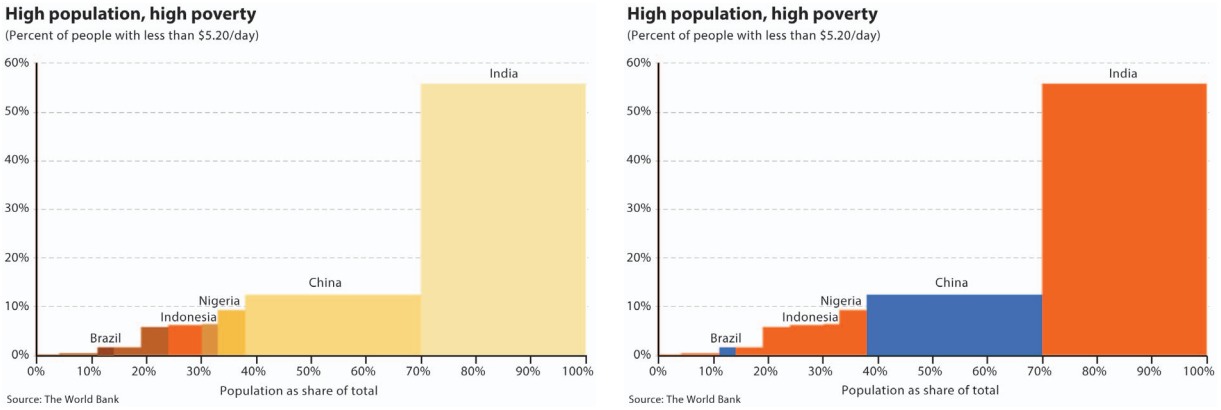
In the standard vertical bar (or column) chart, the data are measured along the height of the vertical axis and the widths of the columns are identical. The Marimekko chart takes that standard column chart and expands the width of each bar according to another data value. The Marimekko chart is an easy way to add a second variable to your standard column or bar chart.

In this Marimekko chart, I show two variables for the ten most populous countries: the share of people with less than $5.20 per day and the share of the total population among these countries. The percent of people with less than $5.20 per day is plotted along the vertical axis



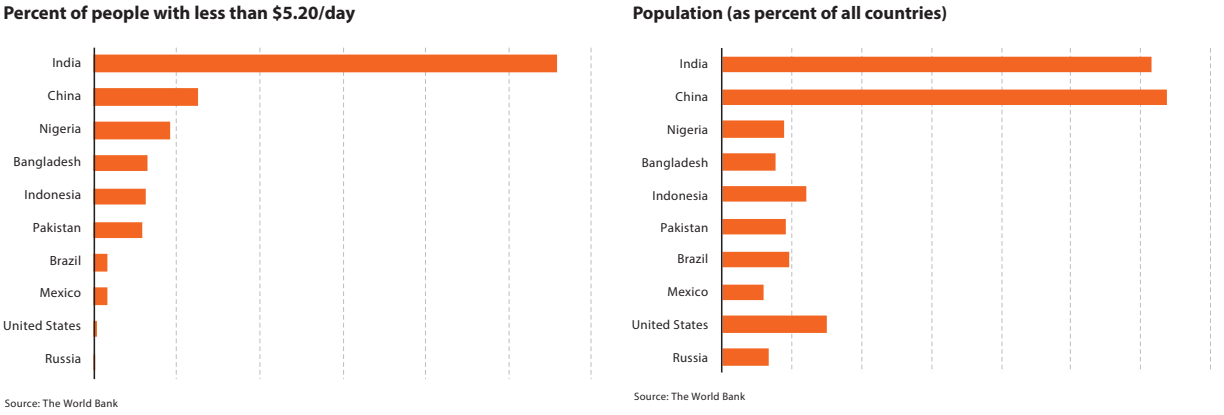
Early Marimekko fabric styles like this one featured straight, oversized, geometric patterns and bright colors.

as in a standard bar chart; the widths of each bar are then scaled according to the share of each country's population summing to 100 percent across these ten countries (an alternative is to show the raw counts rather than percentages). You can see that the most populous countries in this sample—the widest bars—and their distribution of poverty. You can also use color strategically here: if this graph were in an article about poverty in Brazil and China, we might shade all the bars the same color except for those two, as in the graph on the right.



The Marimekko chart (sometimes called a Mekko chart) scales the widths of the bars in a bar chart corresponding to another variable. Color can be used to highlight specific values.

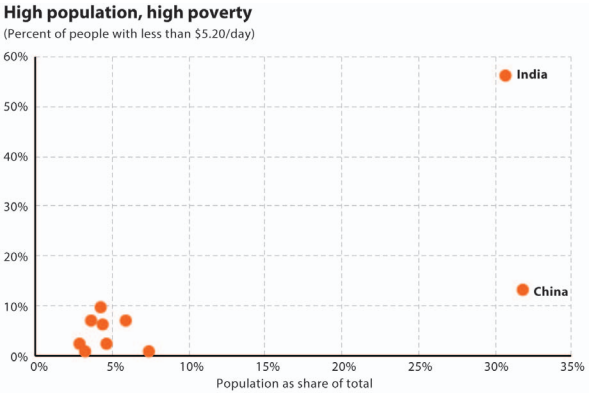
The two variables could also be plotted separately in two bar charts, and while these graphs are familiar and easy to read, they do not communicate the relationship between the two variables as well.

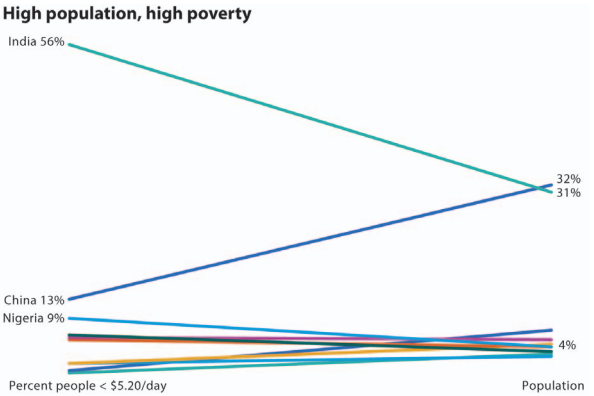


Instead of a Marimekko chart, the two variables could be plotted separately.

Putting two variables on a chart might get you thinking about the association or relationship between the two variables. If that's the case, you could visualize that relationship with other chart types. I've plotted the same data in this scatterplot (see also page 249). You can see how China and India are outliers, especially along the population axis, but the chart doesn't communicate the part-to-whole picture of population. The parallel coordinates plot on the right (see also page 263) similarly shows how many more people live in China and India, and how a greater share of the population in India lives on $5.20 per day. (One potential issue with the parallel coordinates plot is that the lines might suggest a change over time to some readers when instead, in this case, it is being used to compare the two variables.)

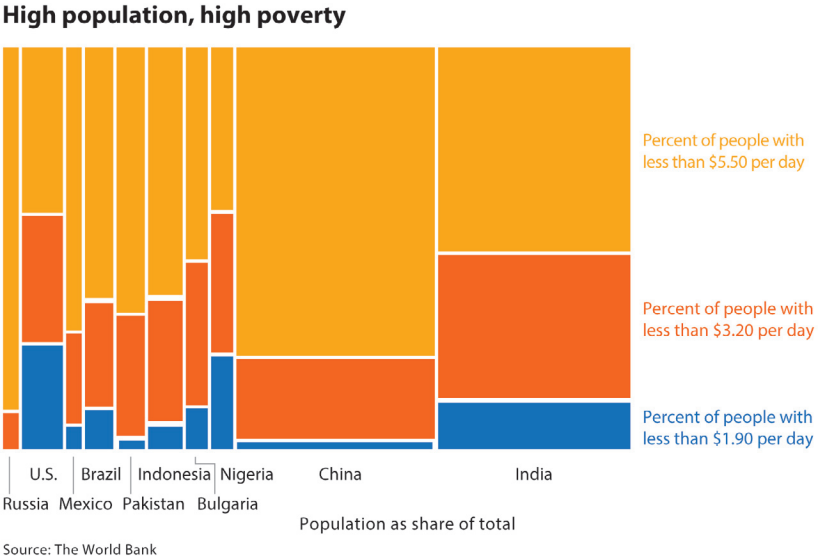
A variation on the Marimekko is to have *both* the heights and the widths of the bars sum to 100 percent. This is sometimes called a mosaic chart, though many people don't differentiate between these two charts and use the terms interchangeably. In this definition of the mosaic chart, you fill the entire graph space and can therefore provide a part-to-whole perspective of the data along both dimensions. In this way, the mosaic chart is also closely related to the treemap (see page 297), but does not necessarily show a hierarchical relationship.



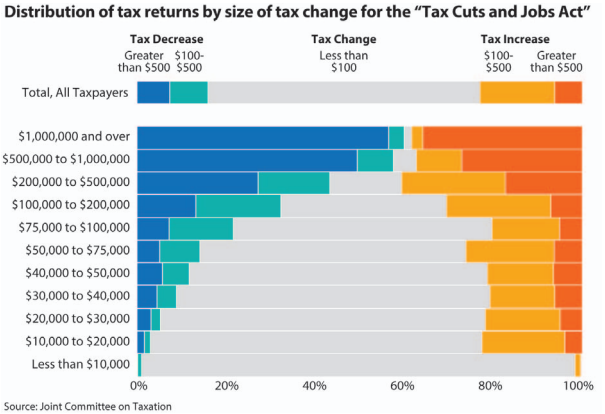


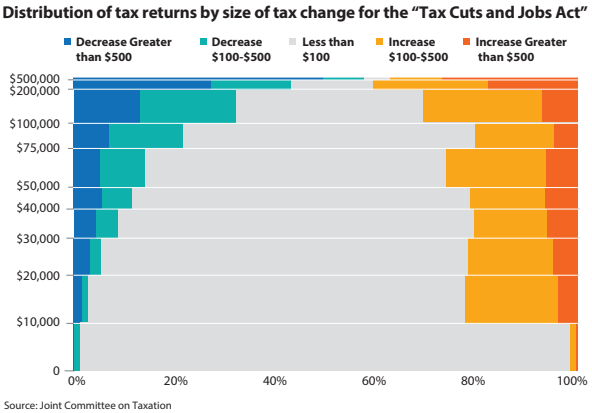
Other ways to plot two data series are a scatterplot (left) or parallel coordinates plot (right). Both are discussed later in this book.

In this example, population is still plotted along the horizontal axis, and the vertical axis now contains three categories for people with low levels of income: share of people with less than $1.90 per day, $3.20 per day, and $5.20 per day.



The mosaic chart is a variation on the Marimekko where both the heights and the widths of the bars sum to 100 percent.





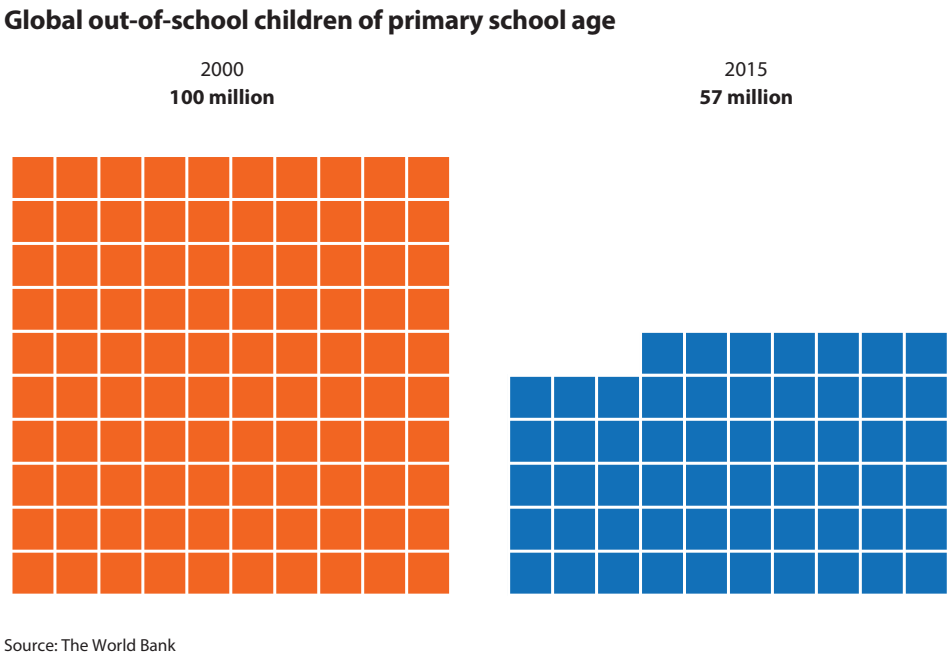
Notice the difference between the stacked bar chart on the left (where all the bars are the same width) and the mosaic chart on the right. While the mosaic chart adds another variable, it is harder to see the details in the top category.

A mosaic chart can also serve as an extension of a stacked bar chart. These graphs show gains from the Tax Cuts and Jobs Act of 2017 for tax units at different points in the United States' income distribution. The stacked bar chart on the left shows five categories of tax gains across eleven income intervals, and each bar shares an equal width. If we scale the widths of the bars to the number of tax units in each income interval—so that the total vertical height of the chart sums to 100 percent—we can create the mosaic chart shown on the right. Notice that the mosaic chart gives a better sense of the distribution of the number of taxpayers in different groups, but because there are relatively few people in the top income group, it's harder to see those values.

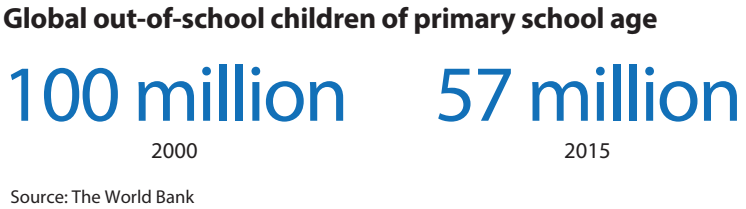
UNIT, ISOTYPE, AND WAFFLE CHARTS

Unit charts show counts of a variable. Each symbol can represent an observation or a number of units. For example, if one symbol represents ten cars and there are ten car icons, the reader mentally multiplies the two for the total of one hundred cars. You can use unit charts to show percentages, dollars, or any other discrete amount. You can arrange them in different directions or break them down into subcategories by using colors or outlines.

Another advantage of these charts is that they can lend themselves to a more human connection. Bar charts, for example, are abstract and impersonal. They collapse all of the people reflected in that data point into a single shape. These charts, on the other hand, offer



Unit charts use symbols to show counts of a variable.



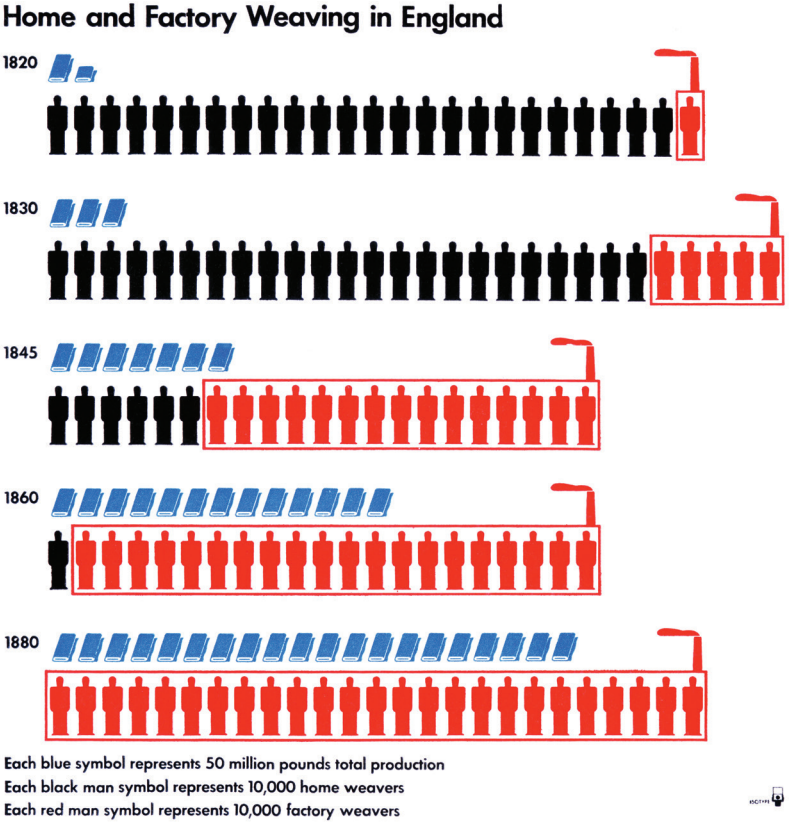
BANs—or Big-Ass Numbers—are a way to just show the values.

an opportunity to connect with the subject by reminding readers of the number of people represented, particularly if each dot represents one person.

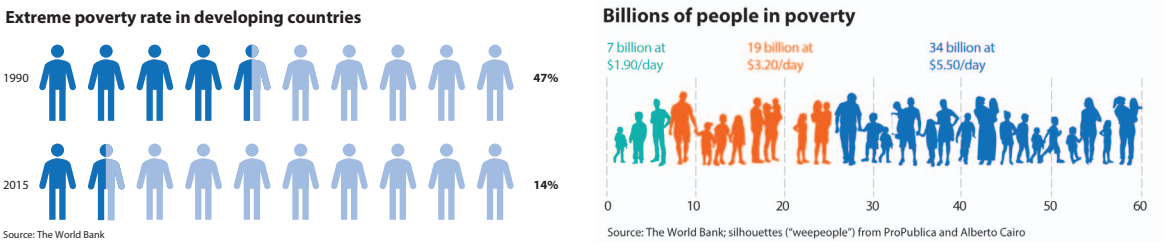
Another simple way to show these kinds of discrete counts is to just show the numbers. In *The Big Book of Dashboards*, authors Steve Wexler, Andy Cotgreave, and Jeff Shaffer call this the BAN approach: "Big-Ass Numbers." BANs might work best in a dashboard, infographic, social media post, or slide deck, but personally, I use them more sparingly in longer reports.

ISOTYPE CHARTS

Isotype charts are a subclass of unit charts that use images or icons instead of simple shapes. The term Isotype—International System of Typographic Picture Education—was coined by German philosopher and political economist Otto Neurath, his wife Marie Neurath, and their collaborator Gerd Arntz in the 1920s. They used the Isotype system to visualize all kinds of data, from workers in different industries, to population density and distribution, to the number of machines used in specific factories. They believed that this kind of visual system would help people communicate demographic, economic, and environmental issues to a broader public regardless of people's educational attainment.



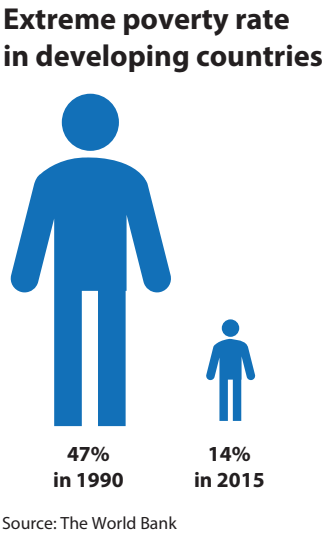
Otto Neurath, Marie Neurath, and Gerd Arntz developed the Isotype chart in the 1920s.



Two different ways to use icons in your data visualizations.

The graphic below is a classic example of their work. Each symbol represents a different count of workers (home or factory) and pounds of production. Aligned along a single vertical axis, it is easy to see how the values change over time.

We can take the same approach with the poverty data we've been using in this chapter. Notice that there's more than one way to use Isotype images in these two charts of extreme poverty rates. The version on the left uses individual icons to show each group of ten percentage points (the lighter icons could be included or not). The version on the right essentially orders the icons atop a bar chart. In either case, the icons connect the subject and content with an immediately recognizable visual image.

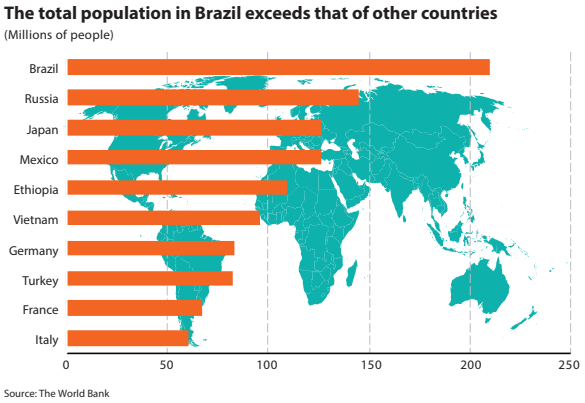


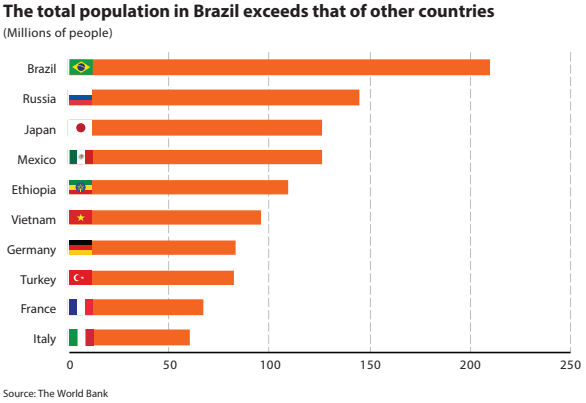
Icons can also be scaled according to their data values, but be careful because it's hard to know whether they are scaled according to the height, width, or area.

Instead of lining up the icons in rows or another arrangement, you can also scale them according to their value. But be careful, as sometimes it's hard to know whether the data are scaled according to the height, width, or area. That may not matter to every audience—it's clear here that the 47 percent is much larger than the 14 percent—but in cases where accuracy is paramount, this icon-scaling approach is inadvisable. Here, the vertical distance represents the data values, but the area of the icon on the left is about 10 times the size as the icon on the right.

These icon-driven graphs may look nice and engage readers, especially for few data values, but they can be difficult for your reader to count and compare. In his 1914 book, *Graphic Methods for Presenting Facts*, Willard Cope Bertin criticized this approach: "Charts of this kind with men represented in different sizes are usually so drawn that the data are represented by the height of the man. Such charts are misleading because the area of the pictured man increases more rapidly than his height." More recently, data visualization author and instructor Stephen Few wrote that unit charts force the reader "to either count, read the numbers, or do our best to compare the areas formed by each, which we can do poorly at best."

But sometimes the downsides of imprecision and slow comprehension may be offset by how memorable the chart is and how it engages readers, an issue that is borne out in recent research. Viewers in one study had clear preferences for graphs that included stacked icons rather than simple bars. They also found that images that sit in the background or are added to a chart but do not depict data are distracting to the reader, so if you choose to use these





Both graphs show the population in ten countries. The graph on the left is cluttered by an unnecessary backdrop of the world, while the one on the right uses flag icons to add identifying detail.

kind of small unit icons or images in your work, be sure to use them only to encode your data and not for gratuitous decoration.

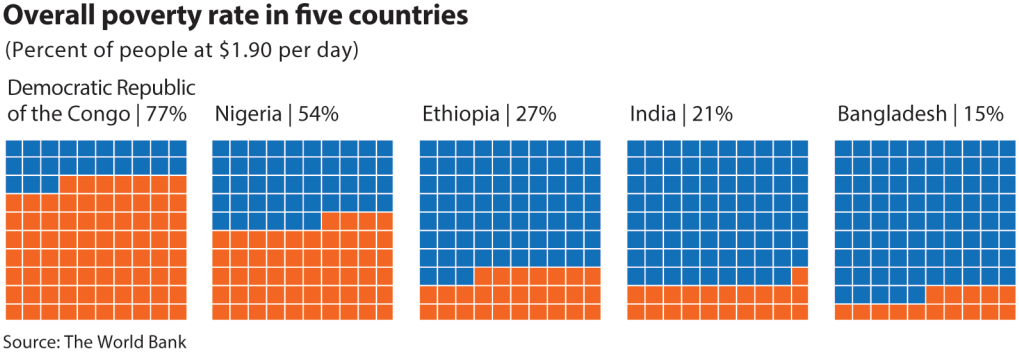
The graph on the left, which shows the population in ten countries, has a backdrop of a world map, a superfluous and distracting decoration. The graph on the right uses flag icons to add identifying detail and some visual engagement to a standard chart type.

Other research suggests that unit visualizations are intuitive and flexible, a way for readers to slowly wade into a visualization. Some have found that "unit visualizations are mostly useful when we want the readers to understand specific data item and encode its value (e.g., a unit, a person, a currency, a region, etc.)." Too much data and too many units, however, can create a visualization that is cluttered and obscures individual data points or the overall argument.

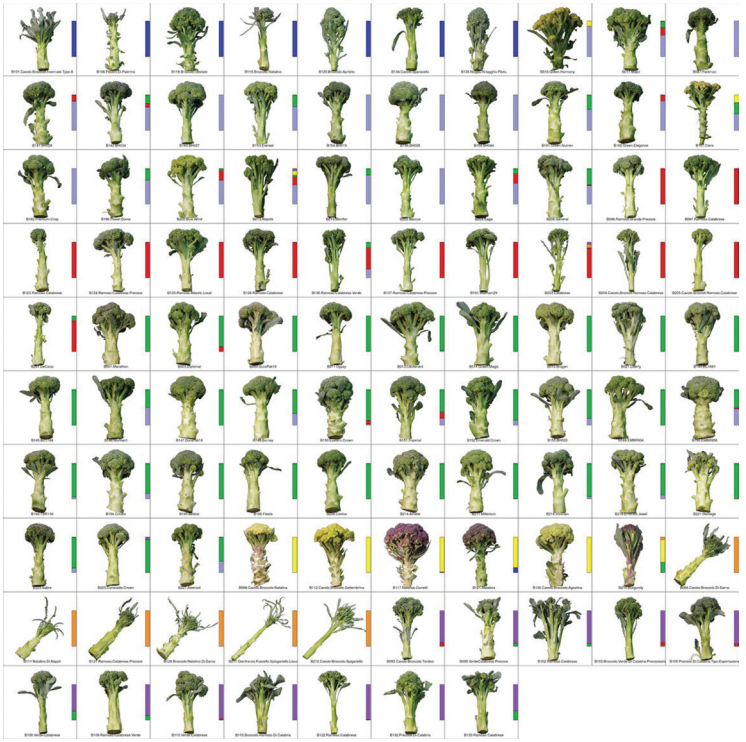
WAFFLE CHARTS

Waffle charts are another subclass of unit charts. They are especially good for visualizing part-to-whole relationships. Waffle charts are arranged in a 10 × 10 grid in which each colored cell represents one percentage point. You can use multiple waffle charts to show separate percentages—so the graph both shows part-to-whole relationships and lets your reader compare across the categories.

When creating unit or waffle charts, especially with icons, be mindful of your audience and how symbols may not appropriately represent your content. If you are visualizing child mortality rates in different countries, for example, an icon of a baby is not appropriate. Using icons of men to represent counts of people may ignore women in your data set. Alternatively,



Waffle charts are a subclass of unit charts and, in this case, arrange the squares in a 10 x 10 grid.



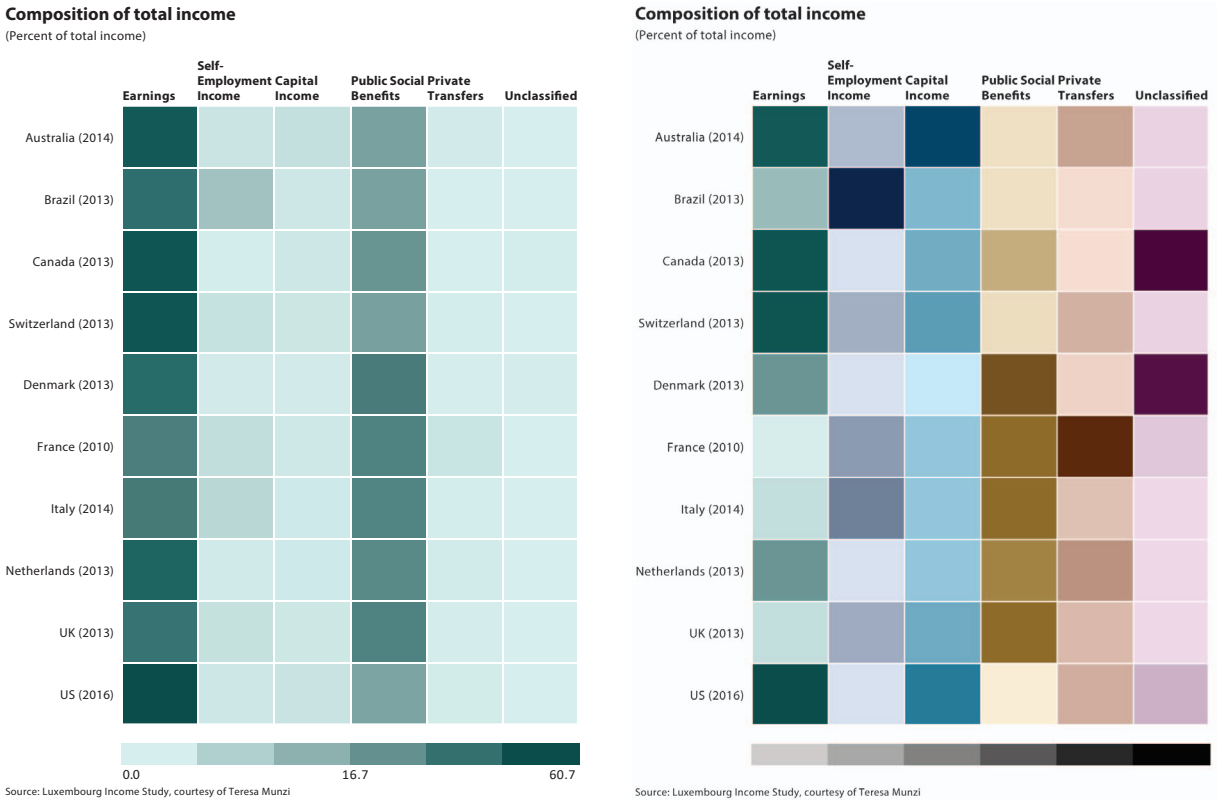
Zachary Stensell created this small multiples visualization of the different types of broccoli he grew in his garden.

if you want to measure the different types of broccoli in your garden, simply line up the pictures, as in this fun project from Zachary Stensell shown above.

HEATMAP

Heatmaps use colors and color saturations to represent data values. Simply put, a heatmap is a table with color-coded cells. They are often used to visualize high-frequency data or when seeing general patterns is more important than exact values.

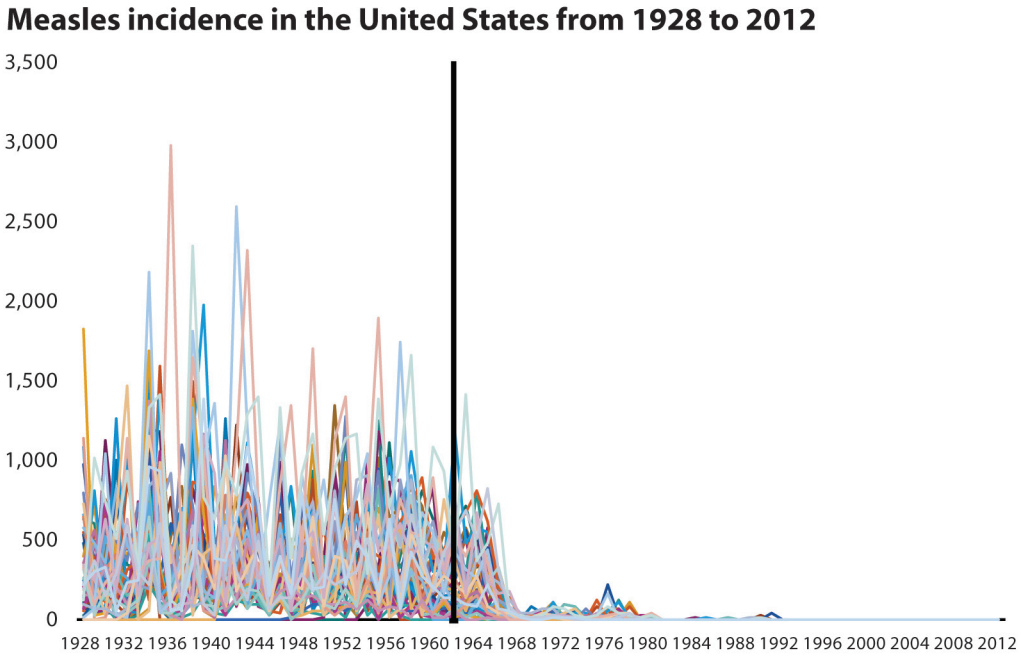
These two heatmaps show the different components of total income for ten countries using data from the Luxembourg Income Study. The version on the left uses the same color scale for all six categories, where lighter colors encode smaller values and darker values



Heatmaps use colors and color saturations to represent data values and can focus the reader's attention along the columns or across the rows.

encode larger values. In this view, you can see that people's earnings account for the greatest share of their total income, and, in most countries, public social benefits appear to be the second-largest share. In the heatmap on the right, each category is assigned its own color scale. Here, you can more clearly see that public social benefits (in the fourth column) account for a smaller share of total income in Australia, Brazil, Switzerland, and the United States. Which one is better depends on your goals. Do you want your reader to compare across all of the values or within each category?

You can also use a heatmap to show changes over time. Imagine a spreadsheet that contains infection rates from the measles disease for every state in America from 1928 to 2008. If the spreadsheet had states ordered along the rows and years along the columns, your first instinct might be to create the line chart on the next page.



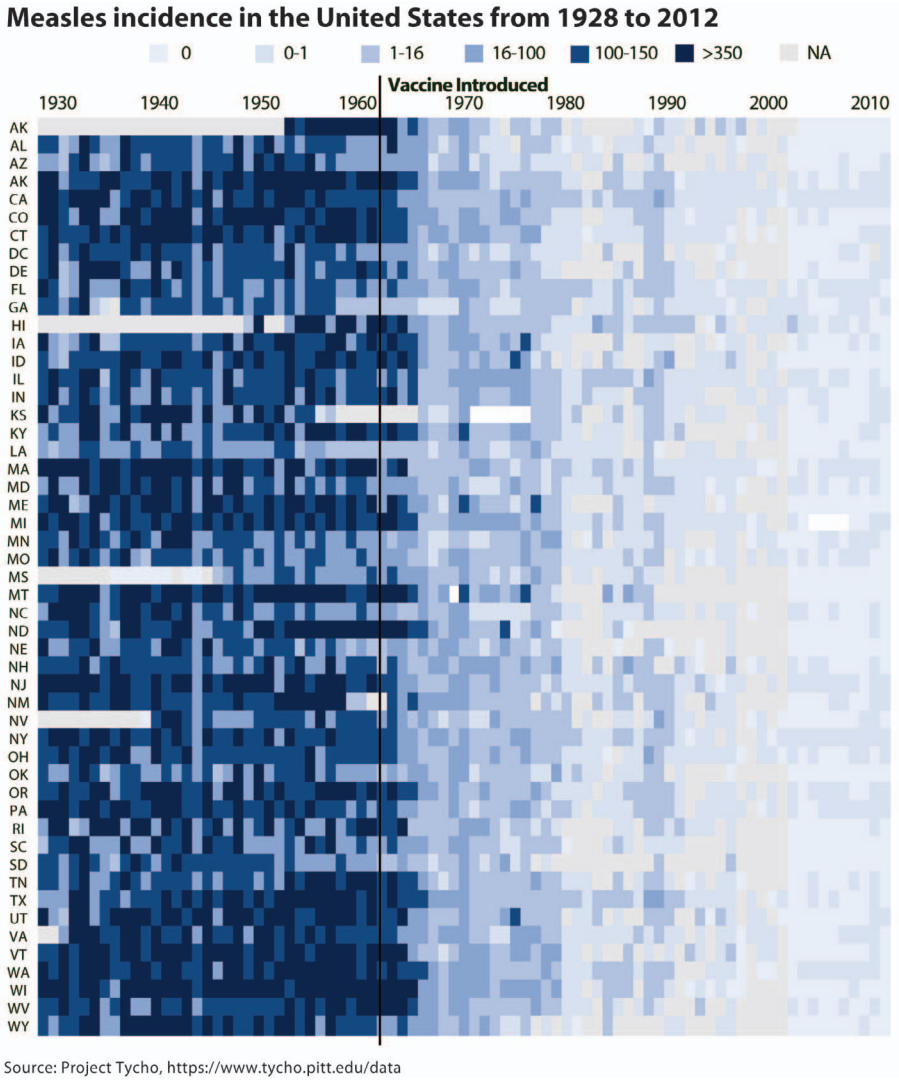
Source: Data from Project Tycho, https://www.tycho.pitt.edu/data

You can see basic patterns in measles infections across the United States in this dense line chart, but it's hard to pick out any specific values.

There's nothing inherently *wrong* with this next line chart—you can see the positive infection rate from 1928 to about 1963 (marked with the black vertical line), the year when the measles vaccine was introduced. Over the next five years or so, infections dropped quickly, and eventually reached around zero within about ten years. What you essentially get from this chart is that there were infections—going up, going down, in a tangle of lines—until about 1963.

The *Wall Street Journal* looked at the same spreadsheet and, instead of creating a dense line chart, they created a heatmap. I've created my own version here using a different color palette and discrete categories of the infection rate. You can see the darker blue cells (mostly above 16 infections per 100,000 people) prior to the introduction of the vaccine, again marked with a black line. After 1963, the colors quickly transition to lighter shades of blue, and ultimately to the lowest rates of infections (zero infections and fewer than 1 infection per 100,000 people).

This chart may not be inherently *better* than the line chart, but it does let you more easily examine each state or year far more easily than picking out a single line from the tangle in the line chart. Also remember that sometimes being different is good in itself. How many



This heatmap may not be inherently better than the line chart at showing measles infections rates, but it does let the reader more easily examine each state or year.

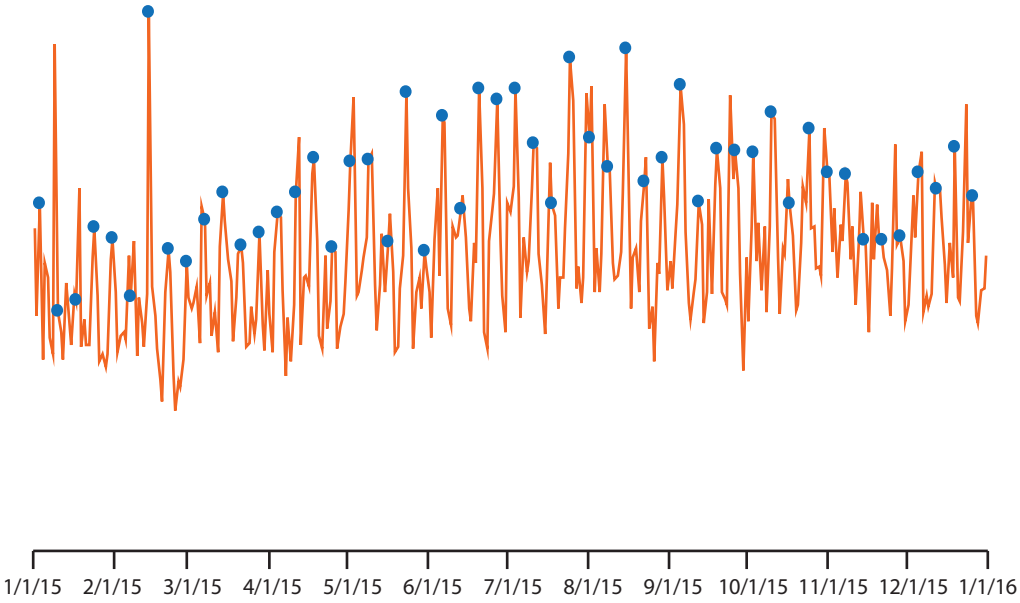
complex line charts have you seen and just immediately skipped over? The heatmap, with its different look and color, can draw readers in. As artist and data visualization expert Giorgia Lupi said, "Beauty is a very important entry point for readers to get interested about the visualization and be willing to explore more. Beauty cannot replace functionality, but beauty and functionality together achieve more."

Another way to use the heatmap is to modify the layout, for example, applying it to a calendar. In this example, vehicle fatalities in 2015 are plotted on heatmaps of months. Notice how easy it is to see the higher fatality rate on Fridays and Saturdays along the right edge of each column.

Source: National Highway Traffic Safety Administration
Note: Inspired by Nathan Yau at FlowingData.com

Another way to use a heatmap is to modify the layout, as in this version that shows vehicle fatalities in 2015.

Vehicle fatalities in 2015



Source: National Highway Traffic Safety Administration

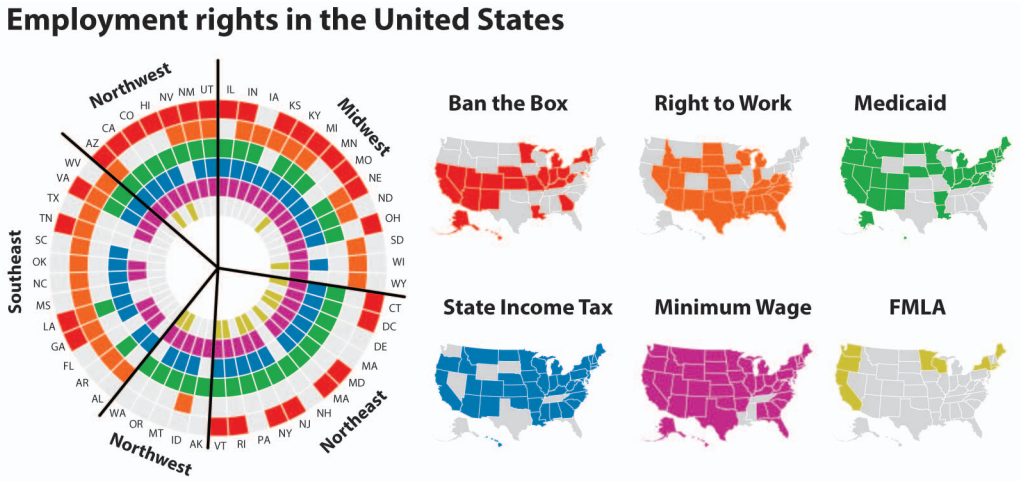
This line chart shows the same data as in the heatmap calendar, but it's more difficult to reach the same conclusion.

By comparison, consider plotting the same data in a line chart. Even with the additional blue circles used to mark Saturdays, it's difficult to reach the same conclusion about more fatalities on the weekends.

Unlike the measles example, where both charts had advantages and disadvantages, in this case the calendar heatmap is a better approach because it does a better job highlighting the important pattern of deaths on the weekends—and it's a more engaging graph placed in a familiar shape.

A final way to modify the heatmap is to change the rectangular layout altogether. On the next page, each of the fifty states is plotted along each radii of a circle, grouped into five geographic regions (separated by the black lines). Each ring represents a different (binary) data type, such as whether the state has right-to-work laws, an income tax, or a minimum wage.

This alternative to a set of six maps has some advantages and some disadvantages. On the one hand, it is a compact representation of six different data series that the reader can quickly and clearly see the categorical and part-to-whole data within and between states. On the other hand, it's not a familiar graph type, and that may turn off some readers. It's also



Heatmaps can be arranged in different ways. The radial layout has some advantages and disadvantages, but so does the six-pack of maps.

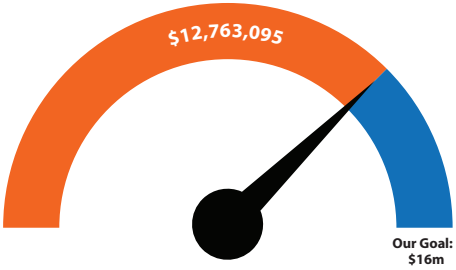
worth noting that the order of the rings can affect our perception of the data because the (red) squares on the outer ring are by definition larger than the (yellow) squares on the inner ring.

GAUGE AND BULLET CHARTS

The gauge chart (or gauge diagram) looks like the speedometer in your car's dashboard. Typically set up somewhere between a half-circle and a circle, it uses a pointer or needle to indicate where your data fall within a particular range. Sections of the gauge are shaded to illustrate sections such as poor, good, and excellent.

I see gauge diagrams most often in financial planning tools because they give an easy, familiar way to visualize targets or progress towards a goal. They also frequently show up in fundraising campaigns where the entire semicircle represents the goal, and the needle and filled area represent money raised so far. This can be a good example of using a familiar shape to support the metaphor of the visualization—everyone understands that "filling up" the gauge means the fundraising effort has reached its goal.

Gauge diagrams do, however, introduce perceptual challenges because, again, people are not very good at measuring and comparing angles. If you want to give your reader a general



Gauge charts are familiar and easy to read.

sense of the values, the gauge chart is a decent choice. But if enabling your reader to discern the specific values and compare those values to the ranges is of utmost importance, then it is not.

Given the familiarity with the gauge and the obvious metaphor it represents, it's perhaps no surprise that they show up in a variety of settings. As just one example, I once received

This series of gauge charts shows four real estate trends in my Northern Virginia neighborhood.
Source: MountJoy Properties, brokered by Keller Williams Realty.

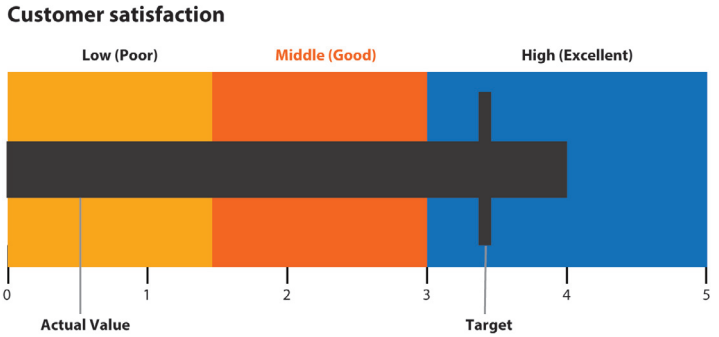
a flyer from Mountjoy Properties (a residential real estate team that serves the DC Metro Area) that consisted of a series of gauge diagrams showing current real estate trends in my neighborhood of Northern Virginia. I could pretty quickly see the current state of the market, but adding more data or more detailed data might be more difficult.

BULLET CHARTS

Because of these perceptual challenges with the gauge chart, author Stephen Few invented the bullet chart, which is a linear, more compact way to show similar kinds of data. The basic bullet chart contains three different data elements:

1. First, there is the actual or *observed value*, shown here as the black horizontal bar. In this illustration, the bar represents an average customer satisfaction score of 4.0.
2. Second, there is a *target value*, shown here as the black vertical line. Here, we were aiming for a satisfaction score of 3.5.
3. Finally, there is the *background range*, which shows grades or bands of success, such as poor, good, and excellent. These sit behind the other two series so the reader can compare the actual and target values. Here, poor scores are 1.5 and below, good scores are from 1.5 to 3.0, and excellent scores are anything above 3.0.

The different components of the bullet charts can vary. There might be a scale of five ranges instead of three, or there may not be a target value. The scales can also show the underlying



The bullet chart includes five separate data values.

Combining different bullet charts is a compact way to let the reader make a series of different comparisons.

distribution of the data—for example, showing quartiles or ranges of percentiles (see Box on page 183 for more discussion of percentiles). Because the bullet chart is so compact, it's easy to create multiple versions and stack them together. The bullet charts above show three metrics you might find in a financial report, but they are more compact than gauge charts and, because the rectangles are aligned, it is easier to compare across the different categories.

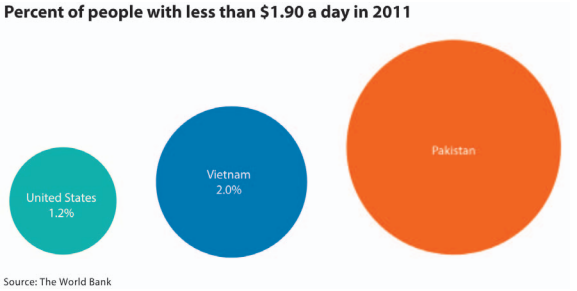
BUBBLE COMPARISON AND NESTED BUBBLES

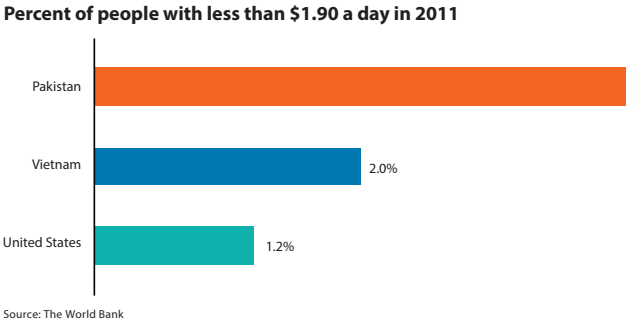
In the basic bubble graph, circles represent values. Like a bar chart, the purpose of these types of charts is to compare values between categories. But unlike the bar chart, humans are not very good at accurately comparing the sizes of circles (remember the perceptual ranking diagram from page 14). Still, circles may be more visually interesting, they can reinforce a visual or metaphor, and they are a good choice when discerning exact quantities is not paramount.

Another drawback of proportionally sized circles is that you cannot visualize negative values. While bars can go in both directions—typically right or upward for positive and left or downward for negative—that's much harder to visualize with circles.

In any case, we are not very good at making accurate estimates from the circles even when they are sized by area. Instead, what I think we try to do is make the comparison based on the diameter of the circle—like in a bar chart—which gives us incorrect conclusions. Take a look at the two graphs on the next page and try to guess the percent of people living on less than $1.90 per day in Pakistan. Do you think that task is easier in the bubble chart or in the bar chart?

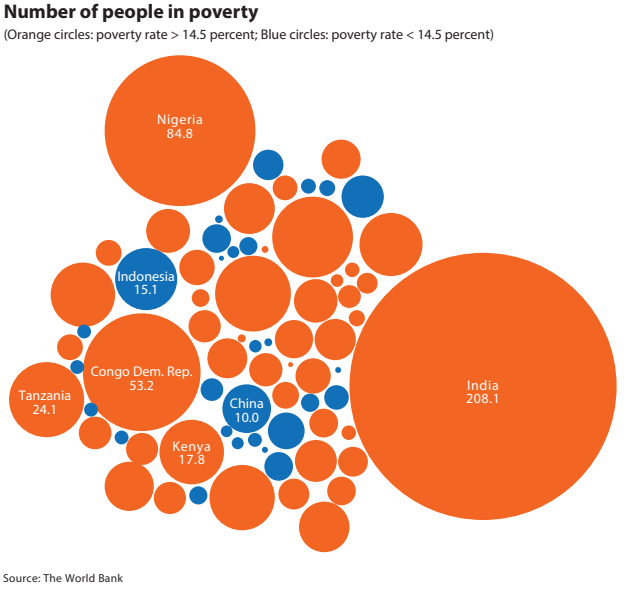
This is not to say that you should never use circles. Remember, again, it's all about your audience. A bubble comparison chart, inserted in a short article off to the side with the





We are better at discerning differences from bars than by areas of circles. By the way, the percent of people with less than $1.90 a day in Pakistan is 4.0 percent.

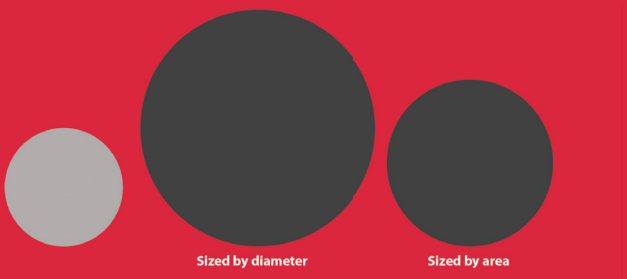
numbers placed prominently in the middle of each circle may be more engaging than a standard bar chart. Too many circles, however, may make it difficult for your audience to discern any quantities or relationships. In this next example, yes, you can see that India, the Democratic Republic of the Congo, and Nigeria have the largest number of people in poverty, but it's difficult to quickly assess *how* different they are or the numbers of the next set of countries.



A bubble comparison chart can be engaging and interesting, but it can also be hard to discern the values.

CALCULATING THE AREA OF A CIRCLE

Remember to size the circles by area, because using the radius or diameter generates circles that overemphasize differences (the radius or diameter scales in a linear way, but the area scales quadratically). The first black circle is sized relative to the gray circle using the diameter while the second black circle uses the area. As you can clearly see, using the diameter skews our perception and makes the difference between the two values look much larger.

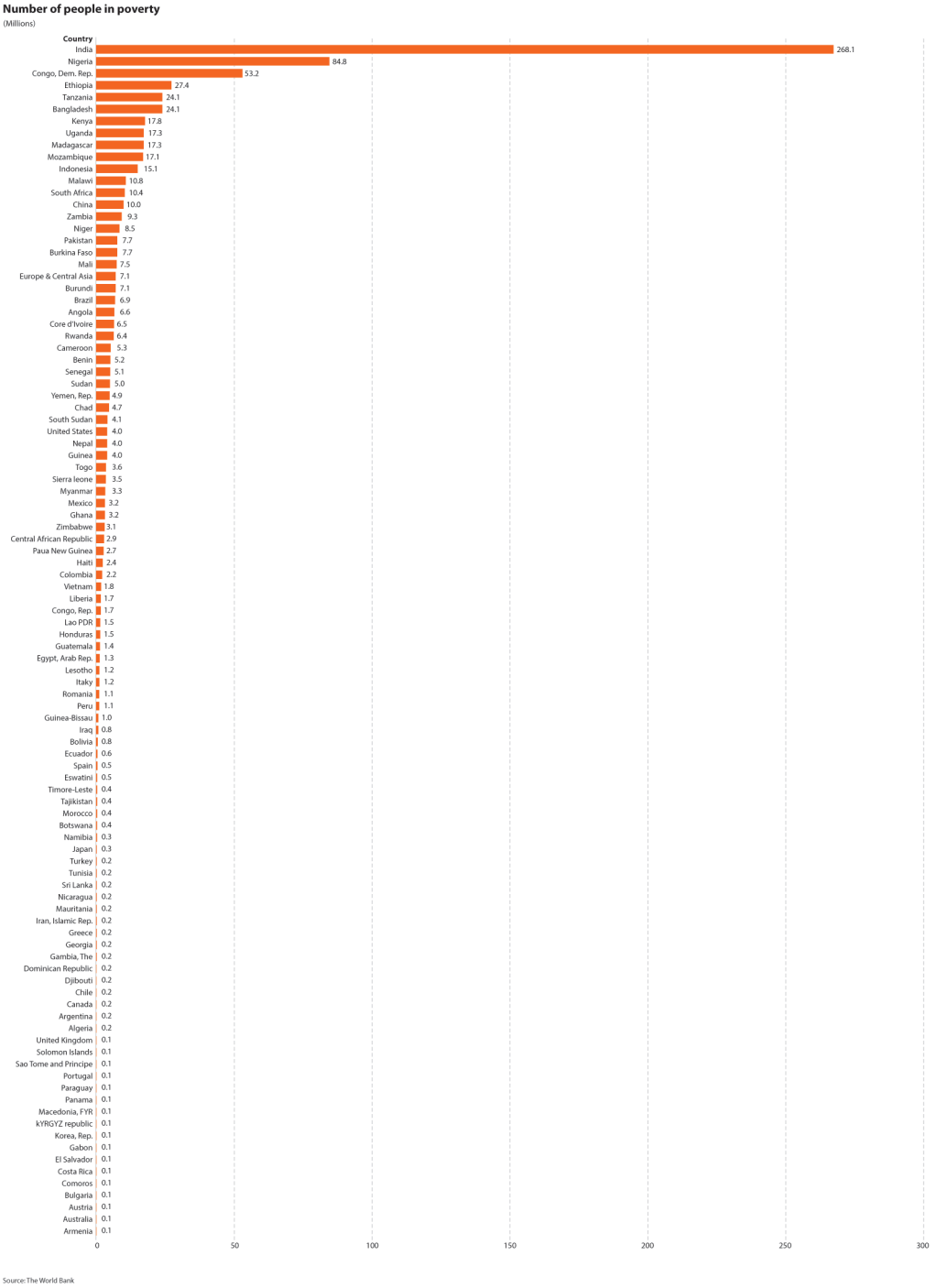


A simple example will demonstrate the importance of using the area rather than the radius/diameter for sizing these circles. In case you don't remember your middle-school math, the diameter is any straight line that passes through the middle of the circle. The radius (r) is half the diameter. And the area (A) is equal to the constant pi (π) times the radius-squared, or $A = \pi r^2$.

So, say the data value for the gray circle is 1 and for the black circle is 2. If we start with the gray circle and set the radius equal to 1, we can find the area is equal to: $A_O = \pi r^2 = \pi 1^2 = \pi$.

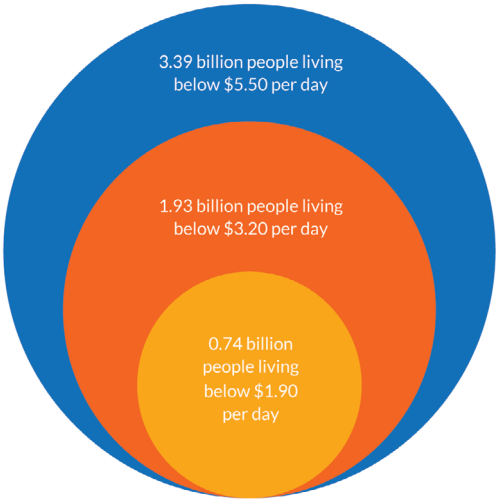
To find the size of the black circle the correct way, we say that the area of that circle is twice the size of the gray circle, corresponding to the difference in their data values. So, if we double the area of the black circle so that it is now 2π, we can then rearrange the formula and find the radius of the black circle to draw it: $r_B = \sqrt{A_B / \pi} = \sqrt{2\pi / \pi} = \sqrt{2}$.

Let's do this the wrong way and use the radius instead of the area. In this case, the radius of the gray circle is still 1, so let's make the radius of the black circle 2, again corresponding to the difference in their values. This makes the area of the black circle now $A_B = \pi r^2 = \pi(2)^2 = 4\pi$. In other words, the area of the black circle sized in this way is four times the size of the gray circle instead of twice the size, as the data suggest.



Source: The World Bank

It's easier to pick out the countries with the most and least poor populations in this bar chart, but it takes up the entire page.



The *nested bubble chart* can sometimes mask circles in the back—but it can also make for easier comparisons.

If we were to use a more perceptually accurate representation of these same data in, say, a bar chart, the visual becomes much larger as on the previous page. Each bar is labeled here, so a reader could find Madagascar, but is that important? Again, is the goal to show all of the countries or just a subset? As always, consider your goals and whether your reader needs a perceptually accurate view to understand your argument.

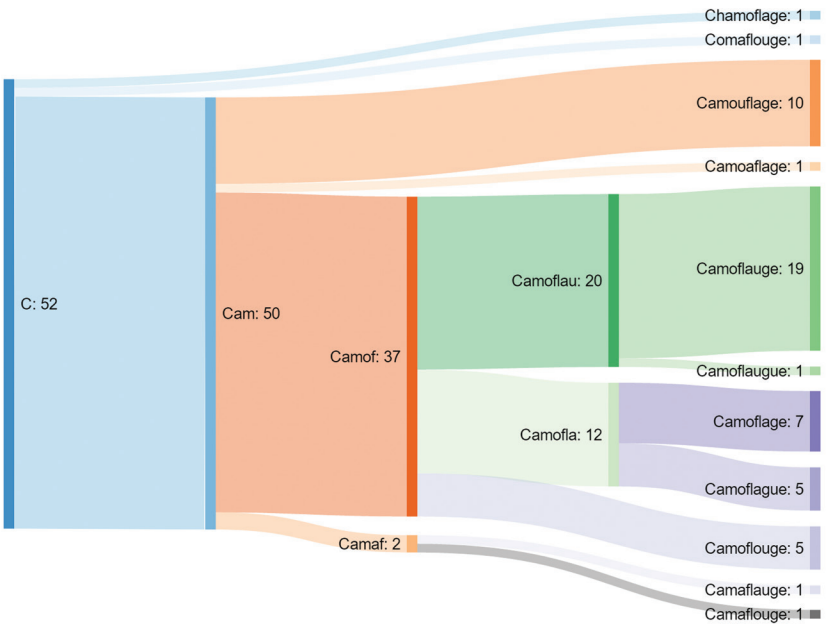
The bubble charts shown above are known as bubble comparison charts. Layering circles on top of each other, as in the graph above, are often called nested bubble charts. The nested bubble chart can sometimes mask circles in the back, but it can also make for easier comparisons.

You can use bubbles to demonstrate correlations (see the bubble chart in Chapter 8) or add bubbles to a map to encode another variable (see the point map in Chapter 7). In general, while there are perceptual issues with using circles and bubbles in data visualization, they can also be more engaging and enjoyable than yet one more bar or line chart. As Amanda Cox, Data Editor at the *New York Times* said, "There's a strand of the data viz world that argues that everything could be a bar chart. That's possibly true but also possibly a world without joy."

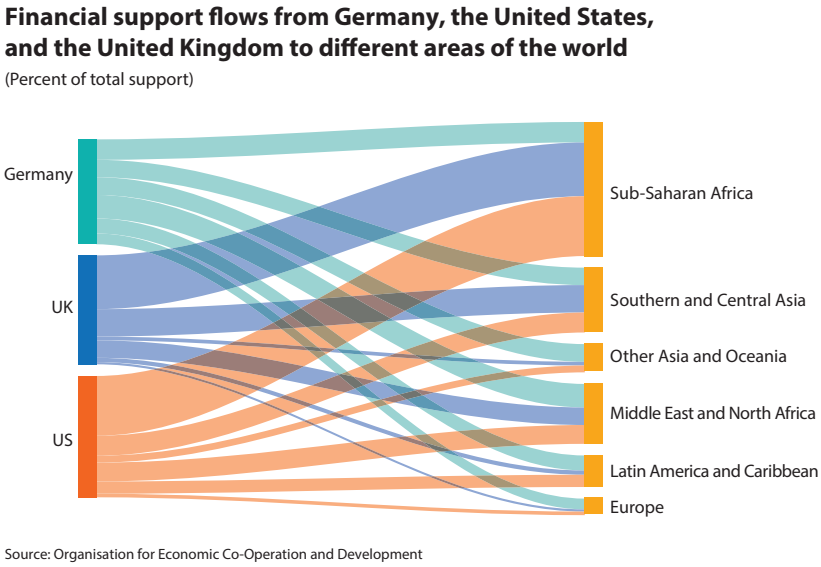
SANKEY DIAGRAM

Sankey diagrams—named for their creator Matthew Henry Phineas Riall Sankey in 1898—are especially useful for showing how categories compare to one another and flow into other states or categories. Arrows or lines display the transition from one state to another, and the width of the lines denote the magnitude of each transition. Changes can occur over time or as comparisons between categories. For example, a Sankey diagram could be used show how different companies in an industry have merged, broken apart, or failed in different years.

This Sankey diagram shows how fifty-two students tried to spell the word *camouflage*. The first blue segment shows that all fifty-two students started with the letter "C", fifty then went to "Cam", followed by thirty-seven with "Camof", and so on. Ten students spelled the word correctly, shown in the orange segment near the top of the graph.



One of my favorite Sankey diagrams—it shows how fifty-two students tried to spell the word *camouflage*. Graphic by Tim Bennett, data collected by Reddit user iheartdna.



This Sankey diagram shows flows in federal aid from the United States, United Kingdom, and Germany to regions around the world.

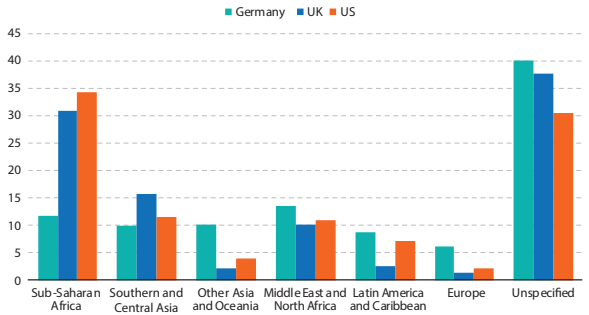
More applicable to the sort of content we've looked at thus far in this chapter, the graph above shows flows in federal aid from the United States, United Kingdom, and Germany to regions around the world. You can see that countries in Sub-Saharan Africa and unspecified countries receive the bulk of the aid and that the United Kingdom and United States contribute more to Sub-Saharan African countries than does Germany.

Presenting these data as paired or stacked bar charts give us a different perspective. In the paired bar chart on the left, the obvious first comparison is across the funder countries for each region of the world. In the stacked bar chart, by comparison you're more likely to compare funding across the recipient regions—that, for example, the greatest share of spending on Sub-Saharan African countries is from the United States. The Sankey draws our attention in a different way, reading horizontally across the page in a way that mixes these two comparisons, privileging neither. Neither view is "right" or "wrong," but may serve different audiences differently, highlight different patterns, and answer different questions.

Sankey diagrams can be layered together with other chart types. In this example, the aid flows from the United States are presented on a world map. This provides a geographic view of the data, but also simplifies things—imagine how cluttered the visual would look if it included flows from the United States, United Kingdom, and Germany.

Financial support flows from Germany, the United States, and the United Kingdom to different areas of the world

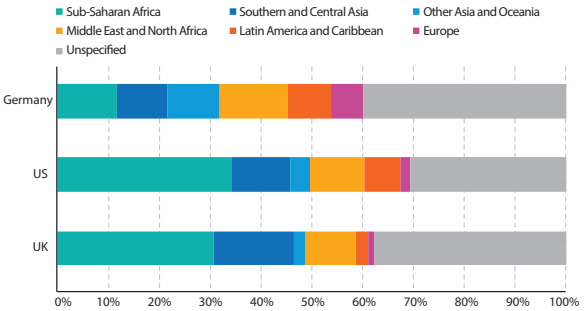
(Percent of total support)



Source: Organisation for Economic Co-Operation and Development

Financial support flows from Germany, the United States, and the United Kingdom to different areas of the world

(Percent of total support)

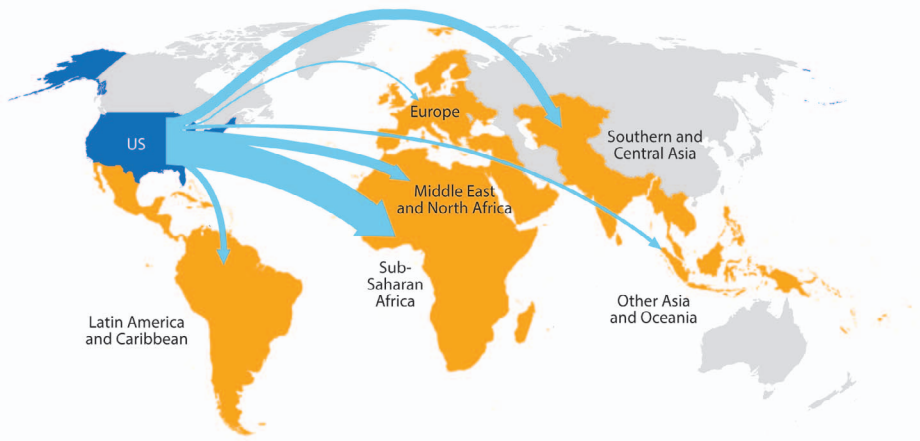


Source: Organisation for Economic Co-Operation and Development

Presenting the financial flow data as a paired or stacked bar chart give us a different perspective than in the Sankey diagram.

Financial support flows from Germany, the United States, and the United Kingdom to different areas of the world

(Percent of total support)



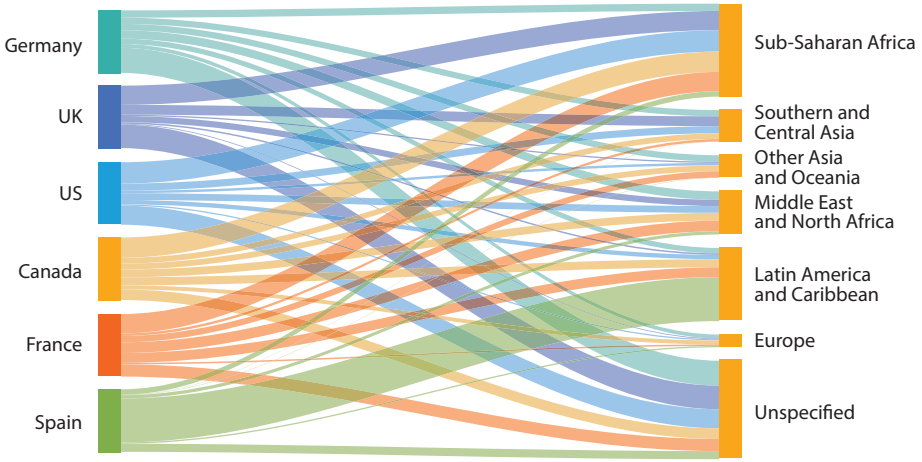
Source: Organisation for Economic Co-Operation and Development

This flow map provides a geographic view of the financial flow data, but it also simplifies things by only showing flows from the United States.

The biggest problem with Sankey diagrams—and many charts for that matter—is plotting too many series, as in this version that includes more countries. With too many groups or too many crossings, the chart becomes difficult to navigate. If you find yourself with too

Financial support flows from Germany, the United States, and the United Kingdom to different areas of the world

(Percent of total support)



Source: Organisation for Economic Co-Operation and Development

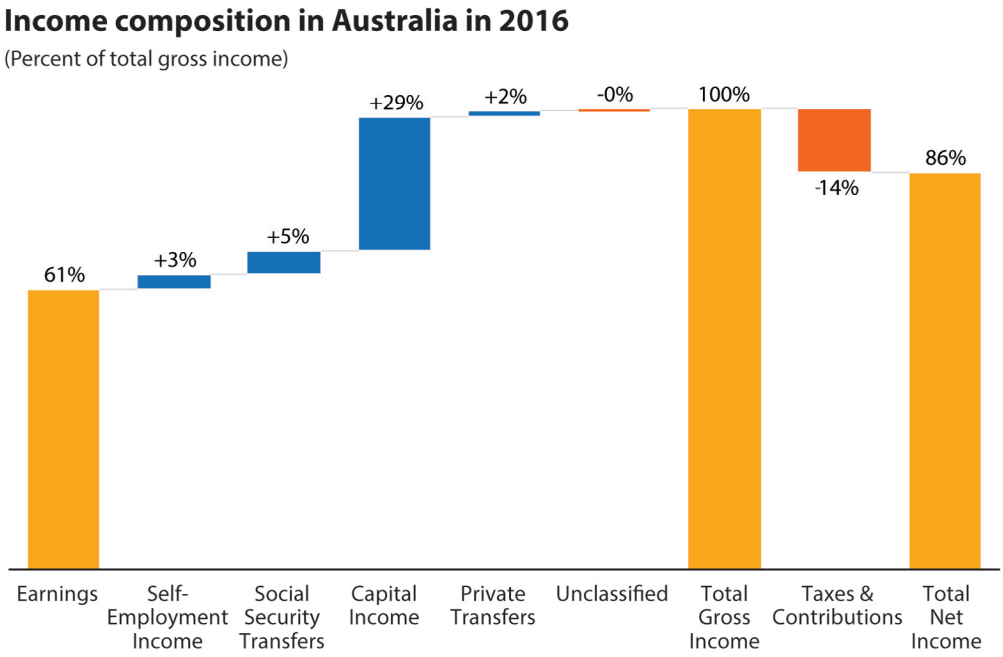
The biggest problem with Sankey diagrams—and many charts for that matter—is plotting too many series makes it difficult to identify any patterns or trends.

many lines or crossings, try simplifying your data, using multiple Sankey diagrams, or using a different chart type.

WATERFALL CHART

A waterfall chart shows a basic mathematical equation: adding or subtracting values from some initial value to produce a final amount. It is essentially a bar chart, but each subsequent bar starts where the previous one left off, showing how they accumulate across the graph. Typically, negative values are given a different color than positive values, and so are the totals at the beginning and end. Including lines that connect the bars can guide the reader through the visualization. Because the lines are guides and not actual data, they should be lighter and thinner than the other elements.

This next chart shows contributions to total gross income and total net income for Australia in 2016. The data are the same as those used in the heatmap example from earlier, but in that approach, I could fit data for ten countries in the same view. Imagine trying to do the



A waterfall chart shows a basic mathematical equation: adding or subtracting values from some initial value to produce a final amount.

same with a waterfall chart—we would need ten different graphs—something that might be useful under certain scenarios, but is certainly less compact than the heatmap.

Waterfall charts can also show changes over time. You might show, for example, contributions to total GDP from one year to the next, and how different values contribute to the change over the course of the year. Any data series that are added or subtracted to one another can be presented in this way, though, again, it is a nonstandard chart type and may require your reader more time to navigate it.

CONCLUSION

From single bars to groups of bars to stacks of bars, the bar chart is one of the most familiar data visualizations for showing categorical comparisons. It also ranks at the top of our

perceptual ranking scale from earlier. But the bar chart also poses certain challenges: too many bars can make the visual seem overwhelming and cluttered, and stacking the series on top of one another makes it more difficult to compare series that are not aligned on the same axis.

The basic bar shape can be organized in many ways. They can sit next to each other or diverge from a central baseline. They can be stacked on top of one another on a horizontal or vertical dimension, or both as in a mosaic chart. They can also be arranged to show simple mathematical equations, as in a waterfall chart. We are generally good at discerning the data values from the lengths of the bars, so many of these chart types will make it easy for your reader to perceive the exact value.

There are other ways to let your reader make comparisons. I'm especially fond of dot plots because they remove a lot of the heavy ink from a standard bar chart and free up space to add annotation and labels. Using icons, squares, or other shapes can engage our audience in ways that standard charts may not, but may be less data dense.

While bar charts sit at the top of the perceptual ranking list, let's be honest: They can be very boring. We see bar charts every day. As chart creators, sometimes our challenge is to find ways to engage our audience, and deploying less common chart types from our data visualization toolbox can do just that. It's up to us to determine where we want to focus our reader's attention, on the level or the difference, the single or multiple comparison, or the relative or total values.

TIME

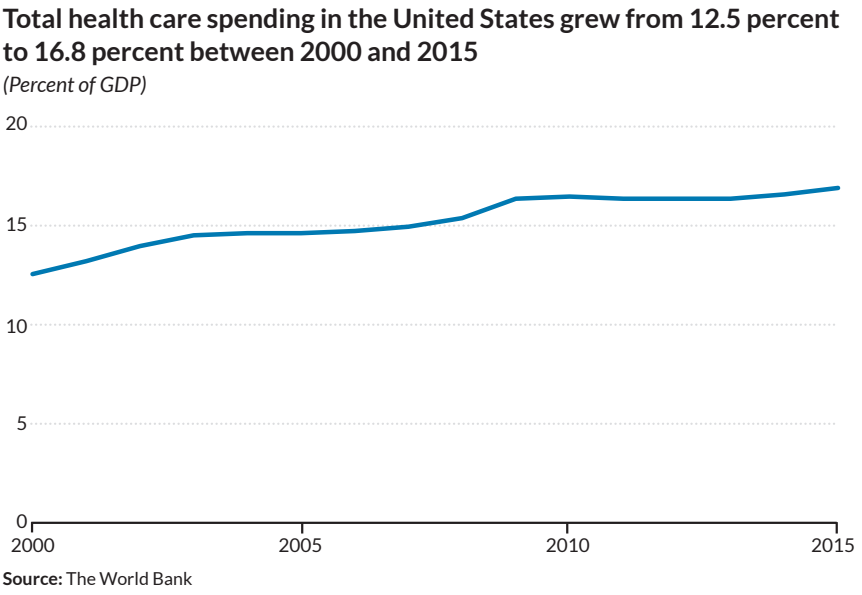
The graphs in this chapter show changes over time. Your readers will most likely be familiar with visuals like line, area, and stacked area charts. But others, like connected scatterplots and cycle charts, may need more labeling and annotation for the reader to navigate them successfully.

Many of the visuals in this chapter are variations on the line or area chart. Some let us include more data on the page than usual, while others allow us to combine changes over time with some other view of the data. With horizon charts and streamgraphs, for example, we can include more data in a single visual, but they are probably not best suited for detailed comparisons. Other graphs, like flow charts and timelines, may use qualitative data or narrative text and visual clues to guide the reader.

Graphs in this chapter are styled following the guidelines published by the Urban Institute, a nonprofit research institution based in Washington, DC. Urban's style guide outlines their color palette, fonts, and guidance for different chart types.

LINE CHART

The line chart and the bar chart may be the most common charts in the world. The line chart is easy to read, clear in its representation, and easily drawn with pen and paper. Data values are connected by lines to show values over a continuous period, tracking trends and patterns.



The basic line chart.

This line chart shows the percent of gross domestic product (GDP) spent on health care in the United States over the sixteen-year period from 2000 to 2015.

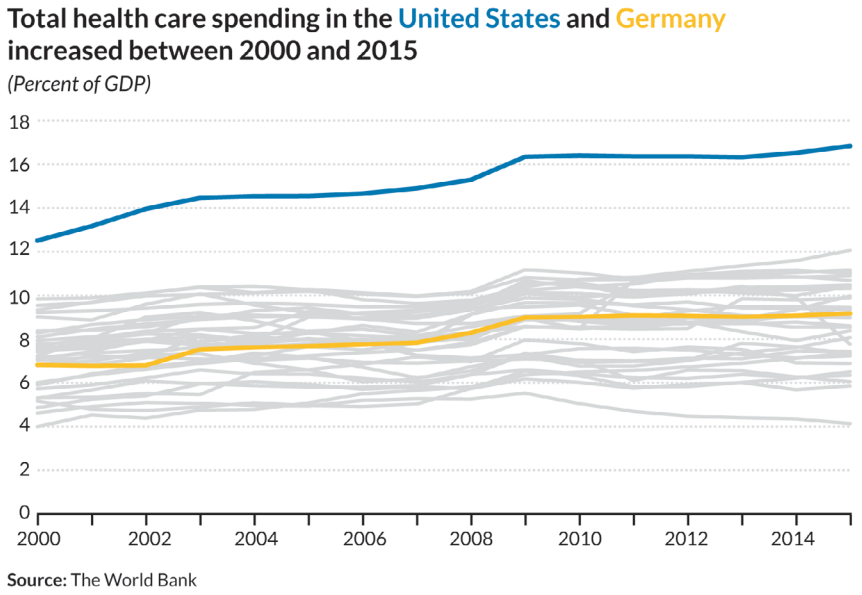
As with the bar chart, the line chart sits near the top of the perceptual ranking scale. With lines relative to the same horizontal axis, it is easy to compare the values to each other and between different series.

As simple as the line chart can be to create and read, there are a number of considerations to take into account, some of which are aesthetic, and some of which are substantive.

THERE IS NO LIMIT TO THE NUMBER OF LINES YOU PLOT

There is no hard rule to dictate the number of series you can include in a single line graph. The key is not to worry about the sheer amount of data on the graph, but instead about the purpose of the graph and how you can focus your reader's attention to it. For example, in a line graph with many series, you can highlight or emphasize a subset of your data.

Say we were interested in showing the share of government spending on health care for the United States and Germany, but we also wanted to show them in relation to the other

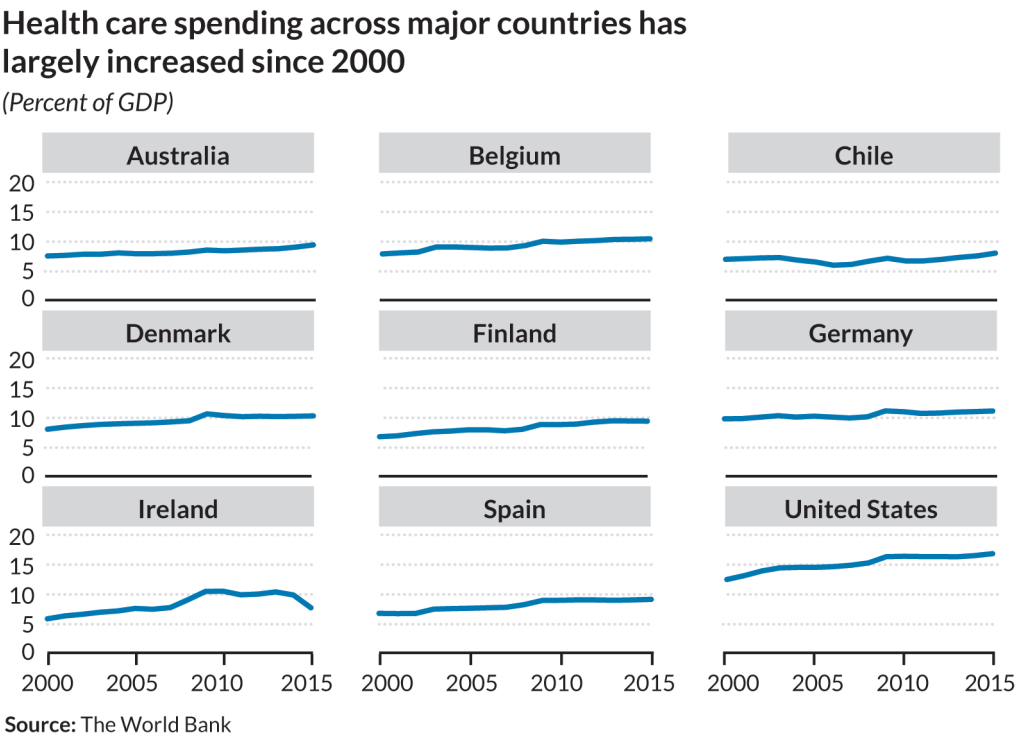


There is no hard rule to dictate the number of series you can include in a single line graph.

thirty-four countries that make up the Organisation of Economic Co-operation and Development (OECD). To do that, instead of giving each line the same color saturation and thickness, we might only color and thicken the lines for the United States and Germany and leave the lines for the other thirty-two countries gray and thin. The "Start with Gray" strategy from Chapter 2 comes in handy here.

Recall the section on preattentive processing: color and line width are two of the preattentive attributes (page 25), thus our attention is drawn to the thicker, colored lines. The advantage of the gray strategy is that the reader can appreciate the general pattern for the entire sample and yet focus on the two lines of interest.

We might also take the line graph and break it into multiple graphs, the small multiples approach. We might include just the line of interest in each small graph, or include all of the lines and use the gray strategy. The set of small multiple line graphs on the next page uses the former approach and shows spending on health care for nine of the thirty-four OECD countries. Instead of forcing all nine lines on a single graph, each country has its own panel. While we might lose some perspective of the *relative* values



The small-multiples approach breaks up a dense line chart into separate components.

of spending in each country, this layout provides more space for each country and thus the opportunity to provide more detail, labels, or other annotation.

YOU DON'T NEED TO START THE AXIS AT ZERO

One of the few rules of thumb of visualizing data is that bar chart axes must start at zero (see Chapter 4). Because we perceive the values in the bar from the length of the bars, starting the axis at something other than zero overemphasizes the differences in values.

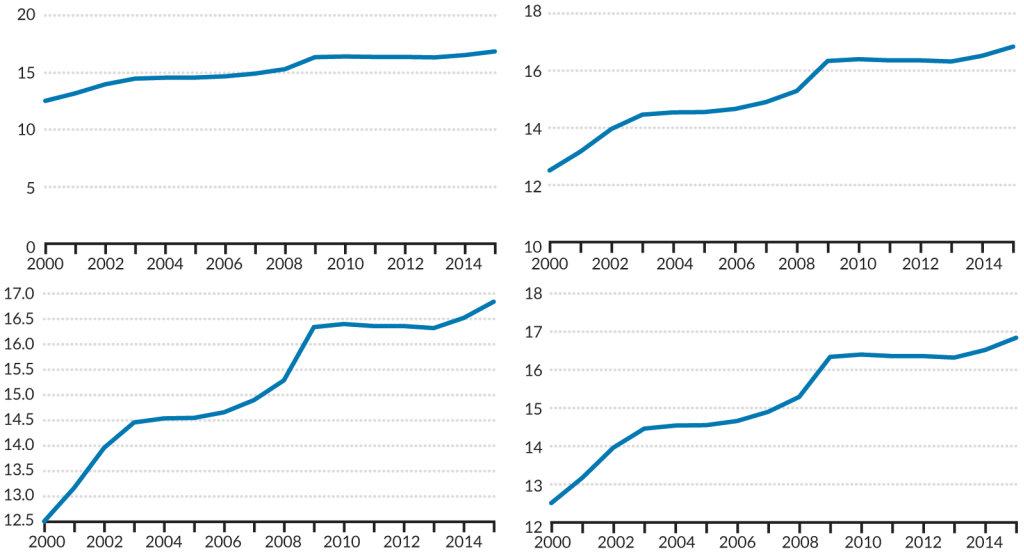
This does not hold true for line charts. The axis of a line chart does not necessarily need to start at zero. As with many aspects of visualizating data, there are complications and different perspectives. If we say the axis does not need to start at zero, what is an appropriate range? Where should we start and end the axis?

To illustrate, let's look more closely at changes in health care spending in the United States. Each of the four charts below uses a different range in the vertical axis. As you can plainly see, those ranges affect our perception of the level and the change in spending. In the top-left graph, where the axis ranges from 0 to 20, we see a slight increase in spending. As you move clockwise through the graphs and the axis range gets smaller and smaller, the change in spending looks increasingly dramatic.

There is no right answer to the choice of the vertical axis; the answer depends on the data and your goal. If you need to demonstrate that the economy will falter if spending reaches 17 percent of GDP, then the bottom-right chart may be best. If you are telling a more general story, then one of the graphs in the top row might be preferable, because it still clearly shows the increase in spending over time. If you need to show a detailed examination of spending in each year, you might want to consider the graphs on the right.

Total health care spending in the United States grew from 12.5 percent to 16.8 percent between 2000 and 2015

(Percent of GDP)



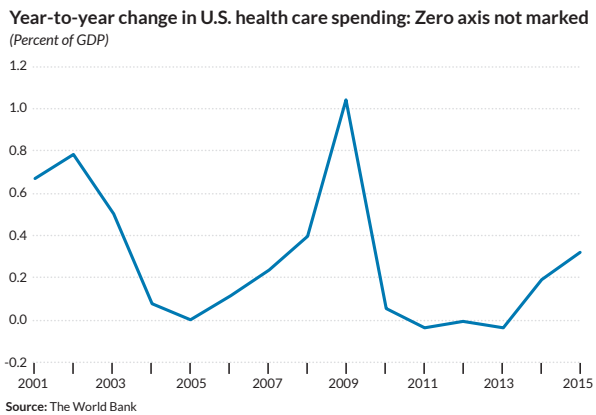
Source: The World Bank

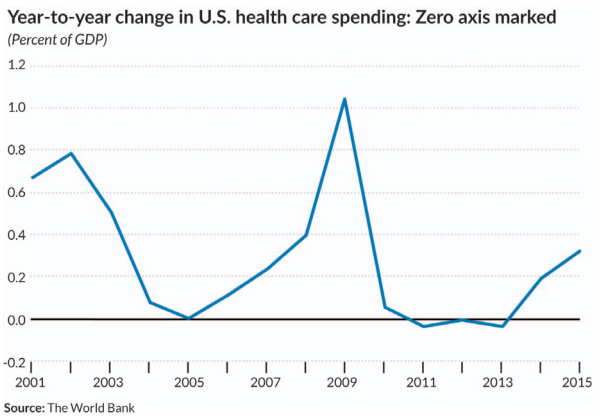
There is no right answer to the choice of the vertical axis. The answer depends on the data and your goal.

Personally, I try to avoid the approach in the bottom-left graph where the axis does not start at zero *and* either the top or bottom value equals the minimum or maximum data value. In this case, the graph feels too "tight" to me and I think it can suggest the data can go no lower or higher than what's pictured, which is rarely the case.

Another way to think about this is how the data, context, content, and units all work together. A change in spending from 12 percent of GDP to 17 percent of GDP is a large change in the context of health care reform. But it's not as important that my kids can beat me seventeen times in a board game now instead of twelve times when they were a bit younger (for me—though it's pretty important for them!).

It's also worth noting that our perception of where zero lies in this space is affected by how we draw the vertical axis. Without looking carefully at a line chart, you are probably inclined to think the bottom of the vertical axis is zero, and in some cases—especially where the data series are both positive and negative—this can be especially important. Take these charts showing the year-to-year *percentage-point change* in health care spending instead of the level of spending as a portion of GDP. In the case on the left, it's not immediately clear that there are any spending declines over this period, because the zero baseline is not clearly delineated. By just darkening that axis line a bit more than the rest as in the graph on the right, it is more evident that there are three years when health care spending as a share of GDP declined year-over-year.



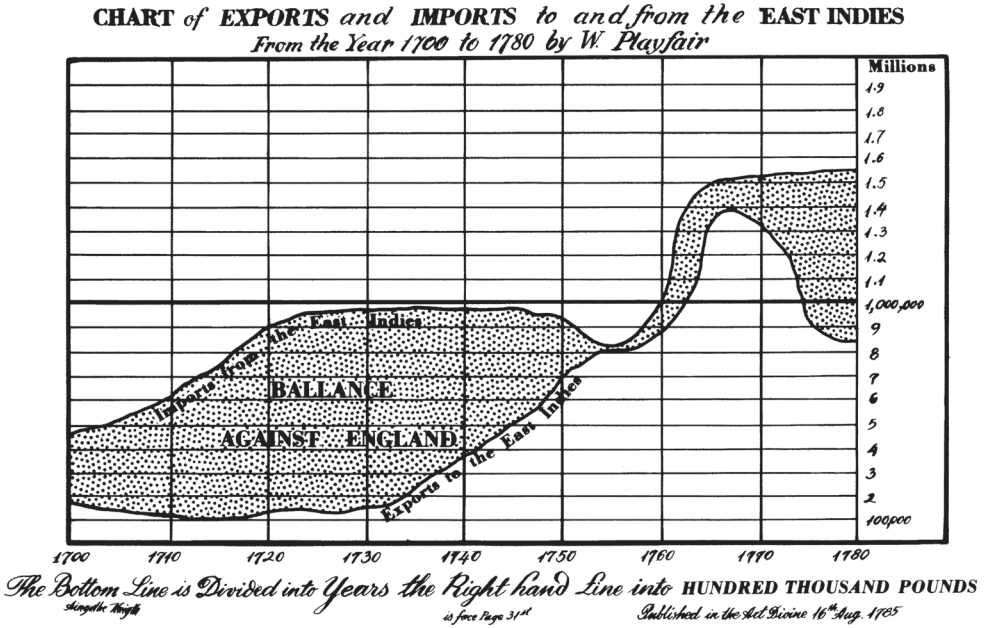


We are inclined to think the bottom of the vertical axis is zero, so using a different color or thickness for the axis makes it clear where zero lies.

BEWARE THE LINE-WIDTH ILLUSION (OR, BE CAREFUL OF THE AREA BETWEEN CURVES)

With line charts (and for other time-series charts, for that matter), we tend to *misestimate* the differences between two curves.

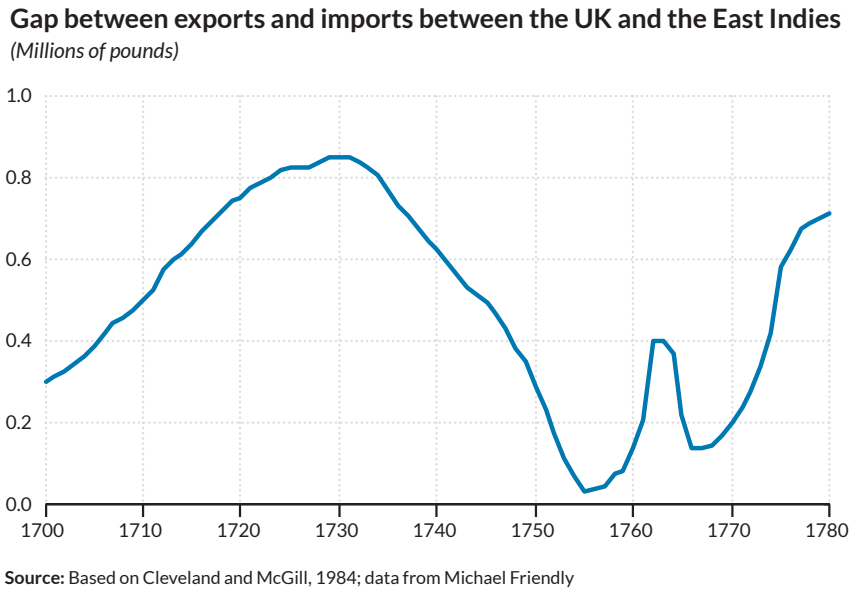
Take this graph from William Playfair, a Scottish engineer and political scientist who is often credited as the founder of graphic methods of statistics. In his chart from 1785, Playfair plots exports and imports (in millions of British pounds) between England and the East Indies from 1700 to 1780. The top line denotes imports, and the bottom line denotes exports. The vertical distance or gap between the two lines shows England's (positive) trade balance with the East Indies. Starting in 1700 on the left, you can see the balance grow over the first thirty years or so. Then, starting around 1730, the gap starts to shrink, reaching its narrowest point around 1755. The trade balance then appears to grow for a time and expands rapidly after around 1770.



In his chart from 1785, William Playfair, a Scottish engineer and political scientist who is often credited as the founder of graphic methods of statistics plots, shows exports and imports between England and the East Indies.

Did you notice the hump in the trade balance after 1760? In the three-year period between 1762 and 1764, imports rose quickly while exports grew more slowly, creating a larger trade balance. Between 1764 and 1766 exports to the East Indies shoots up and brings the trade balance right back down. But the spike between 1762 and 1764 is hard to see in Playfair's original chart. Those changes are much easier to see in this line chart, which plots the gap between imports and exports.

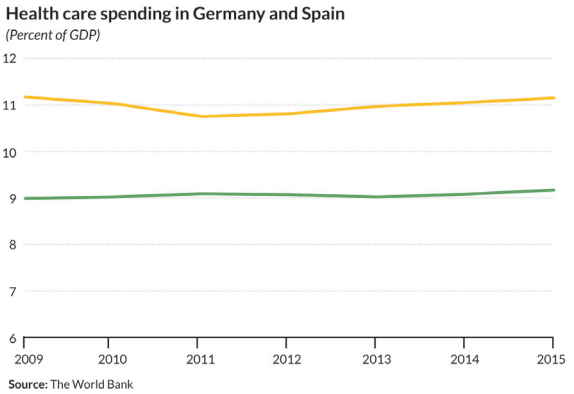
This is the line-width illusion at work: we tend to assess the distance between curves at the closest point rather than the vertical distance. A variety of scholars have demonstrated this effect and have suggested alternative graph types, but the easiest solution may be to plot the *difference* whenever it's the metric of interest.

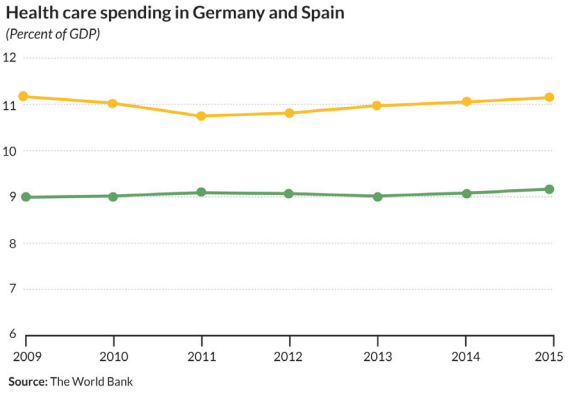


The line-width illusion at work here—the bump in the gap between exports and imports is easier to see here than in Playfair's original.

INCLUDE DATA MARKERS TO MARK SPECIFIC VALUES

Data markers are just what they sound like: symbols along the line to mark specific points in the series. There isn't a right answer to the question of when to deploy them. Personally, I include data markers when I have only few lines or data points, or for specific points I want to label or annotate. The data markers give the graph more visual weight.



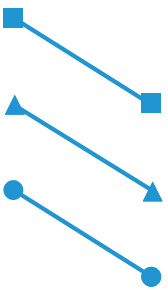


I like to add data markers to my line charts when I don't have a lot of data or when I want to highlight or label specific values.

These charts, for example, show health care spending as a share of GDP for Germany and Spain. There are so few data points and the changes in the series are so subtle that the addition of the circular data markers give the lines more visual heft.

I prefer to make my data markers circles rather than triangles, squares, or other shapes. This is partly an aesthetic preference, but there's also a logic to it. Circles are perfectly symmetrical, and so it never matters where the line intersects the circle. With other shapes, like triangles, the line might intersect the thinner top part or the thicker bottom part.

Other shapes may be necessary if you or your organization are required to comply with certain rules or laws that enable screen readers to differentiate between objects on a screen for people with vision disabilities. In the United States, federal government agencies are required to follow Federal Section 508 regulations that make websites accessible to people with disabilities (see Chapter 12 for more on data visualization accessibility). Even with different colors, most screen readers cannot differentiate between the different series if the shapes are all the same. In these cases, different data markers are a good choice.

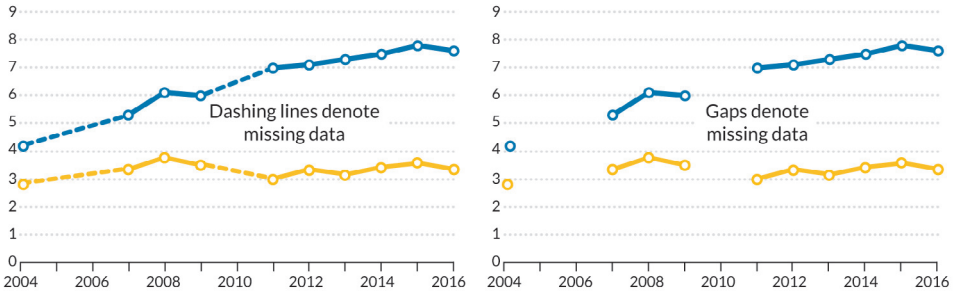


Circles are a symmetrical shape, which is why I prefer to use them as data markers.

USE VISUAL SIGNALS FOR MISSING DATA

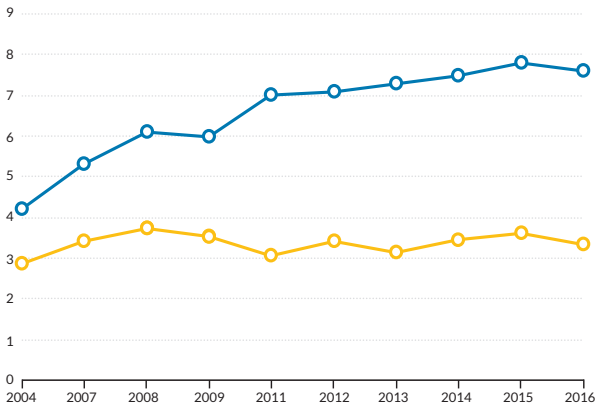
At some level, there is *always* missing data. People change jobs every day, not just when unemployment numbers are published. Lots of things happen in the ten years between the publication of each U.S. Census. Most data are a snapshot in time, but we often treat them as continuous.

Missing data are *truly* missing when a regular series is interrupted because the data were not collected. In these cases, we should make it clear that the data are incomplete. In line charts, we can change the format of the line (for instance, with dashes) or not connect the points at all to signal that those data points are missing. We can also place a note on the chart or below the chart to explain that those data values are unavailable.



Here are two ways to signal missing data: a dashed line or annotation for the gaps.

What we should never do is ignore the missing values altogether and make it appear as though we have a continuous, uninterrupted series.



This chart ignores the missing data points and is misleading. It gives the false impression of a continuous, uninterrupted series.

AVOID DUAL-AXIS LINE CHARTS

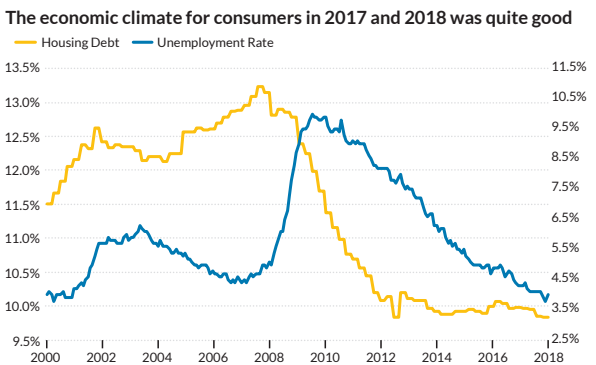
You might be inclined to add another vertical axis to your line chart when comparing changes in two or more series of different units. Resist that urge. Consider this dual-axis line chart that shows the share of income devoted to paying for housing on the left axis and the quarterly unemployment rate on the right axis in the United States from 2000 to 2018. It's not immediately obvious that the unemployment rate is the blue line and associated with the right axis and housing debt is the yellow line plotted along the left axis. The purpose of this graph is to show that the economic climate for consumers in 2017 and 2018 is quite good—low unemployment rates and low housing debt.

But there are three problems with plotting the data like this.

First, they are often hard to read. Did you intuitively know which lines corresponded to which axis? I didn't. Even if the labels and axes were colored to match the lines (which many dual-axis charts don't include), it's hard to discern patterns in the data. They're extra work for the reader, especially when the labeling is not obvious.

Second, the gridlines may not match up. Notice how the horizontal gridlines in this graph are associated with the left axis, which leaves the numbers on the right axis floating in space. At the crossing point in 2009, it's hard to see that the value of the unemployment rate (the blue line) is just shy of 9 percent.

Third, and most importantly, the point where the lines cross becomes a focal point, even though it may have no real meaning. In this graph, the eye is drawn to the middle of

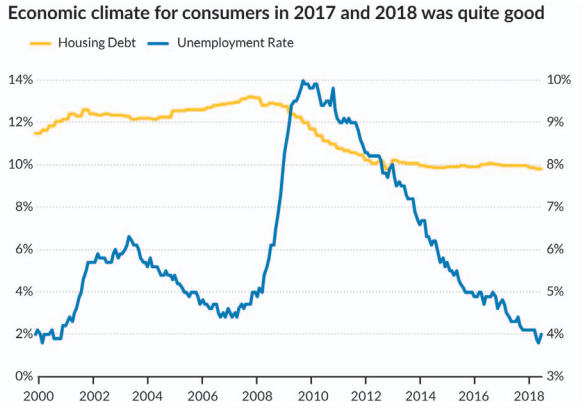


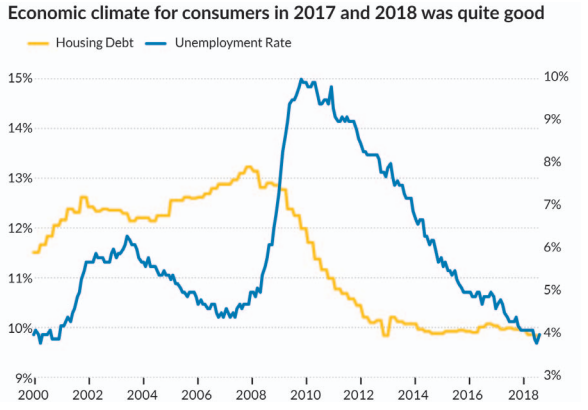
Source: Household debt service ratio, Federal Reserve Board of Governors; Unemployment rate, Bureau of Labor Statistics. Unemployment rate averaged to quarters.

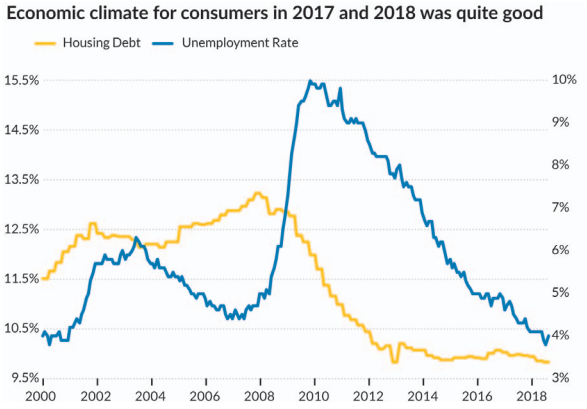
The dual-axis chart introduces a series of perceptual issues, maybe the most important of which is that the eye is drawn to where the lines intersect, even though that may be meaningless.

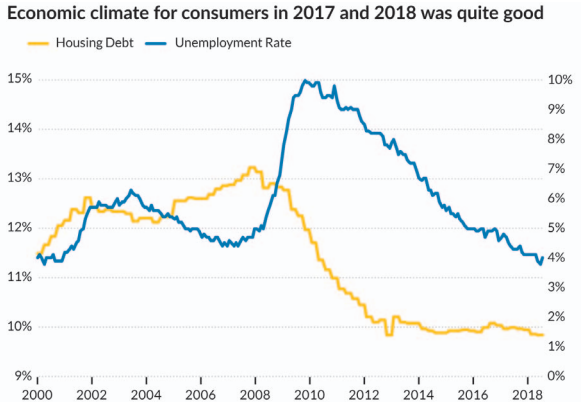
the chart where the two lines intersect, because that's where the most interesting thing is happening. But there's nothing special about 2009, it's just a coincidence that they crossed at that time. The intended takeaway of the chart is how much the economic climate has improved since the 2007–2009 recession, but that's not what draws the eye.

The vertical axis in a line chart does not need to start at zero, so this chart—with the left axis starting at 9.5 percent and the right axis starting at 2.5 percent—is a perfectly reasonable way to plot the two series. By that logic, we could arbitrarily change the dimensions of each axis to make the lines cross wherever we like. And this is the problem with dual-axis line charts: the chart creator can deliberately mislead readers about the relationship between the series.

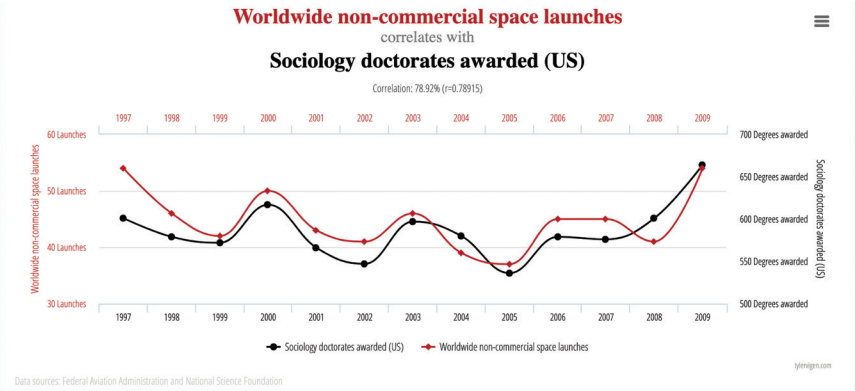








Because there is no hard and fast rule about how to set the dimensions of the vertical axis, we can arbitrarily change the dimensions to make the lines cross wherever we like.



Tyler Vigen's (http://tylervigen.com/spurious-correlations) collection of dual line charts shows how we can imply correlation between seemingly independent data series simply by adjusting the vertical axes.

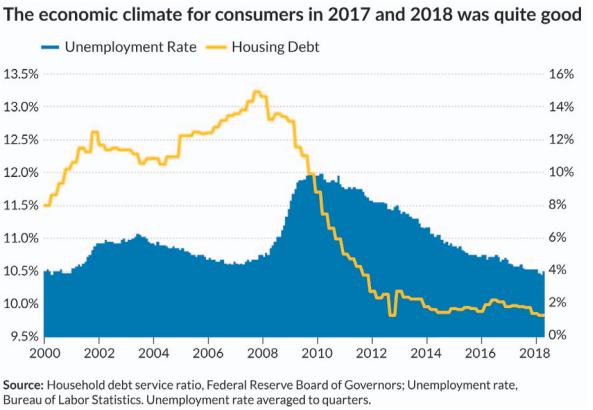
Each of these four graphs are reasonable ways to set the vertical axes, and by manipulating those ranges, I can make the series look like they (a) are closely matched for a few years around 2010 and 2012; (b) cross in the middle and the end; (c) intersect around 2003 and then again a few years later; and (d) are closely related in the first half of the period but then diverge.

By arbitrarily choosing the axes range, we can make different data series look as correlated as we like. On his website, Spurious Correlations, Tyler Vigen shows all kinds of dual-axis charts in which arbitrary vertical axis scales creates erroneous—and humorous—correlations.

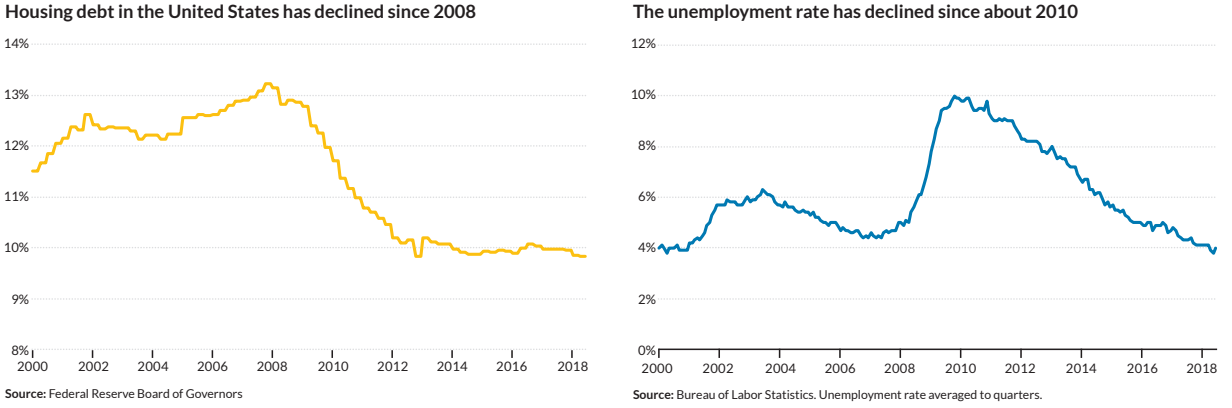
Similar difficulties, though to a slightly different extent, exist in dual-axis charts that combine different graph types. On the next page you can see the same graph of the unemployment rate and housing debt, now with the unemployment rate plotted as an area chart. The right axis starts at zero so the gridlines on both sides match, but it's still not immediately obvious which variable goes with which axis. And though it is more obvious that there are two separate trends being visualized, the same perceptual pitfalls still exist, leading readers to see correlations that might not really be there.

There are a few solutions to the dual-axis chart challenge.

First, try setting the charts side by side. Remember, not everything needs to be packed in a single graph. We can break things up and use a small multiples approach. Although ideally side-by-side graphs should have the same vertical axis to facilitate easier comparisons, we've already determined that approach is impossible here, so splitting them up and using different axis ranges can work.



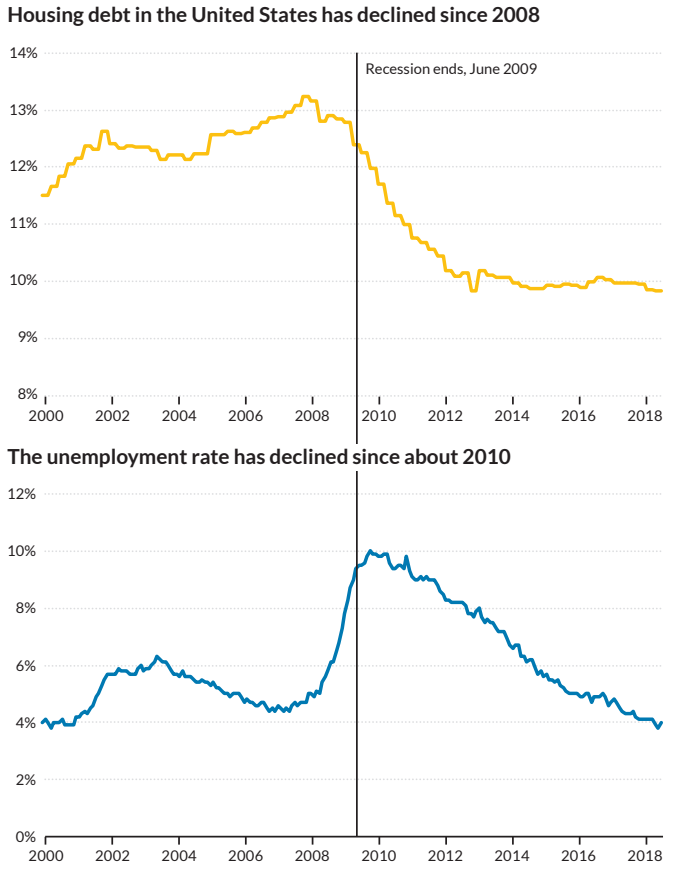
The issue with dual-axis charts is not resolved by combining area charts and line charts.



One alternative to the dual axis line chart is to use two, side-by-side charts.

If it's important to annotate a specific point on the horizontal axis, you could also vertically arrange the two and draw a line across both. This will change the rotation of the final graphic, but is an easier way to label a specific value or year.

Second, we might calculate an index or the percent change from some value or year (see page 148). This way the reader can see the change over time for both series and compare them along the same metric. In the data we've been looking at here, we calculate the difference between each year and 2000, the first year of the period (thus the percentage-point change). The obvious trade-off here is that we lose the *level* presentation of the data and instead present the *change*.

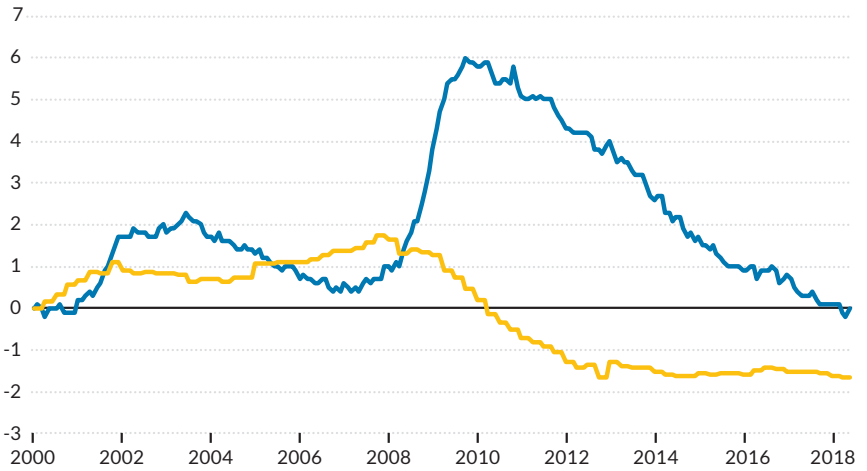


One alternative to the dual axis line chart is to use two charts aligned vertically, which makes it a little easier to mark a specific data point on both charts.

Third, try a different chart type. If showing the changes in the *associations* between the two series is important, try a connected scatterplot. The connected scatterplot—which has its own section at the end of this chapter—is like a scatterplot with a horizontal and vertical axis, but each point represents a different unit of time, such as a quarter or a year. As you can see on the next page, it's easier to see how the relationship has changed over time between these two metrics. You can also see how I have added more labels and annotation (along with different colors) to help the reader navigate the visual.

The economic climate for consumers in 2017 and 2018 was quite good

(Percent pointchange since 2000)

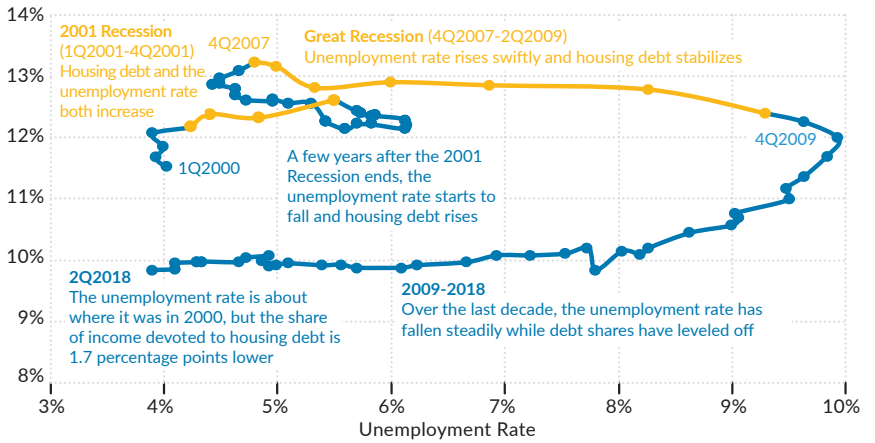


Source: Household debt service ratio, Federal Reserve Board of Governors; Unemployment rate, Bureau of Labor Statistics. Unemployment rate averaged to quarters.

Another alternative to the dual-axis chart is to normalize the data or calculate the percent change from some value.

The U.S. economy appears supportive of the consumer with low-unemployment rate and housing debt

(Household debt service ratio)



Source: Household debt service ratio, Federal Reserve Board of Governors; Unemployment rate, Bureau of Labor Statistics. Unemployment rate averaged to quarters.

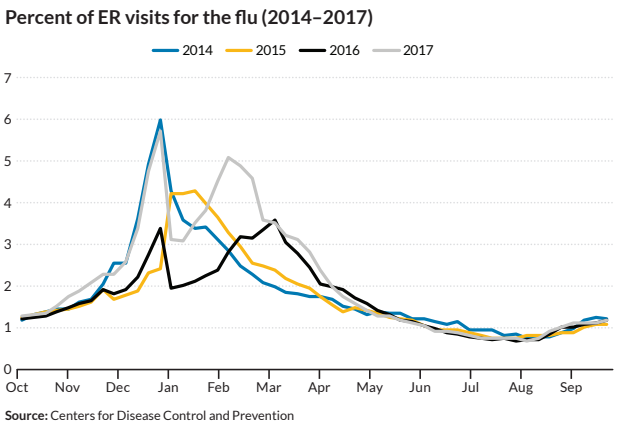
Yet another alternative to the dual-axis chart is to use the connected scatterplot in which one data series corresponds to the horizontal axis and another to the vertical axis.

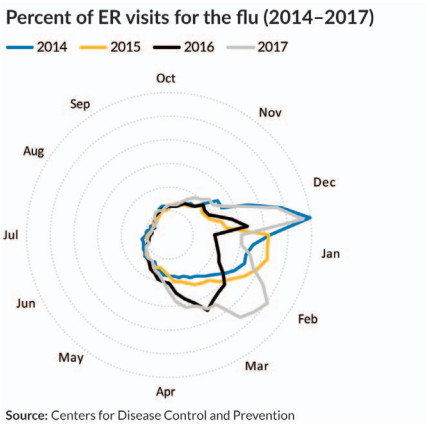
A possible exception to the "no dual axis chart" rule is if you are showing a translation of a single measure, for example Fahrenheit and Celsius temperatures. In these cases, we are not trying to track two different variables but showing how one maps directly onto another. In those cases, the usual pitfalls don't apply.

CIRCULAR LINE CHART

The radial bar chart and circular bar chart in Chapter 4 showed us how to take a bar chart and wrap it in a circle. The same can be done with lines showing changes over time. As before, using a circle may be less perceptually accurate for the reader, but it can be used, for example to improve a visual metaphor.

These two graphs show the percent of hospital emergency room visits for the flu in the United States for each week of the year from 2014 to 2017. Starting at the beginning of the flu season in October, the line chart on the left gives us the standard view: an increase in the flu during the winter months, which fades as we enter summer. The radial chart on the right shows the same data but with a different perspective—the "lean" toward three o'clock on the chart when more infections occur during the fall and winter, and fewer infections during the summer months on the left side of the circle. The radial chart is more compact than the standard line chart, but it is also harder to make precise comparisons because the lines do not sit on a single horizontal axis.





Two ways of showing the same time series data—as a standard line chart or by wrapping the lines around a circle.

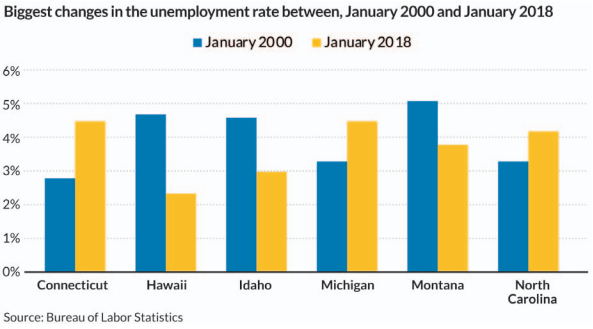
SLOPE CHART

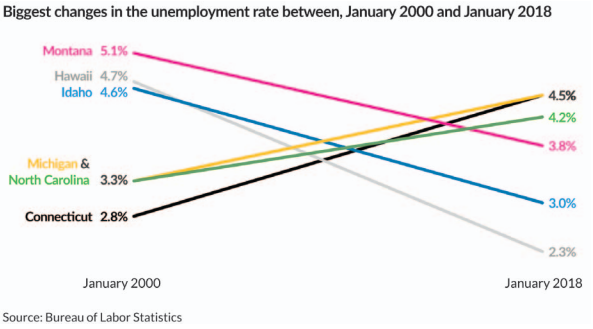
In some cases, it may not be necessary to show *all* of the data in your time series. In these cases, a slope chart—which is really just a simplified line chart—is a useful alternative.

The paired bar chart is a standard way to visualize two data points for multiple observations (also see page 84). As an example, consider these charts of changes in the unemployment rate for six states in the United States between 2000 and 2018. With this kind of visualization, we ask our reader to process the level *and* change in the unemployment rate *between and across* the six states. There is a lot of ink in the graph, and it asks the reader to do lots of mental math. We could, of course, just plot the change between the two time periods, but we often want to show both the level and the change.

The slope chart addresses this challenge by plotting each data point on a separate vertical axis and connecting the two with a line. In this example, the left vertical axis represents the first month of data (January 2000) and the right vertical axis represents the last month of data (January 2018). We can easily see the relative values of each data point. Here, for example, we can see—perhaps even more easily than in the paired bar chart—that Montana had the highest unemployment rate in the first month and Connecticut had the lowest. The line that connects the two data points visualizes the change over time. We can more easily see that the unemployment rate in Montana, Hawaii, and Idaho fell between 2000 and 2018 while it rose in the other three states.

There are many ways to style the slope chart. We can use two colors to denote increases and decreases. We can include or exclude labels for levels and changes. We can even adjust the

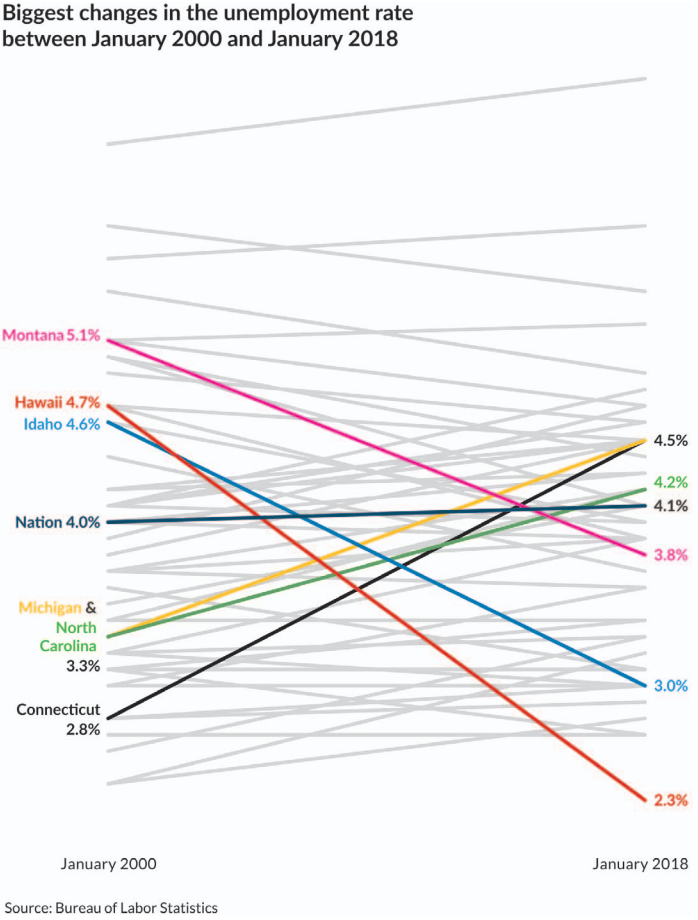




The paired bar chart asks the reader to make several comparisons on their own simultaneously, while the slope chart visualizes these comparisons for the reader.

thickness of the line to correspond to a third variable. We could also rely on the *Start with Gray* strategy and add more data to the basic slope chart. Here, I've included every state in the nation but highlighted and thickened the six states of interest and the national average.

In graphs like these, consider whether a taller chart will make it easier for the reader to see all of the detailed colors, labels, and annotations. As with the dot plot (see page 97), be careful about using the slope chart when a summary of the time series may mask changes in the intervening years. Of course, this is the same consideration when using the paired bar chart.



There are many ways to style a slope chart. This *Start with Gray* strategy can be especially useful here to show many observations while highlighting only a few.

SPARKLINES

There is a specific style of small multiples for line charts called sparklines. Invented by author and statistician Edward Tufte, sparklines are "small intense, simple, word-sized graphics with typographic resolution." They are typically used in data-rich tables and may appear at the end of a row or column. The purpose of sparklines is not necessarily to help the reader find *specific* values but instead to track *general* patterns and trends.

Let's use sparklines with the health spending data. The numbers in the two table columns show spending in 2000 and 2015, while the sparklines show the values for the entire sixteen-year period. This way, readers can see some specific values as well as the patterns over the entire period. Here, for example, you can quickly see that health spending rose for all of these countries except for Turkey, which I've also highlighted so it stands out.

Health care spending in selected countries

Country	2000	2015	2000-2015
Australia	7.6	9.4	
Canada	8.3	10.4	
Finland	6.8	9.4	
Japan	7.2	10.9	
Switzerland	9.3	12.1	
Turkey	4.6	4.1	
United Kingdom	6.0	9.9	
United States	12.5	16.8	

Source: The World Bank

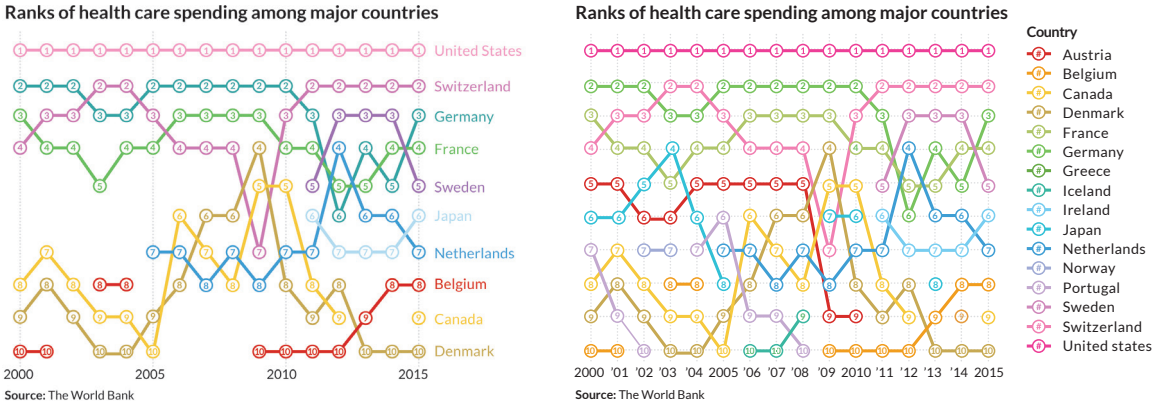
Sparklines are a type of small multiples line chart typically used within data-rich tables.

BUMP CHART

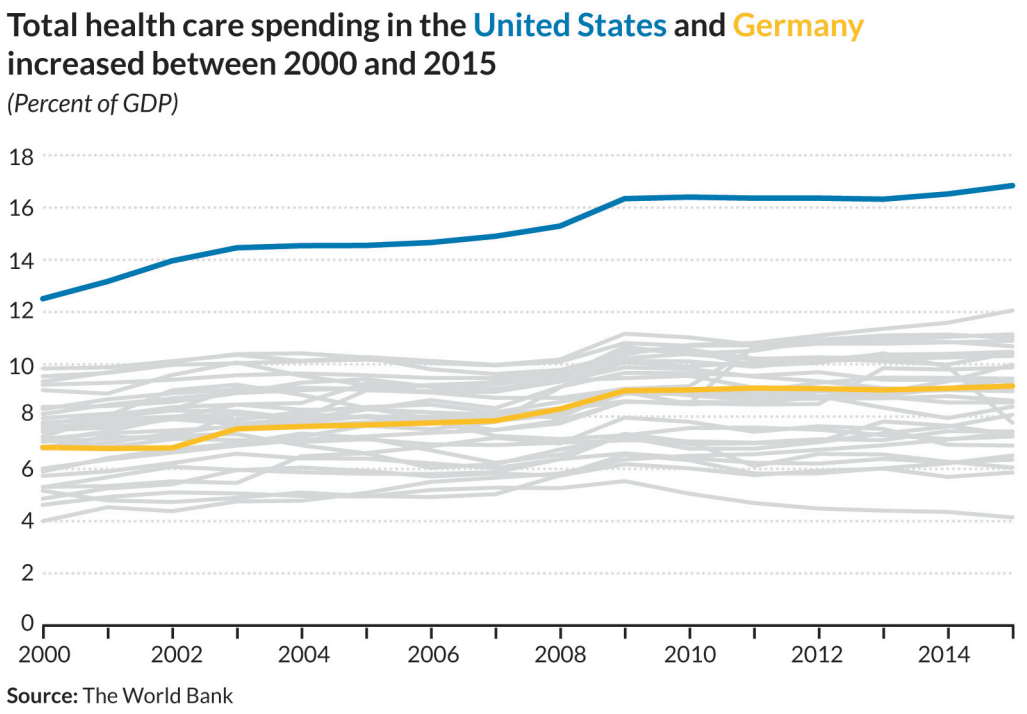
A variation on the line chart is the bump chart, which is used for plotting changes in *ranks* over time, for example, political polling or positions in a golf tournament from hole to hole. When we want to show relative ranks rather than absolute values, the bump chart is a good choice.

A bump chart is, of course, a compromise. It does not show the raw values, which are often preferred, but it can be especially useful if your data have outliers. By plotting the ranks, we abstract from the large differences in magnitude.

These two bump charts show changes in health care spending across the ten countries that have the highest spending on health care as a share of their GDP in 2015. Those countries appear in the far right position of the horizontal axis, above the 2015 label. The difference between the two charts is that that the one on the left shows the patterns and ranks for only these ten countries for every year. You can see some gaps in certain years where other countries would appear in the rankings, but this chart only tracks those countries that end up in the top ten in 2015. The chart on the right, by comparison, includes every country in every year, which requires more labeling so the reader can understand why new countries (with different colors) suddenly appear in the chart. We could emphasize certain countries by changing line colors or the colors of the inside of the data marker circles, or even the thickness of the lines.



Bump charts plot changes in ranks over time.



Use color, data markers, or line thickness to highlight specific data series.

Compare these bump charts to the line chart above that shows all of the OECD countries in gray and highlights the United States and Germany with color. In this case, the United States stands far above the rest of the countries that here appear clumped together in a swirl of lines. As is often the case, there is a tradeoff between the bump chart and the line chart: In the standard line chart, you can see the relative differences between the series, but they're stacked together and hard to disentangle. In the bump chart, by comparison, we can't see the relative differences, but we can see relative ranks.

A modification on the bump chart is a *ribbon effect*. Here, in addition to rank, the widths of the ribbons are scaled according to the actual data values. Like the streamgraph, which we'll see later, this approach has a more organic, flowing look. This chart from the *Berliner Morgenpost* shows the rank, amount, and change in different sentiments around political problems in Germany.

The ribbon effect is a modification on the standard bump chart. The title in this chart from the Berliner Morgenpost translates to "These are the 15 most important political problems in Germany". Arbeitslosigkeit translates to Unemployment; Einwanderung to Immigration; Bildung to education; and so on. It shows the changes in each ranks (and amounts) of these different sentiments.

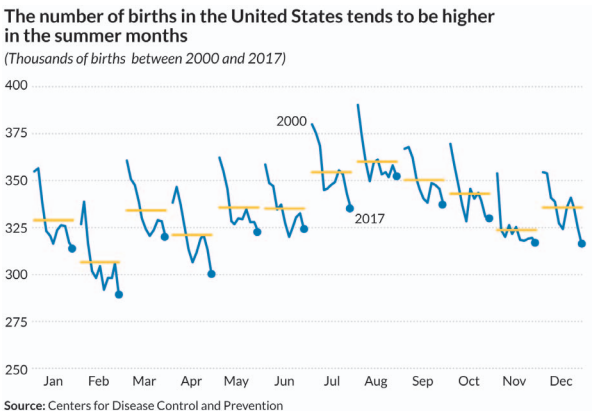
CYCLE CHART

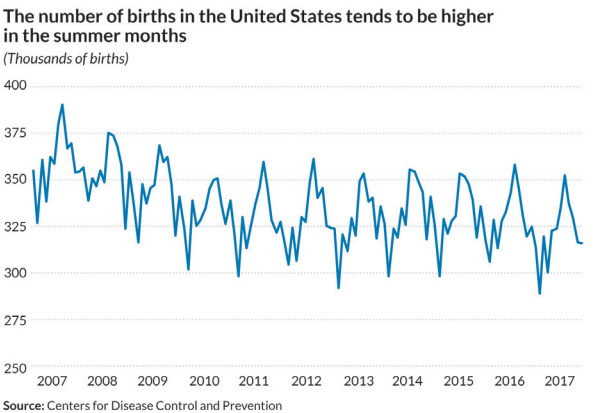
Cycle graphs typically compare small units of time, such as weeks or months, across a multiyear time frame. They are most commonly used to display strong seasonal trends. Here, we see the number of births in the United States in each month from 2007 to 2017. A yellow line marks each month's average value (a general but not necessary characteristic of the cycle chart). We can see the downward trend for births over the decade for every month and the

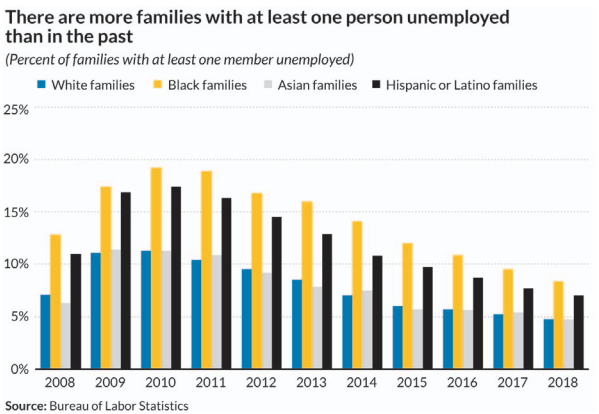
higher birth rate during the summer months, July, August, and September. I've added a dot at the end of each line to mark the most recent year.

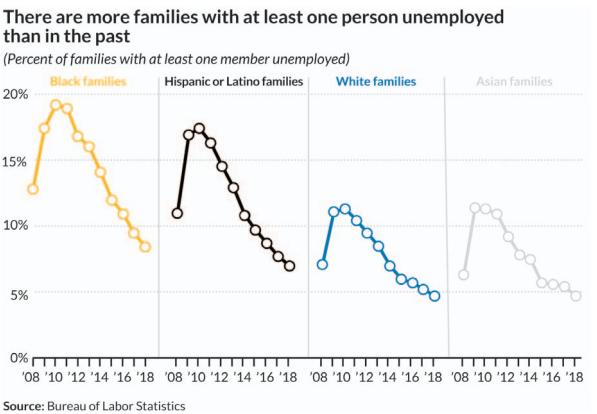
By comparison, this same data displayed as a standard line graph is less clear. We can see a spike in each year, but without more labeling, it's not clear in which month that spike occurs. Even though the cycle chart has more information on it—the average values shown in yellow and the point at the end of each line—it still feels less busy than the standard line chart.

A cycle graph can also split up a dense bar or line chart to give each series more space—something like a small multiples chart. Take this column chart of the unemployment rate for four groups in the United States. This kind of graph—with multiple years for different









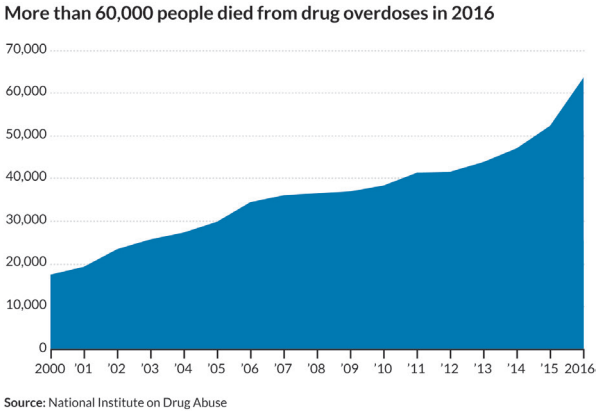
Cycle graphs compare small units of time, such as weeks or months, across a multiyear time frame.

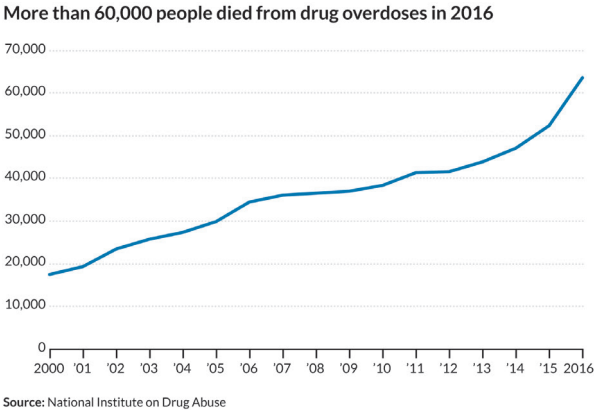
groups pushed together—can be difficult for the reader to make comparisons within or across years. The cycle graph on the right separates each racial group into its own panel, sorted by the value in the most recent year. You could argue that the graph on the right is a small multiples line graph, but the organization and design make it more like a cycle graph.

AREA CHART

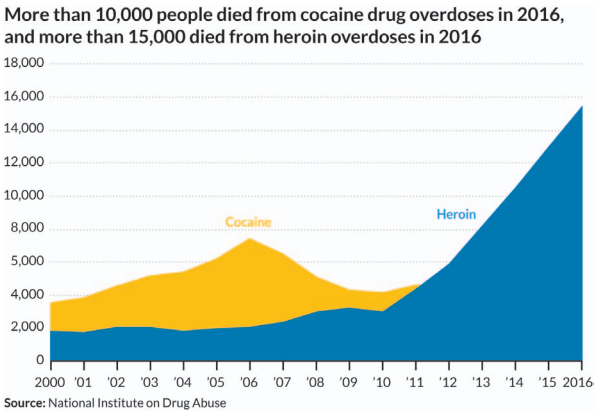
Area charts are line graphs with the area below the line filled in, giving the series more visual weight. The area chart on the left and the line chart on the right both show the number of people who died from prescription opioid overdoses in the United States between 2000 and 2016. You might think of the area chart as a bar chart where the bars have infinitely thin widths and thus, as we saw in the previous chapter, the vertical axis should always start at zero.

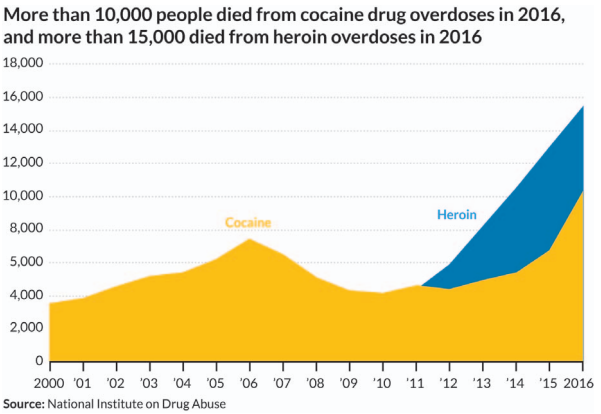
Placing two or more series in an area chart can be difficult because one series can hide (or "occlude") the other, an effect we will see more in the coming chapters. On the next page, the area chart on the left, for example, shows overdose deaths from cocaine and heroin, but the data series for heroin is hidden behind the series for cocaine. Even if the order of the data series are changed so that heroin overdose deaths are in front, now the heroin-deaths series blocks the cocaine-deaths series. Compounding that difficulty is that some readers might





Area charts are line graphs with the area below the line filled in, giving the series more visual weight.

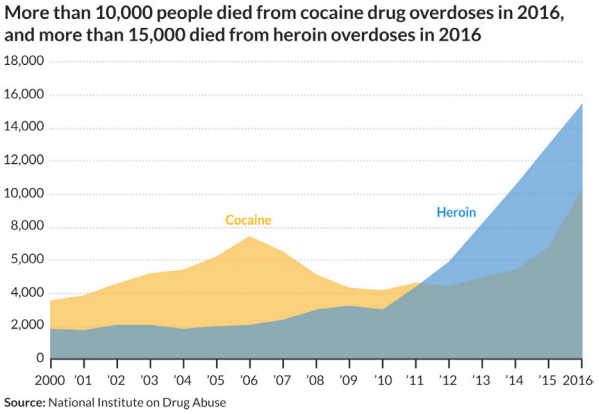


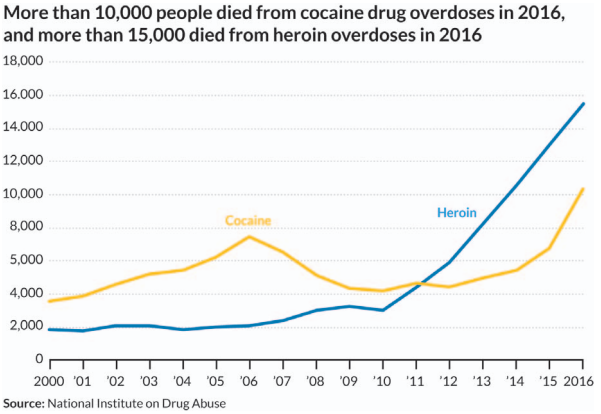


Placing two or more series in an area chart can spell trouble, because one series can hide (or "occlude") the other.

mistake the two as summing to a total rather than as separate data series. Here, it's important to use the title and annotation to make it clear there are two distinct series.

One strategy to address this overlap is to add a transparency to the color of one (or both) series. But be careful: by adding a transparency to only one series, we deemphasize its importance. Another alternative is to use a line chart, as in the graph on the right.





One strategy to address the overlap between series on an area chart is to add a transparency to the color of one (or both) series. Another alternative is to use a line chart, as in the graph on the right.

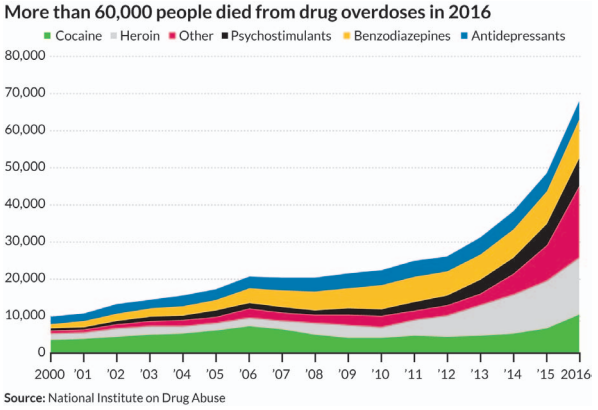
STACKED AREA CHART

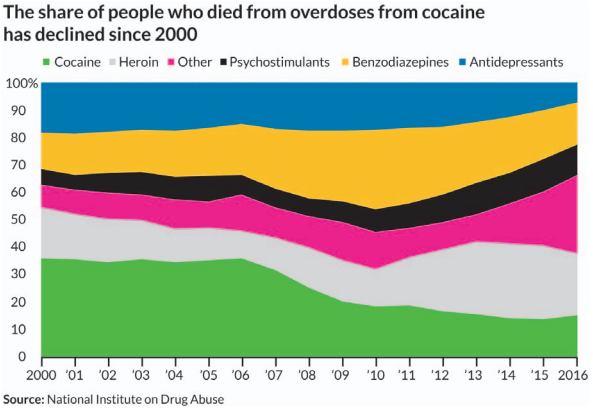
Stacked area charts build on the typical area chart by showing multiple data series simultaneously. Instead of sitting independently of one another as in the previous chart, the data in a stacked area chart sum to a total or a percentage.

The stacked area chart on the left shows the total number of drug overdose deaths between 1999 and 2016. The version on the right shows the same data, but presented as percentages that sum to 100 percent.

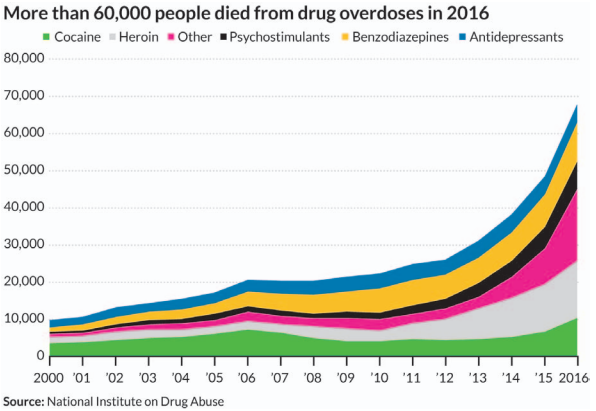
The reader will take away different conclusions from these two representations. In the graph on the left, the eye is drawn to the large increase in overall deaths over the period. In the version on the right, it is drawn to the changes in the distribution of deaths—a decline in overdoses from cocaine, but an increase in heroin, Benzodiazepines (drugs that are often used to treat anxiety, insomnia, and seizure disorders), and other drugs.

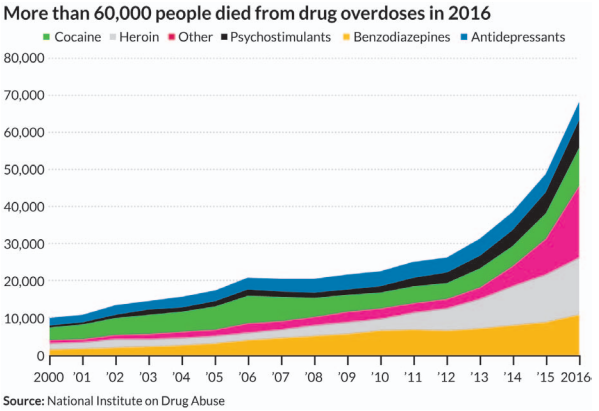
There are three disadvantages with the stacked area chart on the left. First, as earlier, we again see the line-width illusion—we tend to view steep changes as bigger than they actually are. Second, only the bottom series sits on a horizontal axis, so it is hard to accurately compare the changes over time for the other series. (Remember, this is the second row in the perceptual ranking table showed earlier.) Third, the ordering of the data series can affect





Stacked area charts build on the typical area chart by showing multiple data series simultaneously and sum to a total, often 100 percent.

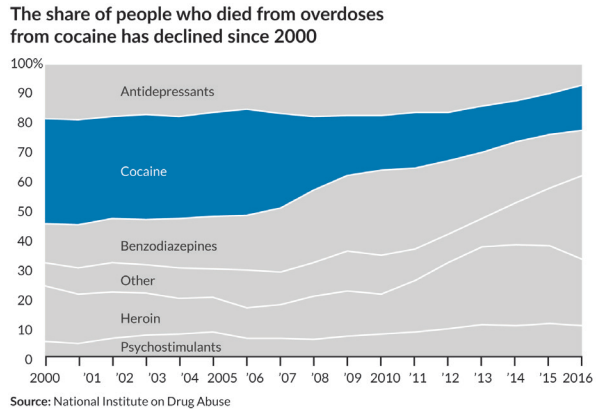


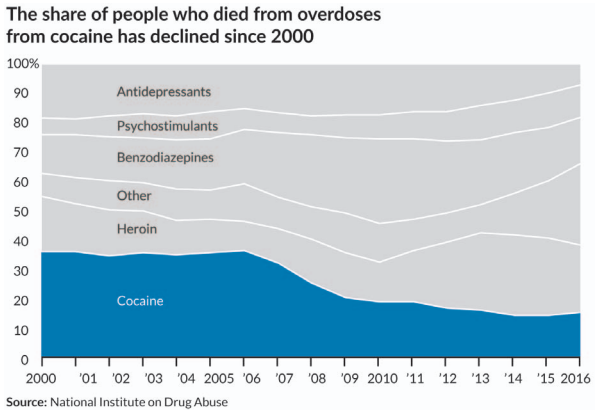


Recall that the perceptual ranking list suggests we can better compare values when they sit on the same axis. That's why it's easiest to compare values for the bottom series.

our perception of the shares of the total and move the reader's attention around from one series to another.

To demonstrate, consider the two stacked area charts above. The version on the left is the same as before while the version on the right changes the order. Notice how in the new version, it is easier to compare changes in overdose deaths caused by Benzodiazepines (the yellow series), because the series sits along the same horizontal axis.



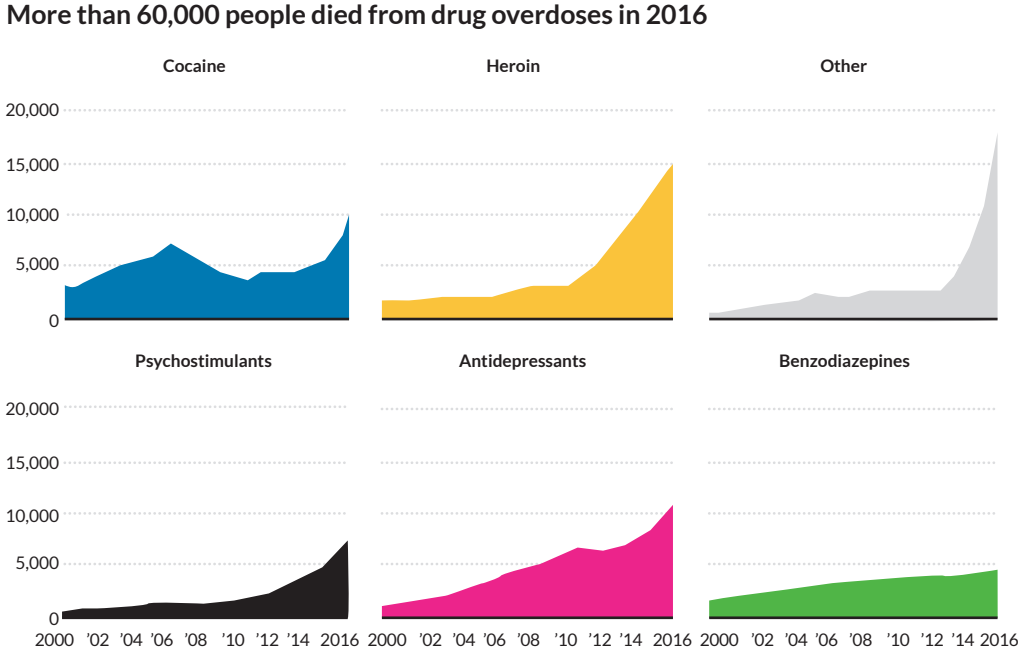


There isn't necessarily a right way to stack the series in a stacked area chart, but how you decide to arrange them will influence how your reader perceives the data.

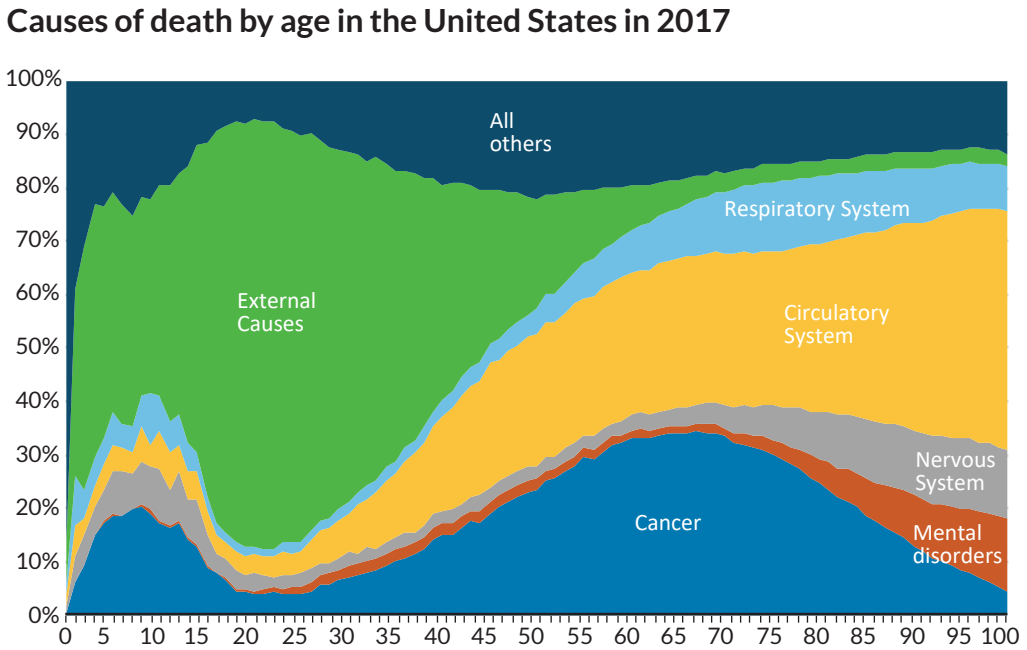
This isn't to say there is a "right" way to stack the data in an area chart, or that the most important data series should sit along the horizontal axis. If, say, you were telling a story about the declining

share of overdose cocaine deaths over this period, you could keep it in the same position as above, but use the "start with gray" strategy and use color in just the cocaine series. Even with the line-width illusion, you can still see the share of deaths has declined. If it's important for your reader to see the exact change in the share, then putting that series along the horizontal axis is a better strategy. By placing the series in the middle of the chart, you can't compare the values to the horizontal baseline and thus less accurately perceive the values. (Also notice how I directly labeled the segments instead of using a legend so the reader can quickly and easily identify the different series.)

A small multiples approach (here, six different graphs) more clearly shows the exact patterns in each category, but does not as clearly show the relative shares of each. You make a couple of tradeoffs here. On the one hand, the stacked chart is more compact than the small multiples—you pack all of the information into a single visualization in which you can see the changes in shares. On the other hand, the small multiples gives you a more perceptually accurate view—because each series sits on its own horizontal baseline—but it is harder to compare across the different categories.



The small multiples approach more clearly shows the exact patterns in each series, but does not as clearly show the relative values of each.



Source: Centers for Disease Control and Prevention

Stacked area charts can also show changes in the distribution of a data series.

Finally, the stacked area chart can also show changes in the distribution of some data series. This stacked area chart, for example, shows all of the different ways people from age zero to one hundred died in the United States in 2017. Instead of years or months along the horizontal axis, this graph shows the number of deaths for each single year of age, a different measure of time. Categorized into fifteen groups, most people who die around age 25 do so of "external causes," (the green series) such as falling or drowning, while most people who die around age sixty die of some form of cancer (blue). As before, we could modify the colors or the arrangement of the data to focus our reader's attention on specific patterns or trends.

STREAMGRAPH

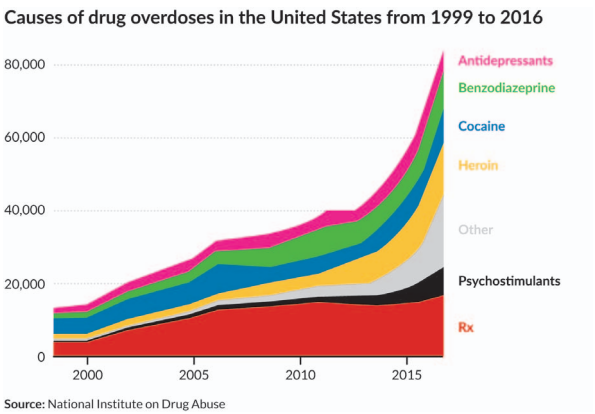
Like the stacked area chart, a streamgraph also stacks the data series, but the central horizontal axis does not necessarily signal a zero value. Instead, data can be positive on both sides of the axis. Together, the streamgraph illustrates fluctuations in data over time in a

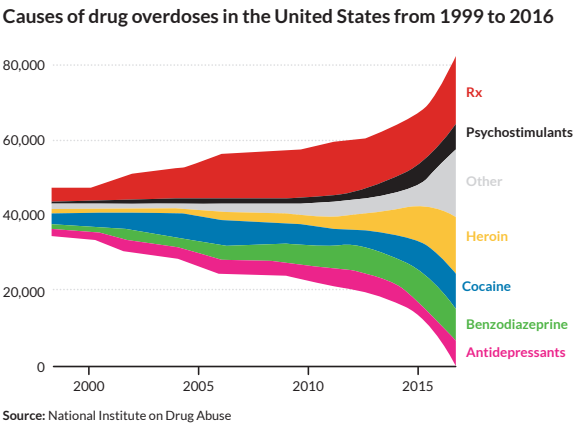
flowing, organic shape. They are therefore best used for time series data when the series themselves have high volatility.

Streamgraphs are well suited for showing patterns that have peaks and troughs. Both the stacked area chart on the left and the streamgraph on the right show the *total number of deaths* rather than shares of deaths, as in the earlier stacked area chart. The streamgraph gives us a slightly different view of the data and may point us more toward overall increases rather than changes in specific series. The idea behind the streamgraph is to minimize the distortion in each layer's baseline that accumulates more rapidly with a stacked area chart.

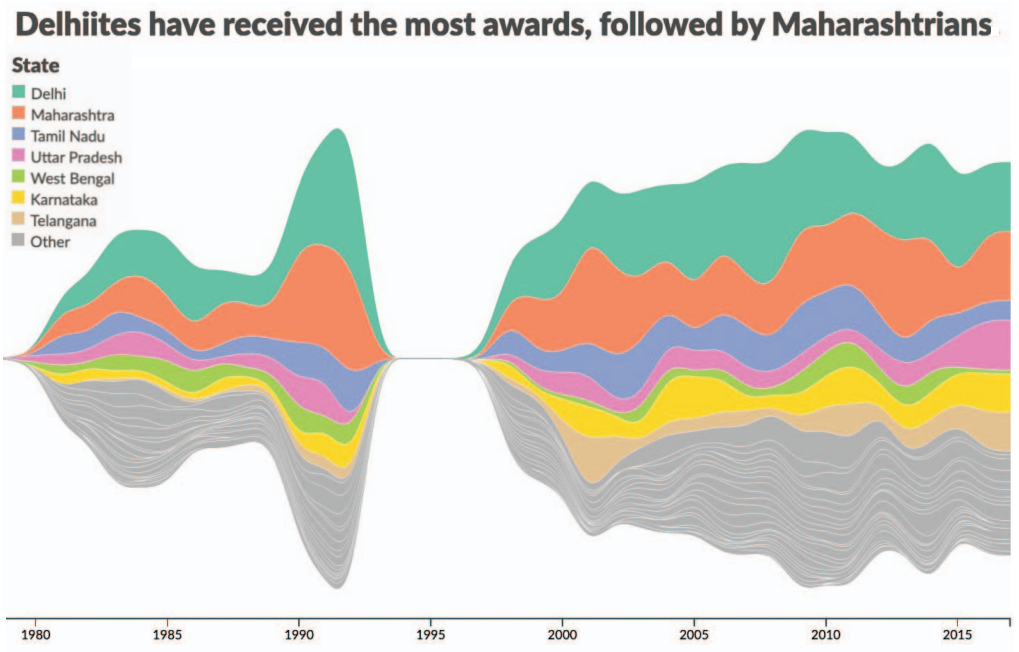
Researchers are aware of how unusual the streamgraph looks, and how it may be more difficult for readers to understand. In a review of a streamgraph published by the *New York Times* in 2008, researchers noted that they "suspect that some of the aesthetically pleasing—or at least engaging—qualities may be in conflict with the need for legibility. The fact that the *New York Times* graph does not look like a standard statistical graphic may well be part of its appeal." Thus, while this kind of graph—or any different-looking graph—may at first confuse or confound readers, they may ultimately find the shapes, colors, and other attributes more interesting and engaging. It all depends on your audience.

On the next page, you can see a more recent example of a streamgraph. This visualization was published by the *Hindustan Times* in 2016 and shows the number and type of the highest civilian awards the Indian government confers. Additional streamgraphs in the original news story showed breakdowns by state, nationality, gender, and discipline.





Streamgraphs are a variation on the area chart and are well suited for showing patterns that have peaks and troughs.



Source: *Hindustan Times*. This image was cropped and condensed from the original for purposes of this book.

This streamgraph from the *Hindustan Times* shows patterns in the number and type of the highest civilian awards conferred by the Indian government.

HORIZON CHART

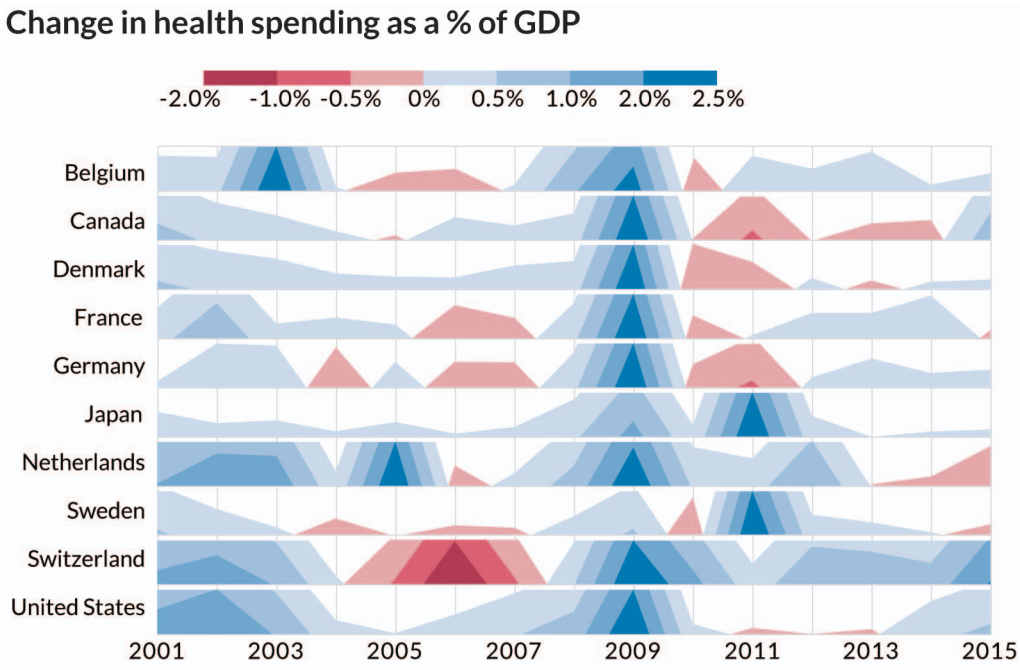
A horizon chart is an area chart that is sliced into equal horizontal intervals and collapsed down into single bands, which makes the graph more compact and similar to a heatmap (page 112). The horizon chart is split into bands with positive numbers collapsed down and negative values flipped above the horizontal axis. Multiple horizon charts—which is how they are typically arranged—can condense a dense dataset into a single visualization. Horizon charts are especially useful when you are visualizing time series data that are so close in value so that the data marks in, for example, a line chart, would lie atop each other. Aligning the charts in this way allows us to include our data in a more compact space than in a series of area charts.

Color is the most important attribute in a horizon chart. Darker colors represent larger values and lighter colors smaller values. Like sparklines and to some extent heatmaps, the purpose of the horizon chart is not necessarily to enable readers to pick out specific values, but instead to easily spot general trends and identify extreme values.

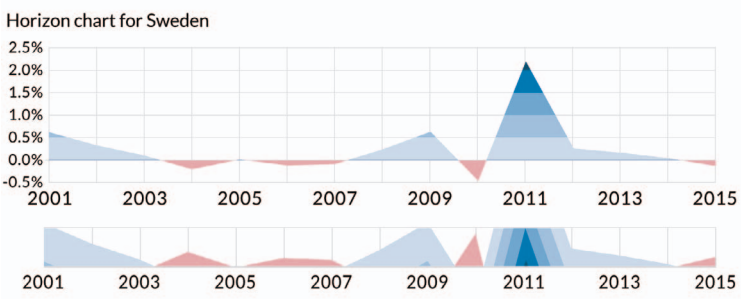
This horizon chart uses the same data we've been using on changes in the percent of GDP spent on public health care. An area chart is built for each country, split and collapsed, and then arranged all together in rows. Notice how much data are packed into the single visualization (ten countries and fifteen years), and—recalling the importance of preattentive processing—see how your eye is drawn to the brighter and darker colors.

I'll use the series for Sweden to show how the horizon chart is built. The change in public health spending for Sweden (the third country from the bottom in this horizon chart) is shown on the next page as an area chart and sliced into equal increments (every 0.5 percentage points). Larger values have darker shades, and negative and positive values have different colors. The negatives are flipped above the horizontal axis and all are collapsed down to the first interval or band.

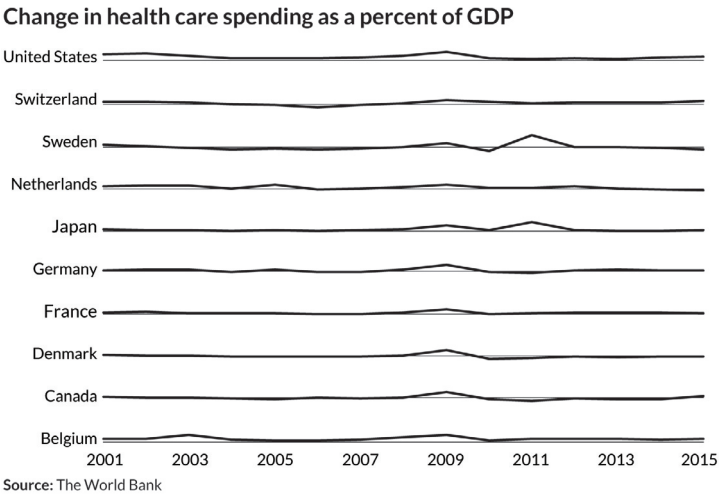
Color is the key here. The same visualization as a series of line charts does not have the same punch. The eye scans the entire visualization for important trends, but no particular region of the visual draws our attention. That could be modified by adding color to the



The horizon chart is an area chart divided into equal intervals and collapsed into a band.



These charts show how the area chart for Sweden is divided and collapsed to create the horizon chart.

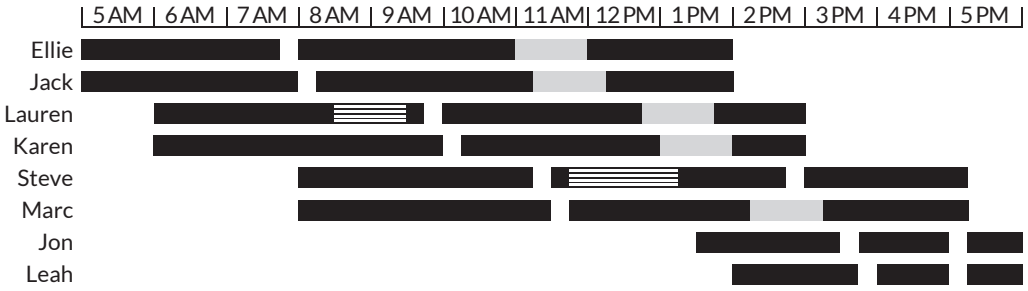


A line chart is sufficient, but the use of color in the horizon chart attracts and directs the reader's attention.

different lines to highlight extreme values, but the horizon chart does a much better job of directing the eye by highlighting values through colors.

GANTT CHART

Another way to show changes over time is to use horizontal lines or bars to show the *duration* of different values or actions. Gantt charts are often used as schedule-tracking devices, for example, to track different phases of a project or budget. Invented by Henry Laurence



Gantt charts often show processes or schedules.

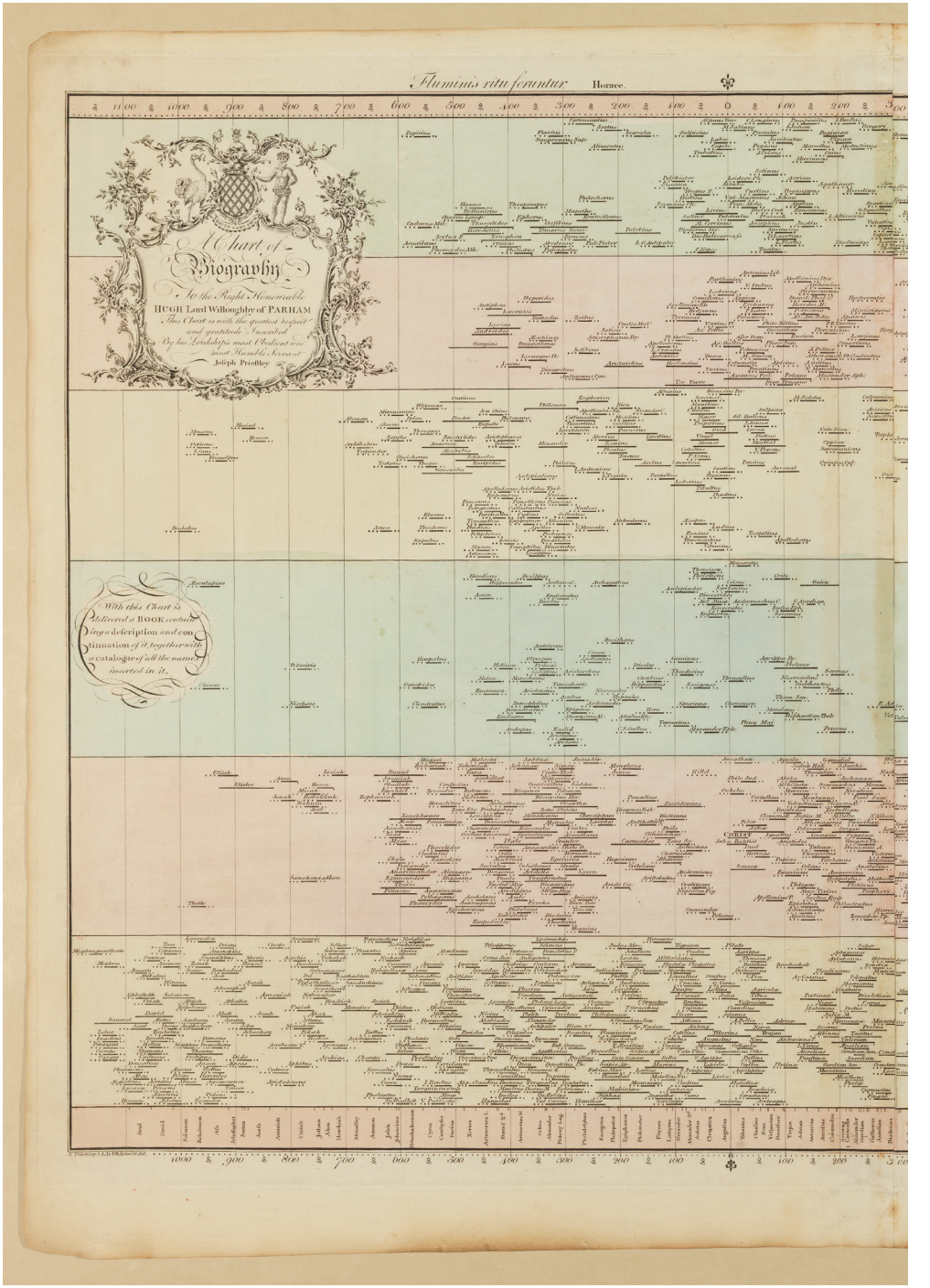
Gantt, an engineer working around the turn of the twentieth century, the charts were first used by production foremen and supervisors to track production schedules.

This Gantt chart shows staffing shifts at a coffee shop over the course of a day, denoting breaks with the white gaps, lunch breaks with the gray breaks, and other time away from the store with stripes.

Gantt charts can be extended by modifying the width of the bars to denote another variable. For example, this Gantt chart modifies the hypothetical chart above to scale the widths according to the pay of each employee.

Joseph Priestley, an eighteenth-century philosopher, chemist, and educator, published *A Chart of Biography* in 1765, showing the lifespans of approximately two thousand statesmen, poets, artists, and other notables who lived between 1200 BC and the mid-1700s. Often called a timeline, Priestley's chart looks more like a Gantt chart because of the use of horizontal bars/lines and the concrete beginnings (births) and ends (deaths).

An extension to the Gantt chart is to adjust the widths of the bars to correspond to another data series.



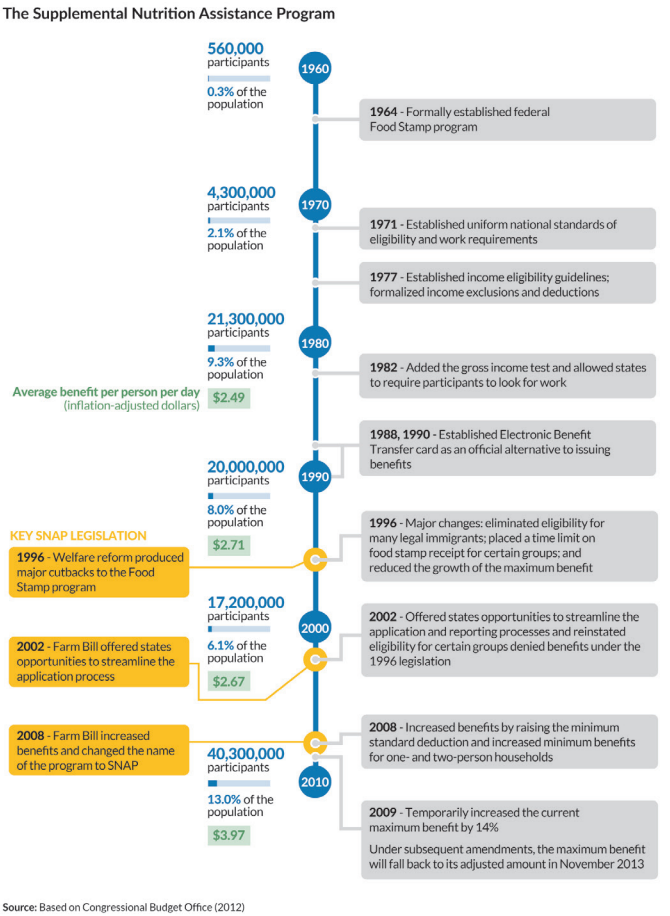
Joseph Priestley's *A Chart of Biography* (1765) shows the lifespans of about two thousand statesmen, poets, artists, and others. It covers an enormous time period,

from 1200 BC to 1800 AD. A horizontal line for each person shows his or her lifespan. Dots indicate wherever such there is uncertainty around those dates.

Source: Library Company of Philadelphia.

FLOW CHARTS AND TIMELINES

Flow charts and timelines are two examples of an array of visuals that can show changes over time or different kinds of processes, sequences, or hierarchies. This class of charts and diagrams can be explicitly tied to data or can be less quantitative and more illustrative, a way



This timeline, based on work I conducted at the Congressional Budget Office, shows major milestones and data for the Supplemental Nutrition Assistance Program (SNAP, formerly known as food stamps).

to demonstrate different structures or processes. In PowerPoint, for example, you can look through the "SmartArt" menu for a wide selection of layouts.

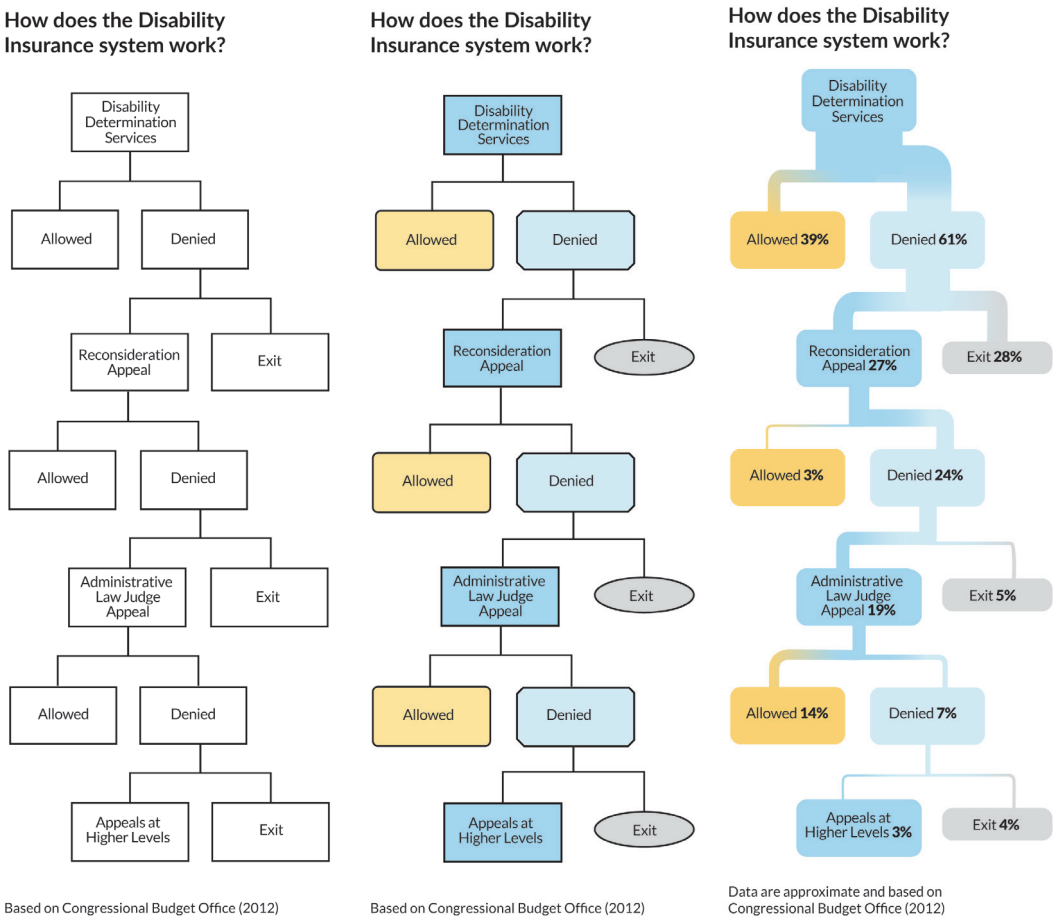
A timeline shows when certain events take place. It can be basic and flag events with a line, icon, or marker, or it can be more involved and include annotation, images, or even graphs. Though horizontal timelines are common, timelines can also be vertical or even a variety of different shapes. This timeline, based on work I conducted at the Congressional Budget Office, shows major milestones and data for Supplemental Nutrition Assistance Program (SNAP, formerly known as food stamps). The text in the gray boxes on the right gives details on specific legislation or program changes, and the information on the left presents changes in spending, number of program participants, and their share of the total population.

Flow charts are slightly different. They are not necessarily tied to time in the sense of days, months, and years, but instead they map a process, often step by step. Flow charts make it easier for readers to understand the paths of a process rather than reading through long passages of text or navigating a convoluted table. The flow chart on the next page shows the process by which people can apply and receive benefits through the U.S. Social Security Disability Insurance program (DI). Applicants start the program at what is called the "Disability Determination Services" stage. If their application is approved, they are "Allowed" onto the program; if not, they can either appeal the decision or exit the process altogether. The program is designed in such a way so applicants may appeal a denied request at each stage.

The shapes in a flow chart may carry different meanings, so we can use them strategically to denote different attributes of the system. For example, in a flow chart with rectangles, other shapes can denote choke points or decision points, and rounded rectangles might signal the beginning or end of a process. Adding different colors can help readers understand and differentiate the parts of the graph from each other.

If, for example, we wanted to highlight the different parts of the DI application system, we could use different colors and shapes, as in the version in the middle. Labels in a flow chart can sit alongside the lines and inside the boxes, but they should be large enough and have enough color contrast to be easily read. We could take this even further and scale aspects of the flow chart according to some data values; in the version on the right, for example, the branches are all scaled according to the respective shares at each stage, similar to a Sankey diagram (page 126).

I place the flow chart in the time chapter because these processes often occur over time, one by one. But that's not always the case. An *organizational chart* or *org chart*, for example,



Shapes color, and other elements can help readers understand the paths of a process in a flow chart or timeline. This chart is based on work from the Congressional Budget Office.

is a type of flow chart that shows the hierarchy or management structure of an organization, and how work flows from the top down. We'll see more examples of an org chart in Chapter 8.

Just as we saw with the line chart with many lines, it's not about the *amount* of content you place in the graph, but what meets the needs of your reader. The flow chart on page 174,

for example, was prepared in 2010 by the Republican staff of the Joint Economic Committee in response to President Obama's Affordable Care Act proposal. The implicit purpose of this flow chart is to show how complex the proposal (and health care in general) is in the United States. In that sense, the chart does its job!

In these types of graphs, the amount of notes, text, icons, and other visual elements should, as always, meet the needs of the reader. Are specific details necessary at each point? Would an image anchor the moment in the reader's mind? Consider what information your reader needs most and provide it as engagingly as you can.

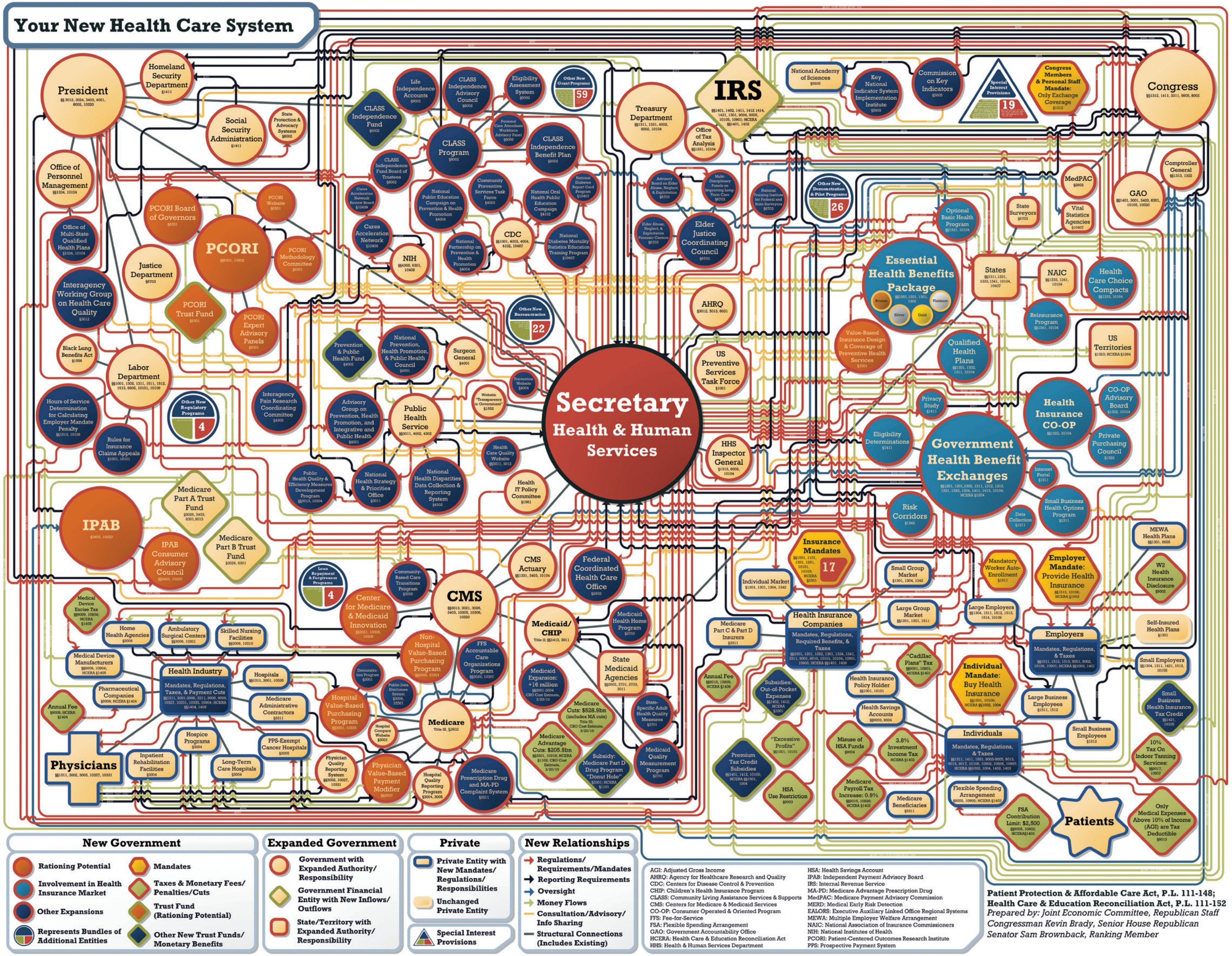
TOTALS VS. PER CAPITA

Totals can tell you a lot about a group, but they can also mislead. Take for example gross domestic product (GDP), a measure that shows up a lot in this book. In 2017, India and the United Kingdom had roughly the same total GDP at around $2.6 trillion. But their populations—and by consequence their *per capita* (or per person) GDP—are quite different.

In that same year, India's population was 1.3 billion people, more than *twenty times* that of the UK, which had a population of only 66 million. Thus, per capita GDP—total GDP divided by population—was $39,720 in the UK and $1,940 in India. If you treated GDP like a box of cash and gave out equal shares to everyone, each person in the UK would get roughly $38,000 more than each person in India.

These adjustments extend to what we call "normalizing" or "standardizing" metrics. We use this all the time when we drive our cars, for example—we drive sixty miles *per hour* and the price of gas is $2.75 *per gallon*. We can see this in all sorts of other areas and metrics, like mortality rates (deaths per 100,000 population) and wages (dollars per hour).

If you are working with totals in your data, consider whether per capita amounts or other adjustments may be a better and more informative measure. Knowing that India and the UK have similar total GDP doesn't tell you as much about their economies or the relative wealth of people living in those countries as does the per capita measure.



The implicit purpose of this flow chart from the United States Congress Joint Economic Committee (2010) is to demonstrate the complexity of the Affordable Care Act proposal.

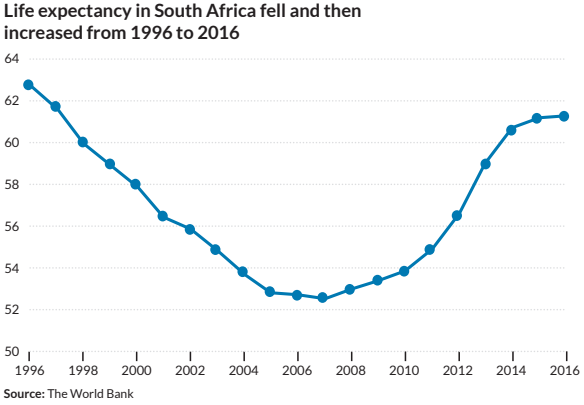
CONNECTED SCATTERPLOT

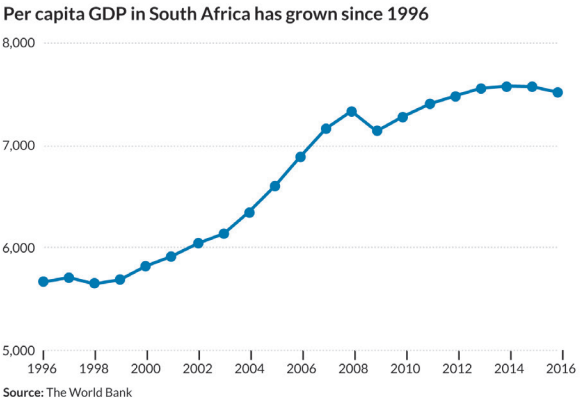
Imagine showing two line charts side-by-side. You may be asking your reader to examine the relationship between the two. Do they move together? Do they diverge or converge? How are they related?

One way to bring two time series together *without* using a dual-axis chart is the connected scatterplot. A connected scatterplot shows two time series simultaneously—one each along horizontal and vertical axes—and are connected by a line to show relationships of the points over time.

As an example, the line chart on the left shows life expectancy in South Africa from 1996 to 2016. Over that twenty-year period, life expectancy followed a U-shape pattern, first falling from about sixty-three years old to fifty-three years old, and then increasing over the next decade, reaching about sixty-one years old in 2016. On the right, per capita GDP is plotted over that same period—that series was flat in the first few years, then increased until about 2008, when it dipped slightly before rising at a slightly slower rate.

With these two charts, we can make some basic visual comparisons—even as life expectancy fell, economic growth continued. When life expectancy started growing again, economic growth flattened out.



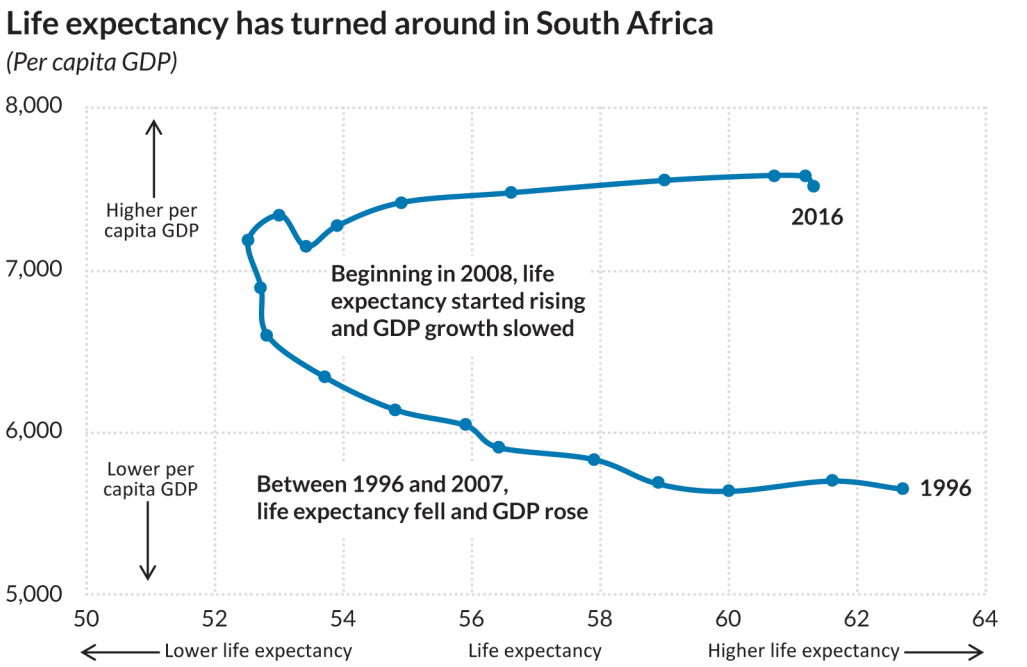


A common challenge is how to clearly show the association between two time series.

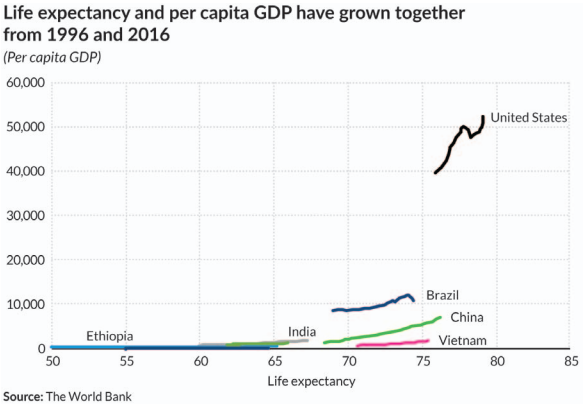
Now, notice what happens when the two lines are combined and plotted in a single graph. Here, life expectancy is shown along the horizontal axis and economic growth is plotted along the vertical axis.

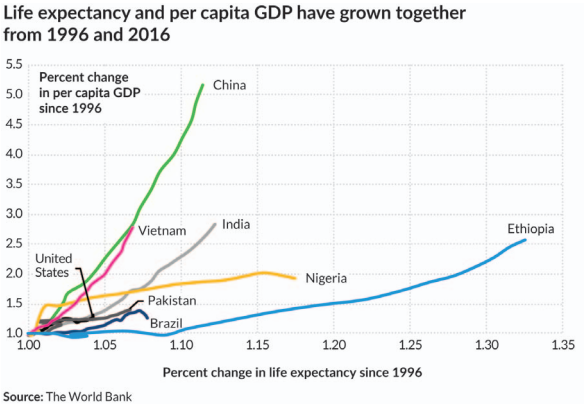
Instead of jumping back and forth between the two charts, we can see that during the first half of the period life expectancy fell (moved to the left along the horizontal axis) and the economy grew (moved up along the vertical axis). Starting around 2006, life expectancy started to increase again (now moving to the right along the horizontal axis) while the economy grew, but at slower rate (the slope of the line along this latter period is flatter than before).

Because this graph is different than standard graph types, your reader will need more time to understand how to read it. Annotation can help: Consider adding more axis labels, arrows and a label for the year on the first and last point. But when your reader understands this graph, it becomes part of their graphic toolbox, just like it has now become part of yours.



The connected scatterplot is one way to show how two time series are related to one another. One series corresponds to the horizontal axis and the other to the vertical axis.





Though less familiar to some readers, the connected scatterplots can be used to show more data series across two metrics.

You can also use connected scatterplots to show more groups. The connected scatterplot on the left shows the *levels* of economic growth and life expectancy for ten different countries. Here, the higher per capita GDP in the United States stands out, but patterns in the other countries are harder to see. The graph on the right shows percent *growth* in both variables since 1996. In this view, the United States is hardly visible, while the large gains in China and Ethiopia, for example, are clearly visible.

Which graph is better? As always, that depends on your audience, your argument, and what pattern, trend, or finding you want to bring to your reader's attention.

CONCLUSION

The graphs we covered in this chapter show changes over time. There are simple and familiar graphs, like line, area, and stacked area charts. But there are also more complex, less familiar, but equally useful chart types.

The line chart may be the most basic and familiar chart type to show changes over time. There is no limit to the number of lines you can plot, but if there are many lines, consider using color and line thickness to draw your reader's attention to the most important ones. Consider using data markers to add nuance to subtle or small data sets or ways to mark important points. As with many graphs we will explore, but perhaps even more so with line charts, use visual cues to signal missing data.

Alternative chart types are useful when you have too many data series to track in a single graph. Try sparklines, a small multiples approach, cycle charts, or horizon charts when you have a lot of data to visualize. For some of these approaches, enabling the reader to discern exact values is less important than showing them the overall trend or pattern.

Other graph types, like flow charts and timelines, have infinite varieties and styles. Horizontal layouts may work for some people, content, and platforms, while vertical layouts may be better online where it matches the natural scrolling motion. Compact layouts are best for mobile platforms.

Whichever graph you use to plot your data, consider how much detail your reader needs and how you can guide them to the point you wish to convey. Many of these chart types are well-known and understood, so our challenge is to make them engaging and interesting without sacrificing accuracy.

DISTRIBUTION

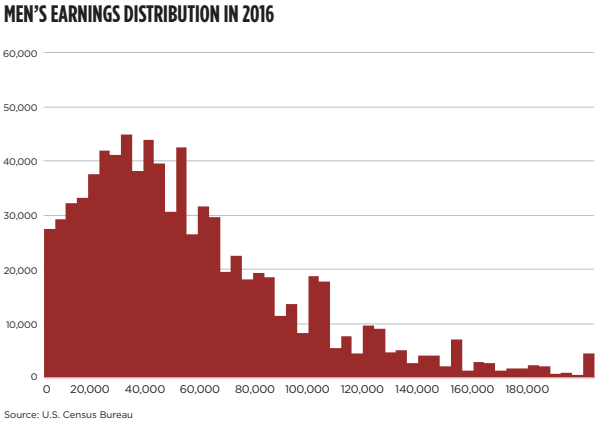
This chapter covers visualizations of data distributions and statistical uncertainties. These may be inherently difficult for many readers because they may not be as familiar with the statistical terminology or the graphs themselves, which may look quite different from the standard graphs they are used to seeing.

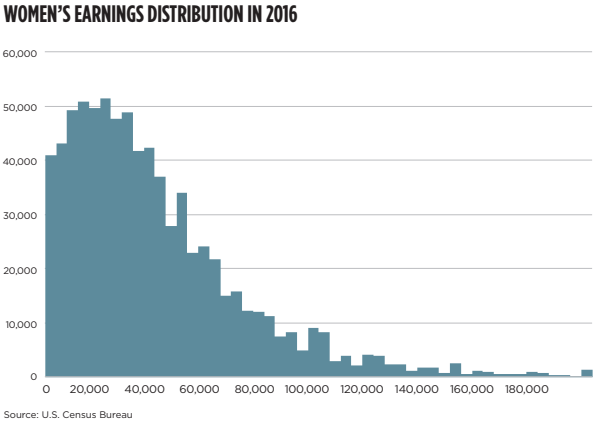
Charts like the fan chart and the box-and-whisker plot show statistical measures like confidence intervals and percentiles. Violin plots, which depict entire distributions, may look so foreign that your reader will need detailed explanations to understand them. This doesn't mean that these charts are inherently *bad* at visualizing data—proper labeling and design can make even the most esoteric box-and-whisker plot interesting—but the hurdle of statistical literacy may make such graphs difficult for many readers.

Graphs in this chapter follow the guidelines published by the *Dallas Morning News* in 2005. The *News*'s guidelines include instructions for specific fonts and colors, as well as ways to design and style different graphs, tables, maps, icons, and a summary of the newsroom workflow. The guide uses two fonts, Gotham and Miller Deck, depending on the size and purpose; I use the Montserrat font, which is similar to Gotham.

HISTOGRAM

The histogram is the most basic graph type for visualizing a distribution. It is a specific kind of bar chart that presents the *tabulated frequency* of data over distinct intervals, called *bins*, that sum to the

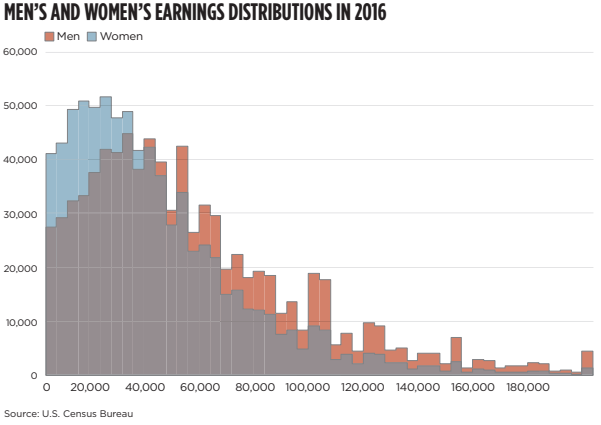


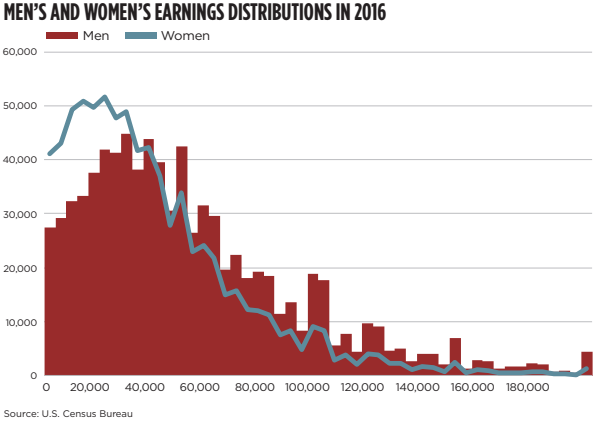


Histograms divide the entire sample into intervals (also called "bins"). The height of the bin shows the number of observations within it.

total distribution. The entire sample is divided into these bins, and the height of each bar shows the number of observations within each interval. Histograms can show where values are concentrated within a distribution, where extreme values are, and whether there are any gaps or unusual values.

We can layer histograms together to show how different distributions compare. The two histograms above depict the distribution of earnings for men and women working in the United States in 2016. We can make some general comparisons between them, but that task is made easier in the next two graphs, where the distributions are placed on top of each other.

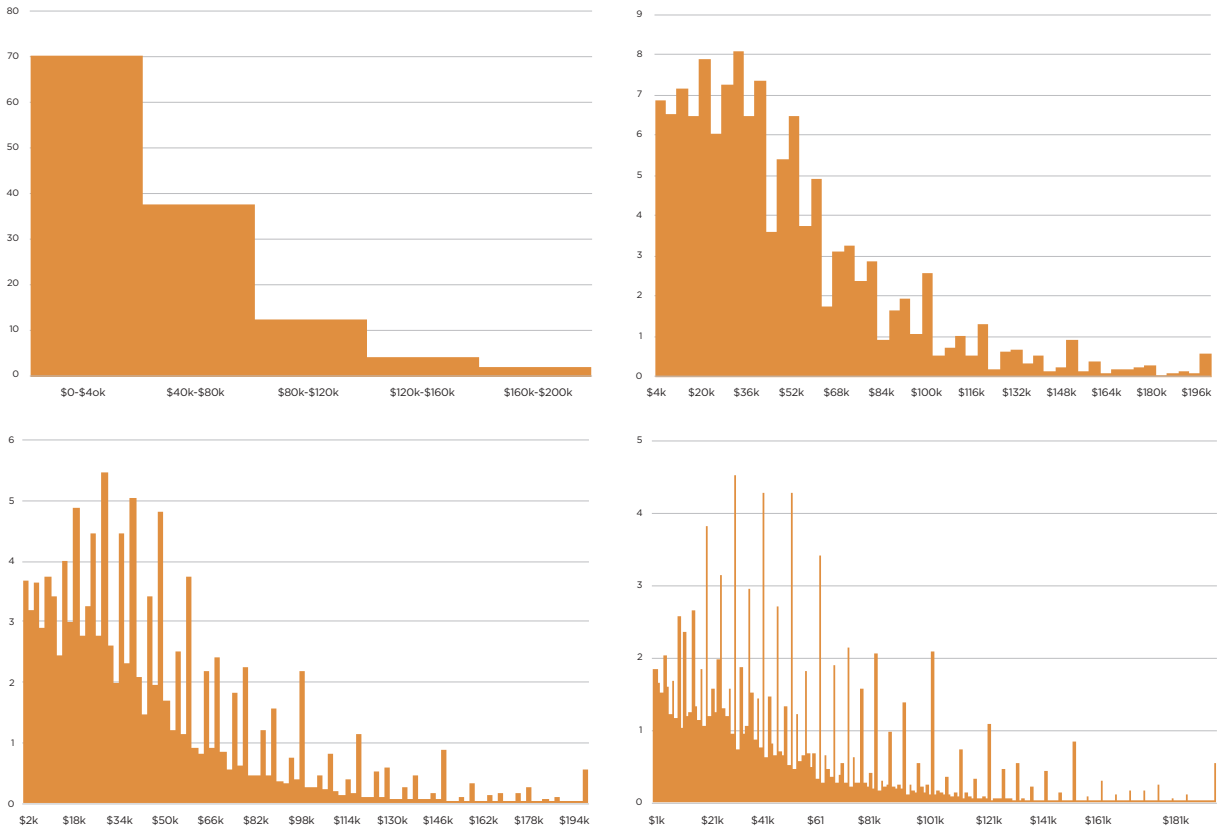




Histograms can be layered together by using saturated colors (left) or different encodings like bars and lines (right).

The graph on the left uses two column charts and transparent colors so both are visible. The graph on the right combines a column chart and line chart, which has the advantage of not using transparent colors, but the balance of how the two groups are presented is now unequal. You might also notice that the line intersects the *middle* of each bin as opposed to spanning the entire bin as the columns do—it's a minor difference but one that you may want to keep in mind.

A key consideration in creating a histogram is how wide to set the bins. Bins that are too wide may hide patterns in the distribution, while bins that are too narrow may obscure the overall shape of the distribution. While there is no "correct" number of bins, there are a number of statistical tests (using, for example, square roots, logarithmics, or cube roots) that



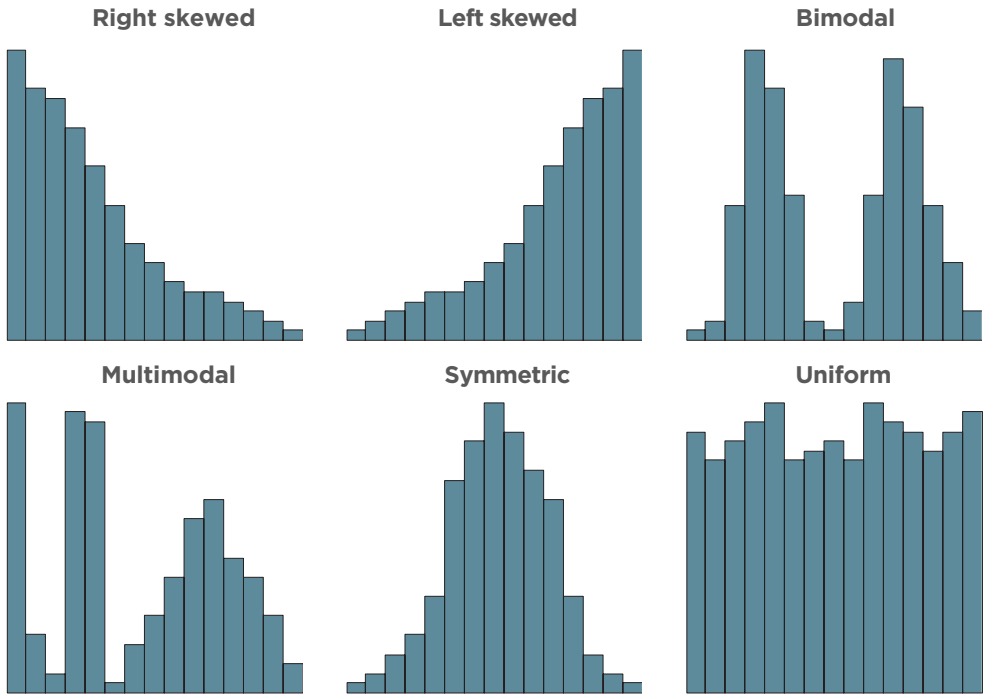
A data distribution will look different depending on the number bins, as it does here as the number of bins increases from 5 to 30 to 50 to 120.

can help determine the optimal bin width. In this example, notice how the distribution looks different as the number of bins increases from 5 to 30 to 50 to 120.

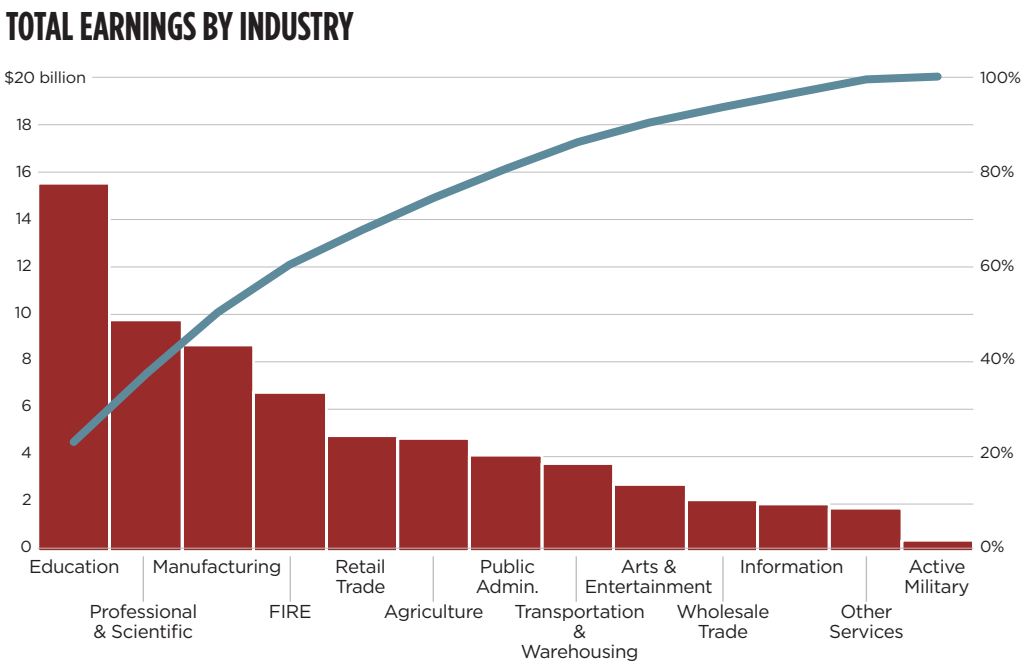
Histograms can help us understand whether our data *lean* to one side or another. A distribution in which more data are pushed to the left is known as *right-skewed*. A histogram with more observations to the right is called *left-skewed*. Distributions that have two peaks are called *bimodal* and distributions with multiple peaks are called *multimodal*. *Symmetric* distributions are those with a roughly equal number of observations on either side of a central value and *uniform* distributions in which the observations are roughly equally distributed.

When we understand the distribution of our data and its possible skew, we're better prepared to conduct more accurate statistical tests. Two completely different distributions may have the same mean and median, but if we don't understand the spread and structure of our data, our results may not paint a complete picture. This is where visualizing our data can be invaluable.

A modification to the basic histogram is the Pareto chart, named after the Italian engineer and economist Vilfredo Pareto. The Pareto chart (next page) consists of bars that represent



Histograms can help us understand the shape of the distribution of our data. Here we see six such forms of distribution.



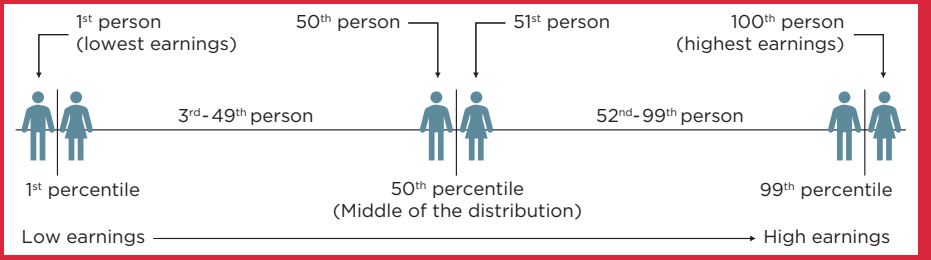
The Pareto chart shows values for each group (usually in bars) and the cumulative total as a line.

individual data points and a line that represents the cumulative total. The Pareto chart may be the exception to the rule against dual-axis charts because its *purpose* is to show the complementary distributions overlaid on each other. Of course, the two metrics are really not two different measures—it's the same metric, one as a marginal distribution (separate values of each group) and one as a cumulative distribution (where the values sum to a total).

This Pareto chart shows total earnings in thirteen different major industries in the United States—the bars show the total earnings in each industry and the line shows how the shares add up to total earnings in the economy.

UNDERSTANDING PERCENTILES

Imagine one hundred people standing on an auditorium stage. You stand in the audience as they line up from your left to right, arranged by their earnings. The person with the lowest earnings stands at the left side of the stage, and the person with



the highest earnings stands at the right. Together, they represent the entire earnings distribution.

The first person in the line has the lowest earnings. At their position, 99 percent of people—all those to their left—have higher earnings. They are said to be in the 1st percentile. Similarly, there is a worker on the other side of the stage in the 100th position. To their right stands 99 other people—99 percent of all people on the stage with lower earnings. They are in the 99th percentile of the distribution, in the top 1 percent. Finally, there is a point in the middle of the stage that splits everyone into two equal groups, 50 percent of the distribution on either side. That point (or more precisely, the *earnings* at that point) represents the 50th percentile or the median of the distribution.

The mathematics of increasing the number of people on stage from 100 to 200 to 1,000 to 150 million does not change—the middle of the ordered distribution is still the median and the person standing at the 10 percent position is at the 10th percentile. Because percentiles are independent of the population, you can compare them across any group such as country or industry.

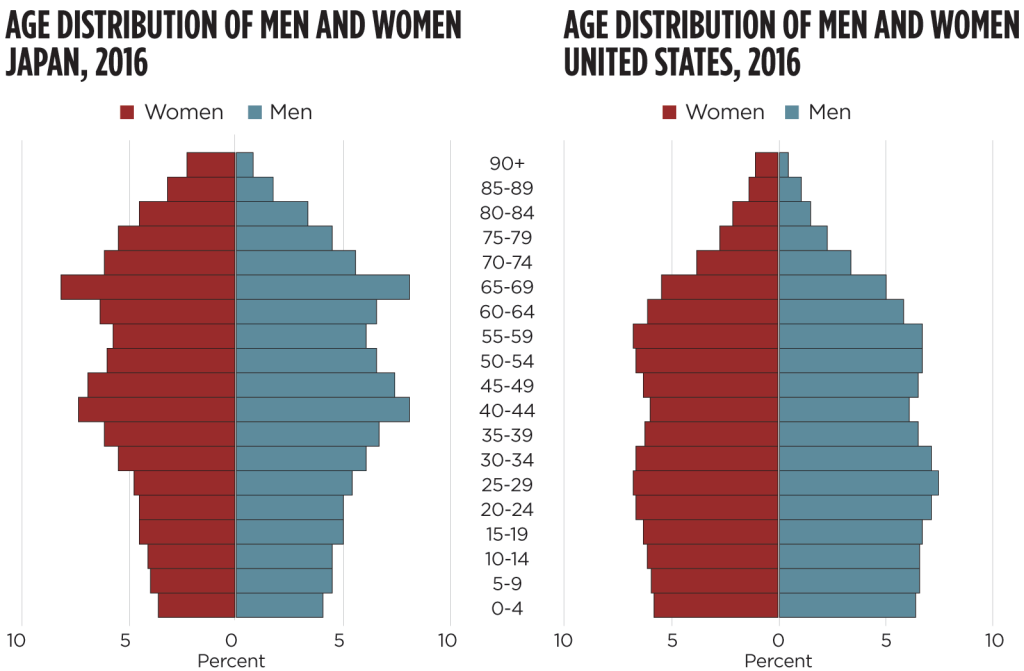
While percentiles identify a specific location in the distribution, there are other metrics that will give you an overall measure of the distribution. The *mean* or *average* is equal to the sum of all values divided by the number of observations. Because we add up all the data, large values can generate a distorted picture of the true distribution. In the example above, the mean would change dramatically if we replaced one of the people on stage with someone who earned $100 million, but—take note—the median would *not* change because that person would still stand on the far right edge of the stage and the rest of the people in line would stay in the same position.

The *variance* is another metric of a distribution and measures how far each observation in a data set is spread out from the mean value. A large variance indicates the values in the data are far from the mean and from each other; a small variance, by contrast, indicates the opposite. A full decomposition of the variance and related formulas are beyond the scope of this book, but if you are working with data and creating data visualizations, it is certainly worth a bit of time and study to understand how they can be used to better understand the shape and scope of your data.

PYRAMID CHART

Most often used to show changes in population-based metrics such as birth rates, mortality rates, age, or overall population levels, pyramid charts put two groups on either side of a center vertical axis. The pyramid chart is a subcategory of the diverging bar chart (see page 92), but the name is reserved for comparing distributions, most often ages. As with the diverging bar chart, the layout may cause some confusion, because your reader may assume the leftward bars represent negative values, and the rightward bars represent positive values.

The advantage of the pyramid chart is that we can assess the overall shape of the distribution because both groups sit on the same vertical baseline. While many pyramid charts use different colors for the two groups, that's not a necessary characteristic.



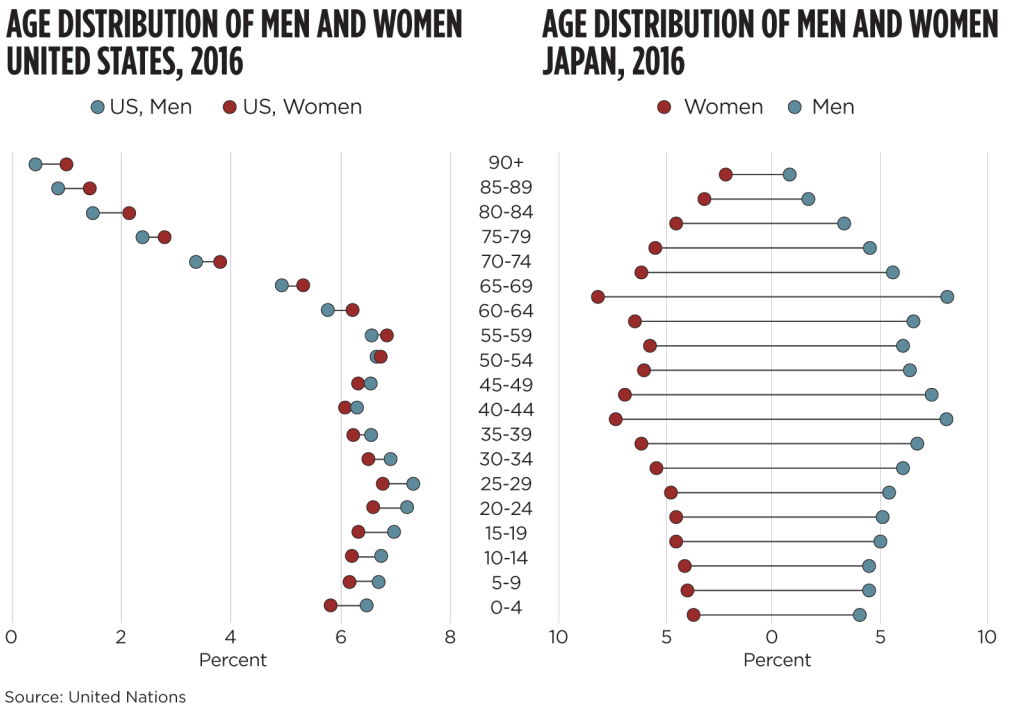
Source: United Nations

Pyramid charts are a type of diverging bar chart, typically used to show population-based metrics like birth rates, mortality rates, age, or overall population levels.

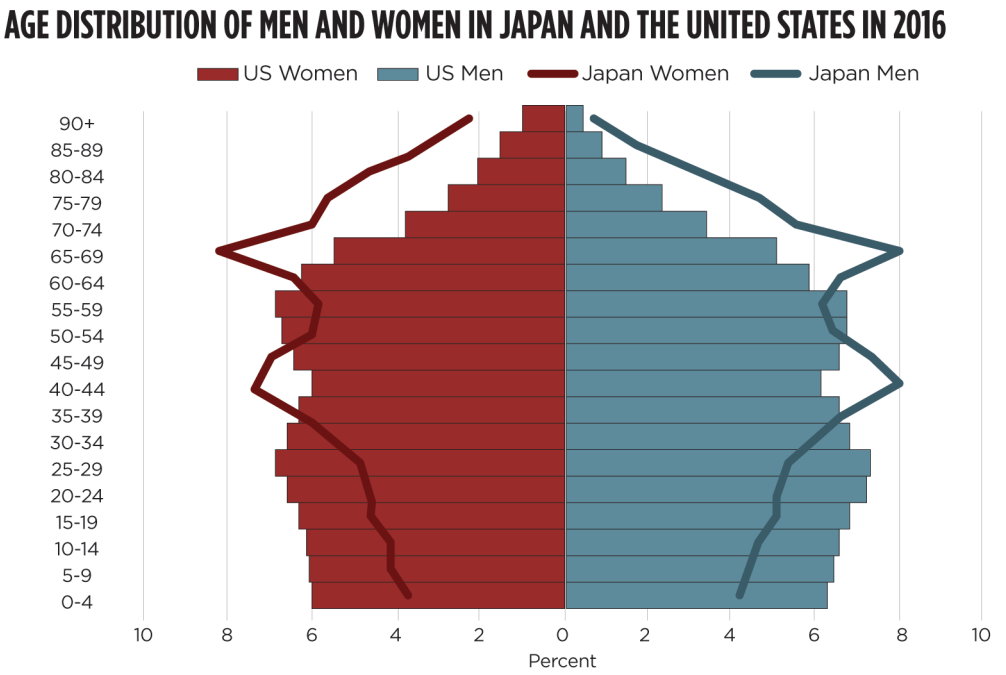
These pyramid charts show the distribution of ages in the United States and Japan in 2016. In both graphs, women are represented on the left branch of the vertical axis and men on the right. Each row represents a different age group: 0–4 years old, 5–9 years old, and so forth. The shape of the chart means we can immediately see that Japan has a greater share of older people, while there is a larger share of younger people in the United States.

Because the bars are not next to each other, it is difficult to compare the total shares of men and women. But again, whether that's a problem depends on the goal of your visual. If you want your reader to compare the shares of the two genders, then a different chart type—such as a paired bar chart or dot plot—would be a better choice. But if you want your reader to see the overall shape of the distribution, the pyramid chart is perfect.

A natural alternative to the pyramid chart is the dot plot or the lollipop chart. Dots for each gender are positioned along the horizontal axis corresponding to the data value and connected by a line. Or, as shown on the right, we could simply use a lollipop chart (also see page 80), replacing the bars with lines and dots. With either approach, we can still use different colors or just use a single color for the entire graph.



Alternatives to the basic pyramid chart are the dot plot or lollipop chart.

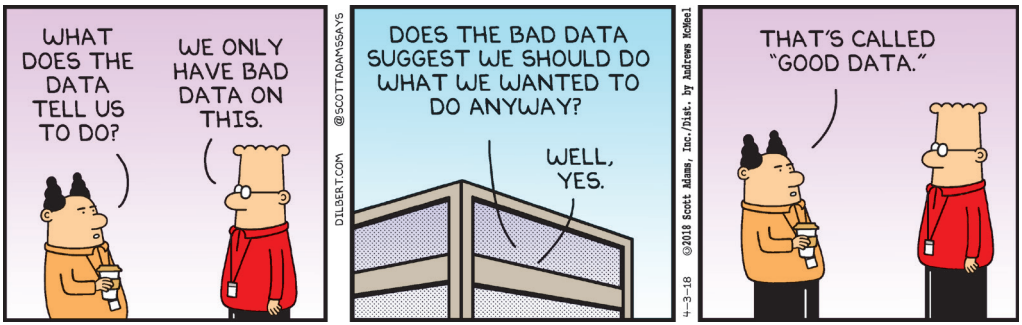


Combining distributions in one chart can be accomplished by adjust colors and combining different encodings.

One challenge with the pair of pyramid charts is to accurately compare the age distributions across the two countries. We can infer the main relationship, that Japan is, on average, older than the United States, but it's harder to make a more detailed comparison. Placing the charts atop each other—a technique we saw earlier with the histogram—makes that task easier. But be mindful that in using this approach the data for the two countries are shown with different encodings (bars and lines), which can risk emphasizing one set over the other.

VISUALIZING STATISTICAL UNCERTAINTY WITH CHARTS

There are lots of types of uncertainty in data and statistics. Before we learn the different ways to visualize uncertainty, it is worth pausing to understand what we mean by the term. Even if you're not a statistician or mathematician, it's important to understand how such uncertainty and measurement error can affect our results and, ultimately our



Source: Scott Adams

visualizations. Accounting for uncertainty—and making it clear you've done so—builds the reader's trust in your work.

We can think of the term uncertainty in two main ways. One is *uncertainty from randomness*, which applies to the statistical confidence in our statistical models and results. As an example, consider the standard margin of error built into political polling data: "Candidate Smith has a 54 percent approval rating with a margin of error of plus-or-minus 4 percentage points." Another kind is what we might call *uncertainty from unknowns*, where our data are inaccurate, untrusted, imprecise, or even unknown. A very simple example might be something like a data set that includes infants' ages in months instead of weeks. Using statistical and probabilistic models enables us to confront the first kind of uncertainty, which we can therefore visualize; the second kind of uncertainty concerns unknowns that can't necessarily be resolved through more data.

A thorough treatment of *uncertainty from randomness* (error margins, confidence intervals, and the like) are beyond the scope of this book. But *uncertainty from unknowns* is something that many readers can easily relate to. To illustrate, let's consider the data I use in this chapter: worker's earnings by industry and state. The data set used for this analysis is the 2016 U.S. Census Bureau's American Community Survey. The survey includes demographic and economic information for about 3.5 million people per year. For the data in this chapter, I examine individual earnings for more than a million people.

Now, imagine all the reasons why someone might tell the Census Bureau the wrong answer when they ask about their earnings. They might lie. They might round their earnings to the nearest dollar—or nearest hundred dollars, or nearest thousand dollars. Maybe they work side jobs they didn't mention. They may be asked about their spouse's or partner's earnings and have to hazard a guess. There are all sorts of reasons they may get it wrong, and recent economic research

shows that reporting error (especially in government program participation) in some of the largest, most trusted government household surveys has been increasing over the past several years.

We must also recognize that for this survey, the Census Bureau only asks a *share* of Americans (so our calculations from these data also suffer from *uncertainty from randomness*). Consider all the reasons why that "sample" may not be truly representative. Maybe some people don't want to answer the survey. Maybe they moved and didn't get the form, or changed their phone number and didn't get the call.

Whenever we work with data, we should consider how these kinds of uncertainty may lead to some "error" around our final estimates. Not mistakes, but uncertainty. This error is fundamental to being careful with data and ultimately visualizing and explaining our results carefully. Line charts and bar charts suggest certainty with their sharp boundaries and crisp edges, but that certainty is rarely actually the case.

In his book *How Charts Lie*, Alberto Cairo remarks that, "Uncertainty confuses many people because they have the unreasonable expectation that science and statistics will unearth precise truths, when all they can yield is imperfect estimates that can always be subject to changes and updates." We should not expect flawless data, and we should be ready to explain those imperfections to our readers as best we can.

▸ ▸ ▸ ▸ ▸

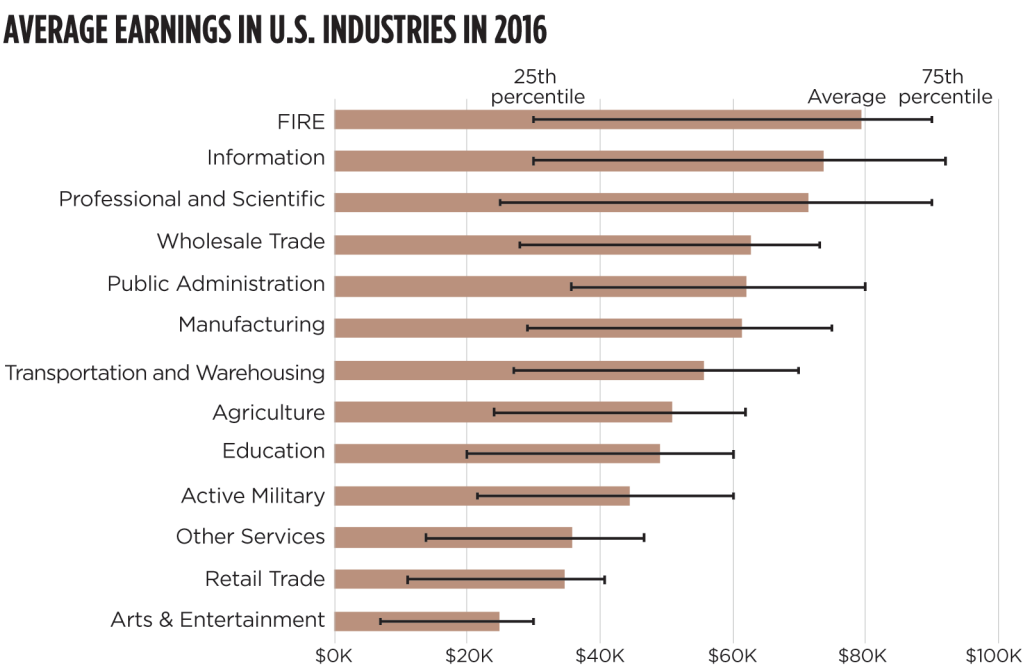
We now move to the specific challenge of conveying uncertainty around data estimates or results from statistical models. This is a common problem: In a survey of ninety data visualization authors and developers, information visualization researcher Jessica Hullman found that graph creators did not include uncertainty in their work for four main reasons. First, they did not want to confuse or overwhelm viewers. Second, they did not have access to information about the uncertainty in their data. Third, they did not know how to calculate uncertainty. And fourth, they did not want to make the data appear questionable. Hullman argues that visualizing uncertainty is important because "a central problem is that authors often omit or downplay information, such that data are interpreted as being more credible than they are." More effectively conveying such uncertainty—especially when making statistical claims—builds trust and credibility.

This section will familiarize you with visual signals of uncertainty. There are many chart types that can be used to show uncertainty around a central estimate, and this section introduces a few of them: error bar charts, confidence interval charts, gradient charts, and fan charts.

ERROR BARS

Perhaps the simplest and most common way to visualize uncertainty is to use error bars: small markers that denote the error margin or confidence interval. Error bars are not really a visualization on their own, but are an addition to other charts, often bar or line charts. The ends of the error bars can correspond to any value you choose: percentiles, the standard error, the 95-percent confidence interval, or even a fixed number. And because error bars can convey these multiple statistical measures, recent research has shown that this can invite confusion on the part of the reader, making incorrect conclusions about the data. We must therefore clearly label the intervals, either in a chart note or, preferably, on the chart itself.

This bar chart shows average earnings in each of thirteen industries in 2016. Error bars denote the 25th and 75th percentiles.



Source: U.S. Census Bureau

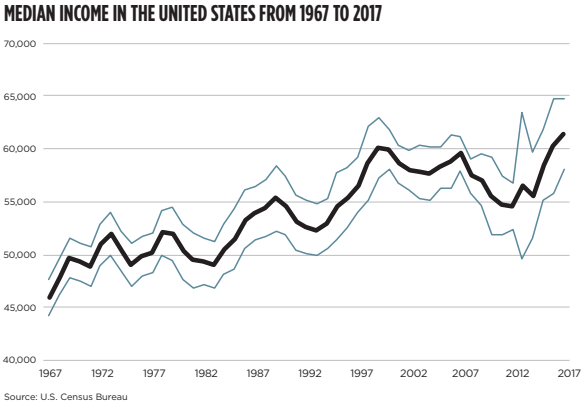
The simplest and most common way to visualize uncertainty or distributions is to use error bars, small markers that denote the margin of error or confidence interval.

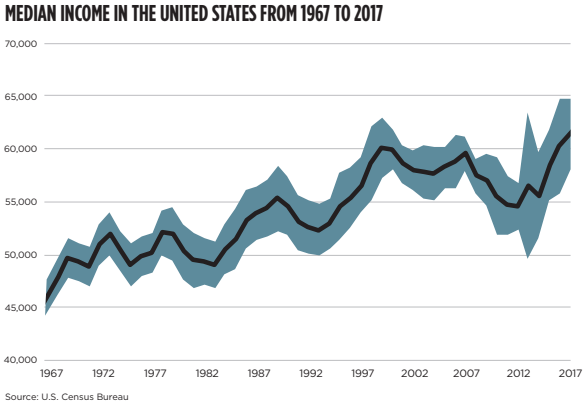
Applying error bars to bar charts raises a potentially interesting complication: some research suggests that we tend to judge the points that fall *within* the bar as more likely than those *outside* the bar ("within-the-bar" bias). In the previous chart, that would mean a reader is more likely to assume that the salary of a worker in the Finance, Insurance, and Real Estate (FIRE) sector is less than $80,000 and not more than $80,000. Other research has found that we can better judge uncertainty and the distribution with other types of graphs, such as the violin plot, stripe plot, or gradient plot.

While it may be a familiar approach for many readers and requires less data than some of these other charts, existing research shows that we are not particularly good at assessing uncertainty through these kinds of visual approaches.

CONFIDENCE INTERVAL

A confidence interval chart typically uses lines or shaded areas to depict ranges or amounts of uncertainty, often over time. The basic confidence interval chart is literally a line chart with three lines: one for the central estimate, one for the upper confidence interval value, and another for the lower confidence interval value (these upper and lower lines can be confidence intervals, standard errors, or a fixed number). The lines may be solid, dashed, or colored, but if the central estimates are the primary numbers of interest, that line should be thicker or darker to highlight it against the confidence interval values.





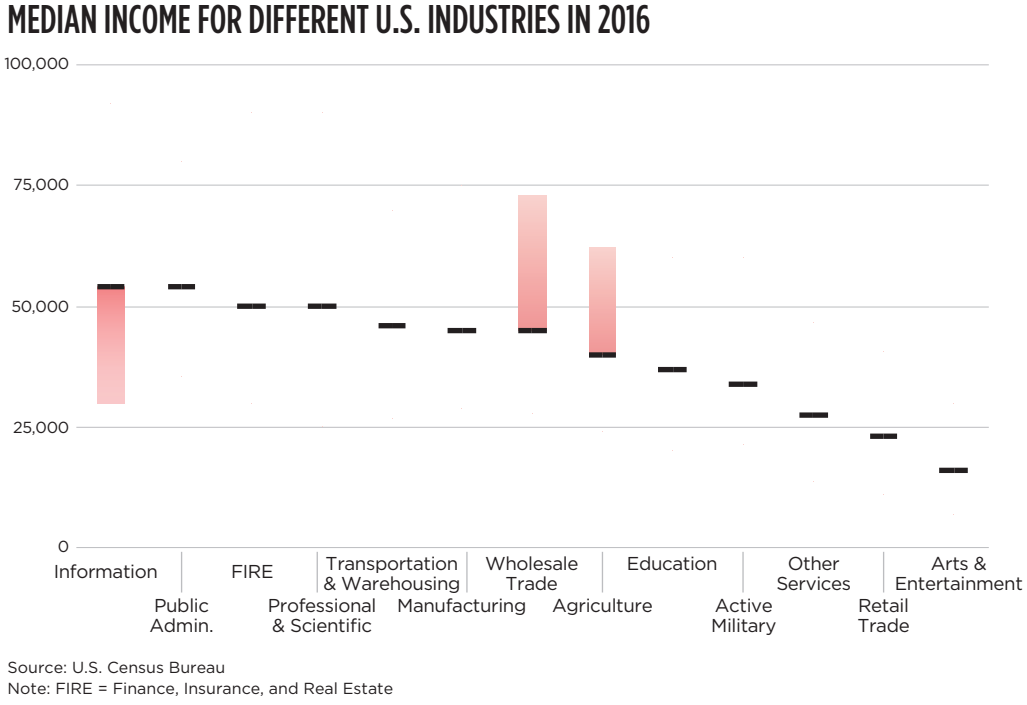
Lines or shaded areas around a central line can visualize ranges of uncertainty.

The two charts on the previous page show median earnings in the United States between 1967 and 2017. The standard error around those estimates is depicted by two separate lines in the left graph and by a shaded area in the right.

GRADIENT CHART

A gradient chart (sometimes called a stripe plot) shows distributions or differences in uncertainty. There are many ways to use the gradient plot, but the basic technique is to plot the primary number of importance and add a color gradient on one or both sides to visually demonstrate the measure of uncertainty around that single point. The plot is named not necessarily after the *shape* of the graph but after the use of the color gradient.

The gradient plot can show changes over time or, as in the graph here, the distribution around individual observations. This gradient plot shows the exact same data as the error

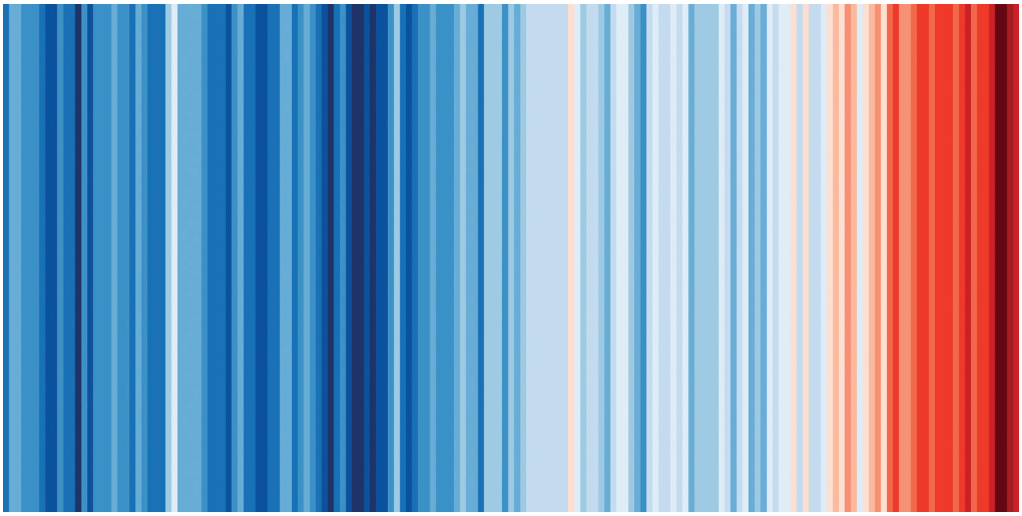


Gradient charts use a color gradient on one or both sides of the primary number of interest to show distributions or uncertainty.

bar chart above, but instead of bars with error lines jutting out in both directions, average earnings are encoded with the dark horizontal line and the 25th to 75th percentiles of the distribution are shown with the gradient. The color gradient might illustrate, for example, multiples of the standard error, which can be a signal to the reader that the outcomes are less certain the further they are from the central estimate line.

Stripe charts can also be an effective way to show changes over time. A strong example of this is the series of stripe charts created by Dr. Ed Hawkins, a climate scientist at the University of Reading, that showed temperature changes from 1850 to 2018. Each bar (stripe) showed a different temperature level, ranging from cooler blues to hotter reds. Together, readers could quickly and easily see the marked increase in temperatures around the world as a whole and in their specific region of the world by using an online tool. These stripe charts were published by a multitude of websites, television stations, and even became a cover of the *Economist* magazine.

In a 2019 interview on the *Data Stories* podcast, Hawkins said that he "was looking for a way to communicate to audiences that aren't used to seeing graphs, or axes, or labels—things that we see day-to-day, but are complicated to them. It may look too mathematical to them, so it turns them off straight away." In a later interview on that same podcast, Jennifer Christiansen, the senior graphics editor at *Scientific American*, described them as, "every region's version of the climate stripe pattern progresses from cool blue to a warm red. No labels

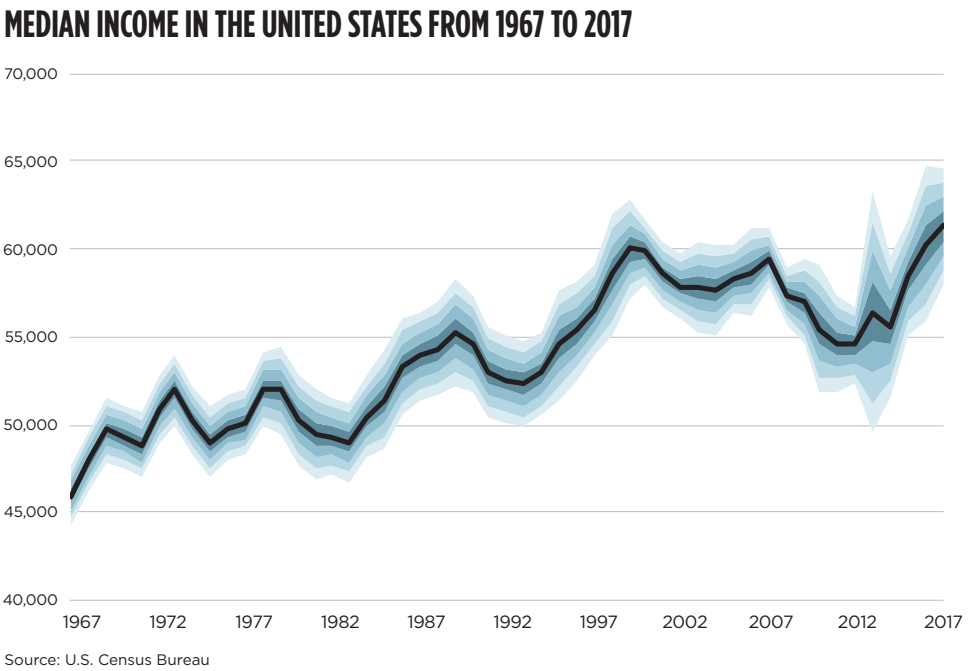


This stripe chart from ShowYourStripes.com shows global temperatures from 1850 to 2018. The simple colored stripes are easy to see and understand.

are needed; no caption is needed. It's a visceral and accessible nod to our warming planet with color representing annual temperature. And it prints legibly on everything from social media profiles, to pins, neckties, magazine covers, mugs, and concert screens."

FAN CHARTS

If the color or saturation of the shaded area between the confidence interval lines changes based on the value, it is often called a fan chart. Fan charts are like gradient plots for line charts, and they are great for visualizing changes in uncertainty over time. In the fan chart, values closest to the central estimate are the darkest and they lighten as the values move outward. The use of color distinguishes the move from higher levels of statistical confidence to lower levels. The advantage here is that it signals to the reader how the estimates become less certain the further they are from the central estimate.

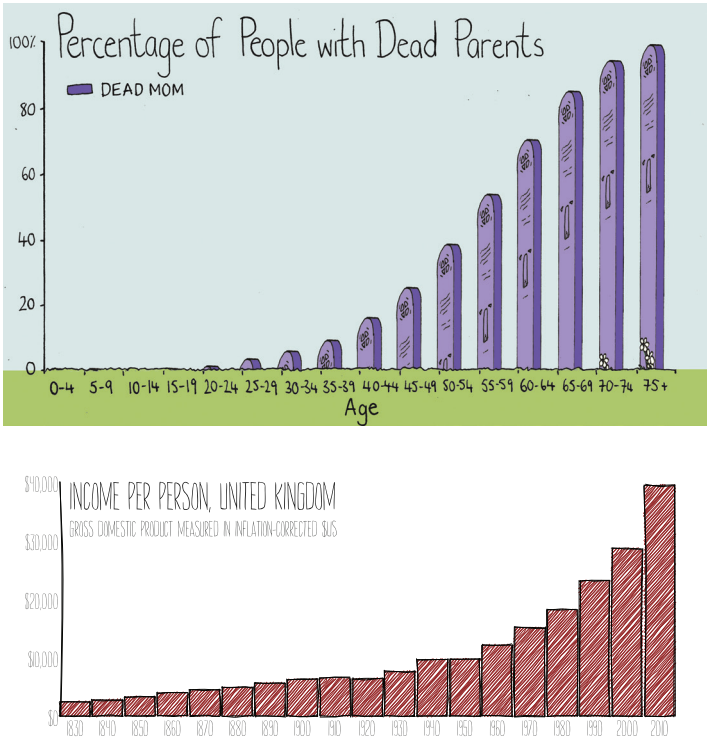


Like the gradient chart, the fan chart shows distributions (in this case, standard errors) around a central estimate.

This fan chart shows the change in median household income over the last fifty years. The color bands show the standard error divided into eight segments, though they could also show bands of percentiles or other measures. Similar to the gradient chart, the change in color saturation can show multiples of the standard error and signal less certainty as we move farther from the central estimate, which is here represented by the black line.

THE HAND-DRAWN LOOK

One last strategy to suggest uncertainty is not a visualization technique per se but a design technique. The hand-drawn, "sketchy," "gooey," or "painty" techniques can be used to add an



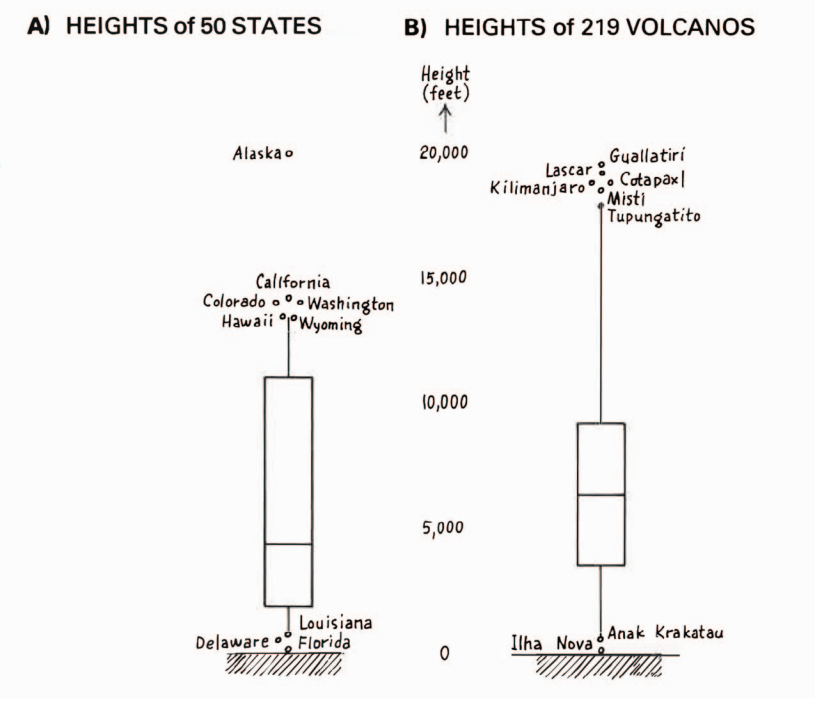
Hand-drawn, "sketchy," or "gooey" design effects use uneven edges to communicate a sense of uncertainty or imprecision.

Sources: Copyright Mona Chalabi (top) and Jo Wood, giCentre, University of London (bottom).

uneven edge or fuzziness to graph objects that will create a sense of uncertainty or imprecision. Research suggests that sketchy graphs generate more engagement in the graph and that we can "tie sketchiness to uncertainty or significance values." These two examples, the first from journalist Mona Chalabi at the *Guardian* and the second from Jo Wood at the University of London, both demonstrate these techniques in action.

BOX-AND-WHISKER PLOT

When you visualize the distribution of your data, you can show the entire distribution or just specific points within it. The box-and-whisker plot (or boxplot), originally called a *schematic plot* by its inventor John W. Tukey, uses a box and line markers to show specific percentile values within a distribution. You can also add markers to show outliers or other interesting data points or values. It is a compact summary of the data distribution, though it displays less detail than a histogram or violin chart.



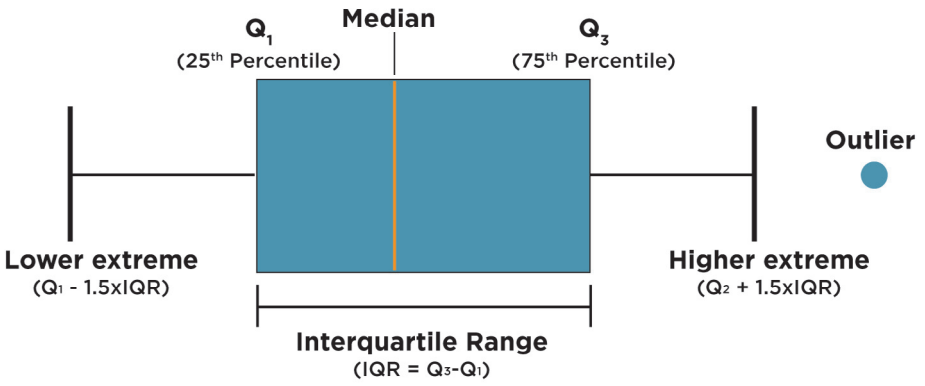
The original boxplot from Tukey 1977

The basic box-and-whisker plot consists of a rectangle (the *box*), two lines (the *whiskers*) that emanate from the top and bottom of the box, and dots for outliers or other specific data points. Most standard box-and-whisker plots have five major components:

1. The *median*, encoded by a single horizontal line inside the box.
2. Two *hinges*, which are the upper and lower edges of the box and typically correspond to the first quartile (or the 25th percentile) and third quartile (75th percentile). The difference between these two points is called the *Interquartile Range* or *IQR*.
3. The higher and lower extremes (sometimes the maximum and minimum) are placed at a position 1.5 times the *IQR* (recall the Box on page 74).
4. Two *whiskers* (the lines) connect the hinges to a specific observation or percentile.
5. *Outliers* are individual data points that are further away from the median than the edges of the whiskers.

Each of these components can vary depending on which parts of the distribution we wish to show. Some creators replace outliers with fixed quantiles such as the minimum and maximum values or the 1st and 99th percentiles. Some use the semi-interquartile range $(Q_3\text{-}Q_1)/2$, which can generate asymmetric whiskers. And some add other descriptive statistics like the mean or standard error. We can also vary the color, line thickness, and how and which parts of the chart are labeled.

As a practical example, the box-and-whisker plots on the next page show the distribution of earnings in our thirteen industries. The vertical line in the middle of each box represents



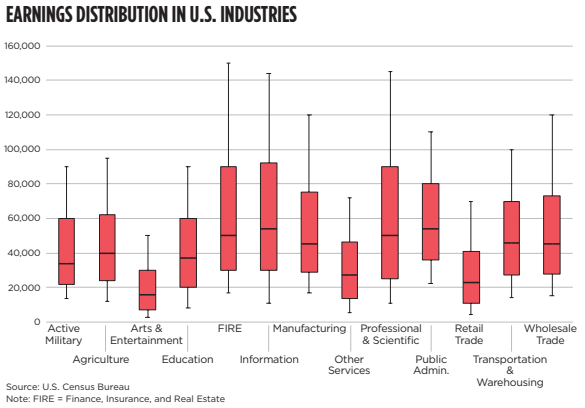
The basic box-and-whisker plot.

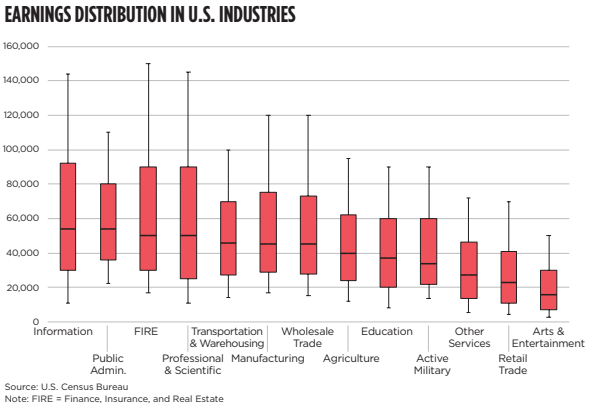
the median and the edges of the box represent the 25th and 75th percentiles. The ends of the lines (the whiskers) show the 10th and 90th percentiles.

The graph on the left sorts the industries alphabetically while the one on the right sorts them by the median value. I generally find it preferable to present data sorted by their values as opposed to alphabetical or some other arbitrary sorting. There may be cases when alphabetical sorting is best—for example, if we were presenting earnings across all fifty U.S. states, readers would find it easier to find individual states if they're sorted alphabetically. If, however, the goal of that graph is to discuss the high/low earnings of a particular state, we would sort the data by their values to make those comparisons easier.

As with all of these visualizations, showing percentiles and statistical or data uncertainty will often depend on the experience, interest, and expertise of our audience. In scientific or research applications, for example, communicating uncertainty is especially important to demonstrate whether a finding is statistically meaningful. But in other cases, where our data only have a single value for each observation—say, a single estimate of per capita GDP in the United States—we may not be able to visualize parts of the distribution.

In the case of the box-and-whisker plot, by plotting these specific percentile points, we are explicitly deciding *not* to visualize the entire distribution. This may not be entirely problematic, especially if other percentile points aren't particularly important, or if the data follow a fairly standard distribution. But we must always fully explore the data we're certain we're not hiding important patterns from our reader—or ourselves!



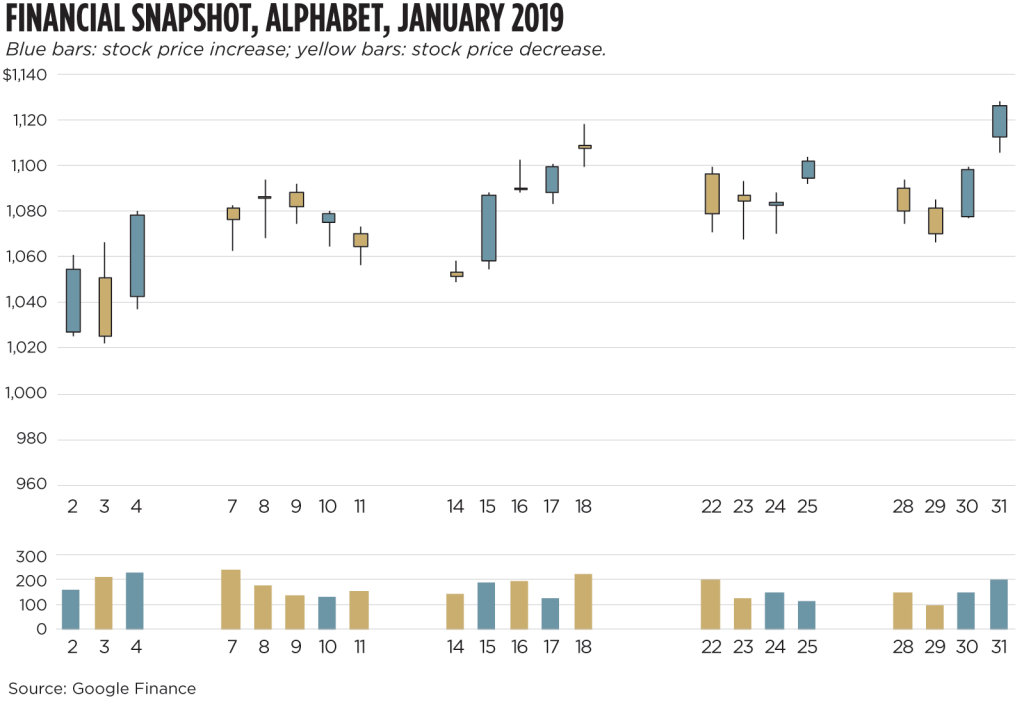


These charts show the distribution of earnings in thirteen industries either sorted alphabetically (left) or by median value (right). The edges of the box show the 25th and 75th percentiles and the whiskers show the 10th and 90th percentiles.

CANDLESTICK CHART

Candlestick or stock charts look like box-and-whisker plots, but they visualize different content. Whereas box-and-whisker plots visualize uncertainty or a distribution, candlestick charts visualize changes in the prices of stocks, bonds, securities, and commodities over time. Bars and lines show opening and closing prices and highs and lows in a day, plotted along a horizontal axis that measures time.

There are two elements of the candlestick chart. The central box—sometimes called the "real body"—shows the gap between the opening and closing prices. The lines that extend upward and downward from the real body—sometimes call the "wick"—show the low and high value for the day. Like the box-and-whisker chart, the candlestick chart includes specific points and does not show *all* of the activity during the day, such as price volatility.



The candlestick chart is like a box-and-whisker chart but is typically reserved to refer to prices of stocks, bonds, securities, and commodities over time.

Specific characteristics of the candlestick chart can vary in some obvious ways: color can be changed to differentiate between a drop in price during the day (i.e., the closing price is greater than or less than the opening price) and icons or other symbols can identify the high and low prices. Although I've placed this chart in this chapter, because of its relation to box-and-whisker plots, it could easily have also appeared in the *Time* chapter or even the *Comparing Categories* chapter.

The candlestick chart on the previous page shows overall daily trading patterns for shares of Alphabet, Inc.—the parent company of the Google search engine—from January to February 2018. The bar underneath shows trading volume. In both graphs, blue bars signal an increase in price over the day and yellow bars signal a decrease. Notice how the two graphs are stacked together as opposed to using a dual axis chart, which might be confusing or just plain cluttered.

VIOLIN CHART

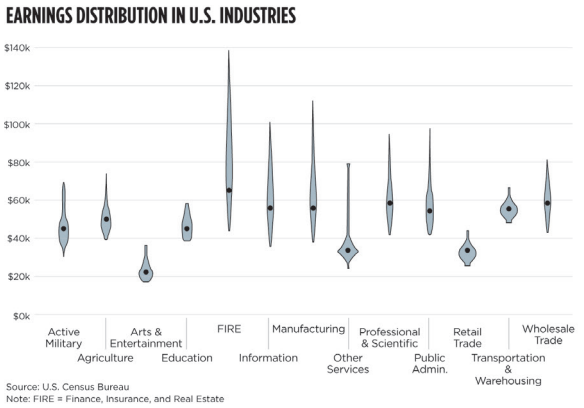
Instead of showing selected percentiles from the distribution, the goal of the next set of charts is to show the *entire* distribution. Unlike the box-and-whisker plot, in which we choose specific points in the distribution, or the histogram in which values are grouped together into intervals, the violin chart shows the shape of the whole distribution.

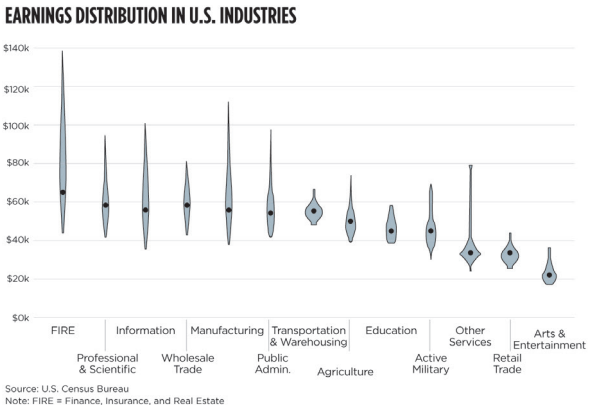
These violin charts use the same data as above, the average earnings in thirteen industries in 2016. The thicker areas mean that there are more values in those sections while the thinner parts imply lower frequency of observations. I've added a dot in the middle to mark average earnings within each industry. Notice again the differences in the view when the chart is sorted alphabetically (on the left) versus by the mean income (on the right).

KERNEL DENSITY

One consideration in creating this chart type is that it requires estimating what is called the *kernel density* of each distribution. Kernel densities are a way to estimate the distribution of a variable—akin to a histogram—but can be smoothed or made to look more continuous using different algorithms. For the violin plot, those density estimates are plotted to mirror each other around an invisible central line.

Think of it this way: A histogram plots a summary view of a distribution along a single axis. The violin plot mirrors a smoothed version of the histogram on either side of that single





Instead of showing select points (percentiles) in a data distribution, the violin chart shows the estimated shape of the entire distribution using kernel densities.

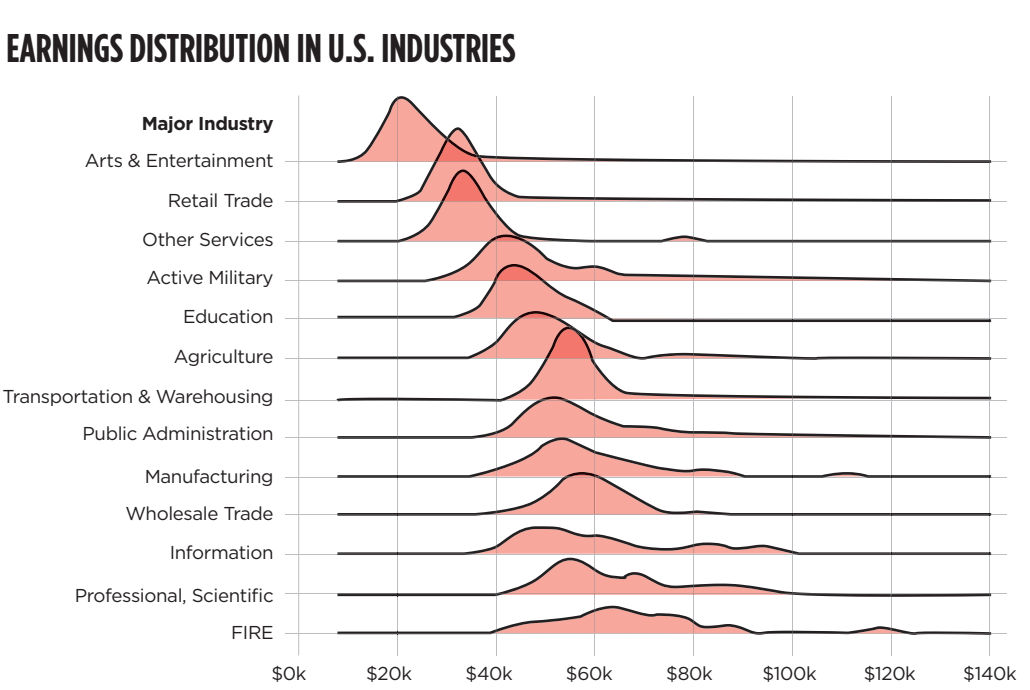
axis. How that smoothing is accomplished will depend on what sort of kernel density estimator you choose, which can vary based on the data, underlying function, and more.

Violin charts, then, are richer than the box-and-whisker plot, but can also be more difficult to create and for our audience to understand. In modern versions of Excel, for example, the box-and-whisker plot is a default graph, while a violin plot requires manually calculating the probability densities and then finding a graphing solution.

RIDGELINE PLOT

The ridgeline plot is a series of histograms or density plots shown for different groups aligned along the same horizontal axis and presented with a slight overlap along the vertical axis. In a basic sense, the ridgeline plot is like a small multiples histogram or a horizon chart where the histograms are aligned in particular way.

The ridgeline plot on the next page shows the earnings distribution across the thirteen different industries. The horizontal axis is shared across all thirteen industries and the distributions sometimes overlap along the vertical dimension. Depending on the color scheme and density of the data, there may be more or less overlap between the series, but as we have seen in other graph types (for example, sparklines and horizon charts), sometimes showing the overall pattern is more important than the reader being able to pick out all of the specific



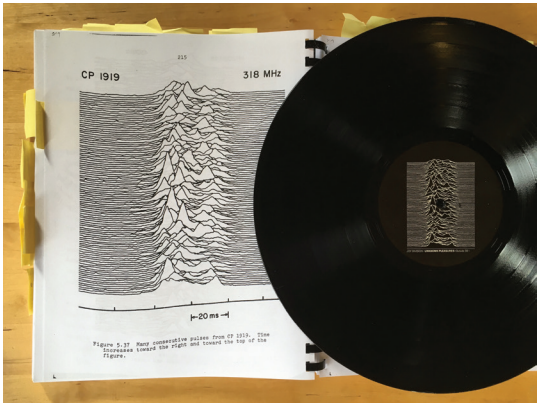
The ridgeline plot is a series of histograms shown for different groups aligned along the same horizontal axis and presented with a slight overlap along the vertical axis.

values. These overlaps can be a problem, but one that is mitigated by how quickly and easily readers see how the different distributions line up with one another along the same axis.

The most famous ridgeline plot is one that you didn't even know was a ridgeline plot: The album cover for *Unknown Pleasures*, the 1979 debut album of English post-punk band Joy Division, which had white lines on a black background with no band name, album title, or other identifiers.

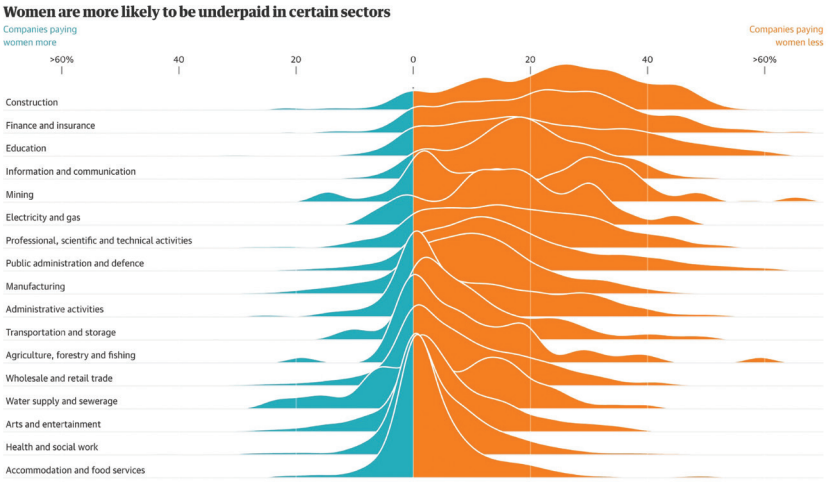
In 2015, Jen Christiansen, the Senior Graphics Editor at *Scientific American*, tracked down the original image to the 1970 doctoral dissertation of Harold D. Craft, Jr., a radio astronomer at Cornell University. The original chart graphed the distribution of consecutive radio pulses emanating from a pulsar, a type of neutron star. The album cover designer, Peter Saville, called it "a wonderfully enigmatic symbol for a record cover."

Data that lend themselves well to a ridgeline plot are those in which the distributions differ from one category (row) to another so that the reader can see the shift up and down the page. With data in hand, variations on color, font, and layout can help engage and interest your reader. The ridgeline plot on the next page was published by the *Guardian* in 2018 and



Source: Photograph by Jen Christiansen (featuring figure 5.37 from "Radio Observations of the Pulse Profiles and Dispersion Measures of Twelve Pulsars," by Harold D. Craft Jr., September 1970).

shows the distribution of the gap in pay between men and women in more than ten thousand companies and public bodies in the United Kingdom. Using different colors on either side of the vertical zero-percent line (perfect equality) and sorting the data from industries that have the highest pay gaps (e.g., Construction) to smallest pay gaps (e.g., Accommodation and Food Services) also helps direct the eye.



This ridgeline plot from the *Guardian* shows earnings distributions for men and women in different industries.

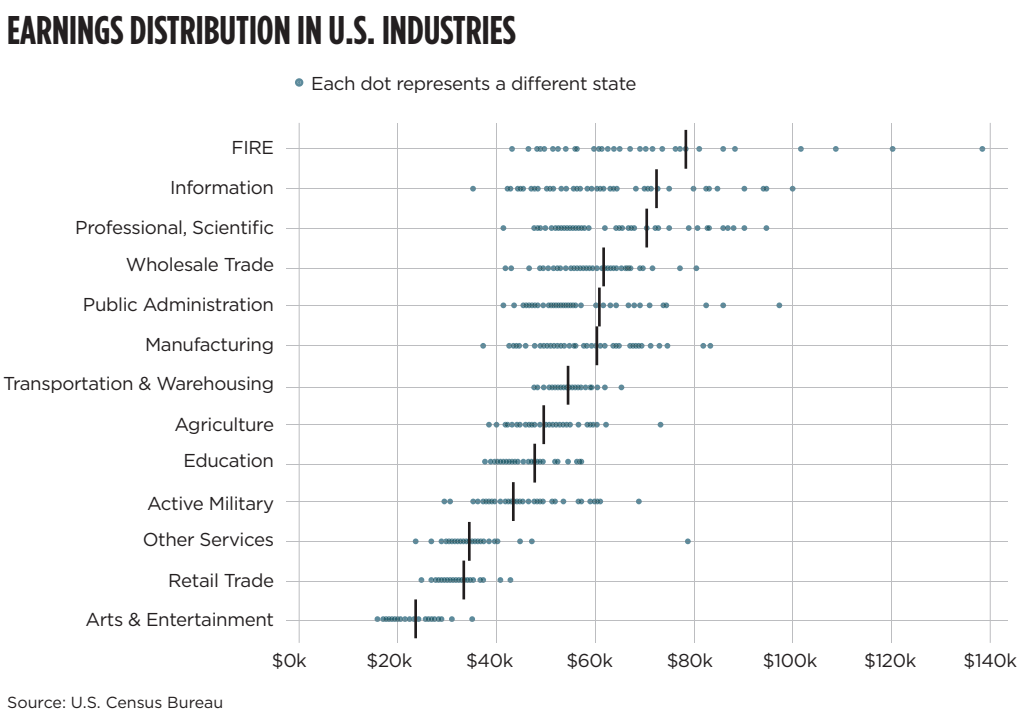
VISUALIZING UNCERTAINTY BY SHOWING THE DATA

The graphs presented in this chapter so far summarize data distributions with lines, points, bars, and color. To different degrees, this is what the histogram, violin, and ridgeline plots all do. Another way to visualize the distribution of your data is to just *show* the data.

STRIP PLOT

The basic way to show your data is with what is known as a *strip plot*. In this graph type, the data points are plotted along a single horizontal or vertical axis.

In this example, average earnings for each state are shown for each of the thirteen industries (the U.S. average is denoted with the vertical black line). We have already seen similar data presented in the box-and-whisker plot and violin chart, but here you can see individual

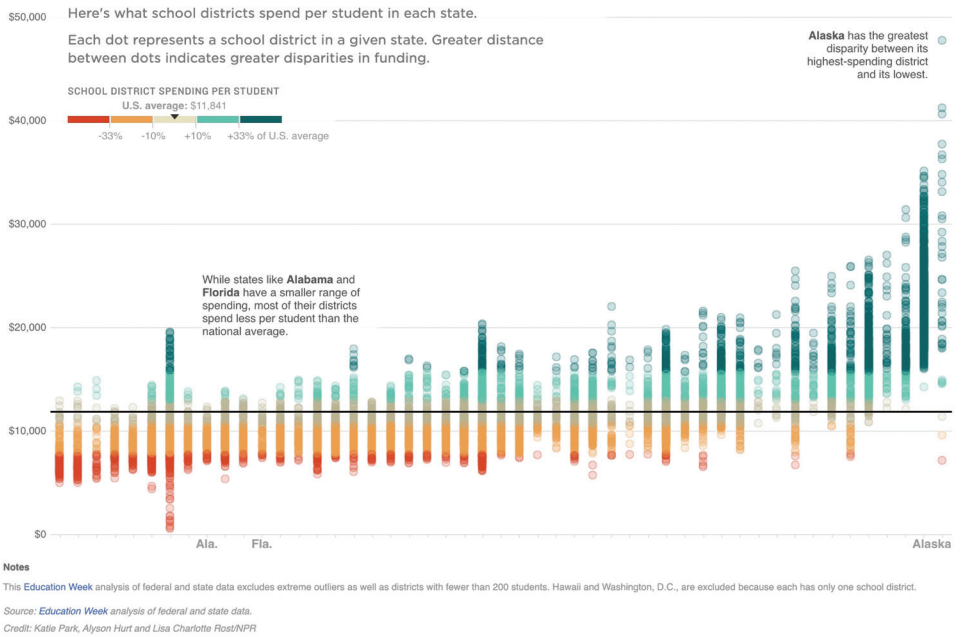


In a strip plot, data points are plotted along a single horizontal or vertical axis. This strip plot encodes the data with circles, but small lines are also often used.

points. Notice, however, that I'm showing average earnings for the fifty states, not for individual workers. Plotting earnings for everyone in my data (some 1.3 million people) would look like a single, dark line. There are too many values.

It's true that some of the data is obscured, but, especially by virtue of the overlapping transparent colors that make the patterns darker, it becomes clearer where the bulk of earnings lie in the distribution. There's no rule for how many data points are *too many*, but as you plot your data, you can always tell when you've passed that threshold.

This static image from an interactive strip plot from NPR is a good example of how this visualization can be richer than a standard bar chart or histogram. Here, they plot a point for every school district in every state. Darker orange dots (and below the black horizontal line) are districts in which spending per student is below the national average. Darker green dots are districts in which spending per student is above the national average. An interesting and useful design choice to make the dots transparent (with a solid border) lets us see where districts are close enough to overlap. Also notice that only Alabama, Florida, and Alaska are



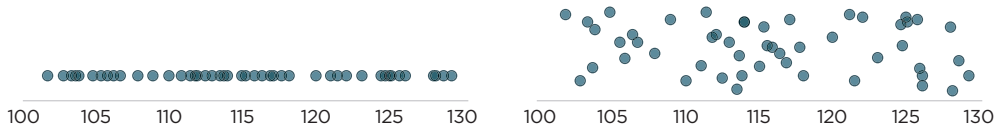
This strip plot from NPR shows the distribution of spending in different school districts around the United States.

labeled along the horizontal axis. Those three states are explicitly annotated in the chart. Other state labels only appear along the horizontal axis when, in the interactive version, the user hovers over a stack of circles.

BEESWARM PLOT

If we want to plot individual data points rather than distributions, one way to make the data more visible is to use a technique called "jittering." This is when we alter individual values slightly so that the data points don't lie on top of one another.

Consider the strip plot on the left, which leaves all of the data along the same horizontal axis. We can see clusters, but not all of the individual values. In the version on the right, the data are jittered along the horizontal and vertical axes to help make each point visible. There are different algorithms and approaches to jittering the data, and the most important consideration is to manipulate the values just enough to make them visible but not so much that it changes the overall view of the distribution. As with choosing a kernel density estimator for violin charts, the choice of jittering technique (for example, do we jitter both *x* and *y* variables and if so, do we jitter each variable independently?) will depend on the data and its underlying distribution.

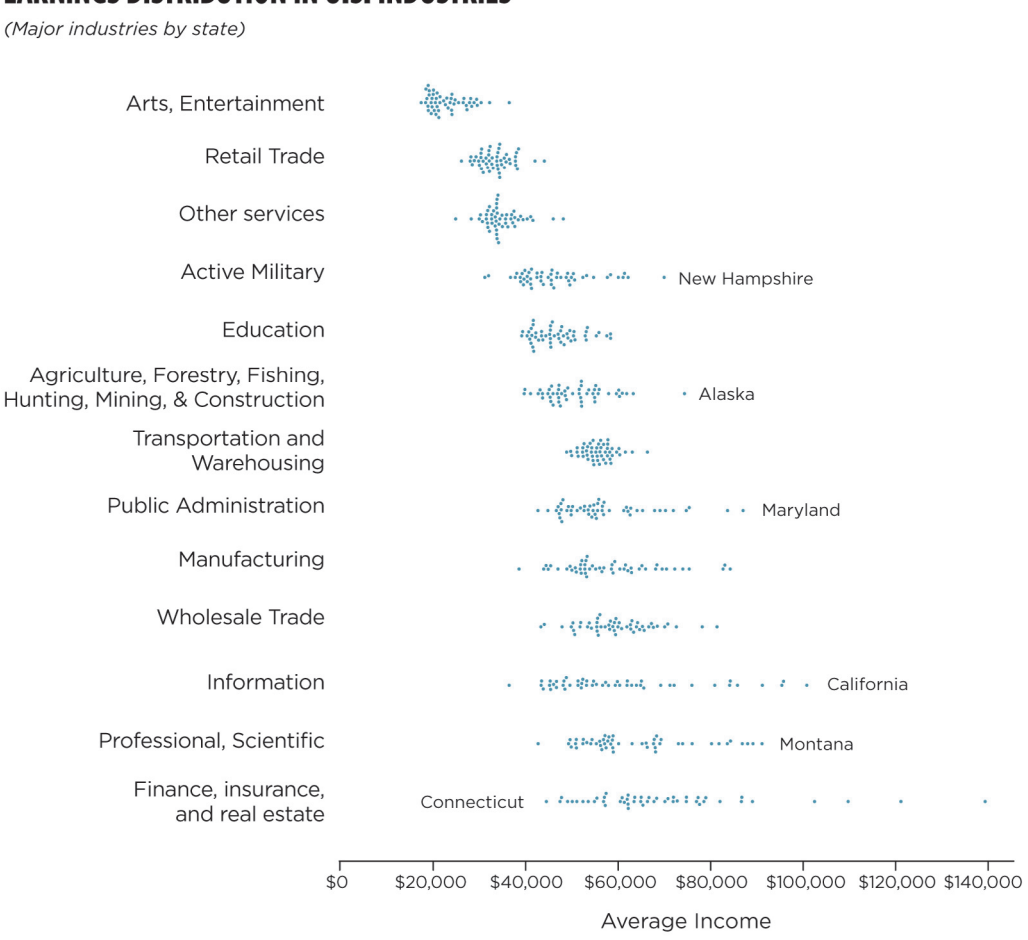


Jittering doesn't work in every case, of course. We are limited by how many points we can plot. With too many data points, jittering would require so much movement that it would modify the underlying distribution. Plotting *everyone's* earnings in each industry, for example, creates a mass of dots that requires so much jittering that it modifies the presentation of the data by moving the points too far away from their true position. But showing average earnings in each of the fifty states across the thirteen industries is not as overwhelming, and you can see where the bulk of the distribution lies in each. Of course, if you're interested in just showing that there are a *lot* of points in your dataset and you can maintain the overall shape of the data, plotting many points may help make your exact argument.

To create what is called a *beeswarm plot*—because the clustering of the data points resembles a swarm of bees—we jitter the values so that they don't overlap and each point is visible. As with other charts in this chapter (and coming up in Chapter 8), there are different calculations

we can use to arrange the dots—for example, arranging the points in increasing order or placing them in a square or hexagonal grid. Here, similar to the ridgeline plot, each industry shares the same horizontal axis so that we can easily compare across the different sectors.

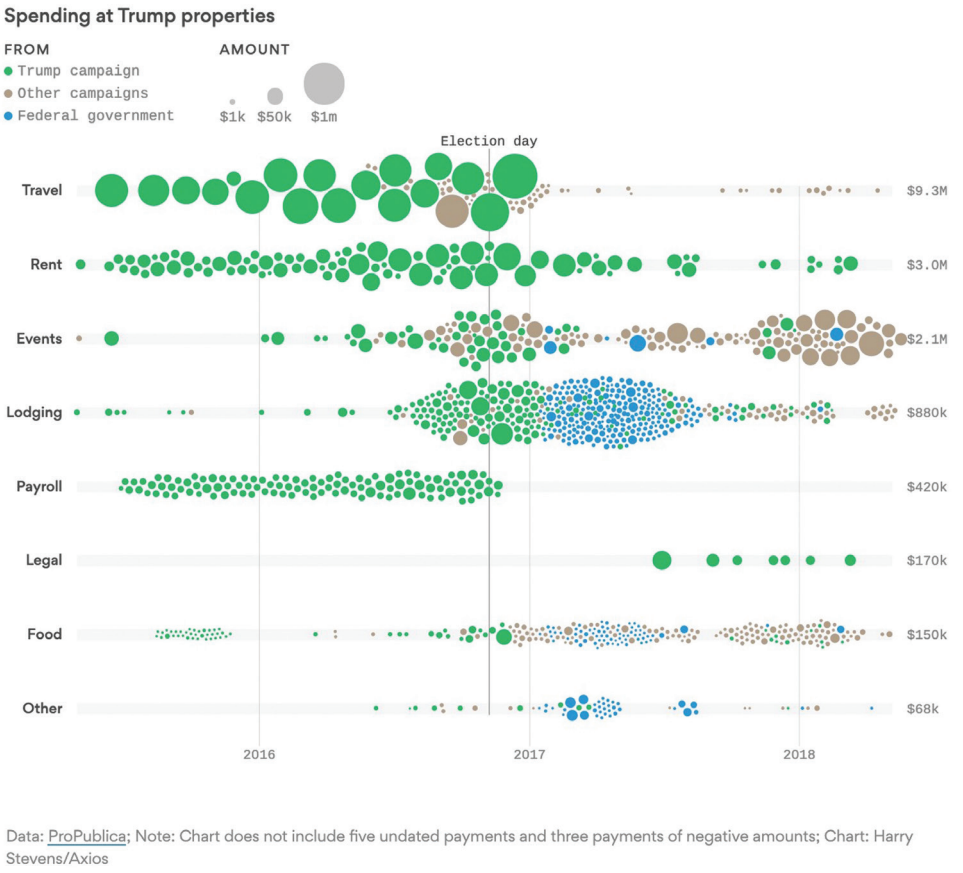
Notice how I have added some simple annotation and labeling to some of the outlier values. These clearly stand out in the graph, and a curious reader will wonder what's going



A beeswarm plot "jitters" all of the points in the data set so they don't overlap and each point is visible.

on with those points. Are they errors? If not, what is that state, and why are earnings so high relative to the rest of the country? I haven't labeled *every* outlier point, but enough to assure the reader I've thought about those values.

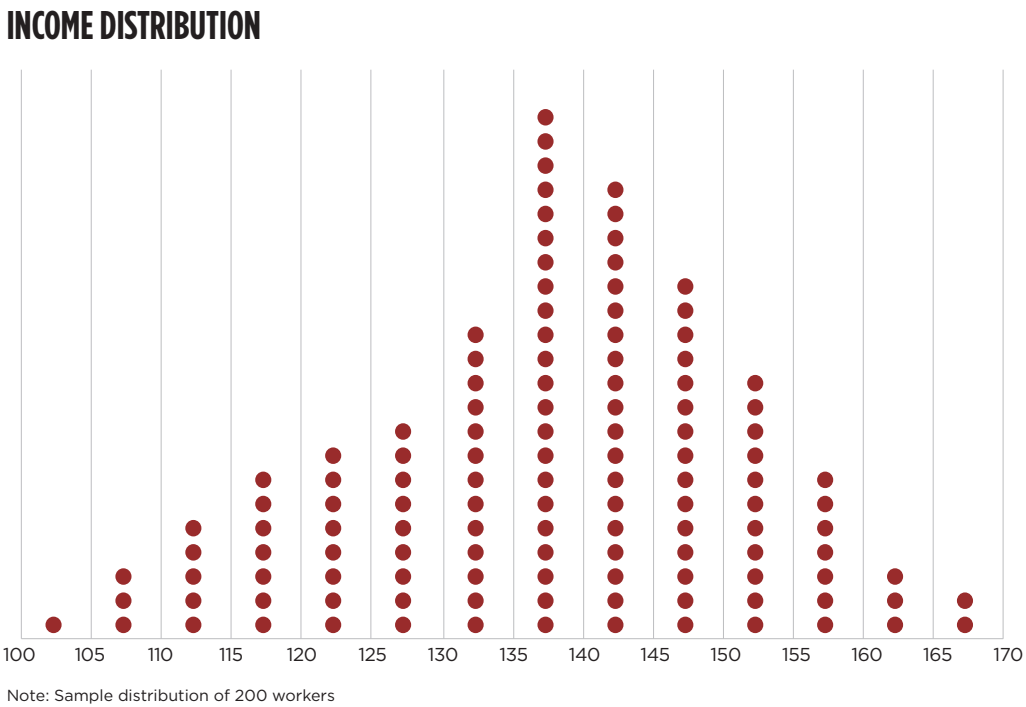
Beeswarm plots can also show changes over time. This beeswarm plot from Axios—really eight beeswarm plots aligned together—shows spending at properties owned by The Trump Organization before and after the 2016 election. The combination of color (spending origin), size (amount of spending), and the density of points (the time dimension) makes this an effective visualization to show the patterns around Election Day.



This beeswarm plot from Axios shows spending at properties owned by The Trump Organization before and after the 2016 election.

WILKINSON DOT PLOTS AND WHEAT PLOTS

The wheat plot, developed and named by Stephen Few, is a richer version of what is called a dot histogram or a Wilkinson Dot Plot (which is named after Leland Wilkinson author of the seminal data visualization book *The Grammar of Graphics*, though Wilkinson himself actually referred to these charts as *histodot plots*, a name that clearly did not stick). A Wilkinson Dot Plot is like a regular histogram except that instead of showing a single bar encoding all of the observations, data points are stacked on top of each other within their relative bins—something like combining a histogram and unit chart. With this approach, the points do not show their actual *value* (measured along the horizontal axis) because they are stacked in a single column. In other words, each dot represents an observation in each bin, not its actual value.

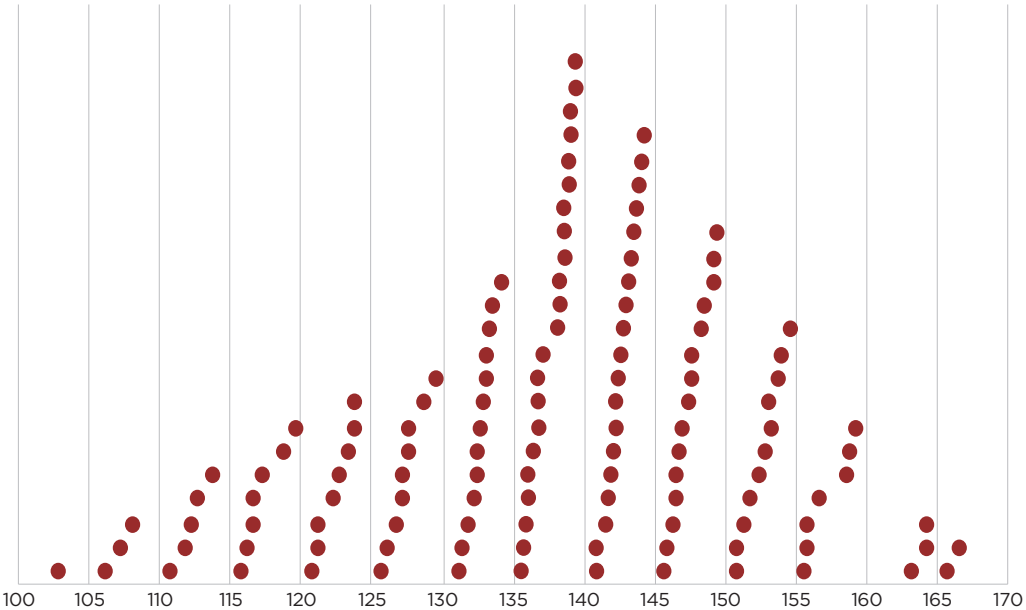


A dot histogram plots individual points within the bins of a histogram. Of course, it can only handle so many individual points.

The Wilkinson Dot Plot can be modified to create a wheat plot in which we show the *actual* data values, stacking them within their separate bins. The actual data values are plotted along the horizontal axis—still grouped into their bins—and stacked vertically to show the total number of observations. Stephen Few writes that, "The curved alignment of the dots is meaningful, for it graphically displays the distribution of values within each interval, based on their positions. Although this looks odd at first glance, it takes only a minute to understand and learn how to read." As with some of the previous distribution graphs, one of the limiting features of the wheat plot is that too many data values may lie on top of each other.

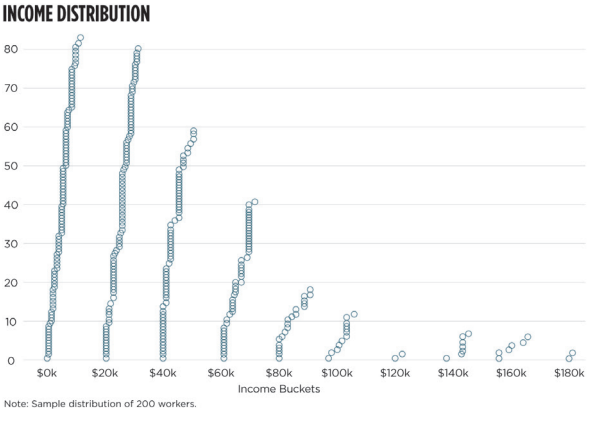
The wheat plot on the next page shows the distribution of earnings for a single industry for about two hundred workers. The histogram on the right is shown as a comparison—you can still see the relative share of observations in each part of the distribution, but not the actual data. There is an obvious tradeoff. On the one hand, the wheat plot shows more detail for the reader to explore and the graph can look more interesting and engaging. On the other hand, the histogram is likely more easily and immediately understandable to readers.

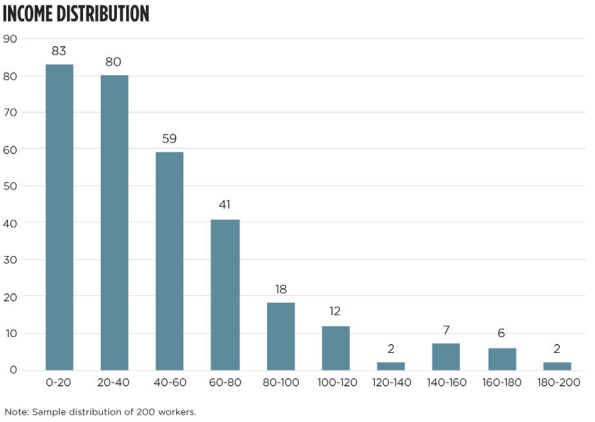
INCOME DISTRIBUTION



Note: Sample distribution of 200 workers.

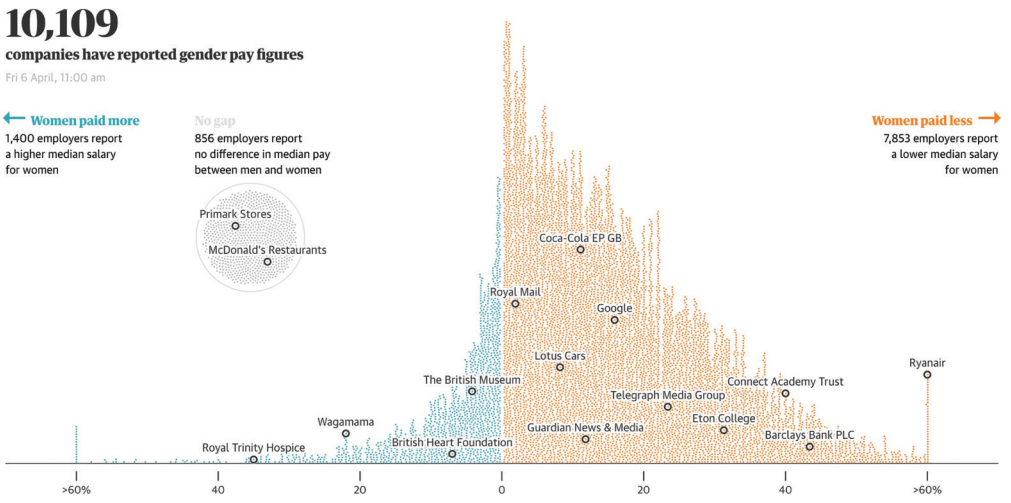
The wheat plot, designed by Stephen Few, adjusts the dot histogram by showing the exact values, still within each bin.





The tradeoff between wheat plots and simple histograms: the wheat plot has more detail but may be harder for people to understand.

We can see the difference between say, a wheat plot and a ridgeline plot in this visualization from the *Guardian*. Leading off the same article that has the ridgeline plot on page 201, this graph includes a dot for every company in their data set. It doesn't quite have the "lean" that a true wheat plot might have (because there are so many points), but you get a good sense of the overall distribution and the greater number of firms towards the right side of the graph. Plotting each individual company also lets the chart creators add labels to highlight



This wheat plot from the *Guardian* includes a dot for every company in their data set.

specific companies, which we couldn't do in a standard histogram or ridgeline plot. It is, however, important to realize that the selected labeled points are arbitrarily chosen along the vertical axis, which is simply stacking the points and not tied to any pay gap data.

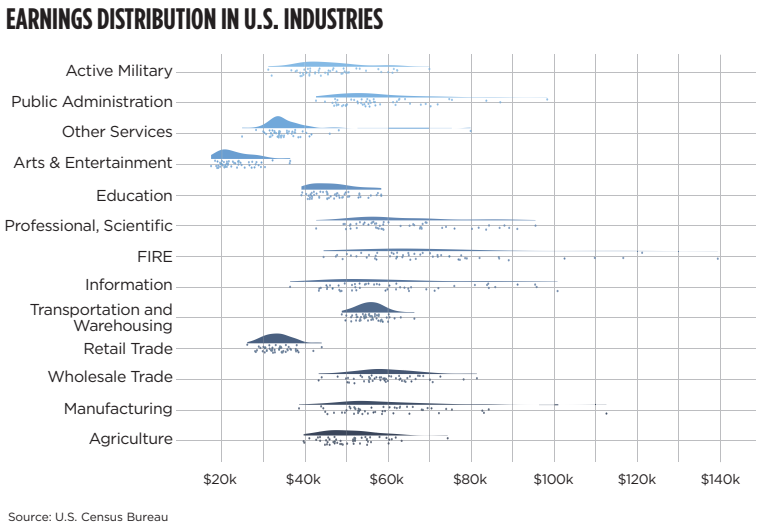
RAINCLOUD PLOT

Sometimes it's useful to show *both* the distribution density of your data and the actual data points. The raincloud plot, perhaps first named by neuroscientist Micah Allan, shows the distribution (think violin chart) with the actual data plotted below. In this arrangement, it looks like a cloud raining data.

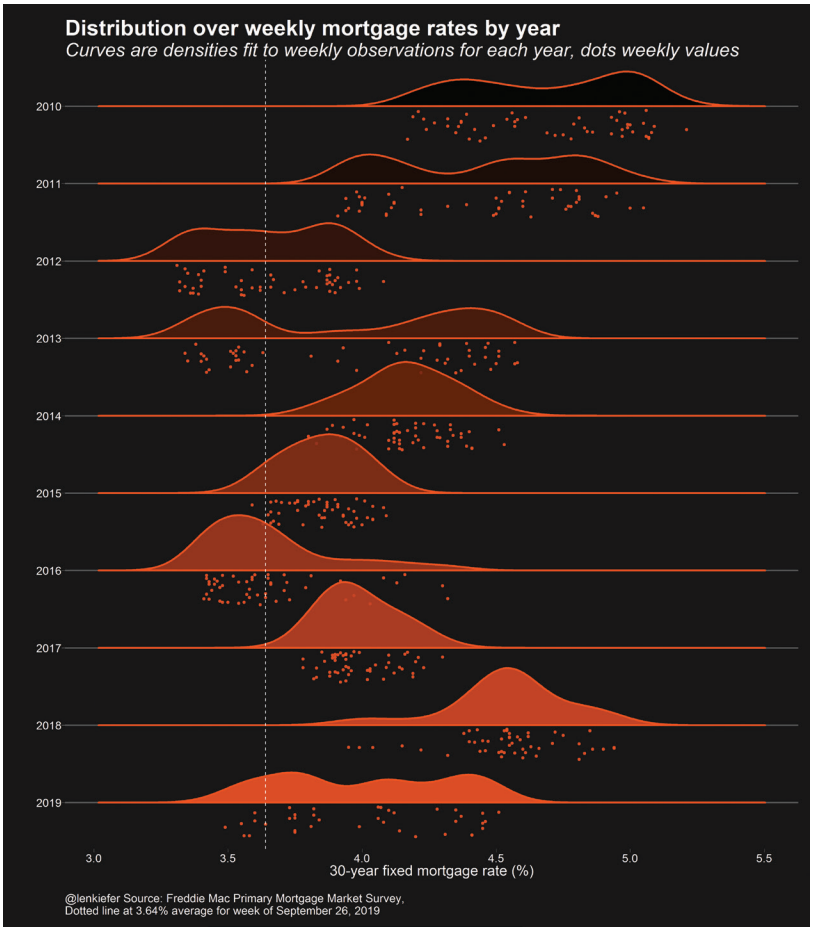
Raincloud plots show us a summary of the data and all the individual data points, so we can spot outliers and patterns. Again, the tradeoff is that this might require more work on the part of the reader to understand how to read the graph.

This raincloud plot shows the distribution of earnings across the fifty states with data values plotted below.

While the raincloud plot may seem like an esoteric chart—and, to be honest, right now it is—there are certainly scenarios and data for which this chart would be a useful choice.



The raincloud plot shows the summary histogram with the actual data plotted below.

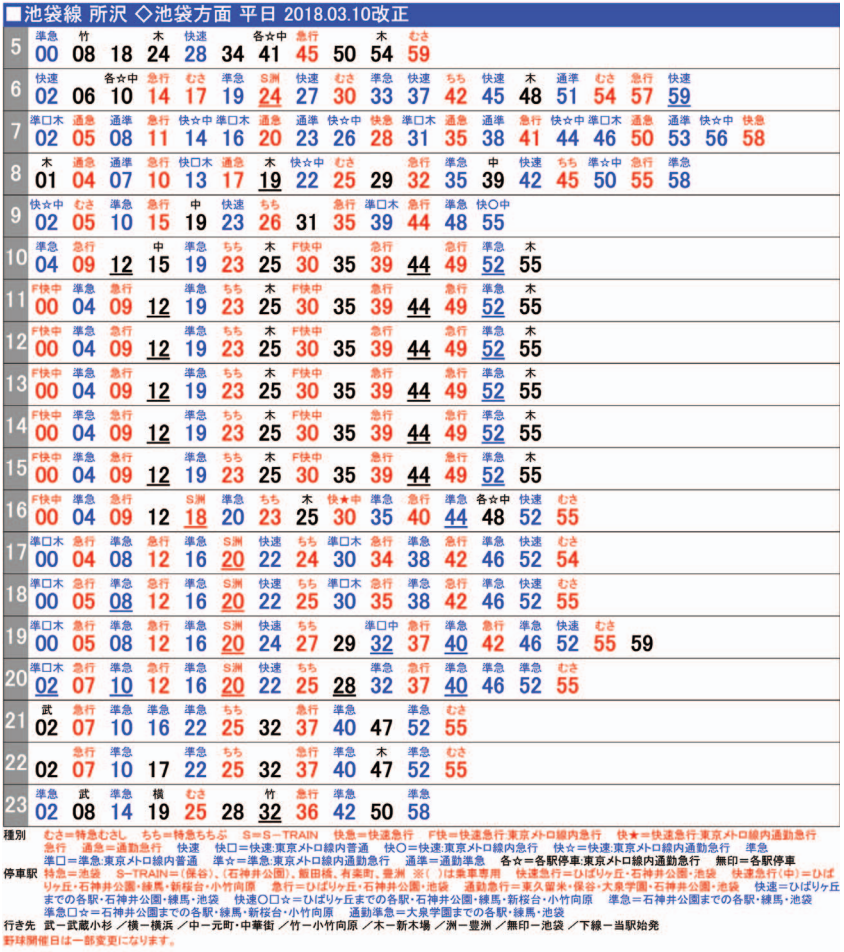


An example of a raincloud plot from Len Kiefer that shows the distribution of weekly mortgage rates for different years. The visualization gives you both an overall summary view of the data and shows the specific data points.

This raincloud plot from Len Kiefer, the Deputy Chief Economist at Freddie Mac, shows the distribution of mortgage rates from 2010 to 2019, with weekly observations shown in the dots just below. This view gives us both an overall perspective of the data and the detailed values in the points below.

STEM-AND-LEAF PLOT

The stem-and-leaf plot is a table that shows the place values of each data value. Values are typically shown listed down the "stem" column with the first digit or digits. The rest of the table is reserved for the "leaf" that shows the last digit (or digits).



■池袋線 所沢 ◇池袋方面 平日 2018.03.10改正

5	準急 00	竹 08		18	木 24	快速 28		34	各☆中 41	急行 45	50	木 54	むさ 59											
6	快速 02		06	各☆中 10	急行 14	むさ 17	準急 19	S洲 24	快速 27	むさ 30	準急 33	快速 37	ちち 42	快速 45	木 48	通準 51	むさ 54	急行 57	快速 59					
7	準口木 02	通急 05	通準 08	急行 11	快☆中 14	準口木 16	通急 20	通準 23	快☆中 26	快急 28	準口木 31	通急 35	通準 38	急行 41	快☆中 44	準口木 46	通急 50	通準 53	快☆中 56	快急 58				
8	木 01	通急 04	通準 07	急行 10	快口木 13	通急 17	木 19	快☆中 22	むさ 25	29	急行 32	準急 35	中 39	快速 42	ちち 45	準☆中 50	急行 55	準急 58						
9	快☆中 02	むさ 05	準急 10	急行 15	中 19	快速 23	ちち 26	31	急行 35	準口木 39	急行 44	準急 48	快○中 55											
10	準急 04	急行 09	12	中 15	準急 19	ちち 23	木 25	F快中 30	35	急行 39	44	急行 49	準急 52	木 55										
11	F快中 00	準急 04	急行 09	12	準急 19	ちち 23	木 25	F快中 30	35	急行 39	44	急行 49	準急 52	木 55										
12	F快中 00	準急 04	急行 09	12	準急 19	ちち 23	木 25	F快中 30	35	急行 39	44	急行 49	準急 52	木 55										
13	F快中 00	準急 04	急行 09	12	準急 19	ちち 23	木 25	F快中 30	35	急行 39	44	急行 49	準急 52	木 55										
14	F快中 00	準急 04	急行 09	12	準急 19	ちち 23	木 25	F快中 30	35	急行 39	44	急行 49	準急 52	木 55										
15	F快中 00	準急 04	急行 09	12	準急 19	ちち 23	木 25	F快中 30	35	急行 39	44	急行 49	準急 52	木 55										
16	F快中 00	準急 04	急行 09	12	S洲 18	準急 20	ちち 23	木 25	快★中 30	準急 35	急行 40	準急 44	各☆中 48	快速 52	むさ 55									
17	準口木 00	急行 04	準急 08	急行 12	準急 16	S洲 20	快速 22	ちち 24	準口木 30	急行 34	準急 38	急行 42	準急 46	快速 52	むさ 54									
18	準口木 00	急行 05	準急 08	急行 12	準急 16	S洲 20	快速 22	ちち 25	準口木 30	急行 35	準急 38	急行 42	準急 46	快速 52	むさ 55									
19	準口木 00	急行 05	準急 08	急行 12	準急 16	S洲 20	快速 24	ちち 27	29	準口中 32	急行 37	準急 40	急行 42	準急 46	快速 52	むさ 55	59							
20	準口木 02	急行 07	準急 10	急行 12	準急 16	S洲 20	快速 22	ちち 25	28	準急 32	急行 37	準急 40	準急 46	準急 52	むさ 55									
21	武 02	急行 07	準急 10	準急 16	準急 22	ちち 25	32	急行 37	準急 40	47	準急 52	むさ 55												
22	02	急行 07	準急 10	17	準急 22	ちち 25	32	急行 37	準急 40	木 47	準急 52	むさ 55												
23	準急 02	武 08	準急 14	横 19	むさ 25	28	竹 32	急行 36	準急 42	50	準急 58													

種別　むさ＝特急むさし　ちち＝特急ちちぶ　S＝S－TRAIN　快急＝快速急行　F快＝快速急行:東京メトロ線内急行　快★＝快速急行:東京メトロ線内通勤急行
急行　通急＝通勤急行　快速　快口＝快速:東京メトロ線内普通　快○＝快速:東京メトロ線内急行　快☆＝快速:東京メトロ線内通勤急行　準急
準口＝準急:東京メトロ線内普通　準☆＝準急:東京メトロ線内通勤急行　通準＝通勤準急　各☆＝各駅停車:東京メトロ線内通勤急行　無印＝各駅停車

停車駅　特急＝池袋　S－TRAIN＝(保谷)、(石神井公園)、飯田線、有楽町、豊洲　※(　)は乗車専用　快速急行＝ひばりヶ丘・石神井公園・池袋　快速急行(中)＝ひばりヶ丘・石神井公園・練馬・新桜台・小竹向原　急行＝ひばりヶ丘・石神井公園・池袋　通勤急行＝東久留米・保谷・大泉学園・石神井公園・池袋　快速＝ひばりヶ丘までの各駅・石神井公園・練馬・池袋　快速○口☆＝ひばりヶ丘までの各駅・石神井公園・練馬・新桜台・小竹向原　準急＝石神井公園までの各駅・練馬・池袋
準急口☆＝石神井公園までの各駅・練馬・新桜台・小竹向原　通勤準急＝大泉学園までの各駅・練馬・池袋

行き先　武－武蔵小杉／横－横浜／中－元町・中華街／竹－小竹向原／木－新木場／洲－豊洲／無印－池袋／下線－当駅始発

野球開催日は一部変更になります。

Stem-and-leaf plots show the place values of each data value. They are sometimes used in transportation schedules, like this train schedule from the Tokorozawa station in Saitama, Japan.

As an example, take a simple dataset with just seven values: 4, 9, 12, 13, 18, 24, and 27. The data are arranged in downward-ascending order with the first digit on the left side and the second (tens) digit on the right. Obviously, for more detailed and complex data, the stem-and-leaf plot may not be a useful approach.

Stem-and-leaf plots are most useful as a reference tool, like a public transportation schedule, or to highlight basic distributions and outliers in a more limited set of data. The Japanese train schedule for the Tokorozawa station in Tokyo on the facing page shows the timing of train arrivals over the course of a day. The hour digit is shown in the far-left column, minutes are shown to the right. The first train begins running at 5:00 am, the next train leaves at 5:08 am, then 5:18 am, and so on. Because the stem-and-leaf plot is a table, it loses some of the advantages of a traditional data visualization, but the leaves illustrate some basic view of the distribution.

CONCLUSION

The collection of graphs in this chapter demonstrates how we can show the distribution of our data or uncertainty around specific values. Some of these charts show summary measures or specific values. We can aggregate the distribution into bins to visualize the distribution in a histogram. Or we might use specific percentiles to generate a box-and-whisker chart, for example, or show stock price variation in a candlestick chart.

With better data visualization tools and faster computers, we can show more data than ever before. Beeswarm charts, wheat plots, and raincloud plots include specific data points. While these visualization types are useful for presenting the full data set to our readers, they have their limits: We can only show so many data points before they begin to overlap and obscure one another.

The graphs in this chapter can introduce challenges to readers who are not familiar with statistical concepts and measures of dispersion. As always, the most important thing you can do when creating your graphs is to remember your audience. If you are a PhD economist presenting your work at your university lunchtime seminar, you don't need to explain the median or variance or even the 95-percent confidence interval. If you were to present the same results to a general audience, however, you would include definitions and annotation. This isn't to say you should avoid presenting these numbers or that you need to dumb things down, but rather that you may need to take time explaining concepts within the visual. The planning, testing, and conceptualizing of your visualization will pay off in the long run as you more effectively communicate your work with your audience.

GEOSPATIAL

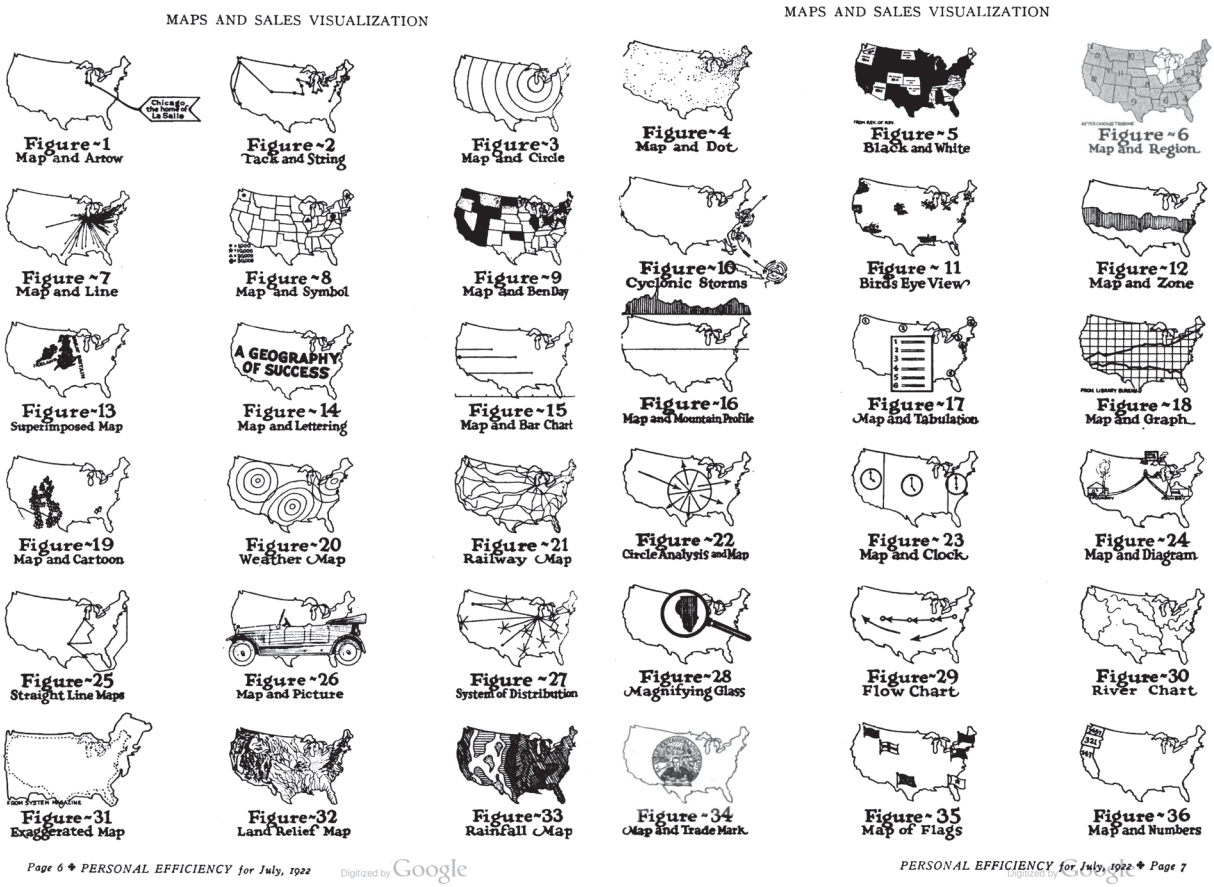
There is an obvious advantage to plotting geographic data on a map—people can find themselves in the data. They can literally *see* themselves in the data, a connection with the subject matter that other visualizations cannot muster. Plotting geographic data may mean adding color to geographic areas likes states or countries, or adding circles, squares, lines, or other shapes on top of a geographic map.

Data-driven maps are not new. In 1922, the "Maps and Sales Visualization" on the next page shows the reader thirty-six different ways to place data on a map. The author writes:

> The use of maps has to do with the visual representation of space. Therefore, in all map work we start with an outline . . . The fact that the Earth is a globe makes visualization of maps a difficult process; arbitrary methods have to be used to get the surface of the ball represented in a flat picture.

This chapter begins with some of the basic challenges of visualizing geographic data and then presents some alternatives to the basic map, the modern version of this 1922 visualization.

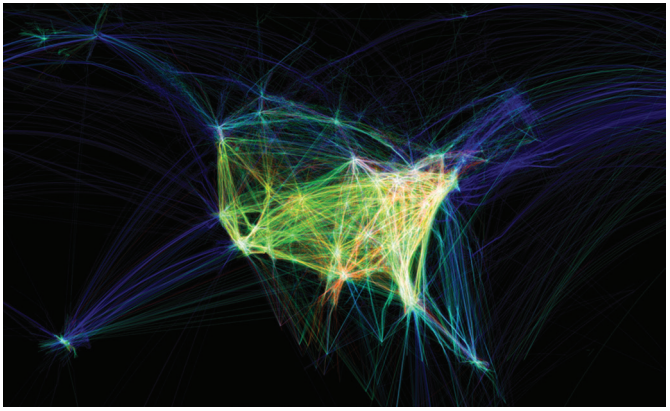
The maps in this chapter use a color palette often used by the *Washington Post*. Maps of the U.S. political system in this chapter use the basic red-blue color palette used by many newsrooms.



E.P. Hermann's (1922) "Maps and Sales Visualization" shows an early example of the multitude of ways to place data on a map.

THE CHALLENGES OF MAPS

Aaron Koblin's *Flight Patterns* is one of those maps that can entrance the reader. Koblin plots all flight paths in the skies above the United States in twenty-four hours. The static version of the map (there is an interactive version in which you can zoom into any area of the country) shows the entire country, major airports, and the activity of the skies above. It's not a visualization that ranks the biggest airports or tells you how to avoid delays, but it quickly shows the traffic patterns in American skies.



Aaron Koblin's *Flight Patterns* project, which consists of static and interactive maps, shows all flight paths in the skies above the United States over twenty-four hours.

There are some distinct challenges when creating maps. The biggest is that the size of a geographic area may not correspond to the importance of the data value. Russia is more than 6.6 million square miles, almost twice the size of Canada, and so it takes up a lot of space on a map. At 270,000 square miles, Texas is roughly the size of California and Colorado put together, but it's actually less than half the size of Alaska (665,000 square miles), which you might not know because most maps of the United States tend to distort it and arbitrarily position it out to sea, south of California. The point is that the data values for Russia, Texas, and Alaska may not correspond to their importance in the data, and the map can distort our perception of the important values being visualized.

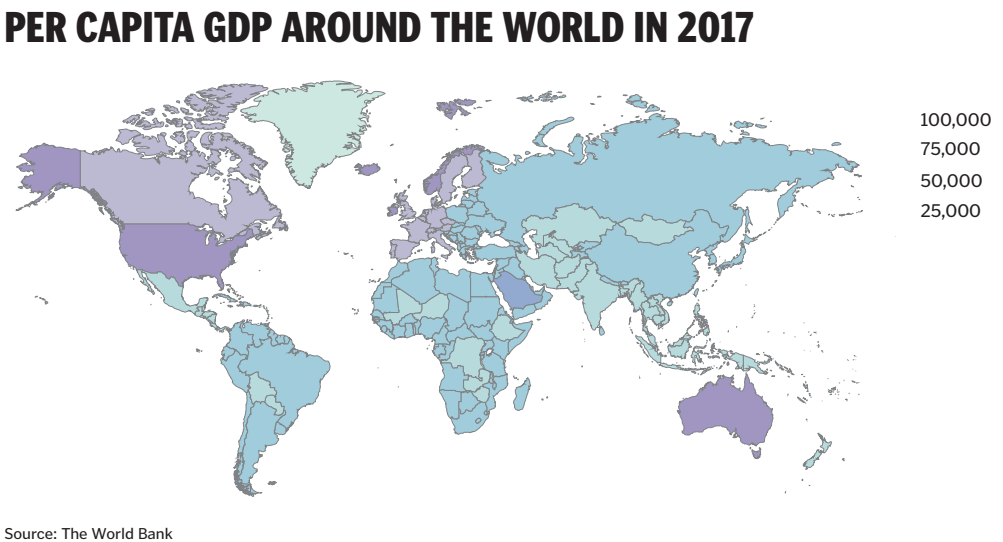
I find that many people should approach their desire to create a map with more skepticism and critical thought. Is a map truly the best way to present geographic data? Or are you just showing where people live? Does it show the relationships we want to explore or are we simply relying on the fact that we have geographic identifiers? This chapter will explore the perceptual issues of maps and why they are not always the best medium through which to demonstrate our points. This isn't to say we should *never* make a map—in many cases we need to make a map to better understand our data—but, especially with maps, we should always take a step back and consider whether it is the right visualization choice.

The message is simple: there are many ways to present geographic data, and there are lots of objects, shapes, and colors you can add to maps. Which kind of map you use to visualize your data will depend on two questions: How important are the geographic patterns? And how important is it for your reader to see a familiar map?

CHOROPLETH MAP

Perhaps the most familiar data map has perhaps the most unfamiliar name: a choropleth map. Choropleth maps use colors, shades, or patterns on geographic units to show proportionate quantities and magnitudes. You likely already know how to read this choropleth map of per capita GDP around the world—it's a simple and recognizable shape that lets you quickly and easily find countries (and, by extension, yourself) in the visualization.

This color palette is also easy to understand—smaller numbers correspond to lighter colors and larger numbers to darker colors (and what is sometimes called a "color ramp"). More often than I care to count, map creators will use an incorrect color palette. For example, for this map, someone might use a diverging color palette in which colors progress outward from a central midpoint. Unless we are comparing per capita GDP above and below some midpoint number, such as the average GDP, the diverging color palette is a bad choice. Instead, we should follow what has become a simple standard of lighter colors to darker colors. We discuss color palettes in more detail in Chapter 12.



The choropleth map is perhaps the most familiar data map. Colors correspond to data values and are assigned to the different geographic areas on the map.

It may be bit difficult to find smaller countries or countries that you don't already know, but the overall shape is familiar, well-known, and easy to understand. Did you know Luxembourg had the highest per capita GDP in 2017 at more than $104,000 per person (for reference, per capita GDP in the United States was $59,500)? It's a country of only one thousand square miles (for reference, France is almost 250,000 square miles) and difficult to find on this map, but it has the highest per capita income.

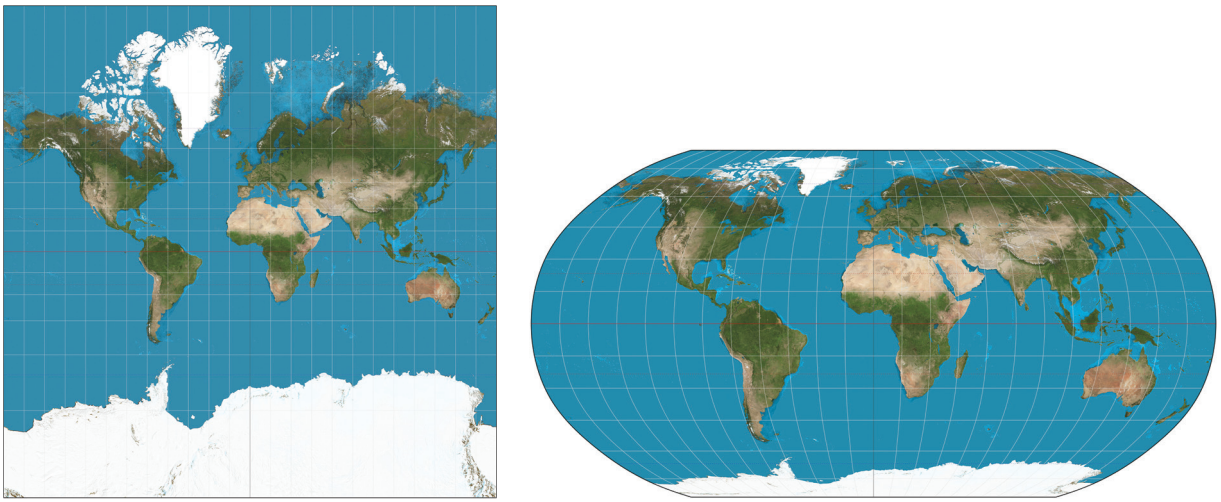
Maps like these introduce a geographic distortion—the size of the geographic area may not correspond to the importance of the data value. Even with this distortion, however, maps are an easy and familiar way to present geographic data to our readers.

There are a variety of alternative map types that correct this geographic distortion. Cartograms, for example, resize the geographic units according to their data values (see page 233) and a tile grid map uses a series of equal-sized squares (see page 238), not to mention the other chart types we can use to plot geographic data, such as a heat map (see page 112). The tradeoff with any of these alternative approaches is that the map is no longer as familiar to the reader as the standard map. But, as Kenneth Field, author of the data visualization and cartography book *Cartography*, once noted, "None of these maps are right and none of these maps are wrong. They are all just a different representation of the truth."

CHOOSING THE PROJECTION

One challenge with mapping data is the map *projection* the creator chooses. The world is a globe, but maps are flat. A mapmaker must choose a map projection to transform the world's sphere into a two-dimensional plane. All maps distort the surface of the planet to some degree and there is considerable debate about which projection does the best job depicting the earth in two dimensions.

The one you are likely most familiar with is the Mercator projection. This is the map used in early versions of Google Maps and is the default in many data visualization tools like Tableau and PowerBI. Developed in 1569 by Flemish geographer and cartographer Gerardus Mercator, it became the standard map projection for nautical purposes. A sailor could draw a straight line between two points and measure the angle between that line (called a *rhumb line*) and the meridian (the vertical line that runs between the north and south poles) to find their bearing. While it may be useful for sailing the seas, the Mercator projection distorts the size of objects as the latitude increases from the equator to the north and south poles. Thus, countries closer to the poles—like Greenland and Antarctica—appear much larger than they actually are. In the Mercator map shown on the next page (on the left), Greenland looks to



Map projections can influence our perception of the map. The Mercator projection on the left, for example, looks considerably different than the Robinson projection on the right.
Source: Wikimedia user Strebe.

be about the same size as South America, when it is in fact about one-eighth the actual size. In the Robinson projection on the right, country areas are closer to their true sizes.

There are three major categories of map projections:

CONIC

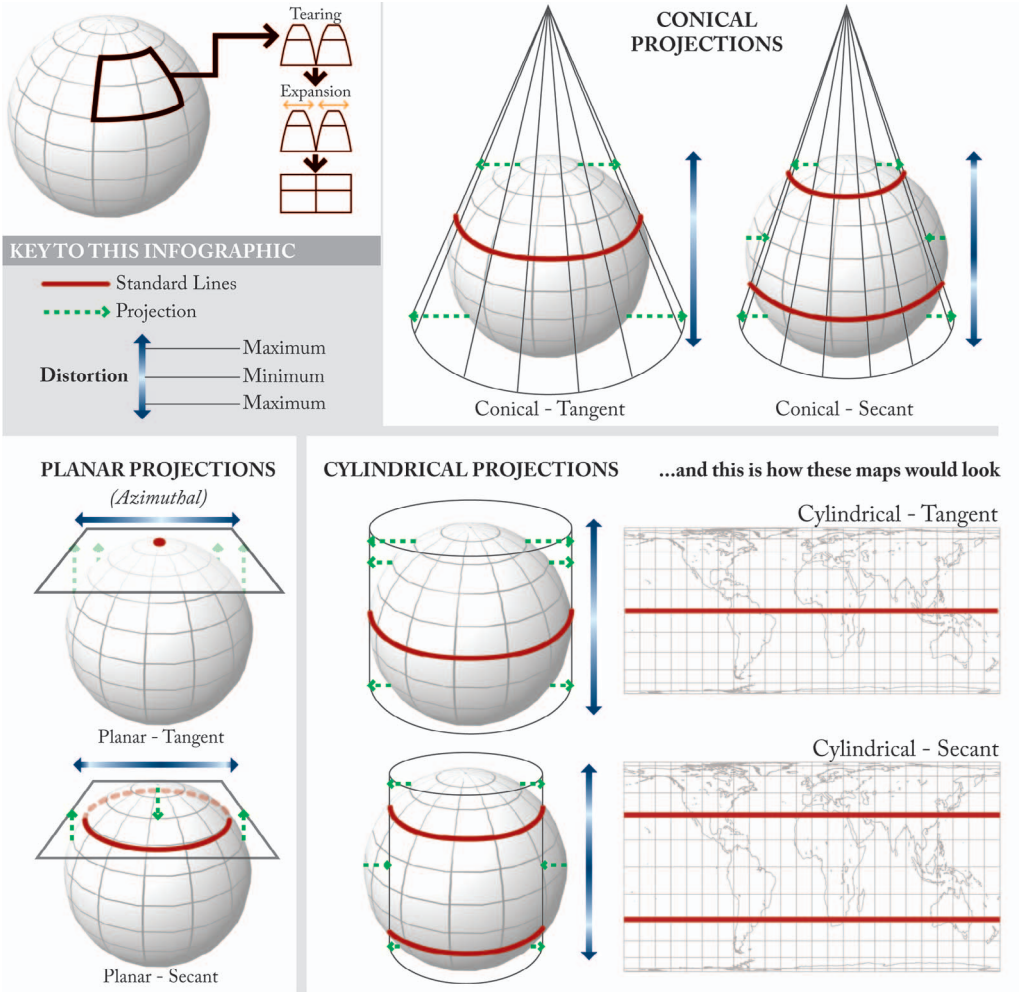
- Conic maps are as though a cone was placed over the Earth and unwrapped. Conic projections are best suited for mapping long east-west geographies, such as the United States and Russia, because the distortion is constant along common parallels. The Albers Equal Area Conic and the Lambert Conformal Conic are two of the more well-known projections.

CYLINDRICAL

- Cylindrical maps work like conic maps but use cylinders instead of cones. Like the Mercator projection, cylindrical maps inflate geographic areas farther from the center.

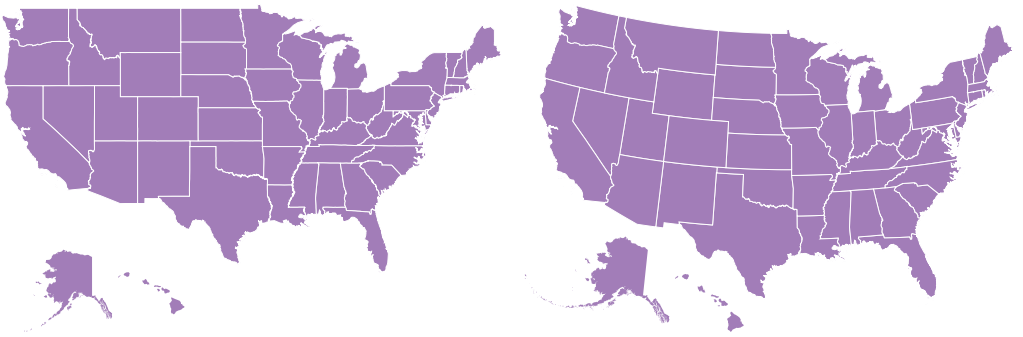
PLANAR (AZIMUTHAL)

- With this approach, the planet is projected on a flat surface. All points are at the same proportional distance from the center point, such as the north pole, but the distortion gets larger as you move further from that center point.



To some degree, all maps distort the surface of the planet. Alberto Cairo's diagram from *The Truthful Art* shows a selection of different types of map projections.

There isn't necessarily a right or wrong map projection, though I'm sure some cartographers would disagree! But each has tradeoffs, and serious mapmakers dig deep into the different properties of these projections and weigh the different options. Many people in the data visualization field shy away from the Mercator projection because of its obvious drawbacks, though it can work well for small areas. For the United States, you can tell a map is using the Mercator projection because the top left border of the country is a straight line. The Albers projection, by comparison, preserves the east-west perspective, which you can see by the slight curvature in the northern border.



Notice how the northern border of the United States looks straight in the Mercator map on the left but is more curved in the Albers projection on the right.

CHOOSING THE BINS

As we start to add data to a choropleth map, our first consideration is the choice of intervals (or "bins") that will shade the geographic units. Placing data into discrete categories is, at its core, an aggregation problem. By combining several states or countries into a single bin, we don't know how different those units are from one another.

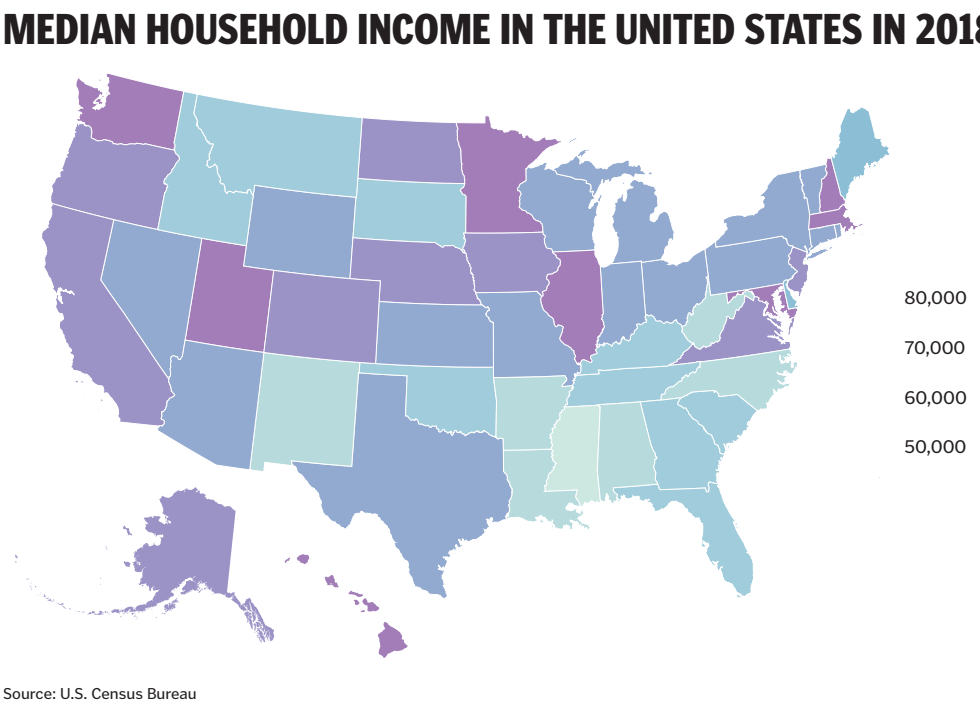
The map of the United States on the facing page, for example, shows median household income in 2018 in each state. How the states are placed into groups (the "bins"), the map shading, and ultimately our reader's perception of the data depends on our choices. Massachusetts ($86,345) and Maryland ($86,223) had the highest median household incomes in 2018 and fall into the highest bin with the darkest color; New Mexico ($48,283) and

Mississippi ($42,781) are at the other end of the distribution and are shown with the lightest shades.

There are four primary binning methods for creating maps.

NO BINS

This is essentially a continuous color palette (or "ramp") in which each data value receives its own unique color tone. On the one hand, this is easy because we don't need to think too much when we create the map—the colors ramp up from the lightest color for the lowest value to the darkest color for the highest value. On the other hand, the resulting color gradient may generate spatial patterns masked by subtle changes in color. In this example, it's hard to distinguish the differences between Iowa ($68,718), Nebraska ($67,515), and Wyoming ($62,539).

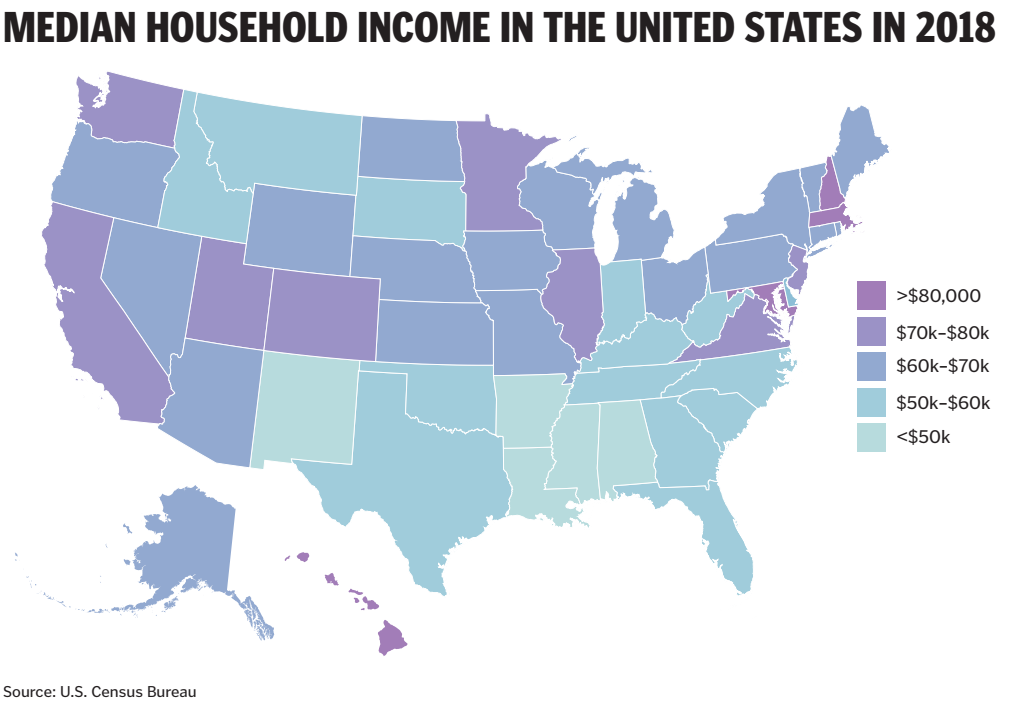


The continuous color palette (or "ramp") seamlessly goes from lighter colors (smaller values) to darker colors (larger values).

EQUAL INTERVAL BINS

In maps with a discrete number of bins, the default approach is to typically divide the data range into an equal number of groups. For example, in a map with four bins and a data range from 1 to 100, we end up with four equal groups (1–25, 26–50, 51–75, and 76–100).

This approach more clearly distinguishes geographic units (such as states) than the continuous (no bins) option, but it can mask the magnitudes of those changes by putting states in the same or different bins. In cases where the distributions are highly skewed, this approach may unevenly distribute the geographic units across the bins. In this map, the bins are split into equal units of $10,000, which results in five states in the bottom category, ten in the next, seventeen in the middle, fourteen in the next, and five in the top category.

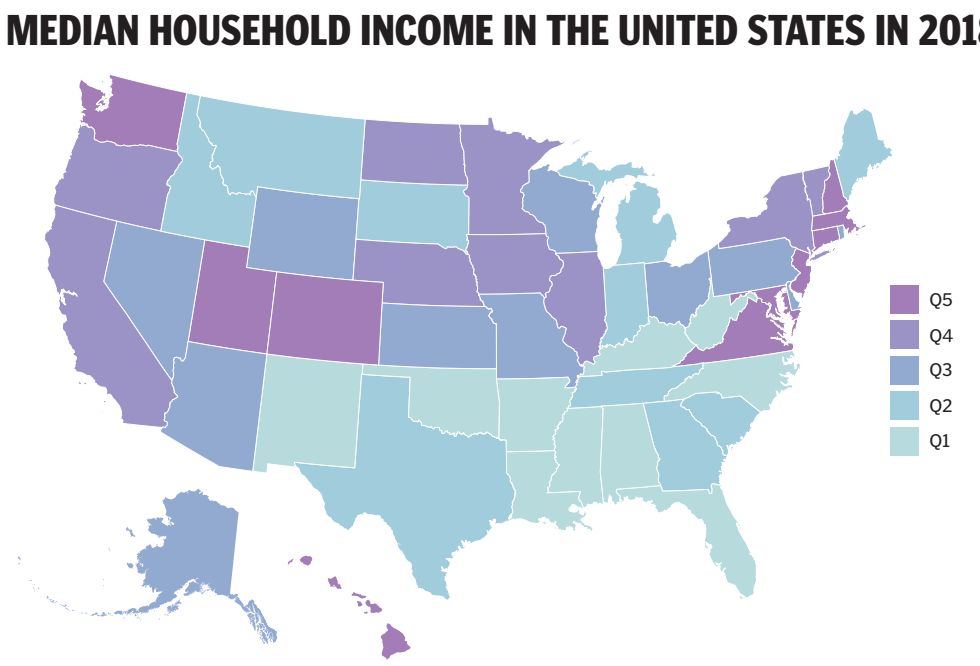


Dividing the data into equal intervals—such as $10,000 gaps—is one way to add color to a data map.

DATA DISTRIBUTION BINS

We could also cut the data into different bins. For example, instead of having a bin at equal intervals, we could arrange the bins to hold the same *number of observations*, such as quartiles (four groups), quintiles (five groups), or deciles (ten groups). Or we could use other measures to collapse the data into groups, such as the variance or standard deviation.

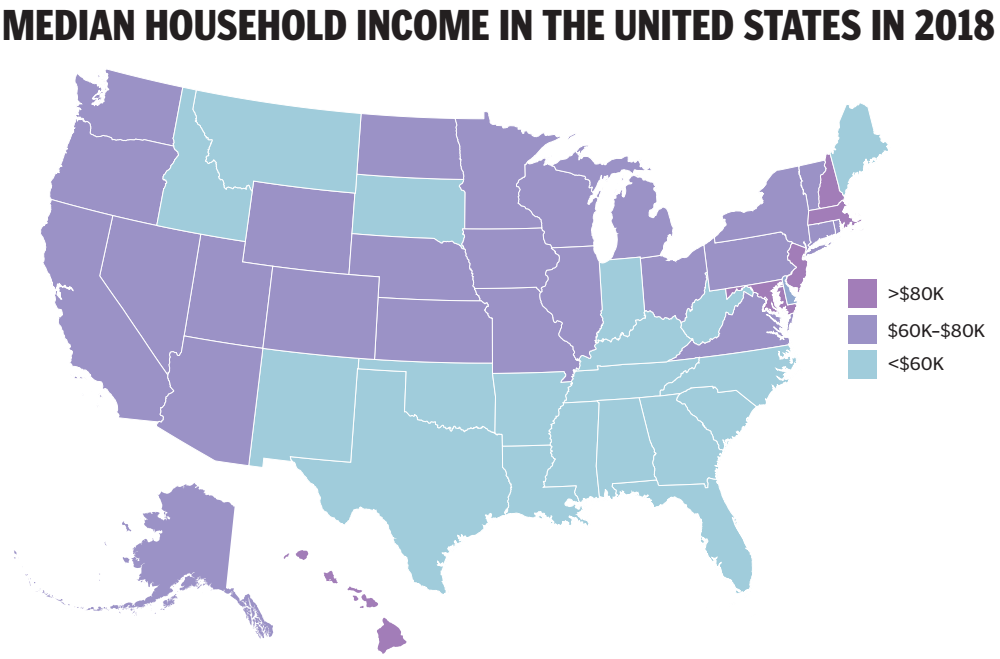
The data distribution approach clearly shows differences between the geographic units, but the created cutoffs may not be numerically meaningful. In this map that divides the country into five equal groups (quintiles), Connecticut is just barely in the top group with a median household income of $72,812, while Minnesota is placed in the next lower bin, even though it has a very close income estimate of $71,817.



Another way to divide the geographic units is to divide the data into an equal number of observations, such as quartiles (four groups) or quintiles (five groups).

ARBITRARY BINS

In this approach, the map creator chooses the bin cutoffs based on round numbers, natural breaks, or some other arbitrary criterion. This method lets us avoid some of the odd breaks that might occur, such as the Connecticut-Minnesota example above, but it can also be misleading. A method in which the selected bins are based on larger groups or round numbers—even without looking at the data—might look like this:



Source: U.S. Census Bureau

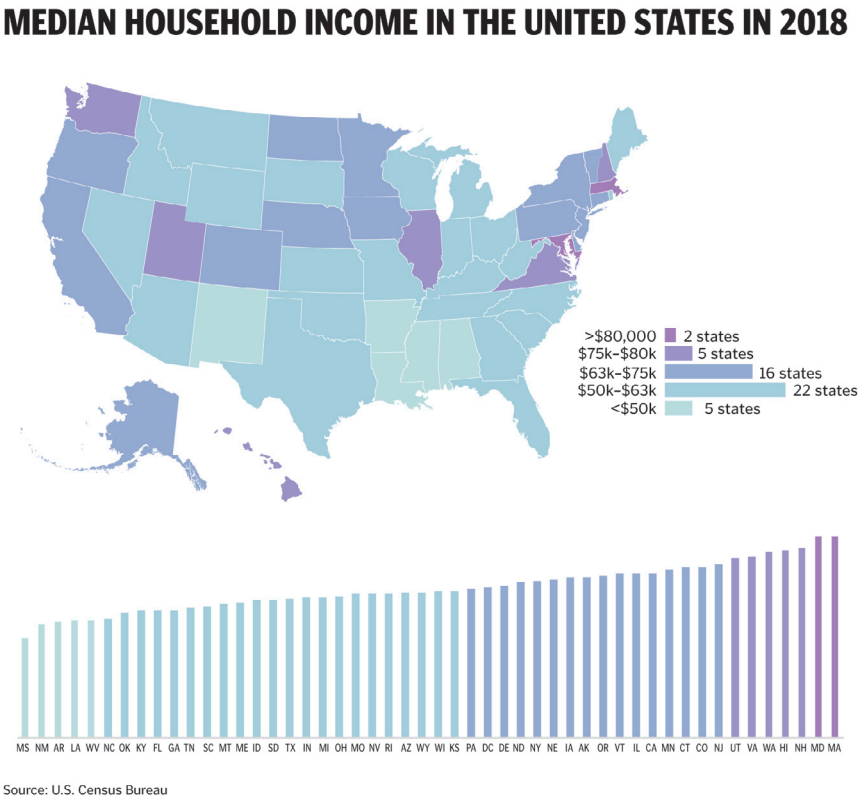
Depending on the goals of your data map, you can also divide the data into an arbitrary number of bins.

ALTERNATIVE OPTIONS

This isn't to say any of these maps are right or wrong, but they highlight the importance of binning decisions in choropleth maps. To make the best binning decision, we can draw on Mark Monmonier's 2018 book, *How to Lie with Maps*. Instead of using equally sized bins arbitrarily or letting the software tool decide which breaks to use, try considering—and showing—the actual distribution. If I add a column chart to the arbitrary bin method map,

the bin breaks and the distinct differences between the values is apparent. Adding another graph takes up more space, but it gives readers a clear picture of the data.

Including the number of observations in each bin somewhere in the visualization can also reveal the distribution to your reader. In this version, the legend is converted to a small bar chart to show the number of observations in each bin.



One way to help readers better understand the data in a map is to pair it with another visualization type, like a bar chart. This visualization has a small bar chart embedded within the legend to make clear how many states are in each group. This is not a necessary component, but one that can help the reader understand the distribution of the data on the map.

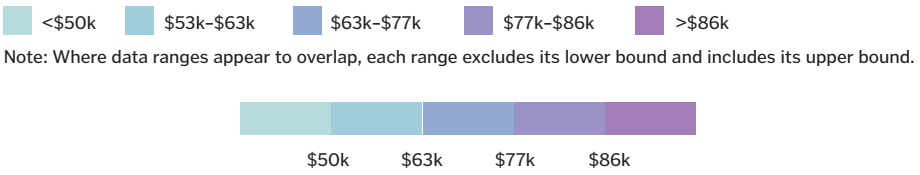
LABELING THE BINS

Another consideration when we create a map is how we *label* the bins. Let's use the visualization we just created of the map and bar chart of median household income.

In that map, the definition of the bins is arbitrary. The top bin is defined as ">$86,000" but because the maximum value in the next category $81,346 (New Hampshire), the top bin label could just as easily be ">$85,000," ">$82,000," or even ">$81,346."

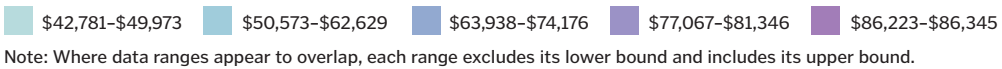
The legend for this map can be defined several ways.

1. Instead of round numbers, we could use the actual income amounts, which still leaves us with arbitrary bins. In this case, it isn't clear whether $86,000 is in the fourth or fifth group. There are a few ways to do this, for instance with separate boxes or a single image with labels just below. (One possible solution is to explicitly note which bins are inclusive or exclusive of the upper and lower bounds.)



2. Another alternative is to define the bins on the actual data values.

   This has the advantage of clearly showing the data values—for example, we can see the gap between $49,973 and $50,573 in the first two bins. The disadvantage is that the legend is overly detailed and complex. As the reader, you might wonder why all this precision is necessary and, depending on the content, you might wonder what's going on in the gap between the maximum of one bin and the minimum of the next.



3. Alternatively, we could create a legend that includes these "gap" bins:

Although accurate and comprehensive, this legend feels busy and obscures the five original bins by adding four non-data bins that connect them.

There is no one-size-fits-all solution to this challenge, but here we have laid out the issues and tradeoffs of creating and labeling bins. Always consider the necessary level of precision (i.e., the number of decimal places), the overall number of bins, and the smoothness of the data (that is, if the data show no visible jumps, round-numbered bins may be fine).

SHOULD IT BE A MAP?

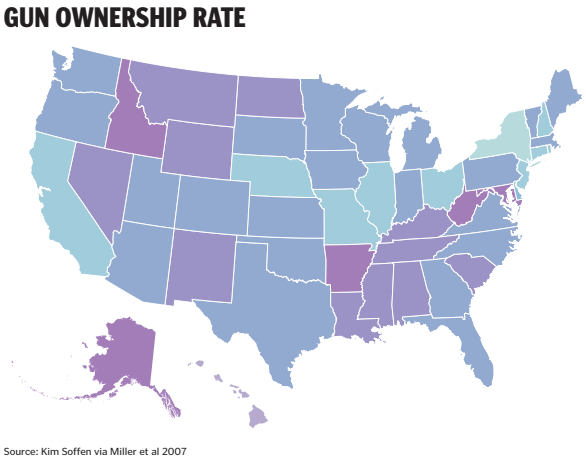
Before we start exploring other types of maps, let's first ask ourselves whether a map accomplishes our goals and best communicates our argument. Many maps are made simply because the creator has geographic data, not because the map is the best medium for that content.

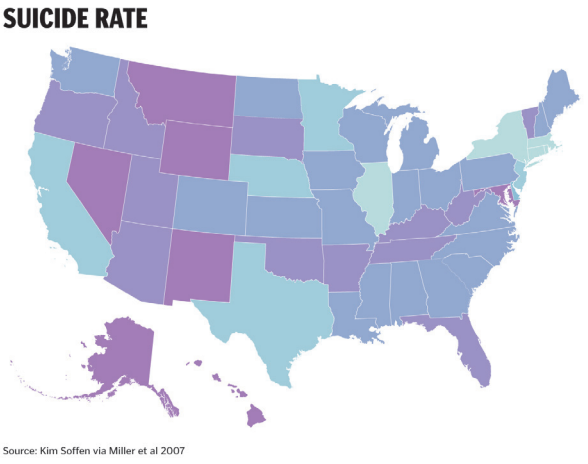
Take this simple example. A 2016 *Washington Post* story examined the relationship between rates of suicide and gun ownership in the United States. The story explains:

> One 2006 study found that from the 1980s to the 2000s, every 10 percent decline in gun ownership in a census region accompanied a 2.5 percent drop in suicide rates. There are numerous other studies that show similar results.
>
> This pattern becomes clear when looking state by state. The states that have higher rates of gun ownership, where people have more access to guns, also have higher rates of suicide. Suicides are twice as common in states with high gun ownership than those with low gun ownership, even after controlling for rates of mental illness and other factors, according to a 2007 study.

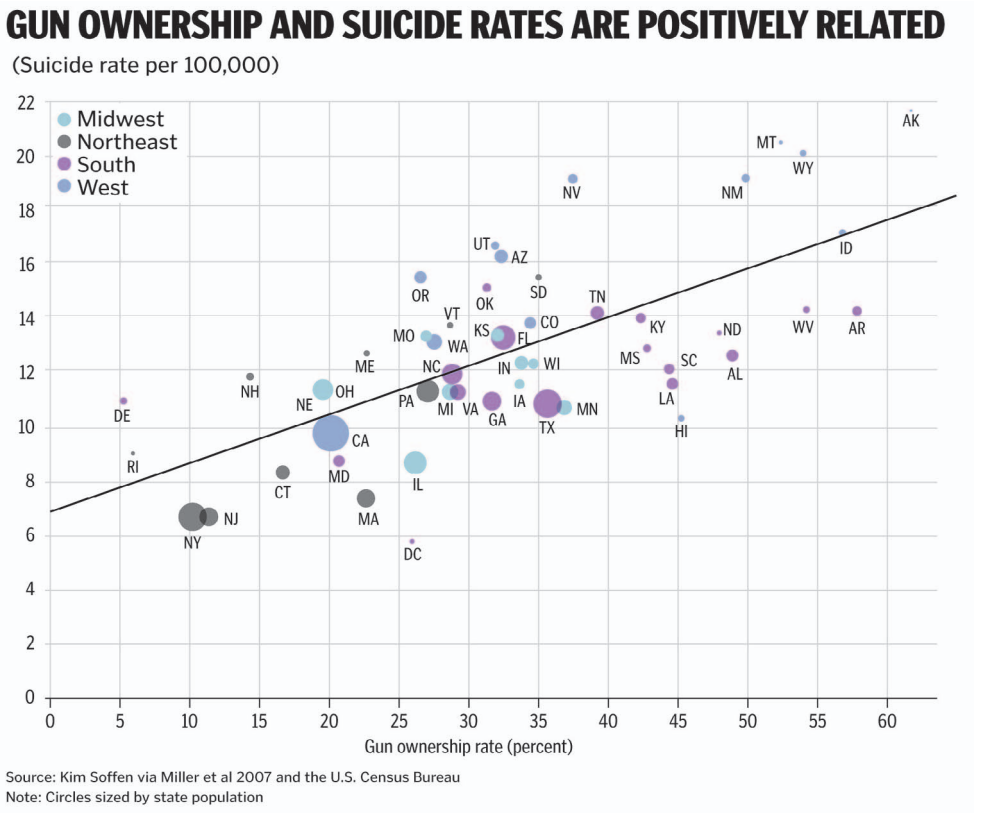
Below these two paragraphs were two maps, which I have recreated on the next page as choropleth maps. Do these maps help you see the positive relationship between the gun ownership rate and the suicide rate? Can you pick out states that have the highest suicide rates and the highest gun ownership rates? I can't. Instead, I found myself jumping back and forth between the two maps trying to identify individual states and regions.

Instead, what if we placed the same data in a bubble plot? We could put the suicide rate on the vertical axis and the gun ownership rate on the horizontal axis, and scale the size of the circles by population. Adding color to differentiate between areas of the country makes it clearer that states in the upper-right area of the graph tend to be western and southern states, where gun ownership is higher and, it turns out, gun-control laws are weaker.





Note: The original maps in the *Washington Post* story were actually tile grid maps—see page 238. Data were extracted from the *Washington Post* story and Miller et al. If you or someone you care for is in distress, suicide prevention and crisis resources are available at the National Suicide Prevention Lifeline in the United States at 1-800-273–8255. Many other countries have similar hotlines.



A scatterplot can be an alternative to a pair of maps.

Bubble plots may be less familiar to readers than simple choropleth maps, but with a clear title and a little annotation, this chart can better demonstrate the relationship between gun ownership and suicide. As you prepare to create your next map, ask yourself, "Is a map the best visualization to communicate my argument?"

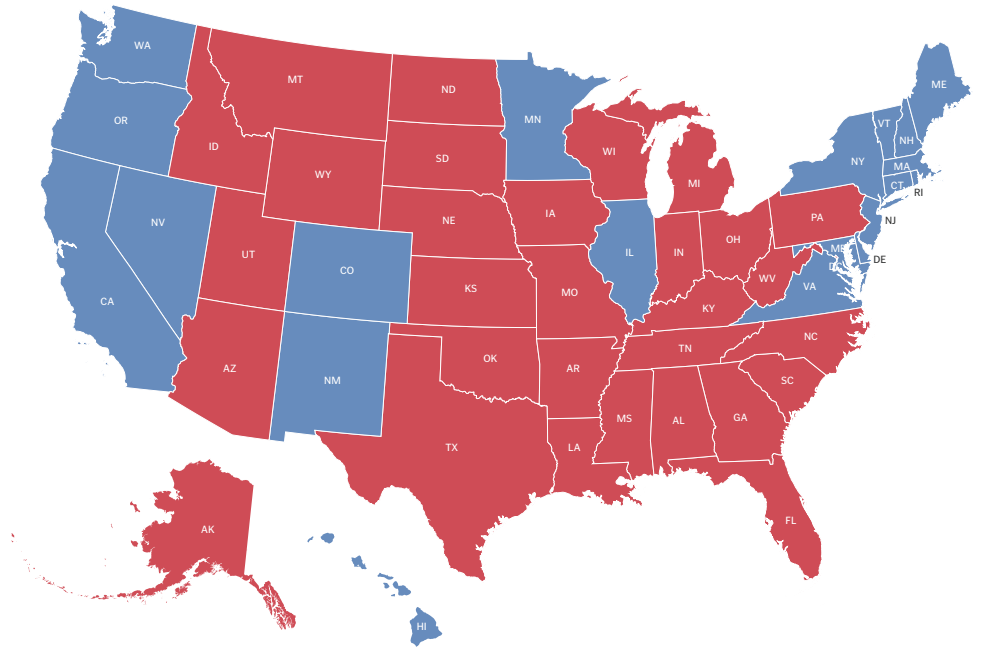
CARTOGRAM

One way to adjust for the geographic distortion of a typical choropleth map is with a *cartogram*, which reshapes geographic areas based on their values. There is an obvious tradeoff here: On one hand, these adjustments more accurately visualize the data because cartograms correlate the data and the geographic size. On the other hand, these graphs are not like the standard maps that we know and recognize. They are therefore not as intuitive for your reader. Your decision about whether to use a standard map or a cartogram will, as always, depend on your goals and your audience.

In his book *Cartography*, Kenneth Field summarizes the purpose of cartograms:

> The intent of most thematic maps is to provide the reader with a map from which comparisons can be made, and so geography is almost always inappropriate. This fact alone creates problems for perception and cognition. Accounting for these problems might be addressed in many ways such as manipulating the data itself. Alternatively, instead of changing the data and maintaining the geography, you can retain the data values but modify the geography to create a cartogram.

There are four primary types of cartograms: *contiguous*, *noncontiguous*, *graphical*, and *gridded*. One of the best ways to demonstrate the value of a cartogram is to examine the U.S. electoral college. In the U.S. election system, each state is assigned a number of electoral votes corresponding to its population, not its geographic size. Thus, states like Idaho, Montana, and Wyoming, which are very large in terms of square miles (325,412 square miles in total) but home to relatively few people, only have ten electoral votes between them. Massachusetts, by contrast, has eleven electoral votes and is 7,838 squares miles, less than 2.5 percent of the size of those three states. In this choropleth map of the 2016 presidential election, Idaho, Montana, and Wyoming take up a disproportionate share of space on the map relative to their electoral votes.

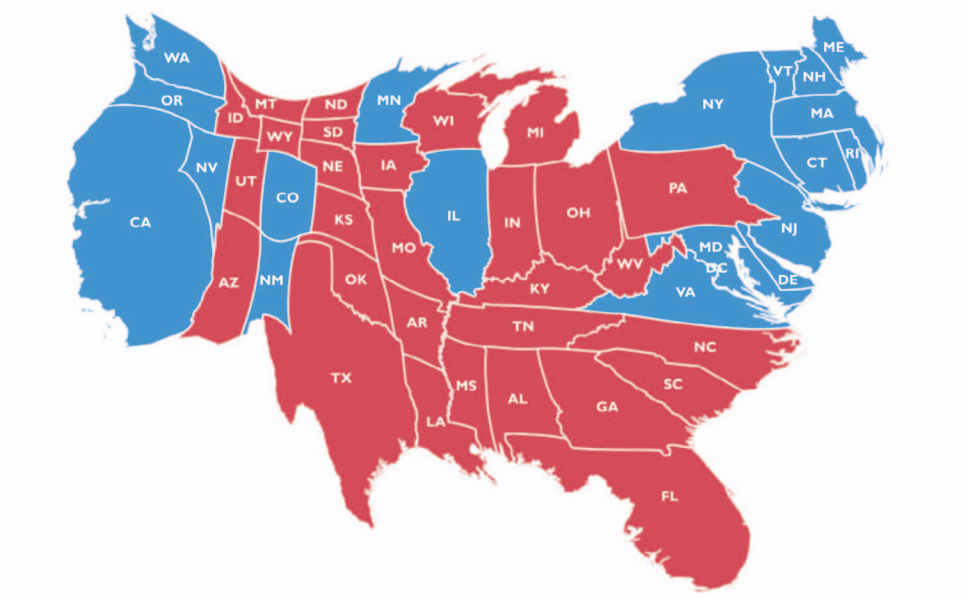


Source: Courtesy Kenneth Field, Esri Inc.

A standard choropleth map, this one shows the results of the 2016 U.S. Presidential election. Notice how much space large, but less populous states, like Idaho (ID), Montana (MT), and Wyoming (WY) in the northwest are compared with smaller but more populous states, like Massachusetts (MA) and Rhode Island (RI) in the northeast.

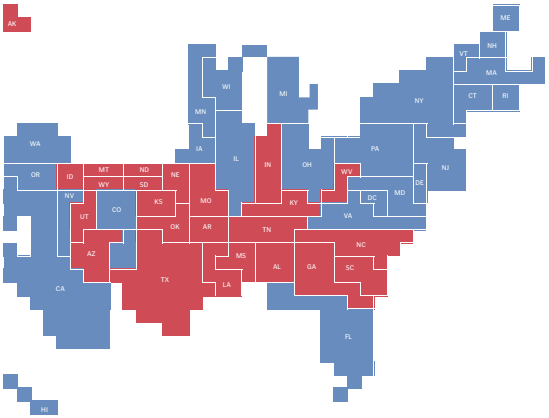
CONTIGUOUS CARTOGRAM

The contiguous cartogram adjusts the size of each geographic unit according to the data. In the map at the top of the next page, for example, each state is sized to its number of electoral votes (or population, if you'd rather think of it that way). The version on the bottom-left of the next page uses squares to scale each state while retaining the original approximate geographic location and borders. The third map scales the counties in each state according to the vote share, which generates a more purple shade to the country, reflecting the split between the two political parties. Each of these approaches distorts the overall shape of the country as the data warps the geography, so they will look somewhat foreign to readers. The tradeoff becomes clear here—we can more accurately scale the states according to their data values, but the geography no longer looks as familiar.

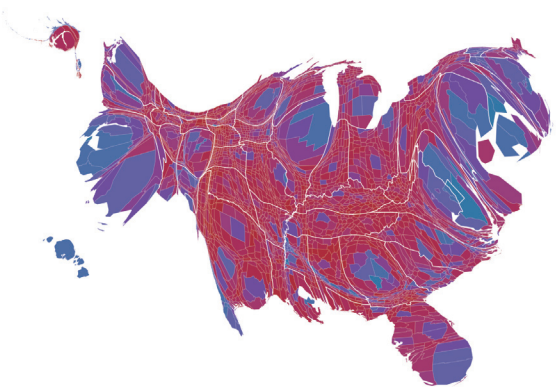


Source: Courtesy Kenneth Field, Esri Inc.

To address the fact that large but less populous states take up a disproportionate amount of space on the map, a cartogram scales the size of geographic units according to another data value. This map scales the states to their number of electoral votes.



Source: Courtesy Kenneth Field, Esri Inc.



Source: Courtesy Kenneth Field, Esri Inc.

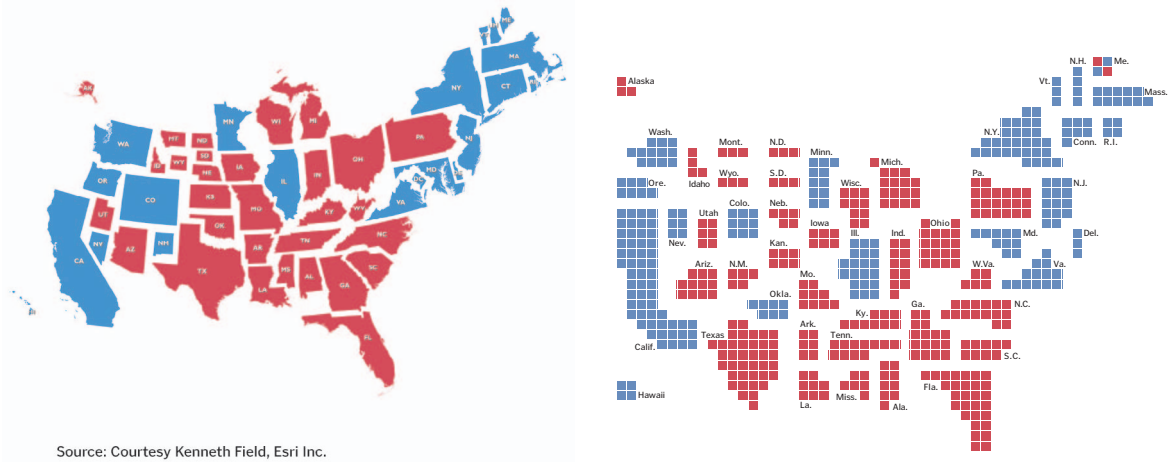
Other contiguous cartograms (using squares on the left or scaling counties on the right) are alternative ways to try to overcome the geographic distortions that occur in the standard choropleth map. But these maps are almost surely to be less familiar to readers than the standard map.

NONCONTIGUOUS CARTOGRAM

Now that you've seen a contiguous cartogram, you can probably guess what a noncontiguous cartogram looks like. In this approach, the size of the geographic units are based on the data value, such as population, but the units are broken apart and not kept adjacent to one another. In this way, we maintain the shape of the individual units but distort the overall view. One advantage of the noncontiguous cartogram is that we can build in more space for labels and annotation.

The map on the left scales each U.S. state according to its number of electoral votes and includes color to denote which candidate won those votes. The exact shape of each geographic unit also isn't necessary—the map on the right uses collections of squares for each state, here scaled to the number of electoral votes. In this version, we can again see how Idaho, Montana, and Wyoming look a lot smaller, while New York becomes much larger.

The noncontiguous cartogram was invented in the mid-1970s by Judy Olson, a geographer then at Boston University. "Probably one of the most interesting aspects of the noncontiguous cartogram," she wrote in her 1976 paper, "is that the empty area between units is meaningful. If the highest-density unit is used as the anchor [in other words, if the most dense geographic unit is used to scale all the other units], then the empty areas reflect the degree of discrepancy between the density in the most-crowded unit and the density in other units. The effect can be quite dramatic."



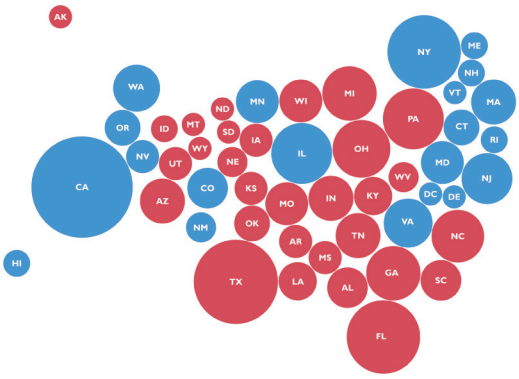
Source: Courtesy Kenneth Field, Esri Inc.

Noncontiguous cartograms like these break up the geographic areas. This helps for adding labels and annotation, but is a more unfamiliar chart type.

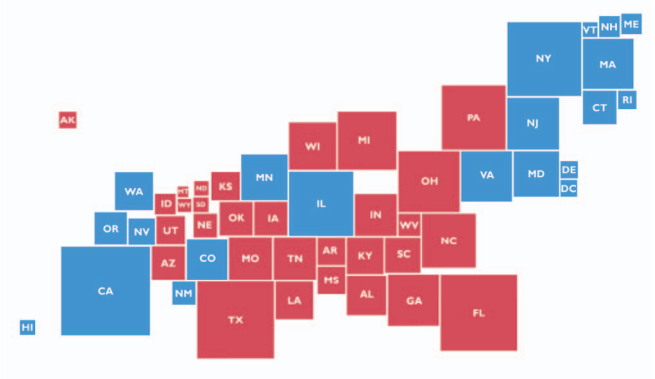
GRAPHICAL CARTOGRAM

Graphical cartograms do not maintain the original shape of the geographic units and instead use other shapes sized to the data values. Perhaps the most well-known graphical cartogram is the Dorling map—named for geographer Danny Dorling at the University of Leeds—which uses circles sized by area to the data.

A variant on the Dorling map is the DeMers Cartogram (or tilegram), which uses squares instead of circles. One advantage with the DeMers approach is that it minimizes the space between the geographic units. A disadvantage is that the entire geography becomes less recognizable. These two graphical cartograms again show the number of electoral votes in each state, and the colors again show which candidate won the state's electoral votes.



Source: Courtesy Kenneth Field, Esri Inc.



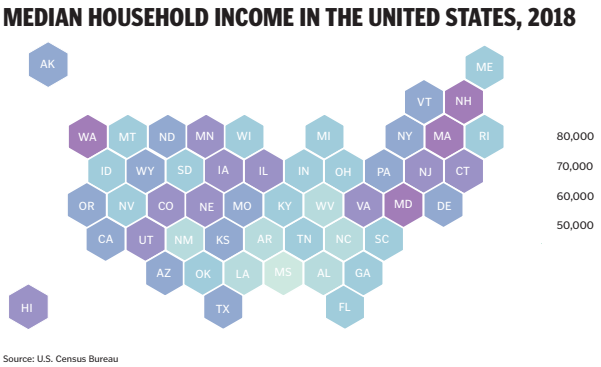
Source: Courtesy Kenneth Field, Esri Inc.

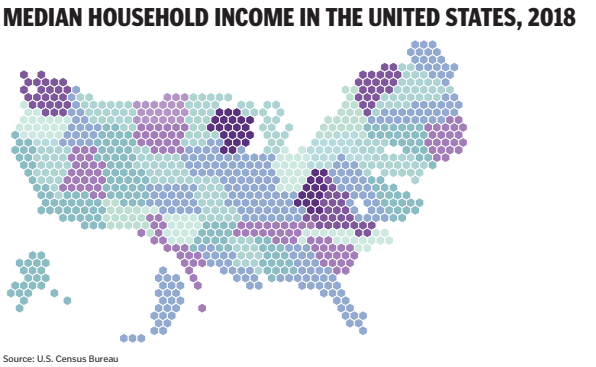
The Dorling (left) and Demers (right) cartograms move further away from the standard geographic maps and use shapes instead.

GRIDDED CARTOGRAMS

The fourth and final cartogram is the gridded cartogram, in which different shapes are scaled to the data and arranged so they maintain the general shape of the major geography. People most often use squares or hexagons to create these kinds of maps.

Take these hexagon grid maps, for example. The advantage of the hexagon over other shapes is that it offers us more flexibility to arrange the tiles closer to the true geography of the country. The map on the left shows one hexagon per state, shaded to encode median

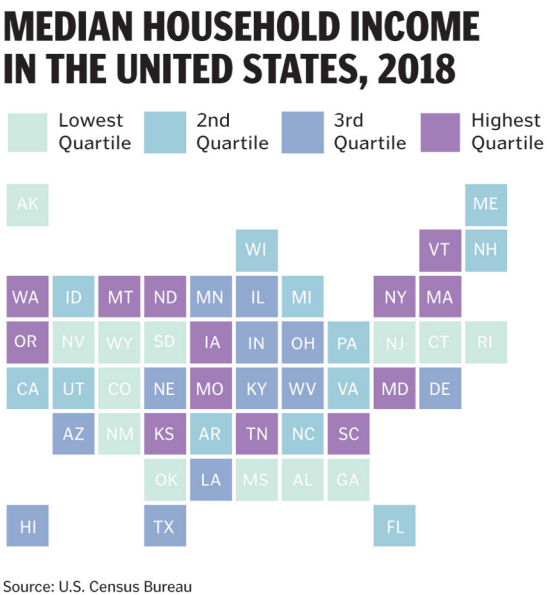




The hexagon grid map is named exactly for what it is: A gridded set of hexagons—either one for each geographic unit or scaled to its data value.

household income. The version on the right uses multiple hexagons per state with the color and number of hexagons corresponding to the data value.

Another popular gridded cartogram uses a single square for each geographic unit. This is often called a *tile grid map*. Here, median household income is divided into four groups (or quartiles).

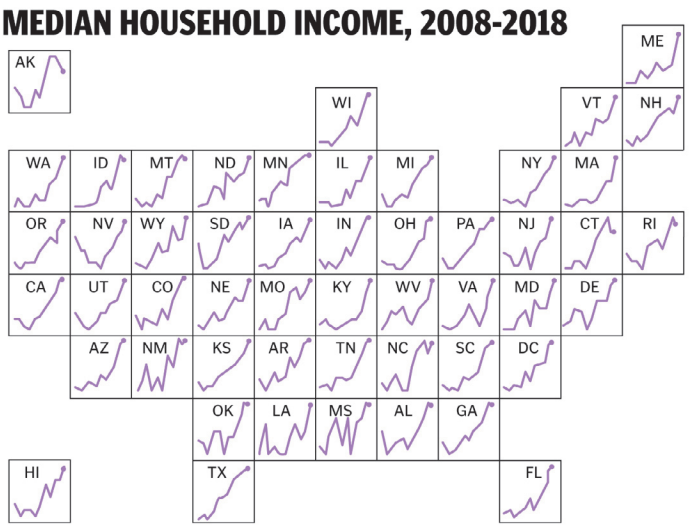


The tile grid map uses a single square for each geographic unit.

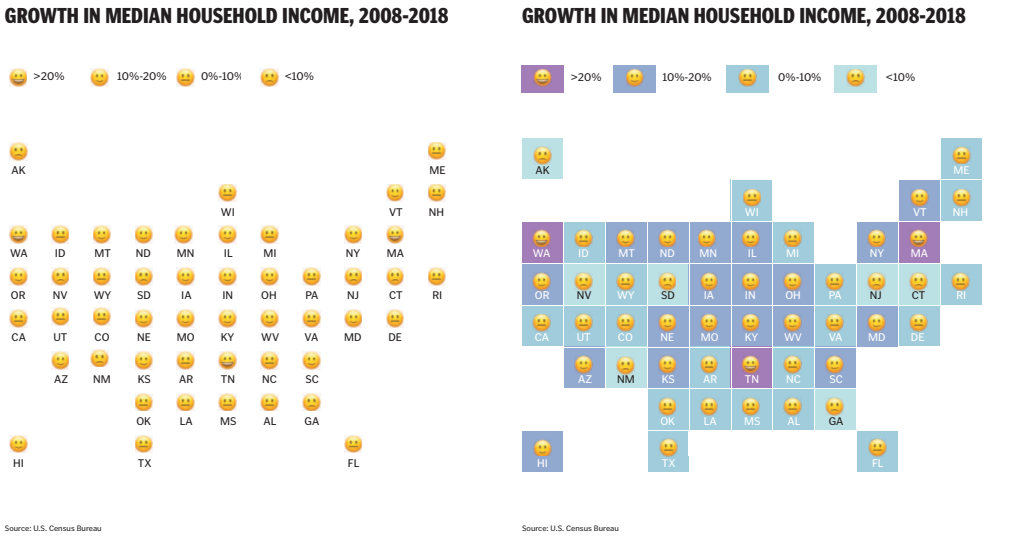
As with all visualizations, there are tradeoffs. The advantage of the tile grid map is that each state is the same size, which abstracts from the geographic distortion. The disadvantage is that the geographic units are now not necessarily in the right place. In the tile grid map on the previous page, South Carolina is located east of North Carolina, California doesn't touch Arizona, and Wisconsin is north of Minnesota, all of which are not their real geographic relative locations. The arrangement of the tiles can be changed of course, but any decision is going to be arbitrary because we have moved away from true geography. But this map can also be easier to construct (it can be made in Excel with resized spreadsheet cells) than choropleths or cartograms.

Another advantage of the tile grid map is that it enables you to add more data in a consistent shape. In this tile grid map of the United States, small lines (or sparklines) are included in the square of each state showing the change in median household income between 2008 and 2018.

Another advantage of tile grid maps (and graphical cartograms, for that matter) is that we can add other shapes to the different geographic areas. Both of the tile grid maps on the next page use emojis to categorize the same household income estimates shown so far. Again, there's a clear tradeoff here: the emojis are a little more fun and a little more visual, but it has



One advantage of the tile grid map is because each state is the same size, we can add small lines, bars, or other graph types to each square.



Another advantage of tile grid maps is that we can use other shapes, such as emojis.

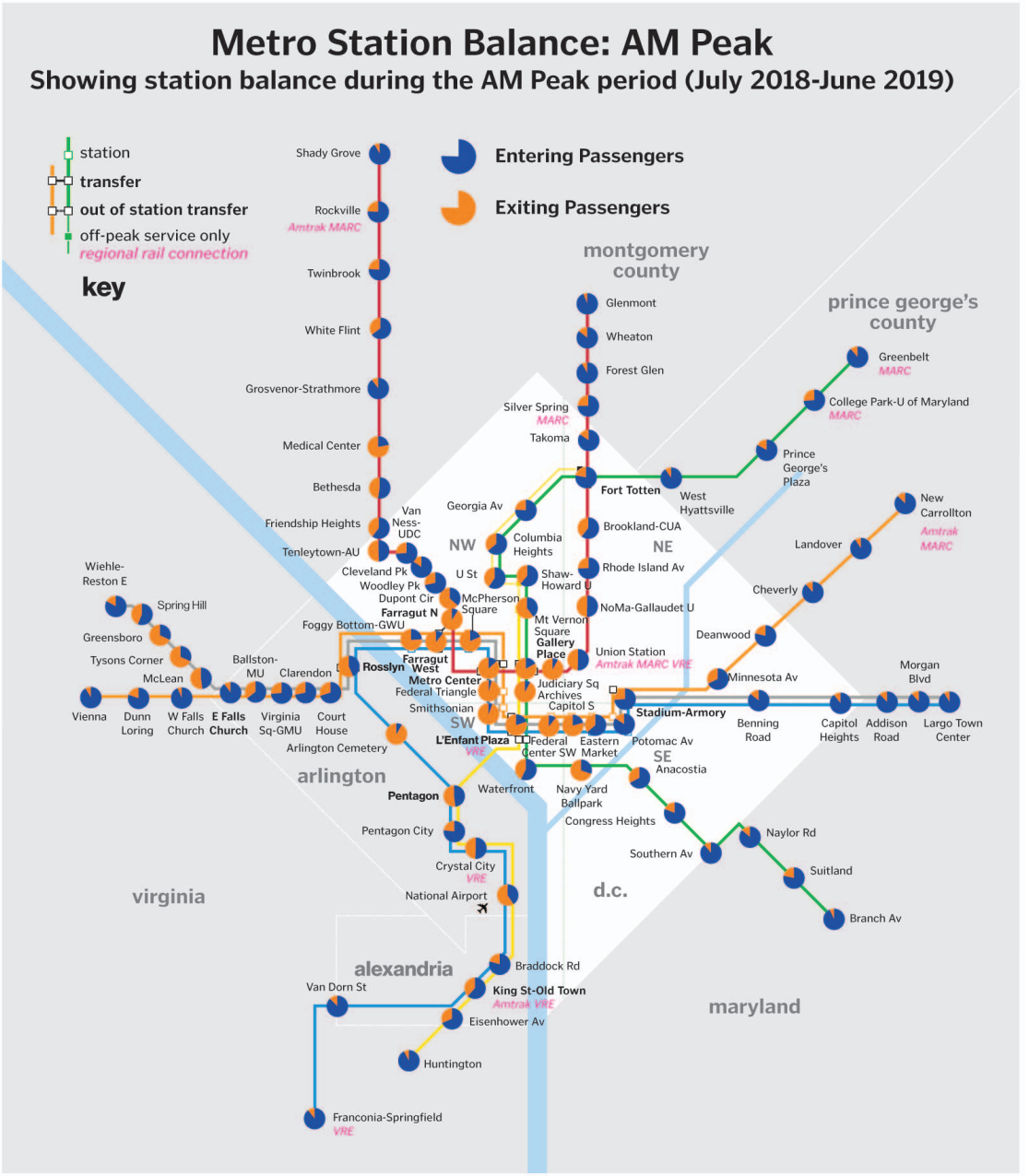
a very different tone and makes it visually more difficult to immediately pick out values for groups of states or larger geographic patterns as in the original map. Adding some boundaries or shading (as in version on the right) merges the two approaches.

NON-AREA-BASED CARTOGRAM

There is one other class of cartograms worth mentioning. Maps do not always have to encode data, and they don't necessarily need to encode them accurately. A non-area based cartogram (or distance cartogram) distorts the physical geography by displaying relative time and distance. This version of the Washington, DC metro (subway) map, for example, shows relatively constant distances between stops, when in fact the distances vary considerably. Check out the western part of the Orange line that runs concurrently with the Silver line. The distance between my metro stop at East Falls Church and Ballston-MU, and between the Ballston-MU and Virginia Square-GMU stops are the same on the map, but the first trip is 2.7 miles long and the second trip is only half a mile long. Stretching the stops out to their actual distances is unnecessary here because the purpose of the map is to provide a compact view of the subway lines so riders can quickly and efficiently plan their trip.

This version of the Washington, DC, metro map from designer Jacob Berman shows relatively constant distances between stops, even though that's not geographically the case.

A familiar map like this also gives us the opportunity to add data. As an example, the next map shows the number of passengers entering and exiting each station in the morning rush hour. Each station is turned into a pie chart in which the blue segment represents the share of people entering each station and the orange represents the share of people exiting. Even if



Source: Map from Jacob Berman; data from the Washington Metropolitan Area Transit Authority. Based on Matt Johnson (2012).

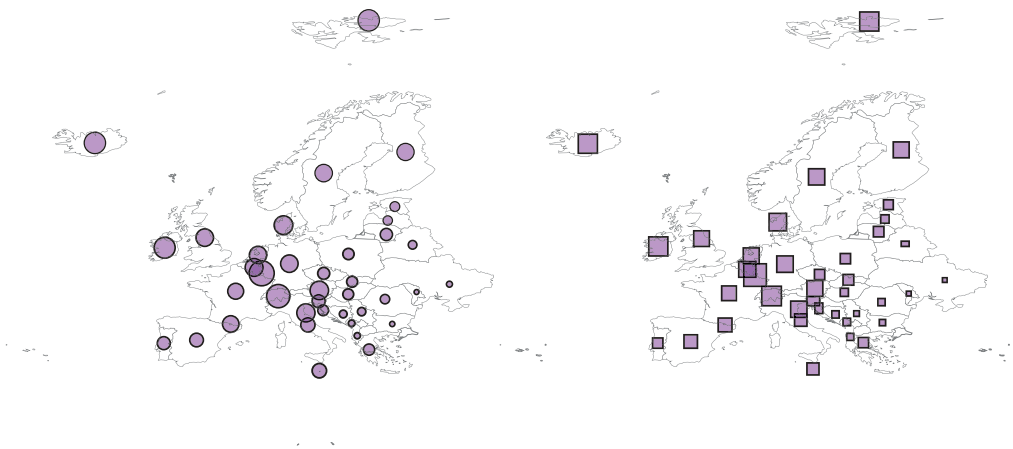
We can add other graphs to maps—pie charts here represent the share of people entering and exiting each metro stop in the morning. Even if you're not familiar with the DC metro, you can see how people tend to commute from the outer parts of the area (more blue areas in the pie charts) to downtown (more orange areas in the pie charts).

you're not familiar with this subway system, you can see the movement from the outer areas to the center city in the morning.

But let's make sure we remember our audience. These maps would be of interest to people who regularly use the DC metro system, but they would not be as valuable if I was an urban planner making an argument in Atlanta or Dallas or Berlin, because those audiences might not be familiar with the shape and arrangement. As always, we must consider the needs, expertise, and expectations of our audience.

PROPORTIONAL SYMBOL AND DOT DENSITY MAPS

Color and size are not the only encoding techniques we can use to visualize data on a map. Different shapes and objects—lines, arrows, points, circles, icons, even small compact bar graphs and pie charts—can all be placed on a map. These are known as *proportional symbol maps*, because the symbols are sized proportionate to the data. Be careful not to clutter the map or the reader will have difficulty identifying the most important information.



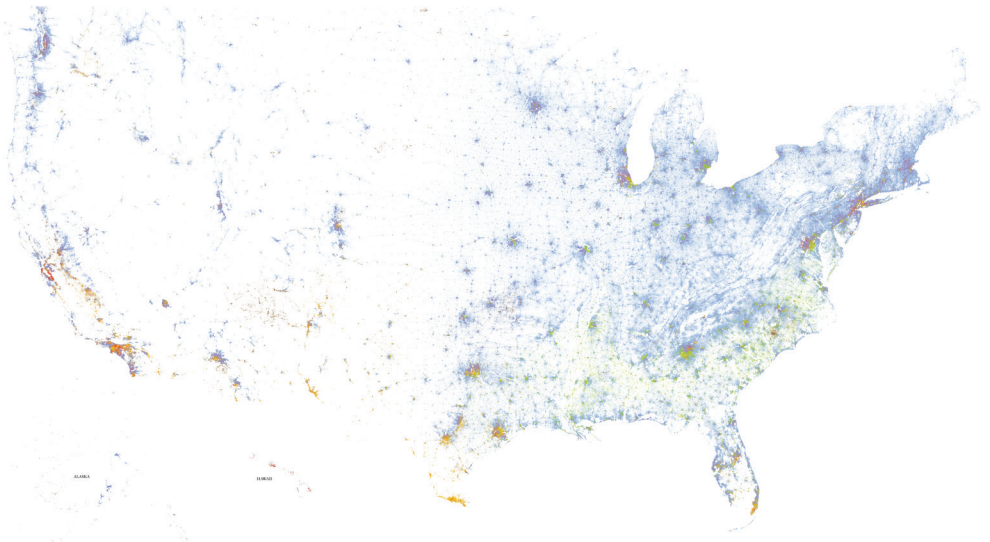
Source: The World Bank

Different shapes and objects—lines, arrows, points, circles, and more—can all be placed on a map, sized according to the data value.

The two maps on the previous page show per capita GDP for European countries encoded with circles and squares, rather than using color to shade countries. Notice the importance of using a transparent color to make overlapping shapes visible, a technique we've seen in previous visualizations. Where there are dense clusters of areas, such as around Belgium and Netherlands, it can be difficult to find the shapes for individual countries. Aside from our inherent difficulty of discerning exact quantities from shapes like circles, the dense clusters can be a barrier to seeing specific geographic units.

A *dot density map* or *dot distribution map* takes the proportional symbol map in a slightly different direction by using dots or other symbols to show the presence of a data value. Symbols can either represent a single data value (*one-to-one*) or a many values (*one-to-many*). These kinds of maps can be data intensive, but they can also illuminate spatial patterns and clusters that would otherwise be difficult to visualize in a choropleth map or cartogram.

Dot density maps are valuable because they quickly and easily show geographic densities through the clustering of the symbols. The primary challenge, however, is that because these



Dot density maps include a dot or other symbol for a single (or many) data values. The "similarity" Gestalt principle helps us see the clusters of people around the country.

Source: Image Copyright, 2013, Weldon Cooper Center for Public Service, Rector and Visitors of the University of Virginia (Dustin A. Cable, creator).

maps require exact geographic locations like addresses or longitude-latitude pairs, which are not usually available (or publishable), the symbols must be placed in a random or arbitrary position within a specific geographic area.

Consider the dot density map of the United States on the previous page. It uses data from the 2010 U.S. decennial census and places a dot for each of the country's 308 million residents in their Census blocks. Colors denote different racial and ethnic groups: blue for White people; green for Black people; red for Asian people; orange for Hispanic or Latino people; and brown for Native American people and people of multiple or other races. As with the map shown in Chapter 2, there is nothing in this map *except for the data*—no state borders, city markers, or other labels. We can still recognize it as the shape of the United States because people cluster in cities, and on borders and coasts.

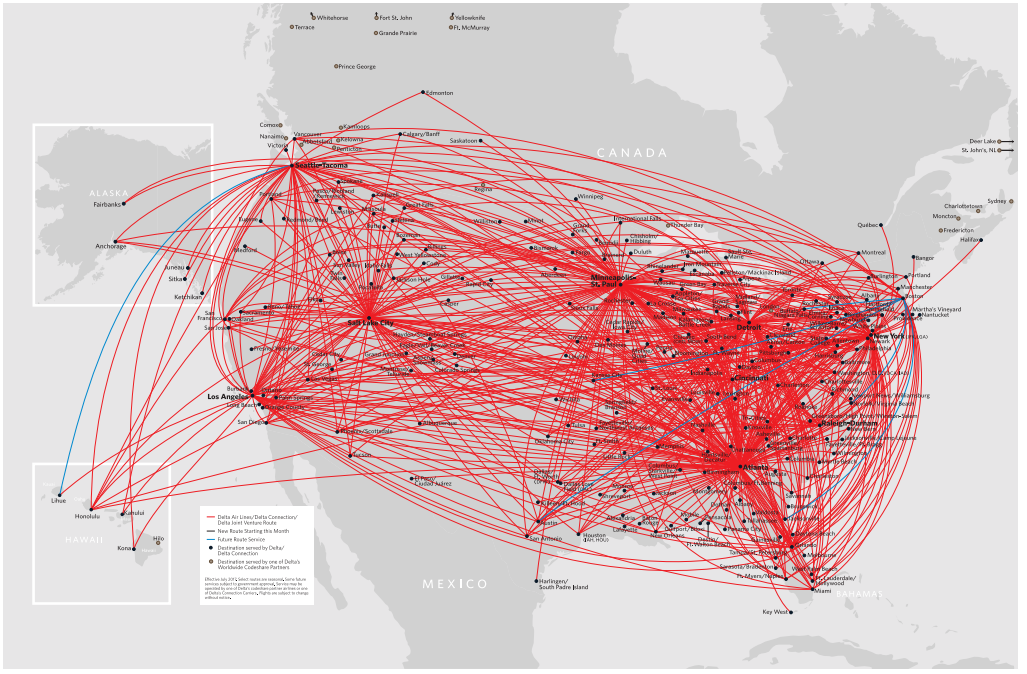
FLOW MAP

Flow maps show movement between places. Arrows and lines denote the direction of the flow, and the width of the line can correspond to the data value. Flow maps can also encode qualitative data, but in those cases the width of the directional symbols may not be scaled to a data value. We saw one such example of a flow map on page 128 to show import and export trade flows between the United States and other areas of the world.

There are different types of flow maps. *Radial flow maps* (also called *origin-destination maps*), show flows from a single source to many destinations. A *distributive flow map* is similar, except that the flow from the single source can fork into many different lines. I like to think of these maps as the ones in the back pages of the airplane magazine, like this one in the back of the Delta Airlines magazine (see next page). This is a distributive flow map because it shows all of the various connections.

As a slight aside here, you might take a look at this map and think, "Wow, that's cluttered! How am I supposed to track my flight here?" But that's not the purpose of the map—instead, the intention of the map (or at least how I infer it) is to demonstrate Delta's domestic "unrivaled coverage"—all the many flights Delta offers around the United States. Showing the tangled web of *all* the flights communicates that point.

Perhaps the most famous flow map (at least in the data visualization field) is Charles-Joseph Minard's 1869 map of Napoleon's Russian Campaign of 1812 to 1813 and the connected, but lesser-known map of Hannibal's 218 BC march through the Alps to Rome. Minard was a French civil engineer who conducted in-depth studies over many decades to



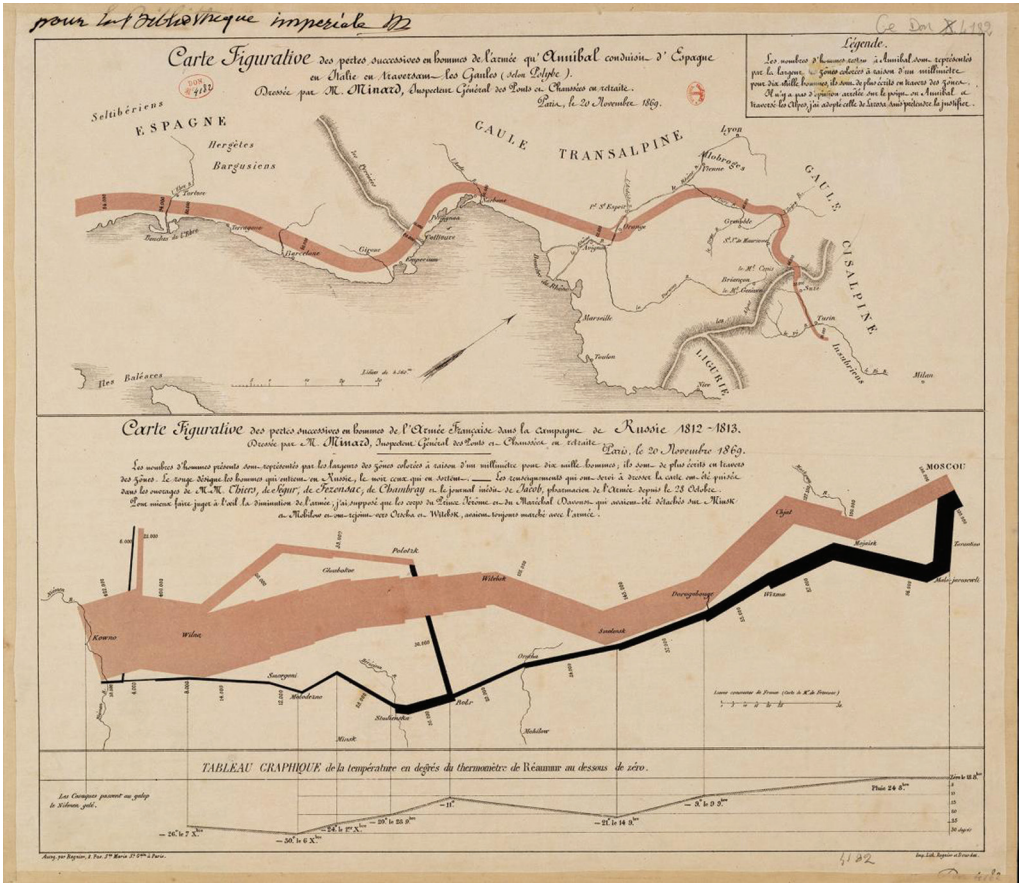
Flow maps show movement between places. Even though this flow map from Delta Airlines looks cluttered, it meets the airline's goal: demonstrating "unrivaled coverage."

create compelling visualizations in support of his research. Minard's Napoleon map—the bottom panel of the image on the next page—shows the "progressive losses in men" that Napoleon's army suffered as it marched into Russia and back.

Starting from the left, 420,000 men invaded Russia in June 1812. By the time the army reached Moscow four months later, only 100,000 troops were still alive. When Napoleon ordered a retreat in the fall, the army was forced to fight through blistering cold (dropping to a low of −30 degrees Celsius), and was ultimately reduced to 10,000 soldiers when it exited Russia in late 1813.

Minard's map integrates six different data values into a single view:

1. The number of troops (thickness of the lines);
2. The distance traveled (scale in the lower-right);
3. The temperature (line graph at the bottom);
4. The time (also included in the line graph at bottom);



Charles Joseph Minard's Hannibal and Napoleon maps show the march of two different armies. Photo courtesy of Ecole nationale des ponts et chaussées.

5. The direction of travel (denoted by color—brown going eastward and black in retreat); and
6. Geography (cities, rivers, and battles—some but not all are included).

In her book on Minard's graphs, *The Minard System*, author Sandra Rendgen writes the following:

> Minard created his visualization more than fifty years after the [Napoleon] campaign. It is a brilliant conceptual transfer: in applying the flow method to a military campaign, Minard shifts

his entire focus to a single variable: the number of people in the flow. This variable sees only one type of variation—*a sharp and steady decline.* It seems to have been this potent and poignant message that made these two maps (and particularly the Napoleon one) so successful in telling a story about the cataclysm of war.

CONCLUSION

In this chapter we surveyed the promises and perils of visualizing geographic data. Sometimes the varying sizes of geographic units may distort the data. Other times sizing the geography to the data may make a familiar geography look foreign.

When working with geographic data, your instinct may be to create a map. But take a moment to consider: Is a map the best way to present your data? Does your reader need to see the exact differences between data values? If so, the aggregation problem inherent in many maps may make that difficult. Are there clear geographic patterns to be seen in the data? If not, then the map may not actually help the reader see your point.

If a data map *is* the right approach, carefully consider the map projection you use and whether the standard choropleth map is the best choice. Maybe some kind of cartogram—even with all its flaws—would be a better fit for your context and reader.

You may also determine that the best approach is to *combine* visualization types. Depending on your final publication type, you might use multiple visualizations, say, a map with a bar chart or table. This approach can help give your readers a familiar visualization type in which they can identify themselves and their location, but also help them gain a better, more detailed view of the actual data.

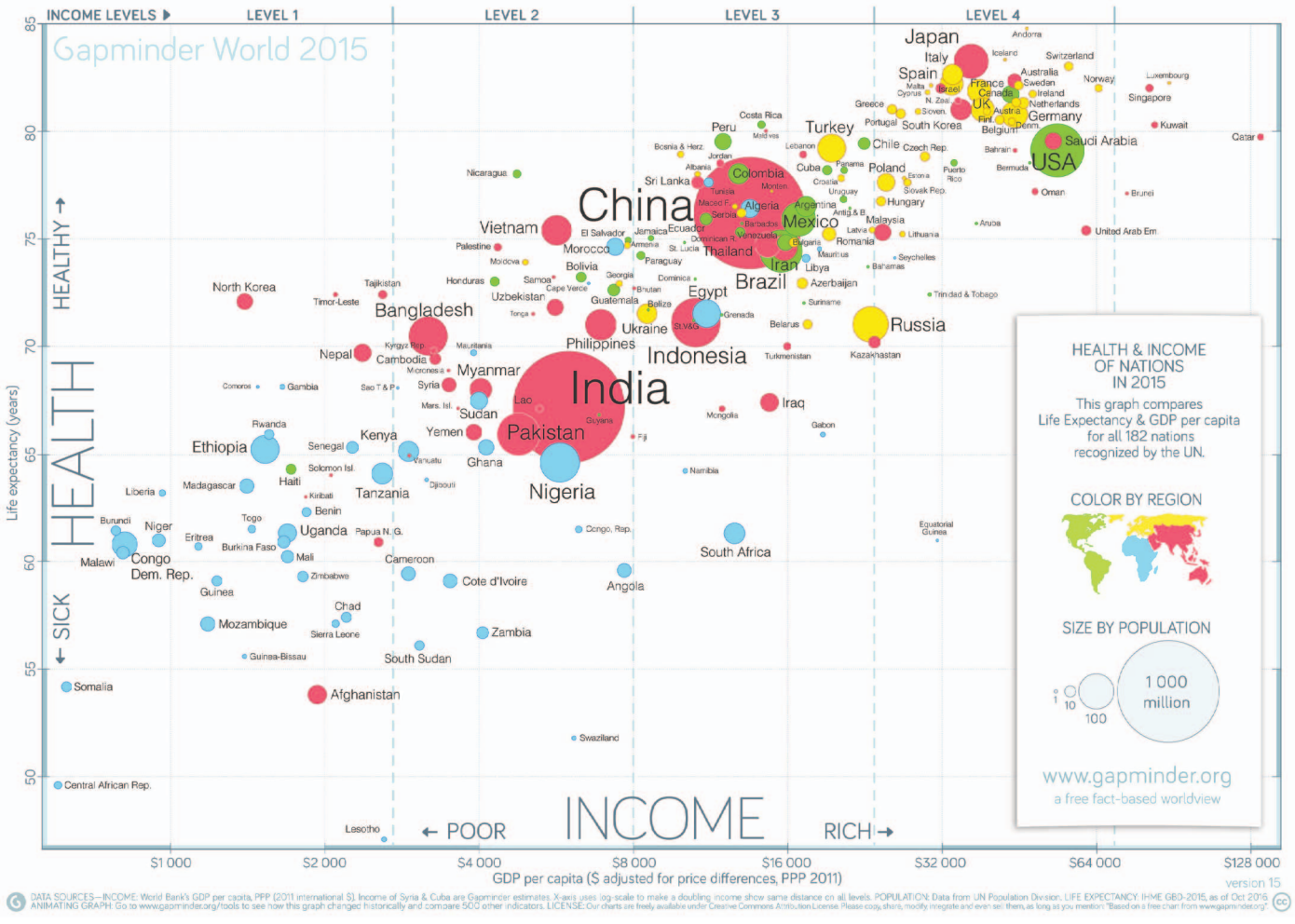
RELATIONSHIP

The charts in this chapter show relationships and correlations between two or more variables. Perhaps the most familiar chart type in this class is the scatterplot, a chart in which the data are encoded to a single horizontal and vertical axis. Other shapes and objects can also be used to visualize the relationship between two or more variables—a parallel coordinates plot uses lines, while a chord diagram uses arcs within a circle. These charts can show the reader correlations and even causal relationships.

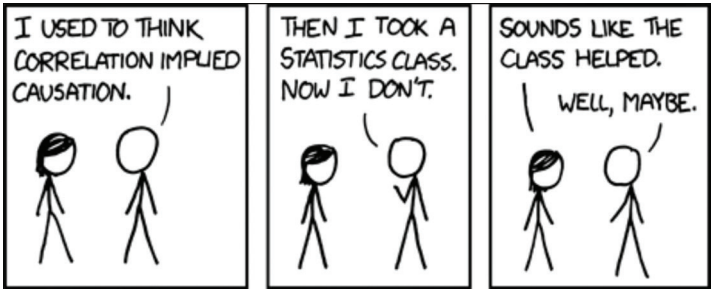
For this chapter, I use the color palette and font from a famous bubble chart created by the Swedish academic Hans Rosling and his colleagues at Gapminder, a foundation dedicated to visualizing statistics. Rosling's Gapminder project didn't lay out a specific data visualization style guide, but the visualizations in this chapter use the basic colors and font (Bariol), with additional styles based on other visualizations from the Gapminder website.

SCATTERPLOT

The scatterplot is perhaps the most common visualization to illustrate correlations (or lack thereof) between two variables—one variable is plotted along a horizontal axis, and the other along a vertical axis. The specific observations are plotted in the created space. Unlike a bar chart, the scatterplot axes do not necessarily need to start at zero, especially if zero is not a possible value for the data series.



Source: Based on a free chart from www.gapminder.org



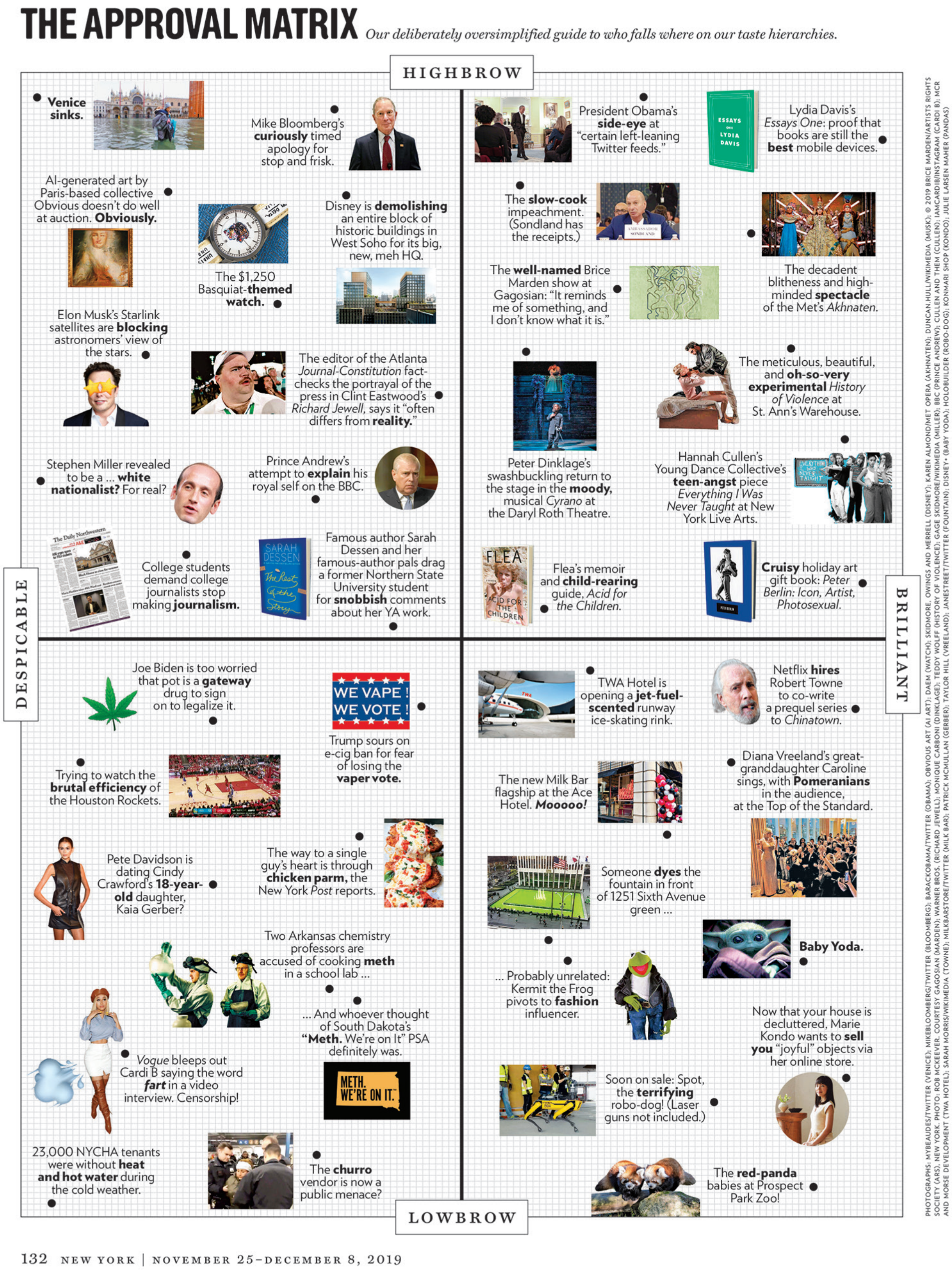
Source: XKCD

One of the most famous graphs that shows the relationship between two variables is the set of scatterplots created by Rosling and his colleagues at Gapminder. A physician by training who spent roughly two decades studying public health in rural areas across Africa, Rosling is perhaps best known for engaging presentations and data visualizations, and for promoting the use of data to explore issues around international development. In Rosling's 2014 TED Talk, he showed an animated scatterplot of the relationship between the fertility rate (number of births per woman) and life expectancy at birth from 1962 to 2003 for countries around the world.

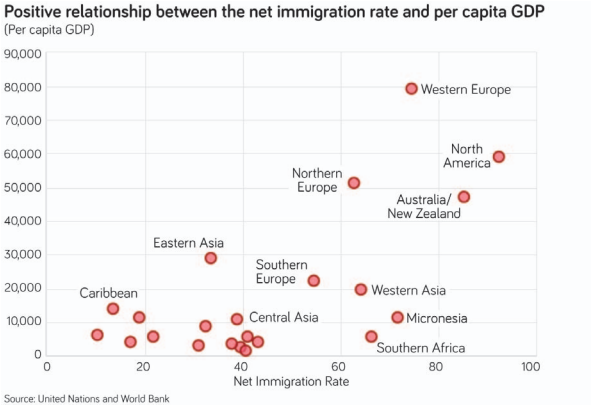
As I've noted previously, there are many nonstandard graph types—scatterplots included—with which your reader may be unfamiliar. This doesn't mean you can't use these visualizations, but it does mean you should be mindful that your reader may be unfamiliar with them, and consider how to prepare them to understand your graphs.

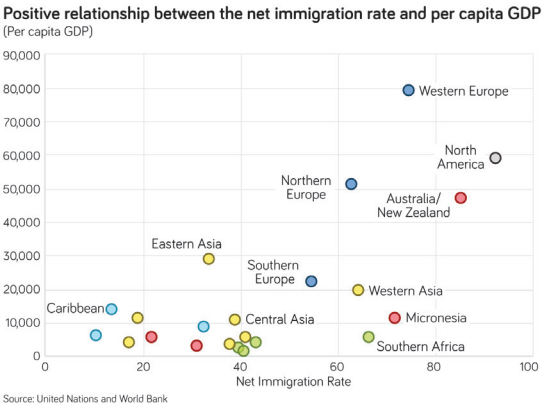
Some readers may be familiar with scatterplots (or other nonstandard graph types for that matter) even if they are not familiar with reading *data* in scatterplots. *New York Magazine's* weekly *Approval Matrix* on the next page, for example, is their "deliberately oversimplified guide to who falls where on our taste hierarchies." Bits of text, images, and icons are plotted in a space defined by the "Highbrow-Lowbrow" vertical axis and "Despicable-Brilliant" horizontal axis. Though it's a light-hearted way to list popular news tidbits, it is, at its core, a scatterplot.

Moving to scatterplots that are a bit more, shall we say, data-driven, these two scatterplots on page 253 show net immigration (defined as the number of people migrating into a region divided by the total number of migrants moving in and out of a region) plotted along the horizontal axis and per capita gross domestic product (GDP) along the vertical axis. The version on the left uses a single color with a slight transparency (or "opacity") so the reader



New York magazine's "Approval Matrix" is a scatterplot. Not truly data driven, but a scatterplot nonetheless.





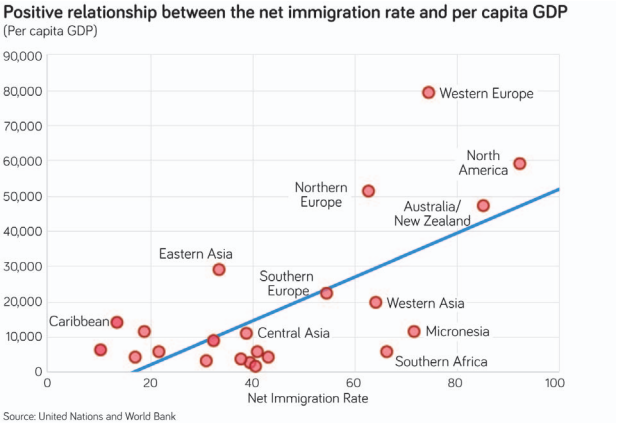
Both scatterplots show the association between net immigration and per capita GDP, using either a single transparent color (left) or different colors for regions of the world (right).

can see overlapping values. The same transparent effect is used in the scatterplot on the right, but this time colors capture the different regions of the world.

A scatterplot can help the reader see whether two variables are associated with one another. If the two variables move in the same direction—to the right along the horizontal axis and up along the vertical axis—they are said to be *positively correlated.* In other words, when both variables get bigger or smaller simultaneously, they are positively correlated. If they move in opposite directions, they are said to be *negatively correlated.* And if there is no apparent relationship, then they are not correlated (see Box on the next page). In the two scatterplots above, you get a visual sense that the two metrics are positively correlated—that net immigration is higher for regions with higher per capita GDP, in particular Western and Northern Europe, Australia/New Zealand, and North America.

One way to make the correlation even clearer is to add what statisticians call a *line of best fit* to the scatterplot. These are also called "regression lines" or "trendlines," and they show the general direction of the relationship. The statistical calculations to create lines of best fit are beyond the scope of this book, but the point is that you can make it even clearer to the reader in what direction (and to what magnitude) the two variables are correlated by calculating and including this line.

While the scatterplot is becoming a more common chart type, readers may still have difficulties reading and understanding it. A 2016 Pew Research Center survey showed that about 60 percent of people could correctly identify what they were seeing in a scatterplot.



A *line of best fit* visualizes the correlation between the two variables.

CORRELATIONS

You have probably heard the old adage that "Correlation does not imply causation." We hear this so often because people regularly assign a causal relationship between variables that is actually coincidental. People eat more ice cream when it's hot outside, but that doesn't mean that more ice cream consumption *causes* the temperature to rise. As you look at your data and visualize the relationship between the observations within, be careful to understand when something may be correlated and when it might be causal. The less we know, the more we observe correlations and not causation.

Correlation is a measure of the strength of the linear association between two quantitative variables. The most common measure of correlation is the *Pearson correlation coefficient*, which measures the linear association between variables and is typically denoted with the Greek letter rho (ρ).

The sign and value of a linear correlation coefficient describes the direction and magnitude of the relationship between the variables. The value of the correlation coefficient ranges between -1 and $+1$. Values of -1 signal a perfectly negative correlation, $+1$ signals a perfectly positive correlation, and 0 represents no linear correlation. Positive correlation coefficients denote a positive correlation, which means that if one variable gets bigger, the other variable also gets bigger; negative coefficients mean the variables move in opposite directions, one variable gets bigger as the other gets smaller.

This discussion is related to these linear associations (or relationships) and it is also possible for two variables to have a *nonlinear* relationship. A linear association is a statistical term that describes a straight-line relationship between one variable and another. A simple example is how we might calculate distance as rate times time. In this case, if we were driving sixty miles per hour for two hours, we would travel 120 miles. The driving speed does not change over time, so the relationship is linear.

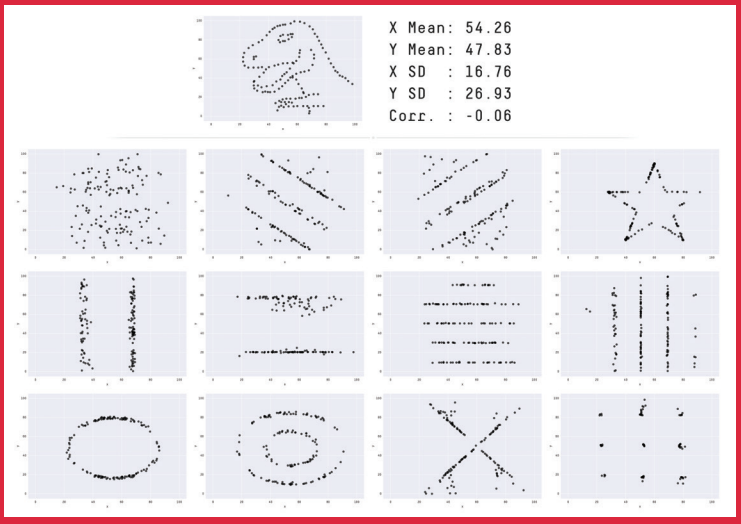
A nonlinear association, by comparison, refers to patterns in data that curve or break from the straight linear trend. Consider, as an example, the profit a company makes from a new product. When it is first released, there is little competition and sales grow. As sales continue to rise, public awareness increases, and profits start to roll in. Competing companies then start making their own version of the product and prices for the original fall to keep up, so profits decline. The company then develops a new version, and the whole cycle starts again.

As you can see in the images below, the two data values lie on a single diagonal line when they are perfectly positively or negatively correlated. When the two values move in tandem, either up or down, they are said to be positively or negatively correlated. The data in the bottom-right graph demonstrates the impact outliers have on this measure of correlation—moving a single point to the top-left part of the graph reduces the correlation from +1.0 to +0.8.

These visuals reinforce the importance of looking at our data as we conduct our analysis. Data visualizations help us not only communicate our work to our readers, but also enable us to explore our data. They can reveal patterns and relationships we wouldn't otherwise see. It's important not to leave this to the end of the workflow.

In 2016, University of Miami journalism professor Alberto Cairo drew a dinosaur with points in a scatterplot and dubbed it the "Datasaurus." His goal was to show the importance of visualizing your data in the exploratory phase. Imagine if you were teaching a data visualization class and asked your students to draw a scatterplot with 142 points, an average *x* value of 54.26, an average *y* value of 47.83, accompanying standard deviations, and a Pearson correlation of −0.06. Do you think anyone would draw a dinosaur?

In a 2017 paper, researchers Justin Matejka and George Mitzfaurice took the "Datasaurus" one step further and generated twelve alternatives that maintained the same summary statistics (mean, standard deviation, and correlation). The message of Cairo's "Datasaurus," Matejka and Fitzmaurice's paper, and Anscombe's quartet from Chapter 1, is that we should never rely on summary statistics alone but also on visuals of the data.



Source: Matejka and Fitzmaurice, 2017

BUBBLE PLOT

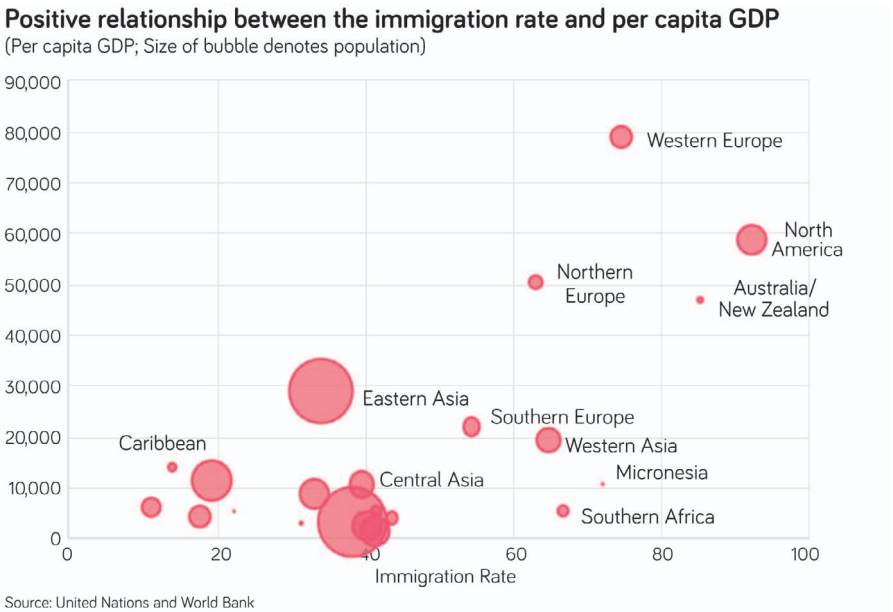
The scatterplot can be transformed into a bubble plot (or bubble scatterplot) by varying the sizes of the circles according to a third variable. The data points don't have to be circles, they can be any other shape that doesn't distort our perception of the data. As mentioned in the section on

bubble charts, the circles should be sized by area, not radius (see page 123). Color can help group or highlight certain points or direct the reader's attention to different parts of the graph.

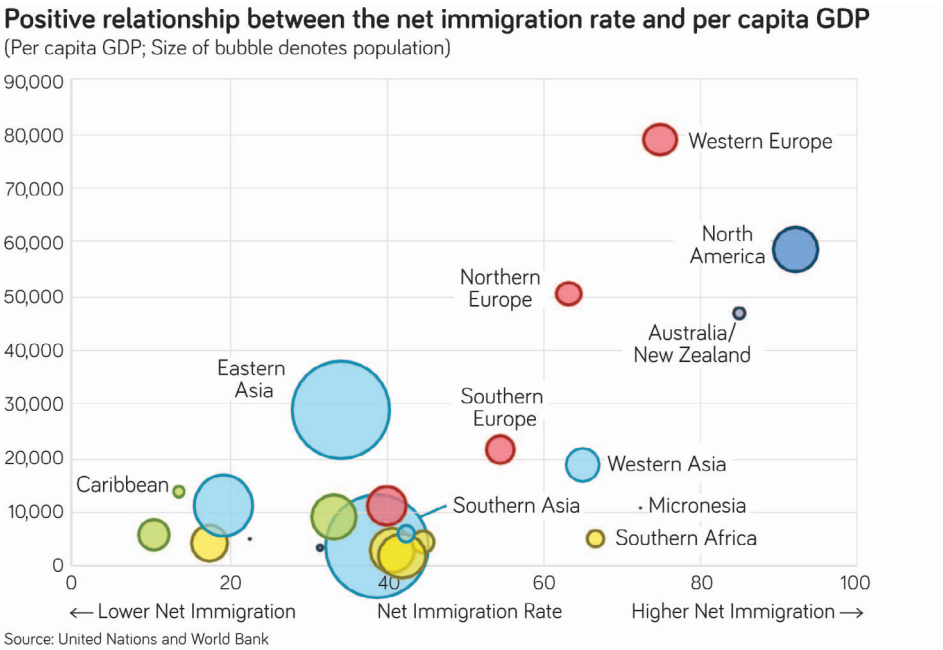
The circles in this bubble plot are scaled according to the population of each region. The same positive relationship is still evident, but you can now see the relative size of each area.

Because it is a more uncommon graph, be especially mindful of how labeling and annotation can guide readers through the chart and its content. One strategy is to label each axis and the direction of the change along the axis. The bubble plot on the next page has a centrally located horizontal axis that reads "Net Immigration" and two other labels, "Higher Net Immigration" and "Lower Net Immigration."

To further guide the reader, we could add a 45-degree line, on which the values are equal. We could also highlight specific points with color or outlines, or we could add text to explain what a point or set of points means. Properly labeled, these elements can lead the reader through the graph and content. Even people who know how to read scatterplots can struggle for a moment to understand what is going on when there are lots of points.



A bubble chart adds a third variable to the typical scatterplot. Here, the size of the circles corresponds to the population in each region.

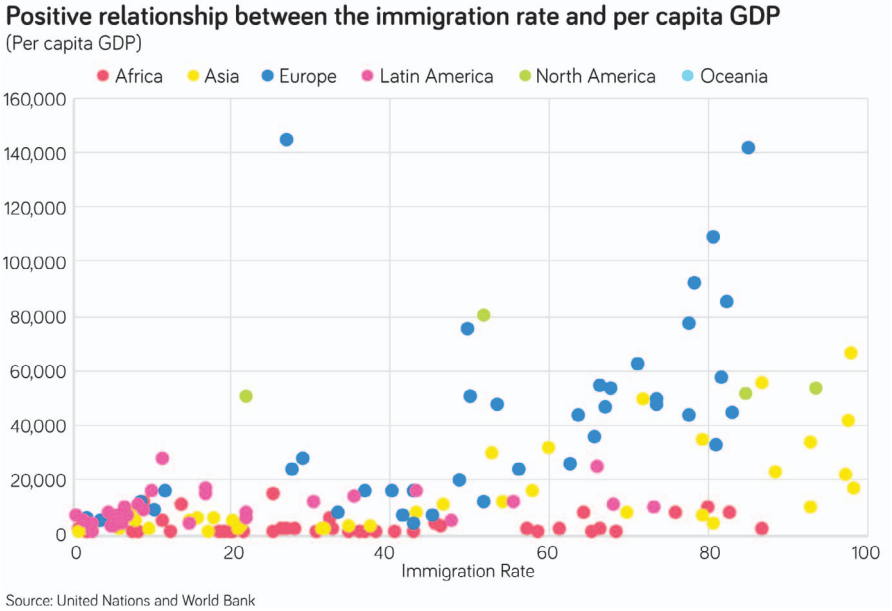


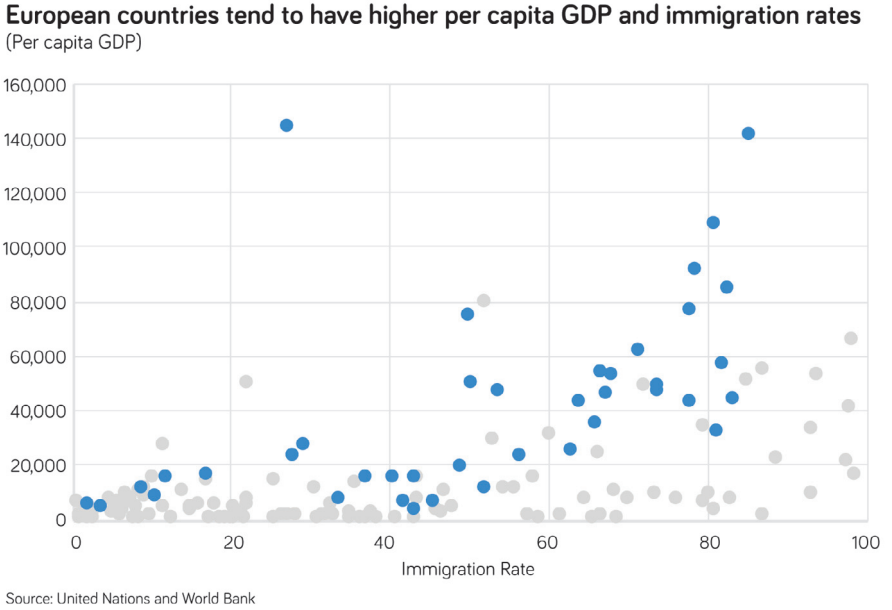
As before, more colors can be added to denote another variable, such as region of the world.

Labeling certain points or groups of points—with text, color, or enclosing shapes—can help the reader navigate the chart and draw their attention. The plot above added color to denote regions of the world. The next two employ the same strategy and include the more than two hundred countries around the world. Using color like this lets the reader identify certain regions. If instead we want to highlight one specific region, we might use a single color for a region of interest and use gray to push the others to the background.

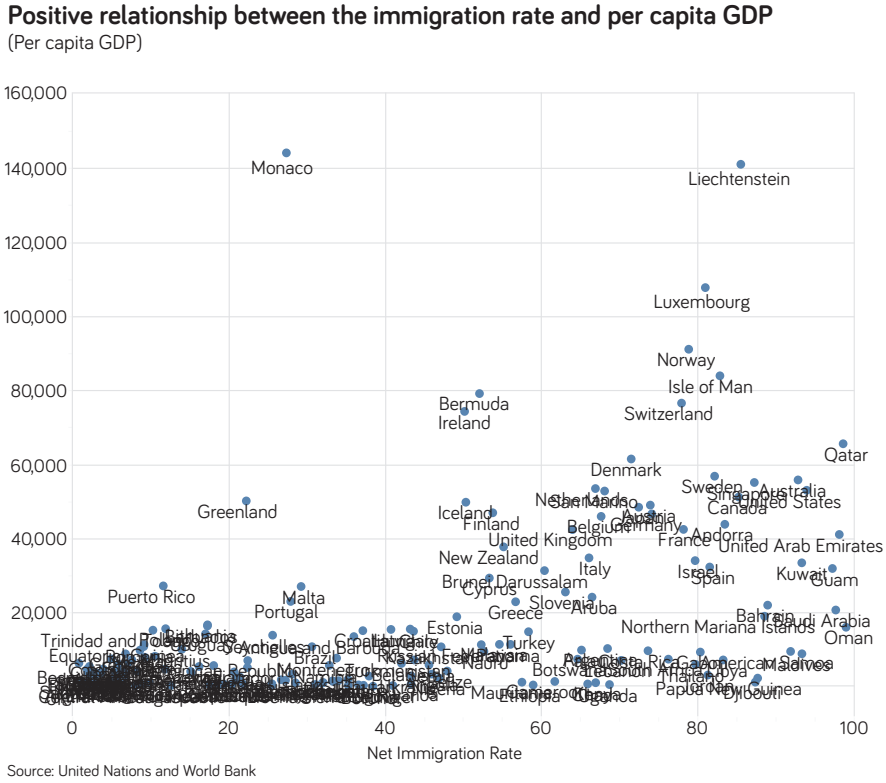
Two final points about scatterplots:

First, you will often see scatterplots that include labels for every single point, like the one on page 260. The end result is overwhelming clutter, with overlapping labels that are impossible to read. Luckily, we are far beyond the time where labels are the only way to convey information. If you believe there are readers who want to know the exact position of some of the points you didn't explicitly label, you can post a data file online, or create an interactive version using a tool like Tableau or PowerBI. Many academic researchers, for example, have an author page or webpage on their university website, as do many academic journals. These are great places to post the underlying data for your graphs.





As we've seen elsewhere, color can be used strategically to highlight different groups (for example, regions of the world as in the top graph) or to highlight a single group or data point (as in the bottom graph).

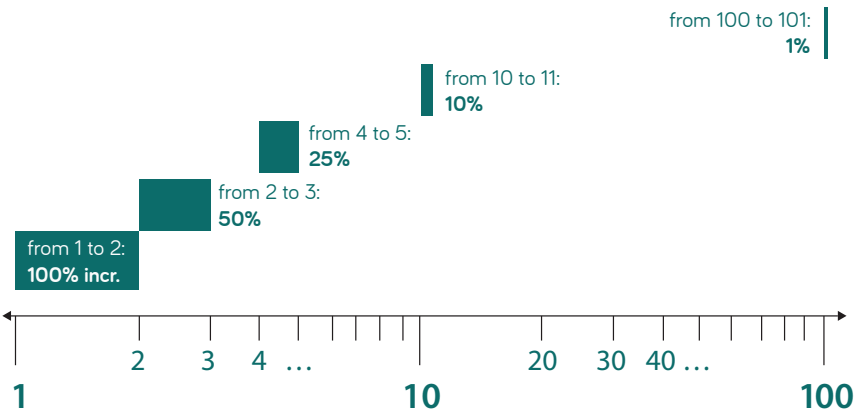


There's rarely a case where labeling all of the data points is necessary. Reduce clutter like this so your readers can better see the data.

Second, there may be times when normalizing your data, calculating percent change, or taking the logarithm of your data may improve the visual clarity of your graph. This is especially true when our data are clustered too densely in a visual. The logarithm (or log for short), is, simply put, an exponent written in a different way. Using mathematical laws of exponents, the log transformation shows relative values instead of absolute ones. Visualizing log data can make highly skewed distributions appear less so.

In a log scale, the fact that 101 minus 100 and 2 minus 1 are the same doesn't matter. Instead, what matters is that going from 100 to 101 is a 1 percent increase and from 1 to 2 is a 100 percent increase. Thus, on a log scale, going from 100 to 101 is about 1 percent of the distance as going from 1 to 2.

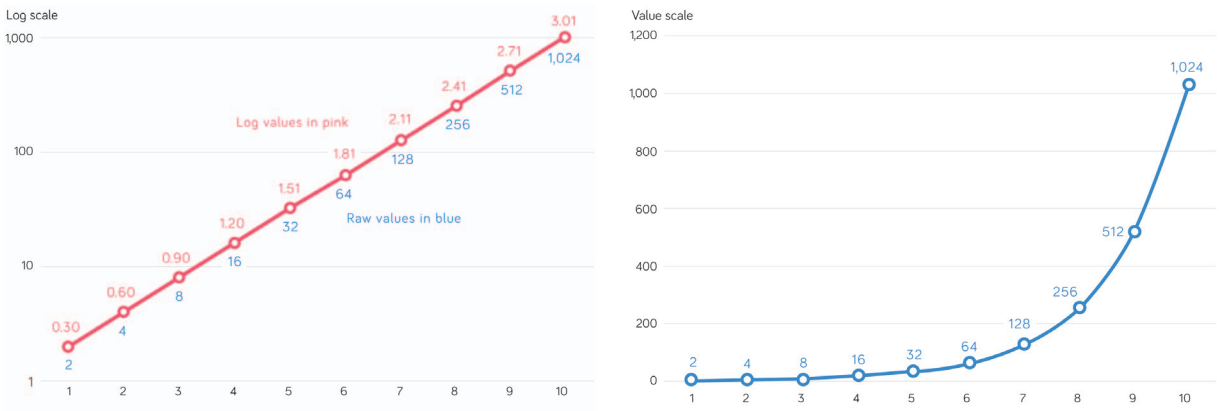
Another way to understand the differences between absolute (level) and relative (log) values is to graph some sample data. The graph on the left shows a simple doubling of each



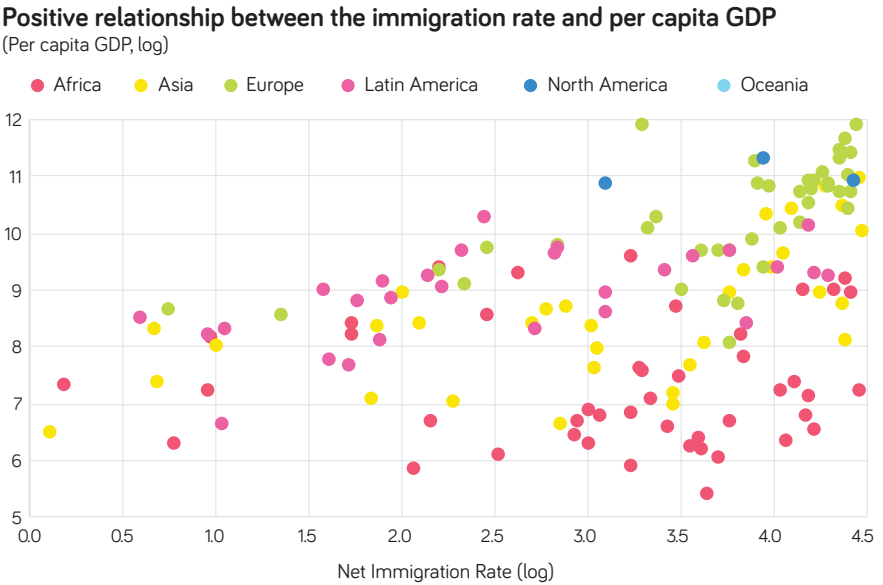
On a log scale, going from 100 to 101 is about 1 percent of the distance as going from 1 to 2. Based on Lisa Charlotte Rost (2018).

number—2, 4, 8, 16, 32, and so on. In this linear scale, we see what is called an "exponential" curve as the difference between each sequential value gets farther and farther apart. By contrast, in a logarithmic plot (the graph on the right), each gridline represents a tenfold increase over the previous one: 1, 10, 100, and 1,000. In this representation, the same numbers appear as a straight line as opposed to a curved one, even though the growth rate is the same.

In the GDP-immigration scatterplot, there are many countries clustered around the origin, with both low per-capita GDP and net immigration rates. By taking the logs of both variables, the data are more spread out across the graphic space. The tradeoff is that logs (or other transformations, for that matter) are not immediately intuitive. Knowing that per



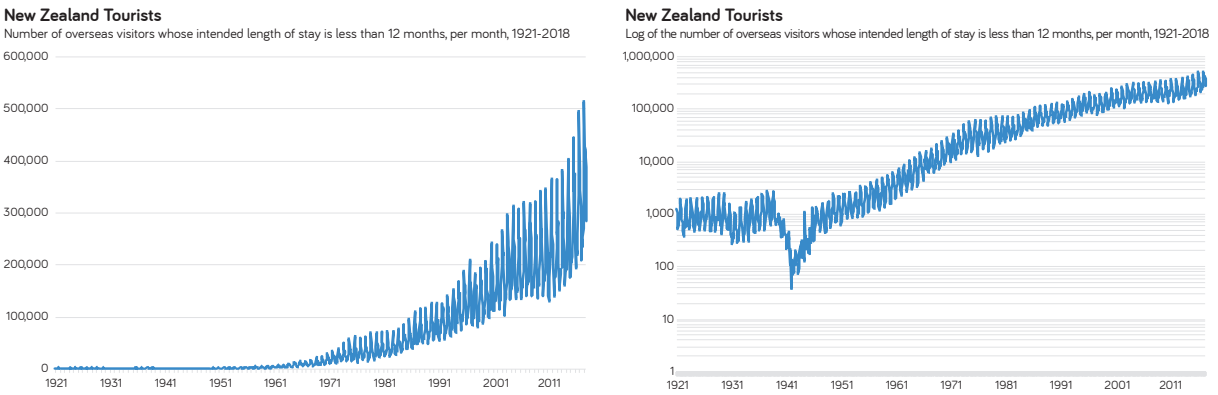
Logarithmic values are useful to show relative values rather than absolute values.



By taking the logs of both variables, the data are more spread out across the graph.

capita GDP in Luxembourg is $107,865 (in U.S. dollars) is a number we can all understand, but we don't as easily grasp that if we write it as the log per capita GDP: $11.59.

Here's another example of how to use a log scale. This one uses time series data so we can see *relative* changes over time. Lisa Charlotte Rost, a designer and blogger at the online data



We can't see the drop in the *number* of visitors to New Zealand during World War II, but when we convert the data to logs, the *change* becomes apparent. Based on Lisa Charlotte Rost (2018).

visualization tool Datawrapper, used New Zealand tourism data to demonstrate how the log transformation works and how it can affect our view and understanding of data showing changes over time. In the line chart on the left, she's plotted the number of tourists per month from 1921 to 2018. We can see the relatively flat pattern from the beginning of the period to about 1970, when the number of visitors starts to rise. The version on the right uses log values and thus shows the *relative* number of visitors. Here, we can see a clear drop in the early 1940s during World War II. In the graph on the left, there isn't as clear a decline in *absolute* numbers (the number of tourists fell from about 2,000 in early 1939 to fewer than 100 in 1942) but there was a sharp decline in *relative* numbers.

Whether it's appropriate to transform your data is largely a function of the question you want to answer. Are you after relative or absolute values? Percent changes or levels? There is no right or wrong answer to this question, but each has their tradeoffs.

PARALLEL COORDINATES PLOT

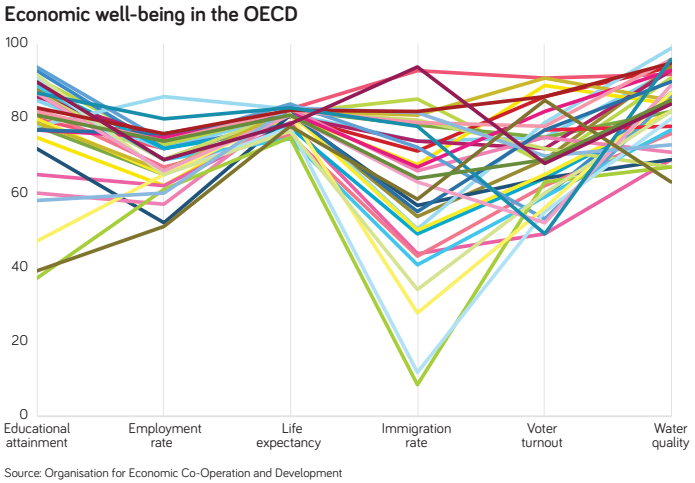
A scatterplot has data along two variables represented by a horizontal axis and vertical axis. Sometimes, though, we have more variables to visualize. That's where the parallel coordinates plot comes in.

In these charts, the data values are plotted along multiple vertical axes and connected by lines. As in the scatterplot, the axes can have different units of measurement, or they can be normalized—for example, as percentages—to keep the scales uniform. Thus, instead of visualizing a single correlation between two variables, the parallel coordinates plot permits multiple correlations within a single view.

As an example, the parallel coordinates plot on the next page shows correlations across six different variables related to migration for thirty-two countries around the world. Each vertical axis represents a different variable—like educational attainment, employment rate, and life expectancy—and each line represents a different country.

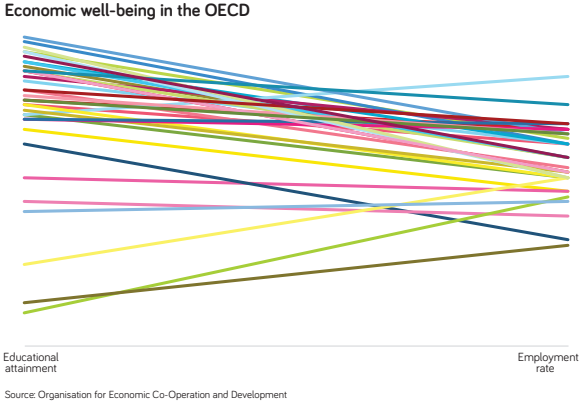
But this graph is really, really hard to read! There are too many lines, all different colors, and all crossing at different places. But before we try to address the challenges of the full parallel coordinates plot, let's try simplifying.

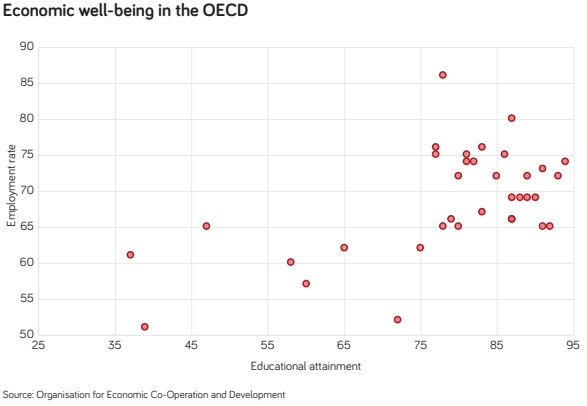
If we zoom in for a moment on the first two axes, we see perhaps the simplest parallel coordinates plot, one that resembles a slope chart. (I differentiate this from the slope chart in Chapter 5 because slope charts show changes over time, while parallel coordinate plots



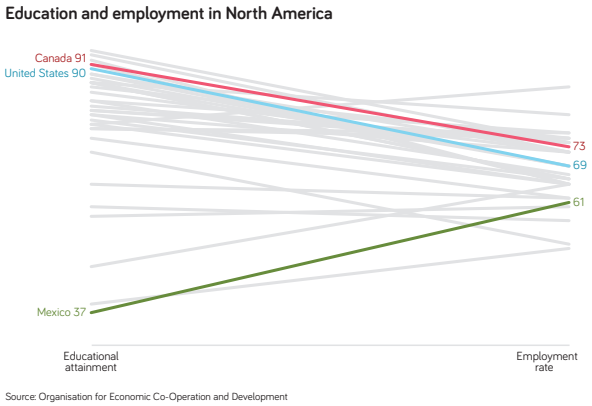
The parallel coordinates plot shows correlations between two or more variables across multiple vertical axes.

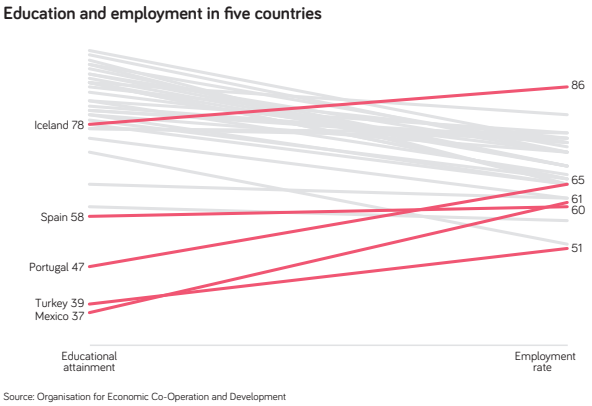
are used to compare different variables.) Here, I plot the relationship between educational attainment and the employment rate. Because the lines (countries) at the top of the left axis (education) are also near the top of the right axis (employment rate), these two variables are positively correlated. (It's not that the lines slope down, but the relative position of the points on each axis.) This is also clear in the scatterplot shown to the right. As always, which chart you use depends on your purpose and audience.





It's much easier to see the relationship between two variables in a parallel coordinates plot with just two axes (similar to a slope chart). An alternative visual approach is the scatterplot.





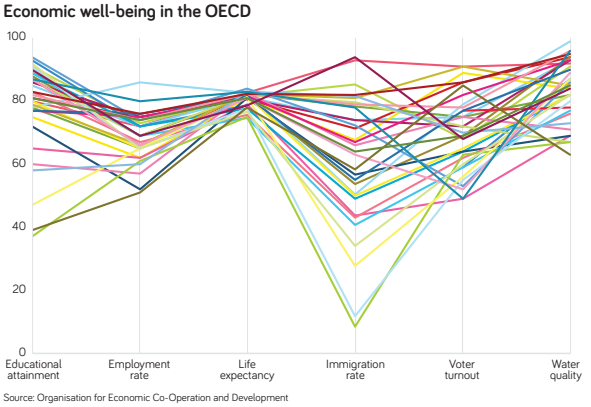
As with the slope chart, you can use different colors, line thicknesses, or other visual elements to highlight areas or values.

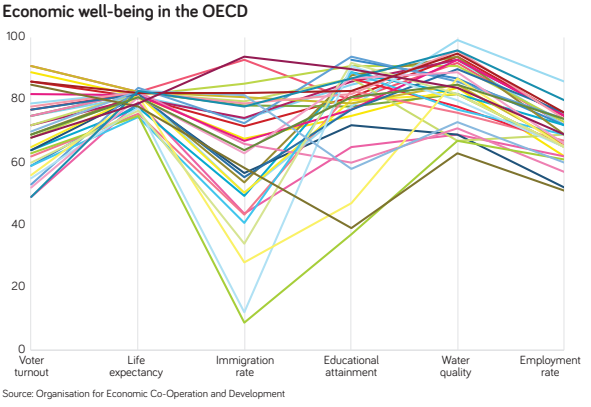
As with the slope chart, you can use different colors, line thicknesses, or other visual or textual elements to highlight certain areas or values. You could, for example, highlight the North American countries (left graph) or maybe just those lines that are upward sloping (right graph).

Back to the parallel coordinates plot with all six variables shown at the top of the next page. Now that you understand how to read the chart, you can see the positive correlation between education and the employment rate in the first two axes. You can also see the positive correlation between the employment rate and life expectancy in the second and third axes. Our view of the data and the specific correlations we can most clearly identify are a function of how we organize the axes. The plot on the right changes the order of the vertical axes so we can now see the positive correlation between voter turnout and life expectancy in the first two axes, which we could not see in the original plot.

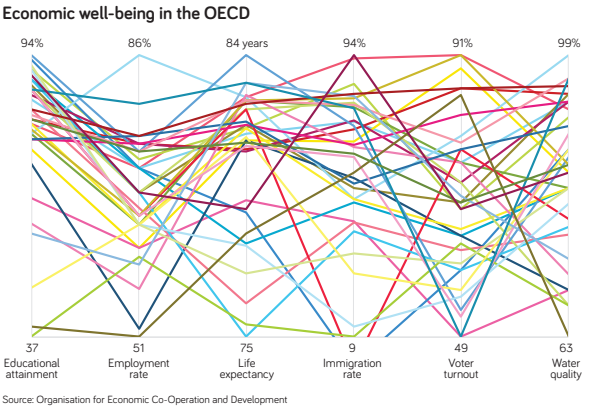
Placing all six metrics on the same vertical range also has the effect of suppressing the range (or variance) in some of these measures. For example, life expectancy varies only slightly, from 74.6 years to 83.9 years, while net migration varies more widely, from 8.6 percent to 93.9 percent. Allowing the ranges along the axes to fluctuate from one to the next (as in the middle chart on the next page) requires more labeling along each axis, but it also gives a better view of the data. The advantage of the two plots at the top is that you don't need to label every line. The disadvantage is that you suppress the variation within each variable. Notice, however, that this parallel coordinate plot—in which the range of each axis differs—looks more variable.

In sum, the challenge with many parallel coordinates plots is that they quickly become cluttered. With lots of observations (lines) and multiple axes, readers may have trouble finding the

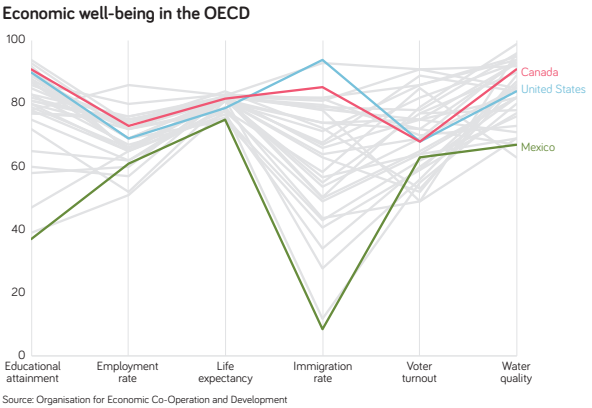




Parallel coordinates plots with too many observations quickly become cluttered.



The axes in parallel coordinate plots can differ based on the minimum and maximum of the metric.



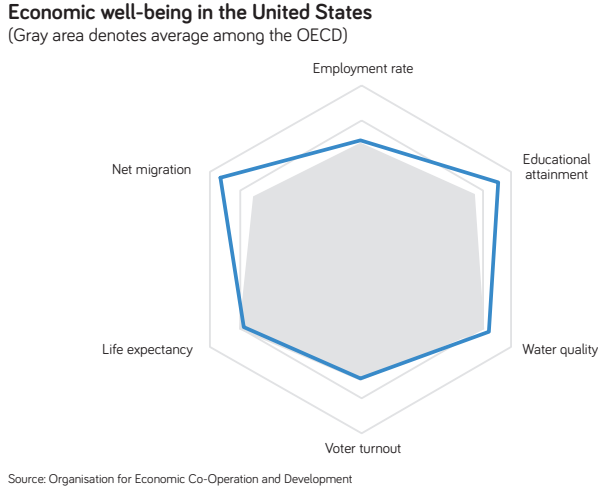
As we've seen previously, one way to simplify dense graphs like these is to use the "start with gray" strategy and add color to only a select number of observations.

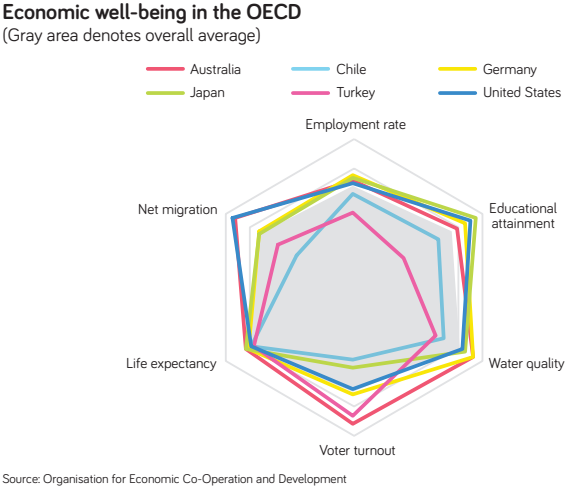
correlations and picking out specific values. One way to alleviate this difficulty is to remember the "Start with Gray" guideline: Color a group of lines gray and highlight just a subset of the data.

RADAR CHARTS

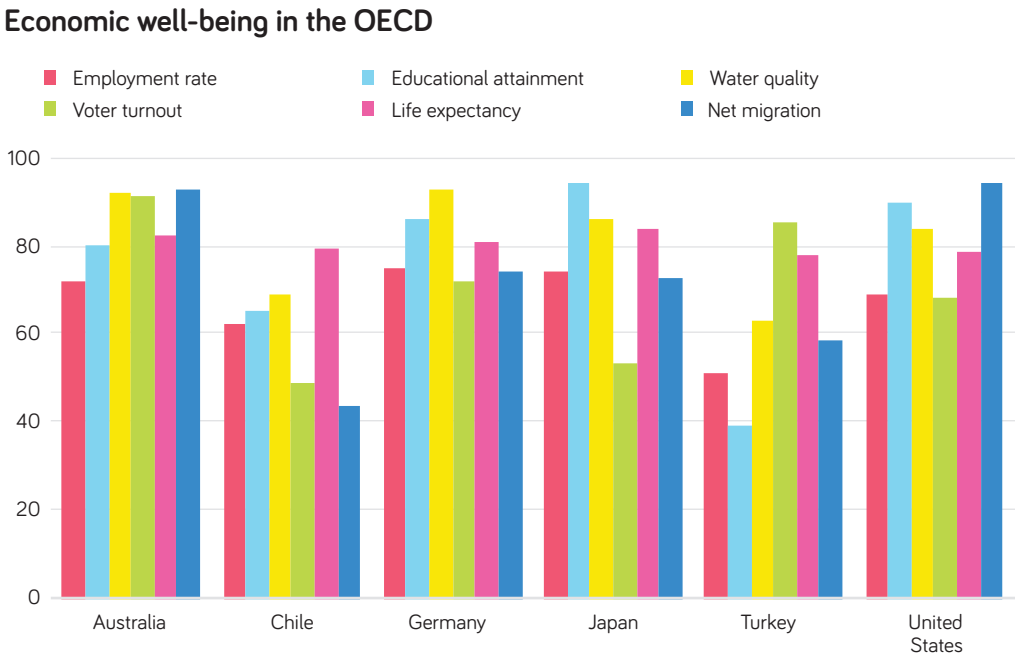
Radar charts are like parallel coordinate plots, but the lines wrap around a circle instead of being arranged parallel to one another. These are also sometimes called *spider charts* or *star charts*, and they're a good way to show multiple comparisons within a relatively compact space. Data values are plotted along separate axes that radiate from the center (the axes themselves may or may not be shown) and are connected by lines or areas to show the relationships between the different variables.

The radar chart on the left shows the same six variables used above—the line for the United States and the gray area behind it the average for the thirty-two countries shown earlier. The version on the right shows the same variables for those six countries as well as the overall average in the gray. Both charts are compact and especially good at highlighting outliers. You can quickly and easily see the shape for Turkey (the pink line) is markedly





Radar charts are like parallel coordinate plots, but the lines wrap around a circle instead of being arranged parallel to one another. The gray interior area represents the overall average for each metric.



Too many bars in a bar chart like this make it hard to pick out specific observations or patterns.

different than the other countries. It's much harder to make that observation when the data are arrayed in the paired bar chart above.

As with many charts, the radar chart gets more complicated as more lines are added, and the crossing pattern around the circle can make perception even more difficult. As was the case with the parallel coordinates plot, plotting different metrics can also make perception difficult because it requires some normalizing or other modification of the data values—in the multi-country radar chart above, again notice how the values for life expectancy bunch together.

Another strategy is to use the small multiples approach, in which a separate radar chart is created for each country or group. In this case, the small multiples version takes up more space than the original and doesn't necessarily allow easy comparison across specific countries. But it is easier to see the values for each country relative to the overall average.

A small-multiples approach lets us see the values for each country relative to the overall average.

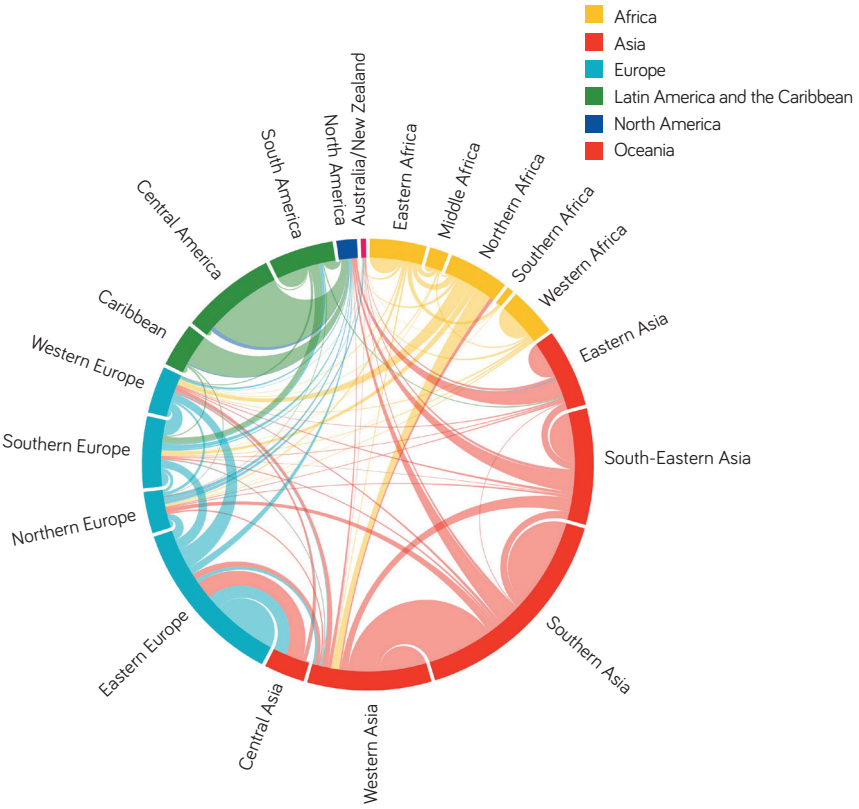
CHORD DIAGRAM

Like the radar chart, the chord diagram is another way to show associations or relationships between observations arrayed in a circle. It is perhaps best used to show how observations

have shared characteristics. In chord diagrams, observations (called *nodes*) are located around the circumference of the circle and connected by arcs within the circle to illustrate connections. The thickness of the arcs—often also differentiated by color or the transparency of the color—represent the degree of the connection between the different groups.

This chord diagram uses the same migration data used so far in this chapter to show migration flows between major regions of the world in 2017. Each region is placed along the circumference of the circle and the bands emanating from each correspond to the number of migrants entering or leaving each region. There are more than 110 values plotted in this

Migration around the world

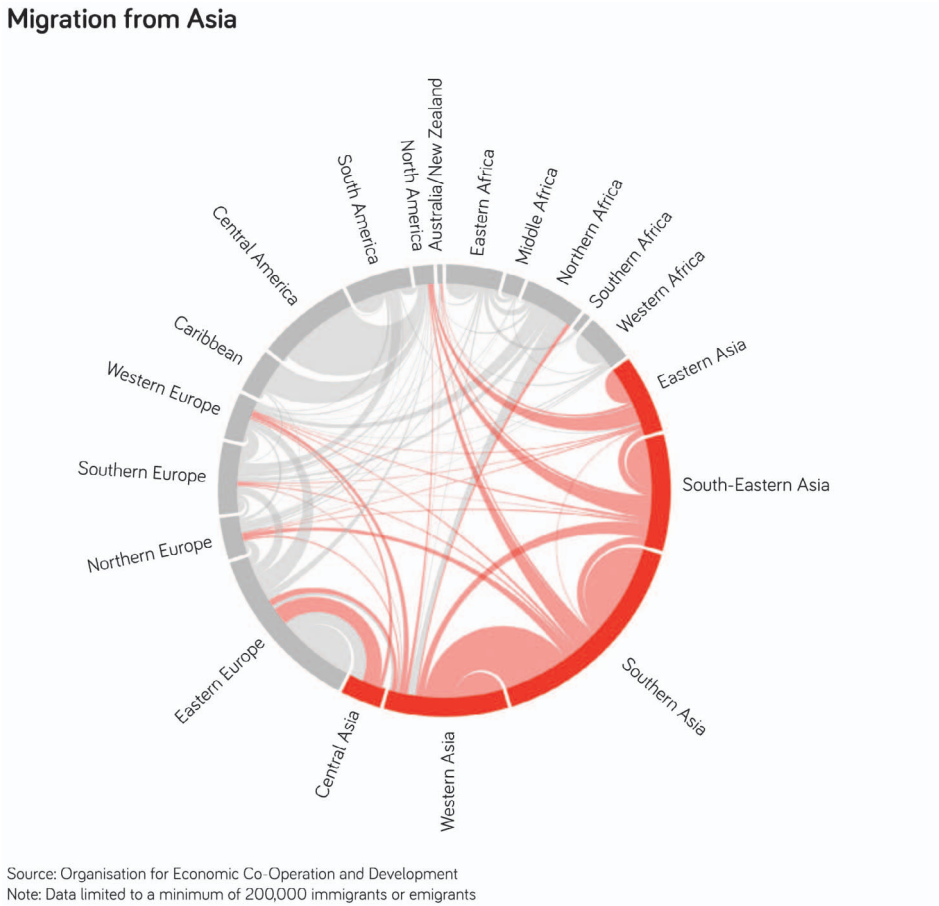


Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

In chord diagrams, the observations are located on the circumference of the circle and arcs within illustrate the connections.

single graph—though you could pick out specific values in a (very large) table, the chord diagram is clearly more visual and spatially efficient.

In the first chord diagram, you can see the large migrant flows within Asia (the red areas), and the movement between Central and North America (the thick green line to the blue segment near twelve o'clock in the circle). One danger is that the graph can quickly become cluttered and hard for the reader to easily see relationships. Again, we can use the strategy of highlighting specific groups with colors or lines. I've done that in this chord diagram in which the Asian region is red with all other regions in gray. The complexity of the chord diagram (and



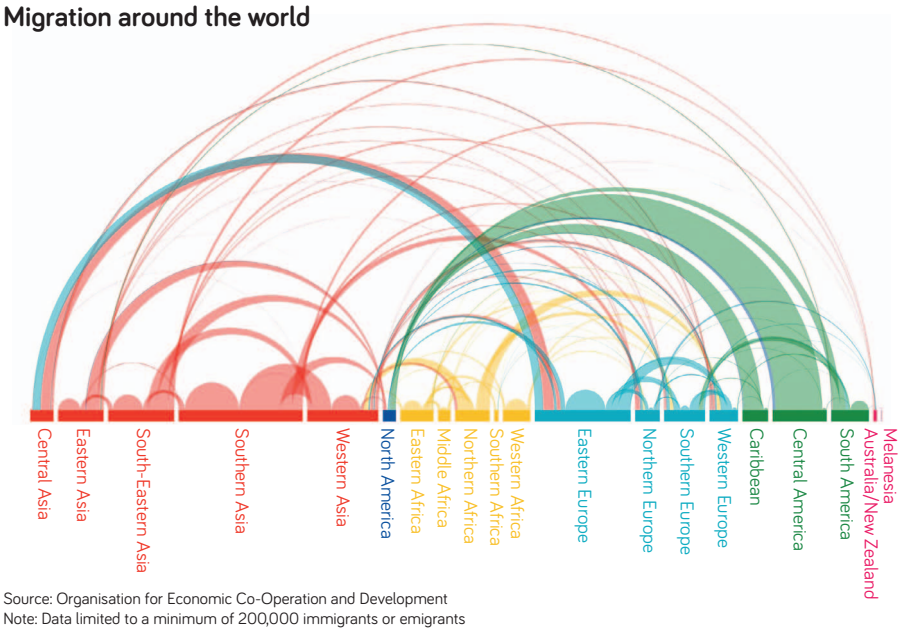
Using color strategically—especially with the color gray—can draw attention to groups or points.

its relative compactness), make them visually intriguing and invites the reader to explore the data in more depth.

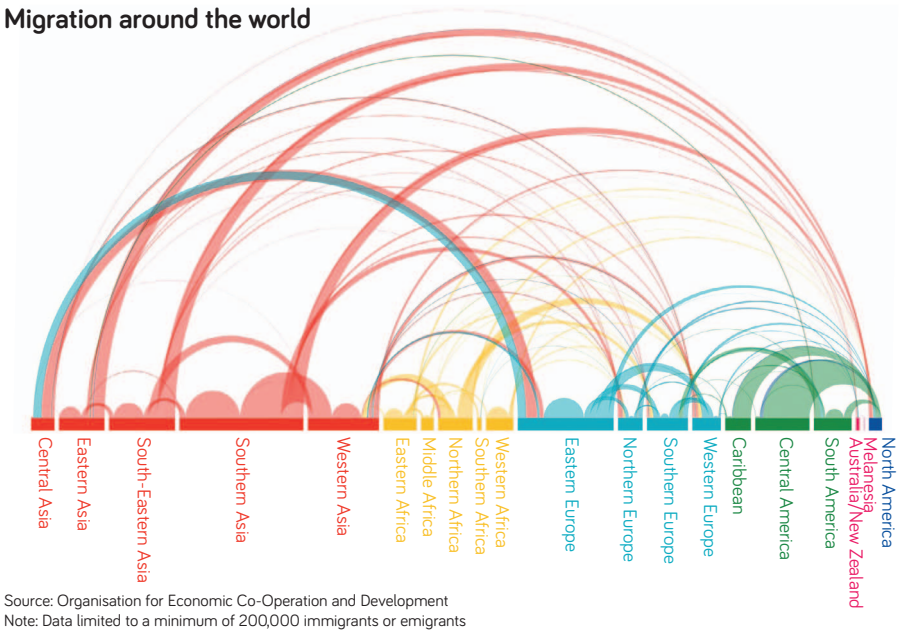
ARC CHART

Stretch a chord diagram out along a single horizontal axis and you have an arc chart. In this case, the nodes are placed along a line and are connected by arcing lines. The lines can vary in height, thickness, and color to illustrate the strength of the relationship or correlation. This arc chart shows the same migration flows between regions of the world as in the chord diagram.

A major consideration of the arc chart—and which also applies to many charts in this section—is that the order of the data can influence our perception of the results. Notice the high, wide green arcs stretching from North America to countries in the Caribbean, and Central and South America. If, by contrast, North America is placed to the far-right side of the graph and next to the other countries in the Western Hemisphere, the visualization is



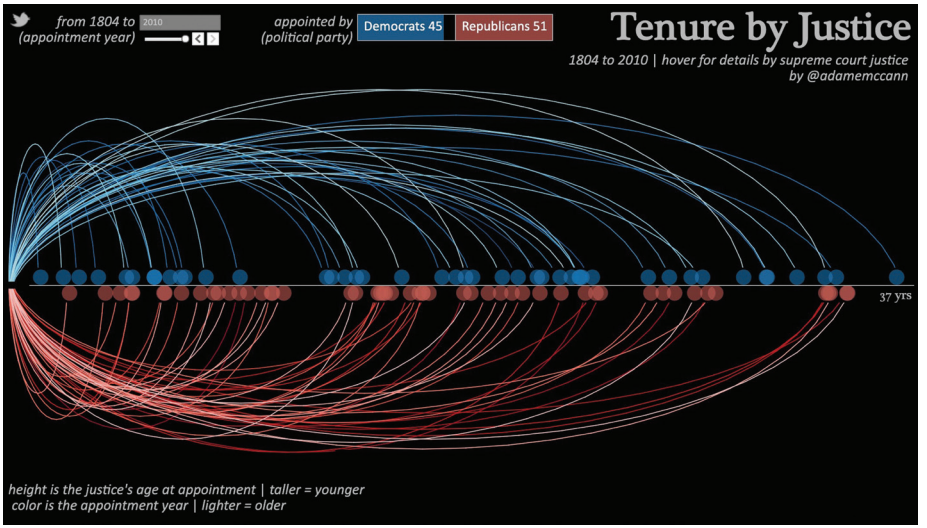
The arc chart is like a chord diagram stretched along a single horizontal axis.



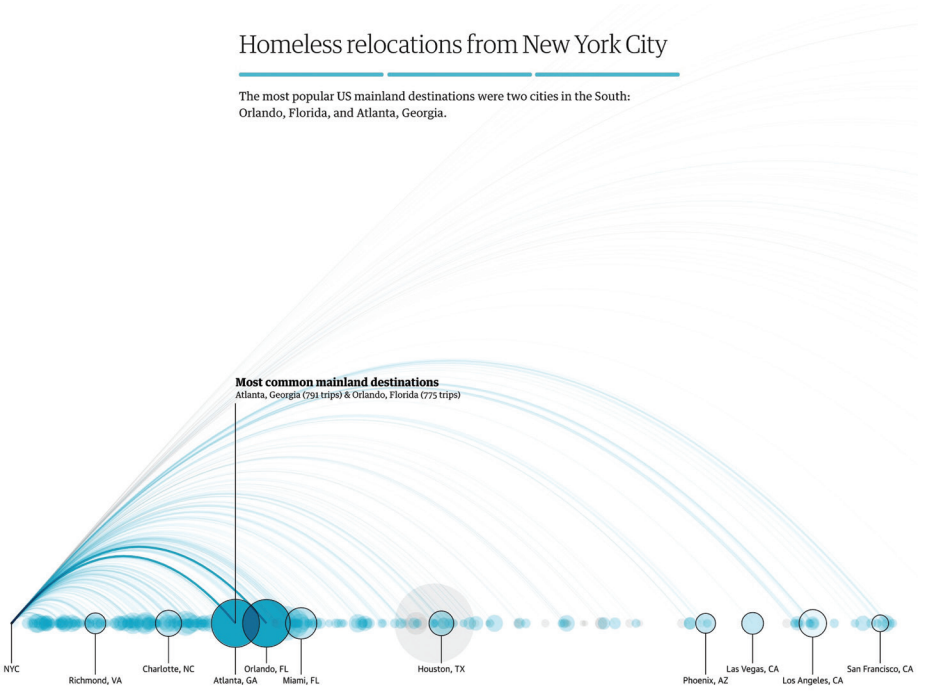
Like many graphs in this chapter, the organization of the data along the horizontal axis can affect our perception of the data. Compare this arc chart to the previous arc chart—same data, different shape.

substantially changed. There is no longer a tall green arc dominating the view, but a series of red bands that reach across the entire space between North America and countries in Asia. Some of this is obviously a function of the colors used, but the arrangement of the countries also matters. It is worth taking time to experiment with color and node placement to arrange the arc chart in the way that best communicates your argument.

A variation on the arc chart is an *arc-time chart* or *arc-connection chart*, in which connections over time are plotted in the same way. Instead of illustrating the correlation or relationship between two distinct variables, the nodes denote time. The arc-connection chart can also be thought of as an alternative to a timeline or flow chart, which we saw in Chapter 5. On the next page, the arc chart from Adam McCann shows the tenure of all Supreme Court justices in the United States since 1804. The origin (the left-most point) shows when each judge started his or her tenure, and the arc stretches to their retirement age. The height of each arc represents the age at the time of appointment (taller is younger) and the color represents their political party and year they were appointed (lighter shades are earlier years). An



Adam McCann used an arc chart to show the tenure of all US Supreme Court justices since 1804.



Arc charts can also show geographic data, as in this one from the *Guardian* showing where New York City sends their homeless population.

alternative chart type, like a bar chart or a heatmap, could be used to show the same changes, but there is something arresting about the shapes in this view.

Another way to use the arc chart is to plot distances. The arc chart from the *Guardian* shows where New York City sends their homeless population. The cities are organized by distance from New York—Richmond, Virginia, on the left and San Francisco, California, on right. The height and thickness of the arcs and the size of the circles show how many people go to each city.

CORRELATION MATRIX

A correlation matrix is a table with the variables listed along the horizontal and vertical axes. Numbers in each cell represent the strength of that relationship, often as a Pearson's correlation coefficient (see Box on page 254).

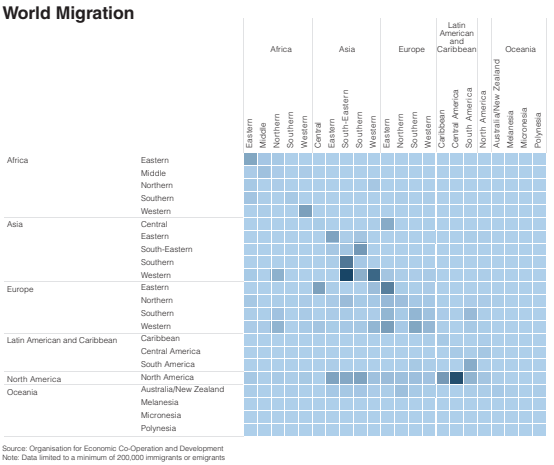
The *correlation matrix graph* uses the same layout but instead of numbers it uses shapes—often circles—to show the strength of the correlation, and sometimes color and shades to organize the table. The correlation matrix is a cousin of the heatmap, and we can also think of it as a way to add a visual element to a standard table, a topic we will visit in Chapter 11.

World Migration

		Africa					Asia					Europe				Latin American and Caribbean			North America	Oceania			
		Eastern	Middle	Northern	Southern	Western	Central	Eastern	South-Eastern	Southern	Western	Eastern	Northern	Southern	Western	Caribbean	Central America	South America		Australia/New Zealar	Melanesia	Micronesia	Polynesia
Africa	Eastern	49.0	10.0	7.1	0.6	0.1	0.0	0.1	0.6	0.0	0.1	0.0	0.1	0.1	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Middle	3.9	17.0	3.7	0.5	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Northern	7.2	1.1	3.2	0.0	0.4	0.0	0.1	0.1	0.4	9.0	0.2	0.2	0.3	0.8	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0
	Southern	14.0	2.0	0.2	7.1	0.5	0.0	0.5	0.8	0.1	0.2	0.4	1.7	1.1	1.8	0.0	0.0	0.1	0.3	0.2	0.0	0.0	0.0
	Western	0.0	1.5	0.6	0.0	58.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.4	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0
Asia	Central	0.0	0.0	0.0	0.0	0.0	4.9	1.0	0.1	0.0	1.6	44.0	0.1	0.1	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Eastern	0.0	0.0	0.0	0.0	0.1	0.3	53.0	1.8	12.0	0.0	0.2	0.6	0.1	0.3	0.0	0.0	3.6	2.2	0.4	0.0	0.0	0.0
	South-Eastern	0.0	0.0	0.0	0.0	0.0	0.0	9.2	13.0	68.0	0.2	0.0	0.6	0.1	0.2	0.0	0.0	0.0	0.9	0.4	0.0	0.0	0.0
	Southern	0.0	0.0	0.0	0.0	0.0	0.1	2.1	110.0	8.6	1.6	0.0	0.6	0.1	0.1	0.0	0.0	0.2	0.5	0.0	0.0	0.0	0.0
	Western	6.0	0.1	38.0	0.2	0.3	0.9	0.3	170.0	40.0	130.0	12.0	1.7	2.9	5.4	0.0	0.0	0.6	1.4	0.1	0.0	0.0	0.0
Europe	Eastern	0.0	0.0	0.2	0.0	0.1	56.0	1.5	0.4	1.0	21.0	100.0	4.7	2.5	3.9	0.0	0.0	0.1	0.5	0.0	0.0	0.0	0.0
	Northern	8.5	0.7	2.0	2.4	4.5	0.4	5.0	23.0	5.9	8.8	26.0	20.0	8.9	8.8	2.4	0.3	2.7	4.2	2.2	0.1	0.0	0.0
	Southern	2.0	2.4	15.0	0.3	5.3	0.5	4.1	6.2	2.3	3.2	32.0	5.8	31.0	16.0	3.8	1.7	25.0	1.9	0.6	0.0	0.0	0.0
	Western	4.2	4.0	34.0	0.5	6.5	12.0	4.8	8.7	8.5	31.0	60.0	7.7	49.0	29.0	2.0	0.7	7.3	3.6	0.6	0.0	0.0	0.0
Latin American and Caribbean	Caribbean	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.2	0.3	1.4	7.1	0.2	1.0	2.6	0.0	0.0	0.0	0.0
	Central America	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.1	0.0	0.1	0.1	0.1	0.5	0.4	0.5	6.5	2.1	9.8	0.0	0.0	0.0	0.0
	South America	0.0	0.1	0.1	0.2	0.1	0.0	1.8	0.1	0.1	0.6	0.5	0.5	8.0	1.6	0.8	0.5	42.0	1.3	0.0	0.0	0.0	0.0
North America	North America	8.2	1.3	6.7	1.5	8.1	1.2	53.0	47.0	53.0	17.0	23.0	18.0	19.0	15.0	66.0	160.0	34.0	12.0	1.6	0.7	0.2	0.4
Oceania	Australia/New Zealand	1.6	0.1	0.8	2.5	0.2	0.0	8.8	8.3	10.0	3.1	1.9	18.0	6.8	3.7	0.1	0.2	1.3	2.1	7.4	1.4	0.0	1.5
	Melanesia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.1
	Micronesia	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0
	Polynesia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.2

Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

The basic correlation matrix is a table with numbers that show the strength of the relationship between observations.

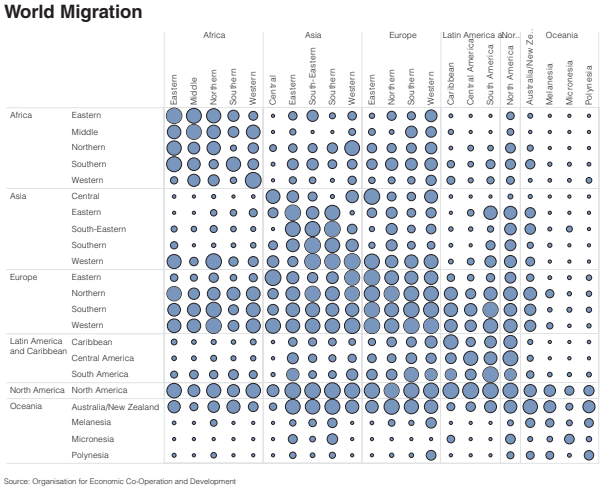


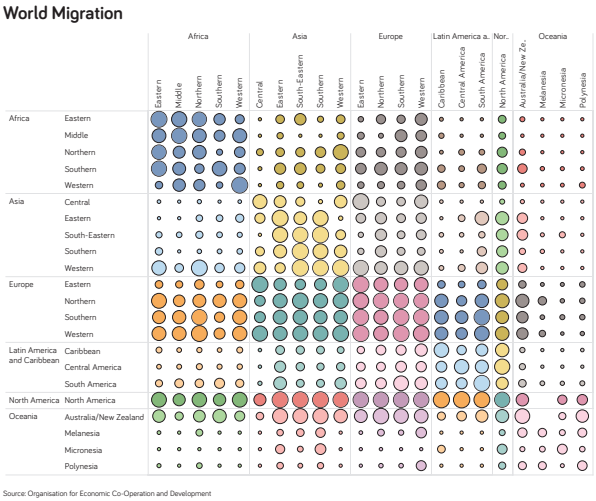
A heatmap approach to the matrix makes the patterns clear without (in this case) showing the numbers.

These two matrices show the relationship between immigrants (those entering each region) and emigrants (those leaving each region) across the world in 2017. The view on the left is the standard correlation matrix shown as a table. This gives us all the detail we would need to understand the exact correlation between different variables, but it's difficult to navigate. There are a lot of numbers, and the important values don't stand out. The matrix on the right is a heatmap (see page 112), which loses the detail of the table version in favor of highlighting the stronger (positive) correlations, especially within Asia. We could include both the colors and numbers, but the view might end up looking cluttered and busy.

The next two visuals display the data as standard correlation matrix graphs. Circles represent the strength of the relationship, and color (in the version on the right) helps organize each area, though in this case there may be too many colors. It can be hard for the reader to clearly see differences because humans are not very good at assessing quantities from the sizes of circles. In both cases, the circles are sized to fit within each cell, but that doesn't necessarily need to be the case. We could make the circles larger to fill the entire space and use transparent colors when they overlap.

One final consideration with any correlation matrix or table is that the values along the diagonal are, by definition, equal to one. That is, migration between Eastern Africa and Eastern Africa is the same. This means they are often left out because they can visually dominate or clutter the visual.





An alternative to the correlation matrix table is to use circles or other shapes, to which color can be added to visually organize the space.

NETWORK DIAGRAMS

We now enter a class of graphs for which I use the term *diagram* instead of graph, plot, or chart, largely because some of the decisions about layout and structure are not always determined by math or the data but by what looks best and is most clear. These diagrams are used to show hierarchies and connections within and across groups and systems. The thickness of the lines and size of the points can be sized according to data values to signal the strength of those relationships, and arrows can visualize movements inside groups and communities. Consider a family tree: the lines show links between parents, siblings, spouses, and children, but the connecting lines and the pictures or names of family members are not scaled according to a data value.

We start with the standard network diagram, which shows connections between people, groups, or other units. Generally speaking, the points in a network diagram (called *nodes* or *vertices*) denote the individual person or observation, and the lines (called *edges*) link them together and show the relationship. The position of the nodes and the length (and sometimes thickness) of the linking lines illustrates the strength of the relationship. While nodes are often depicted as circles, you could also use icons, symbols, or pictures.

The ultimate appearance and organization of a network diagram depends on the kind of network we want to visualize and the method with which we arrange the nodes and edges. When creating a network diagram, we must be careful about how edges cross and nodes overlap. In general, we want to achieve some kind of visual harmony in the visualization by finding a uniform and meaningful length of the edges and some symmetry for the entire graph.

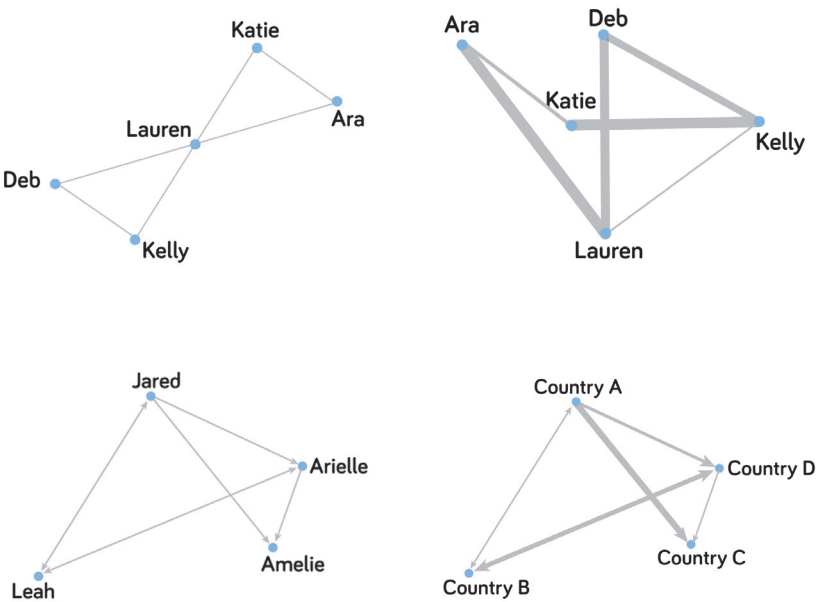
To start, we can distinguish between four different kinds of network diagrams:

1. *Undirected and Unweighted*

   Lauren, Ara, and Katie are friends. Lauren is also friends with Deb and Kelly.
2. *Undirected and Weighted*

   Researchers in this diagram are connected if they published a paper together. The thickness of the line is the number of times they have published together.
3. *Directed and Unweighted*

   Jared follows Leah, Amelie, and Arielle on Twitter, but only Leah follows him back. Leah and Arielle follow each other, and Arielle follows Amelie. The connection is not weighted—they are either connected (in one or more directions) or not.

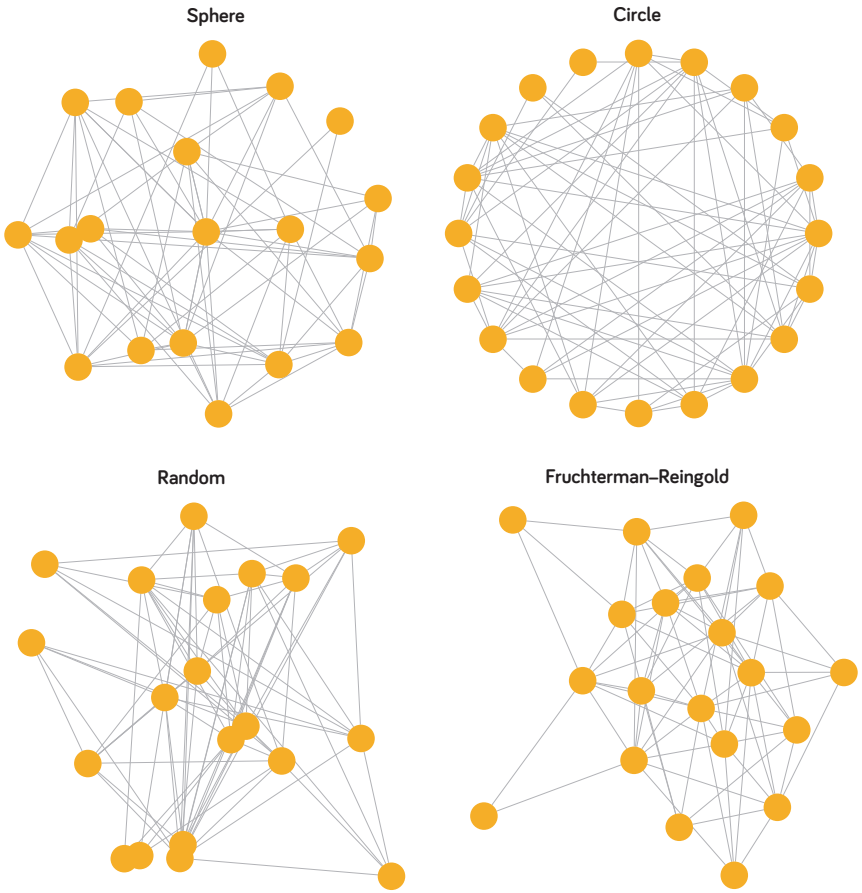


Four kinds of network diagrams, clockwise from top-left: Undirected and Unweighted; Undirected and Weighted; Directed and Unweighted; Directed and Weighted.

4. *Directed and Weighted*

People migrate from one country to another. The thickness of the line is the number of people migrating and the direction is the destination.

There are many algorithms we can choose from to lay out the nodes and edges in a network diagram. Usually, network algorithms try to minimize how often the edges cross one another and prevent overlap of the nodes. Generally, we want the edges in a network diagram to be of roughly uniform length and the vertices to be distributed evenly. Using example data

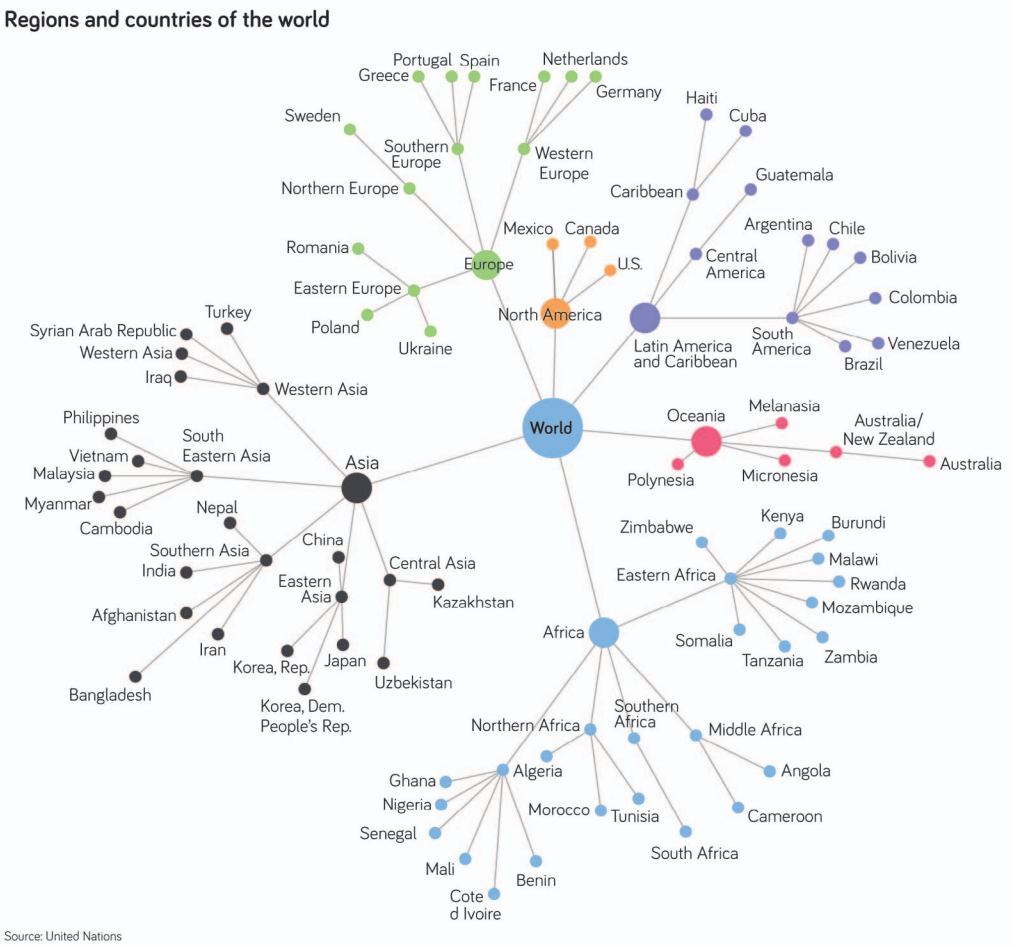


Four types of algorithms to create a network diagram.

Source: Based on the R Graph Gallery

around twenty points, these four network diagrams demonstrate how select organizing algorithms will generate different views of the network and the relationship between the points.

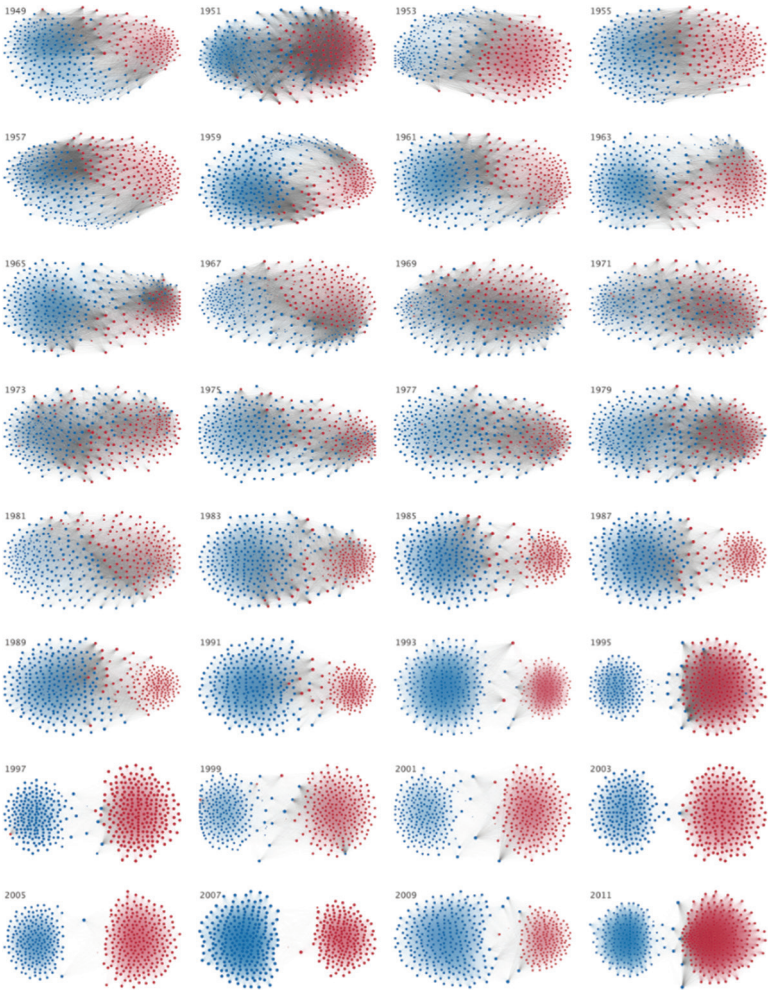
This network diagram shows the relationship of the seventy-five or so most populous countries in the world (those with more than one million people) within their different geographic regions. I'm not arguing that this network diagram is a better visualization than a standard geographic map, but I show it here because you can easily understand the content



A simple network diagram that shows the arrangement of countries with more than one million people.

and see how the diagram works. Imagine showing the links between people in your Twitter or Facebook network, grouped by family, friends, and coworkers.

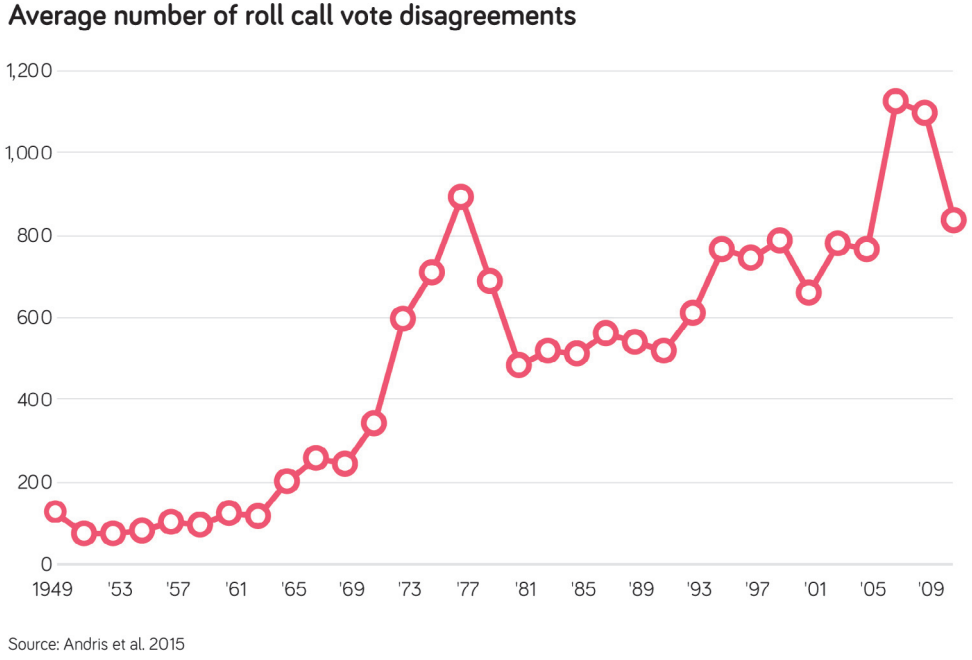
Network diagrams are ideal for showing the structure and relationship between different agents in a system. In some cases, groupings or concentrations become clear as specific nodes cluster near one another. We can use color or other shapes to highlight specific groups within the larger network.



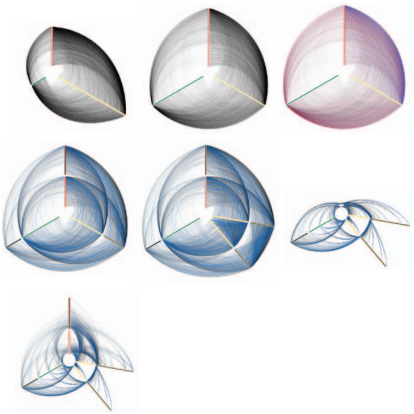
This small multiples set of network diagrams from Andris et al. (2015) shows voting behavior in the United States Congress.

As with all other charts, including too much data in a network diagram can clutter it and make them difficult to read. But unlike many charts, sometimes the goal of a network diagram *is* to show the dense clustering. The set of thirty-two network diagrams from Clio Andris and her collaborators shows the polarization of voting behavior in the United States Congress. The authors created network diagrams for each U.S. House of Representatives from 1949 (top-left) to 2011 (bottom-right). Republican members are represented by red dots, and Democrats by blue dots. The lines denote how often members of Congress voted with one another. Even though each network diagram is very dense, the use of these small multiples makes it clear that the two parties were much less likely to vote together in 2011 than in the past.

By comparison, the line chart below shows the measure of disagreement between the parties in a simple, straightforward way. Though it's immediately informative, it's not as visually stunning as the network view.

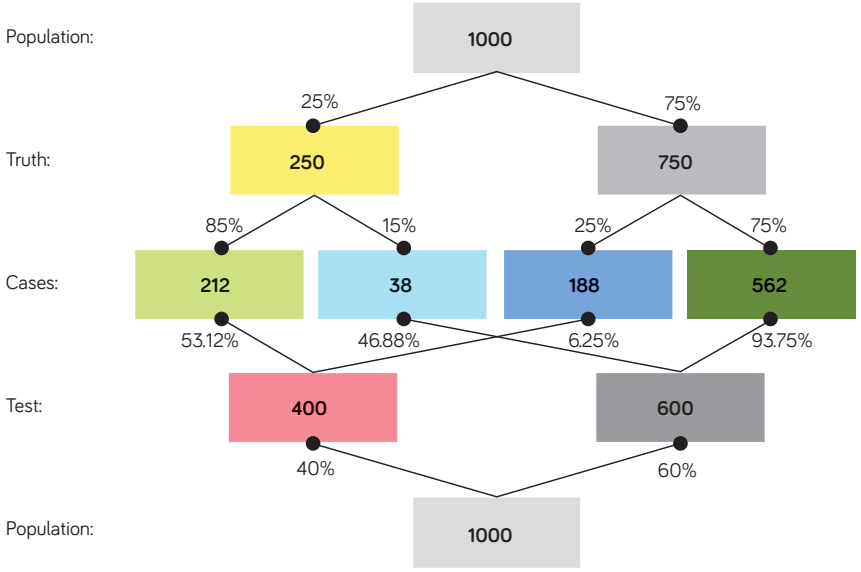


We can make a similar case about voting behavior in the United States Congress using summary data from Andris et al. (2015), but the line chart probably doesn't grab you the way the small multiples network diagrams did.



Martin Krzywinski's hive charts are another way to show networks.
Source: Canada's Michael Smith Genome Sciences Center.

Because network diagrams can look like hairballs with too many edges and nodes to make the visualization readable, some researchers have developed alternative visualization types. The *hive plot*, for example, first organizes the space along linear axes emanating outward from a single central point. Nodes are placed along three or more axes (possibly divided into



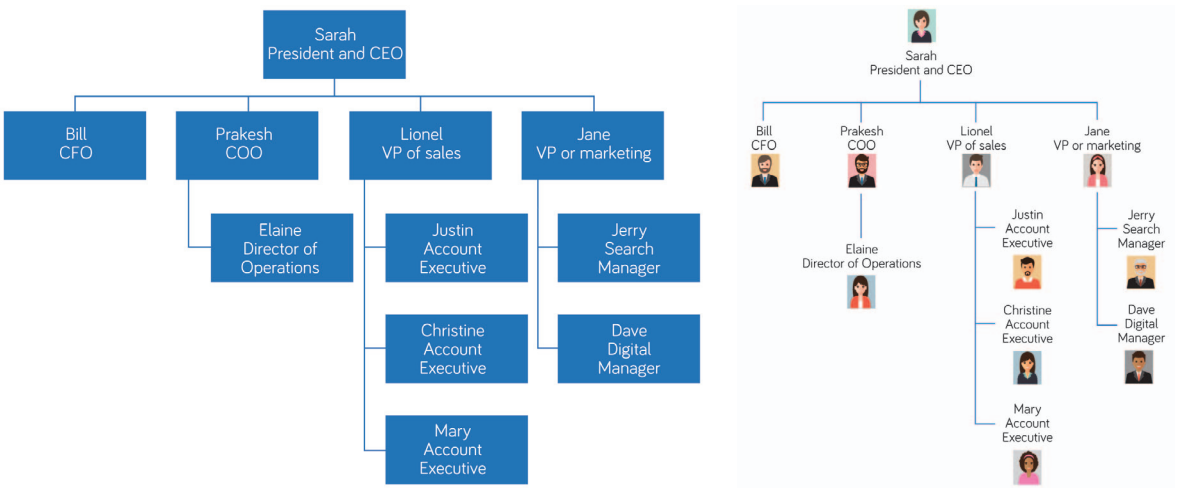
Some fields use network diagrams to visualize a process.

segments) and edges are drawn as curved links connecting the points. Martin Krzywinski, inventor of this visualization, writes that, "The hive plot is itself founded on a layout algorithm. However, its output is not based on aesthetics but network structure. In this sense, the layout is rational—it depends on networks features that you care about."

It's also worth noting that some fields refer to network diagrams in different ways. Instead of plotting how individual values correlate to one another, some network diagrams show flows or processes, similar to a flow chart or timeline. Examples might include how a computer network, a staff directory, or even a logic model of probabilities, as shown on the previous page.

TREE DIAGRAMS

Like the flow chart in Chapter 5, tree diagrams show levels of a hierarchy in a system or group. To imagine the basic tree diagram, think of a hierarchical organizational (or "org") chart. Nodes branch outward from an initial root connected by lines called *links*, *link lines*, or *branches*. The initial node is called the *root* and is the *parent* to all other nodes, some of which have child nodes of their own. Nodes who are not parent nodes are called *leaf nodes*.

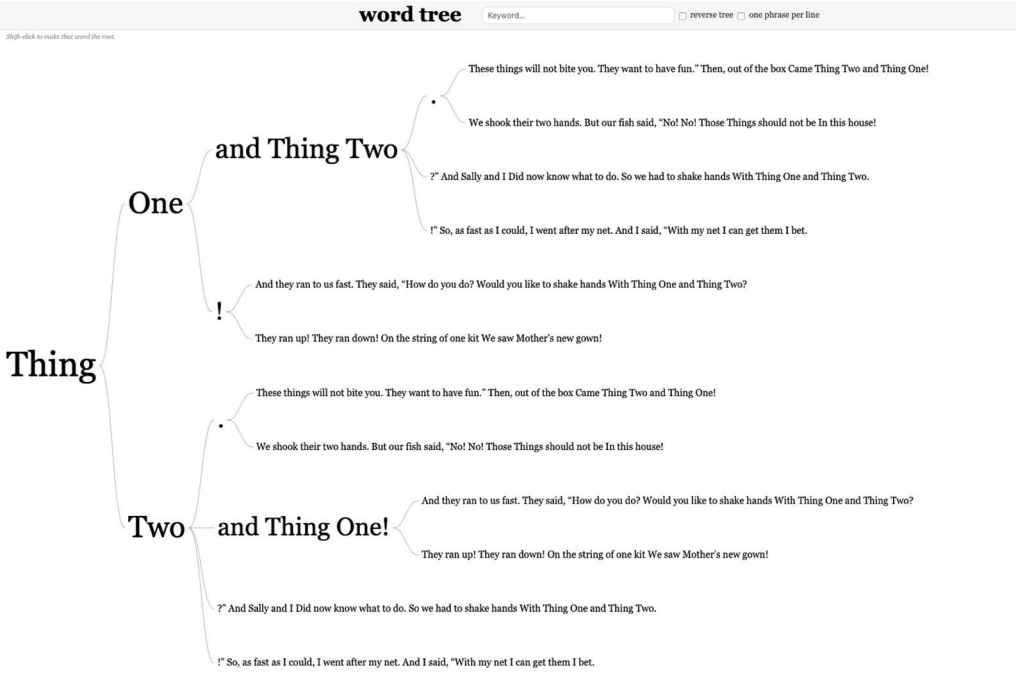


Tree diagrams describe hierarchies. These two org charts could be used for different purposes, depending on whether we think the reader would like a more designed version.

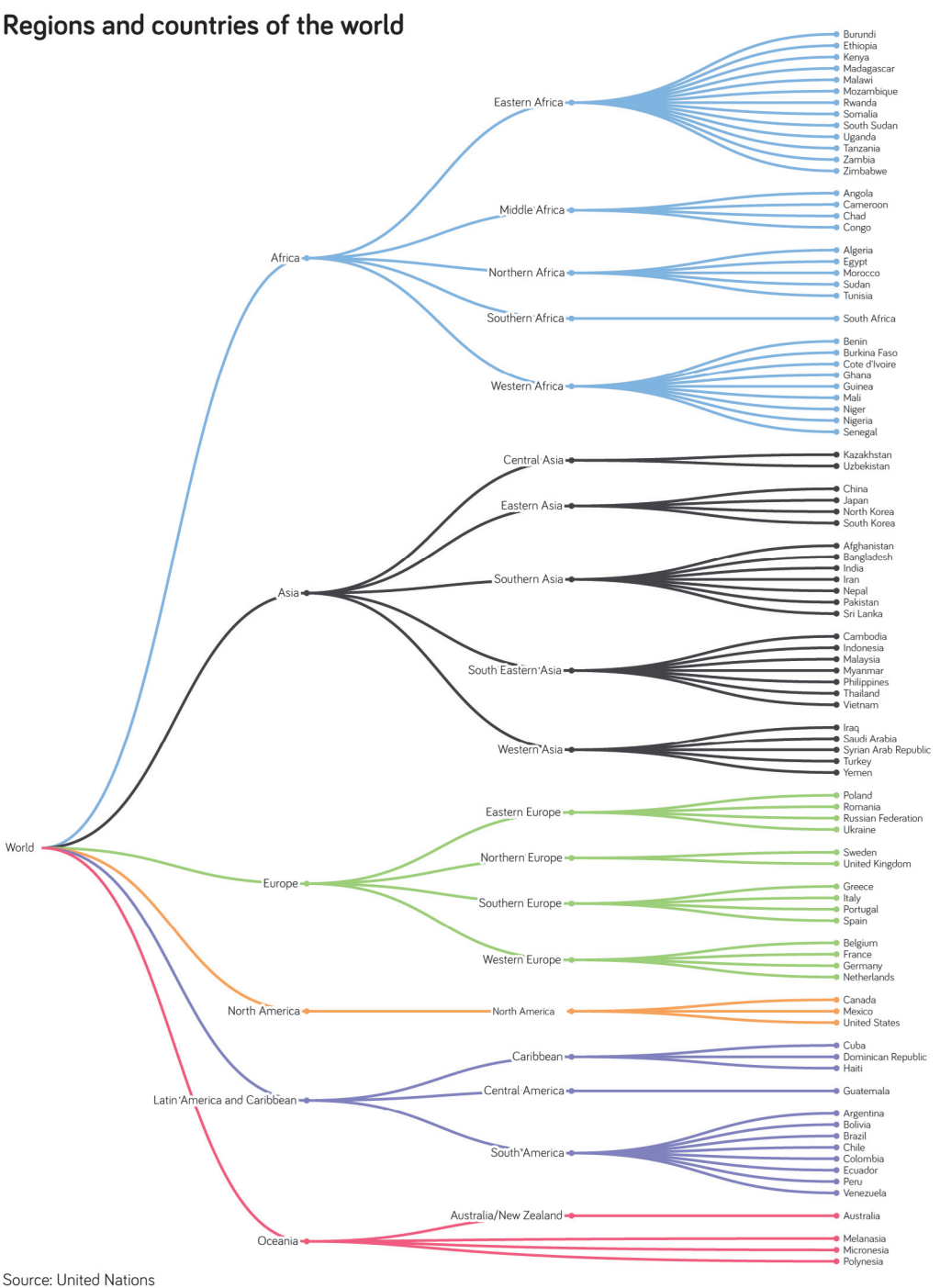
As with many of the visuals we have discussed so far, design is especially important here because there may not be much or even any data to define the elements in the diagram. Take these two imaginary org charts, for example. The chart on the left consists only of names, while the one on the right includes icons. Which one you would use depends entirely on your purpose. The one on the left might work well for the company board meeting or formal presentation; the one on the right might work better in a marketing campaign or website.

While the basic tree diagram will often branch downward, starting with the CEO or president at the top, there are lots of other ways to show these kinds of relationships. We could create a family tree that branches upward instead of downward, or a horizontal arrangement to show a different kind of hierarchy or taxonomy.

Another type of tree diagram is the word tree. Developed by Martin Wattenberg and Fernanda Viegas in 2007, the word tree is a visual representation of text in a book, article, or other passage (also see Chapter 10 on qualitative data visualization). The visualization is



This word tree is a visual representation of the text of Dr. Seuss's book, *The Cat in the Hat*.
Source: Jason Davies



Source: United Nations

A simple tree diagram that shows the breakdown of regions into countries.

typically ordered horizontally with a word on the left or right that branches out to show the different contexts in which it appears. These contexts are arranged in a treelike structure so the reader can uncover themes and phrases. Individual words, which act here as the nodes, are often sized by the frequency in which they appear.

There are lots of different kinds of trees to visualize quantitative or qualitative data. They can show complex data, like the human genome, or simple data, like the breakdown of countries into regions. What kind of tree you create and the design touches you include will—as always—depend on your purpose and audience. The tree on the previous page, for example, shows the same data as the network diagram shown earlier. On the one hand, it's not particularly efficient, but on the other hand, it might be easier to navigate than the network diagram.

CONCLUSION

In this chapter, we surveyed charts and diagrams that visualize relationships between variables, individuals, or groups. We often want to understand how two or more things are related, but remember that because two variables might be correlated does not mean there is a causal relationship. Clearly understanding how elements in your data are related before presenting them to your reader or audience is of utmost importance.

This class of graphs uses different strategies and shapes to communicate these relationships, and there are advantages and disadvantages to each approach as they trade off between clarity, order, and compactness. Scatterplots have a single horizontal and vertical axis; bubble plots add a third variable. Parallel coordinate plots are defined by using two or more vertical axes. Radar charts pull the axes together and radiate outward from a center point of a circle while a chord diagram wraps everything around the circumference of a circle. An arc chart then stretches everything out along a single horizontal axis and a correlation matrix uses a square or rectangular format. Network and tree diagrams can be used to show relationships between individuals or groups or passages of text.

As with the graphs in previous chapters, some of the graphs in this chapter may be unfamiliar, even difficult for you or your readers to understand. This doesn't require you to dumb things down or leave things out, but it should prompt you to consider how to best communicate the content of nonstandard graphs. Use labels, annotations, active titles, and helpful pointers to teach them how to read the graph so that they can more easily understand the content.

PART-TO-WHOLE

This class of charts shows how the shares of some amount relate to the total. The most popular and familiar graph in this class is the pie chart, which introduces a variety of perceptual challenges, as we'll see. Other charts in this class like the treemap and sunburst have different perceptual issues and, as always, we must ask ourselves whether we must show *all* of the components and how they sum to the total. Graphs in this chapter can also be used to visualize hierarchical data—data that can be grouped into layers where natural groups exist—which we have already seen in cases such as the tree diagram.

Graphs in this chapter are based on the online style guide from the *Texas Tribune*, a digital-first publication based in Austin, Texas. In addition to the basic styles outlining colors and fonts, the *Tribune's* style guide also includes instructions for the online elements of their website.

PIE CHARTS

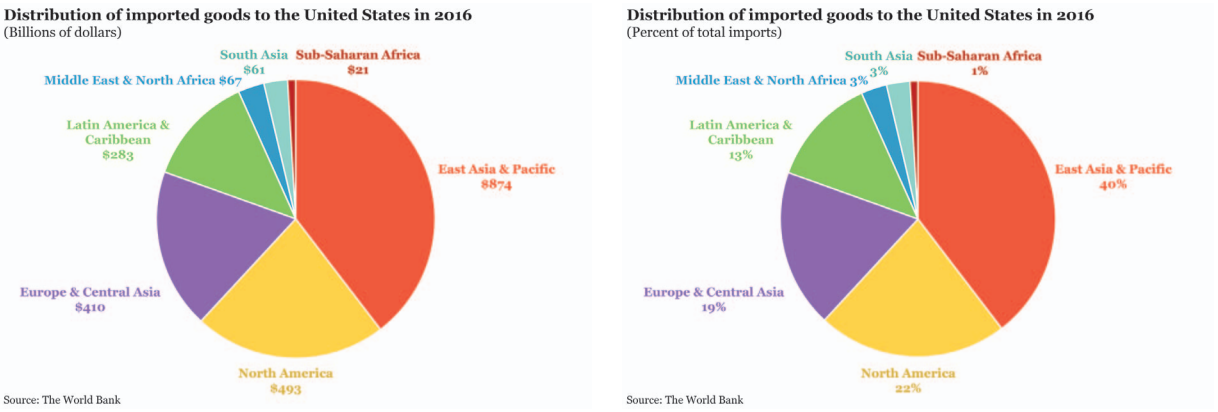
Disdain for pie charts pervades the entire field of data visualization. The most often-cited reason is that pie charts are a poor visualization choice because we have a hard time discerning exact quantities when they are visualized as slices of the pie. If we return to the perceptual ranking chart, we'll see that pie charts fall below the middle of the ranking. While the pie chart gets its fair share of complaints, it's also a very familiar chart type for many people,

and familiarity can be useful. Research has also shown that people are more attracted to curves than to objects with sharp points, and author Manuel Lima has presented evidence that affinity for the circle shape dates back millennia in human evolution.

These pie charts show the distribution of imports (the dollar value of goods flowing into a country or region) in seven areas around the world. Notice that the version on the left shows the percentages, while the version on the right shows the dollar amounts (in billions). Either approach works, just be sure to use correct labeling and note it in the title.

The most important rule for pie charts is that the slices must sum to 100 percent or at least some sort of total. You cannot leave segments out or—unfortunately more common—include segments that sum to more than 100 percent. When you arrange the slices of a pie chart, a good rule of thumb is to order them from largest to smallest beginning at the 12-o'clock position. This is often the best way, but sometime it isn't always possible or natural. For example, if you were visualizing shares of a total by age group, it would be better to start with the youngest group at 12 o'clock, then the next oldest group, and so on. In this case, it is more natural for our reader to comprehend the data when they are ordered by category rather than by value.

In situations where you might use more than one pie chart—something I would only recommend in very specific cases—always order the slices in the same way in all the charts. It becomes far too difficult to compare the slices in multiple pie charts when the slices are in different positions.



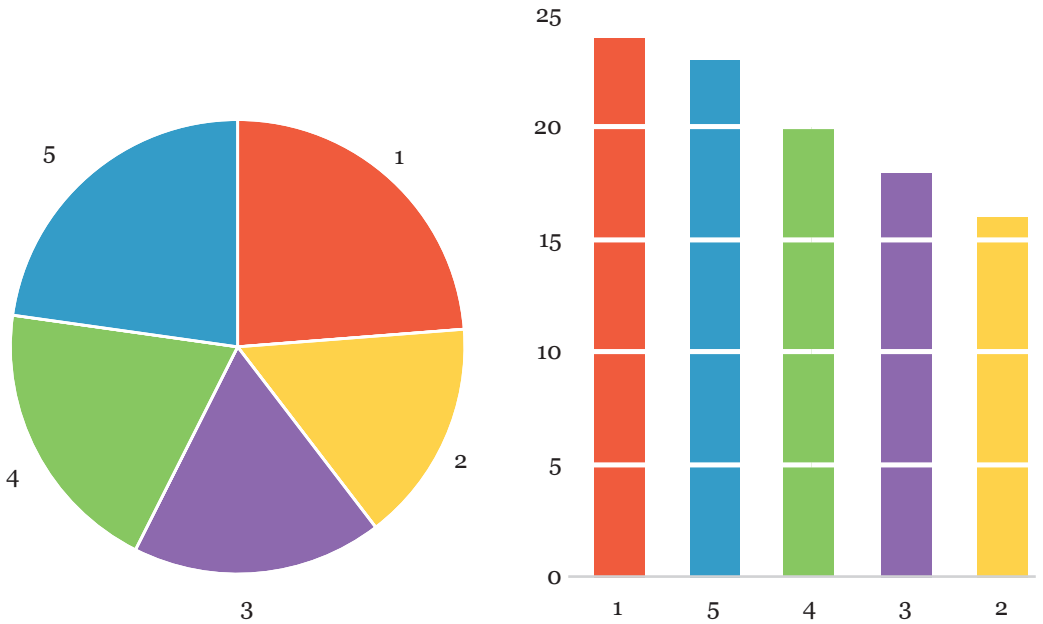
Pie charts show part-to-whole relationships. These two show the distribution of imported goods to the United States, either in dollars or percentages.

THE CASE AGAINST PIE CHARTS

The trouble with pie charts is that humans cannot easily compare differently sized slices. Here we have the same data arrayed as a pie chart and as a bar chart. See how it's easier to rank the values in the bar chart and how small differences are imperceptible in a pie chart?

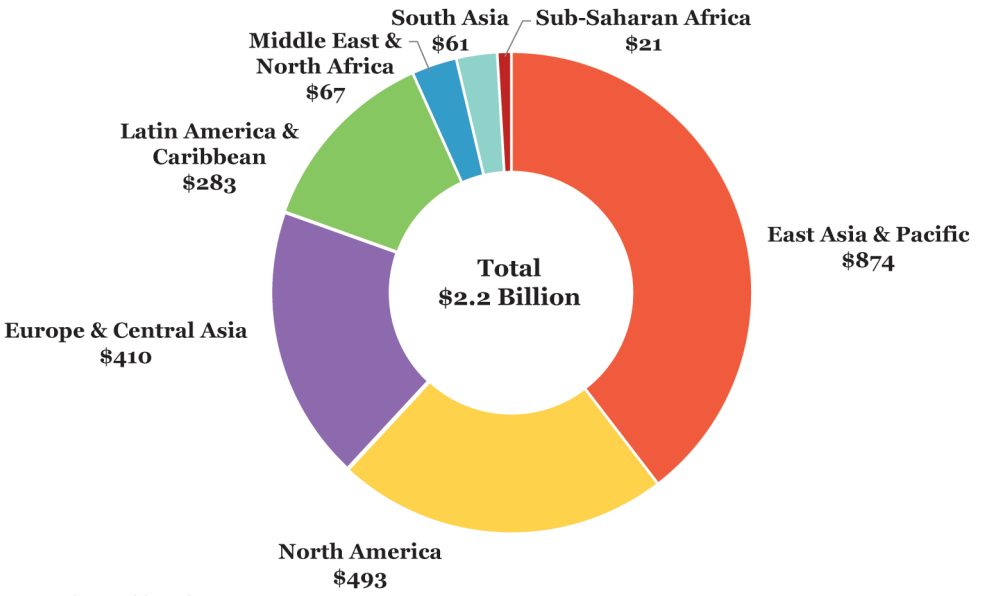
Even if you may have correctly ranked the values, we can agree that it is easier and faster to rank them in the bar chart version. It's also easier to see how much the values differ from each other. The pie chart simply does not let us accurately discern the values, and if your goal is to help your reader make clear and accurate determinations about the data, the pie chart is not the best choice.

It's also worth noting that it's not really clear how we perceive the quantities in the pie chart. Is it the angle at the center? The area of each slice? The arc length? A pair of research papers in 2016 suggested that the angle of slice meeting at the center is not the main way we read pie charts and that area and arc length—the segment of the circumference—were better predictors of people's ability to read values than angle alone.



Comparing the values is easier to do in the bar chart on the right than in the pie chart on the left.

Distribution of imported goods to the United States in 2016
(Billions of dollars)



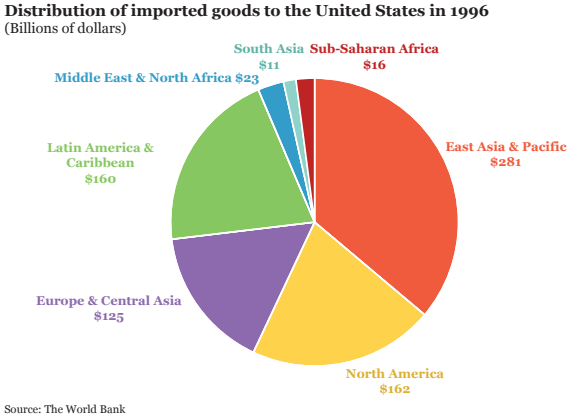
Source: The World Bank

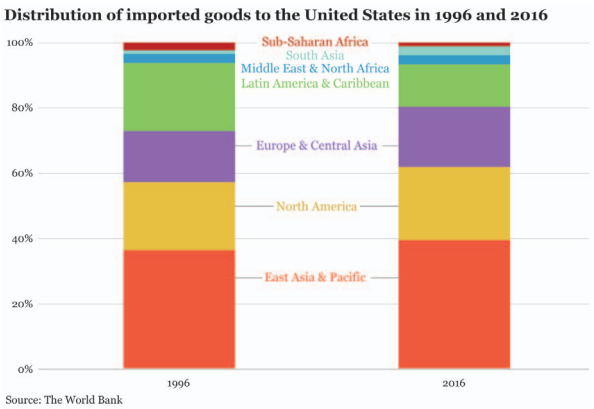
The donut chart punches the middle out of a pie chart, which leaves room for an additional title or text.

If we don't perceive the values in a pie chart from the angle at the center, it would mean that a donut chart—a pie chart with the center removed—is a viable, even preferable alternative. But if it's the case that we discern the quantities in the pie chart by the *angle* of the slices, then the donut chart is an even worse choice than the pie chart because the angles are held in the center, which is missing from the donut chart. No one has decisively proved it either way. Still, the primary advantage of the donut chart is that you can include a number or statement in the center of the chart.

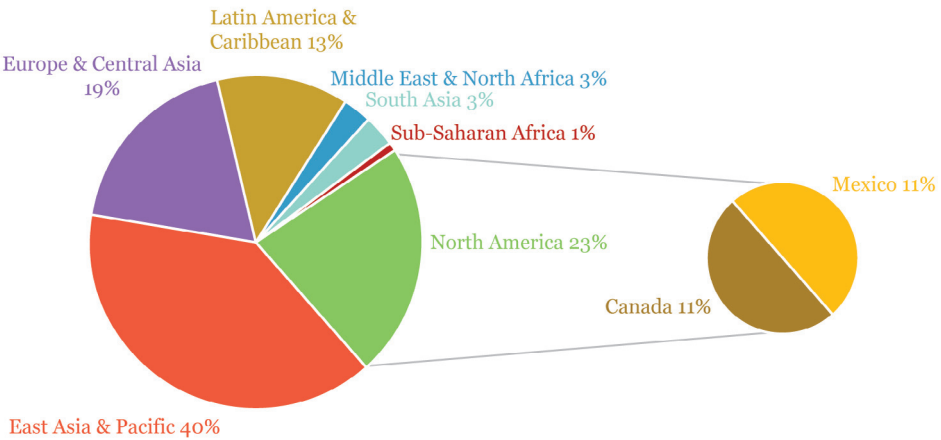
I recommend avoiding using pairs of pie charts, even when they show only a few slices. The reader must look back and forth between the two pie charts to see whether the values of the different groups changed. That task is made faster and easier with a stacked bar chart or slope chart, though with the slope chart, we move away from the part-to-whole comparison and instead focus on the change over time.

Especially avoid pie charts that show too many slices. Five is already too many—more becomes incomprehensible. Also avoid the "breakout" pie chart that removes a single slice





Distribution of imported goods to the United States in 2016



Source: The World Bank

Pairs of pie charts are rarely useful to show changes over time. That task is made faster and easier with other visualization types, like a stacked bar chart or slope chart.

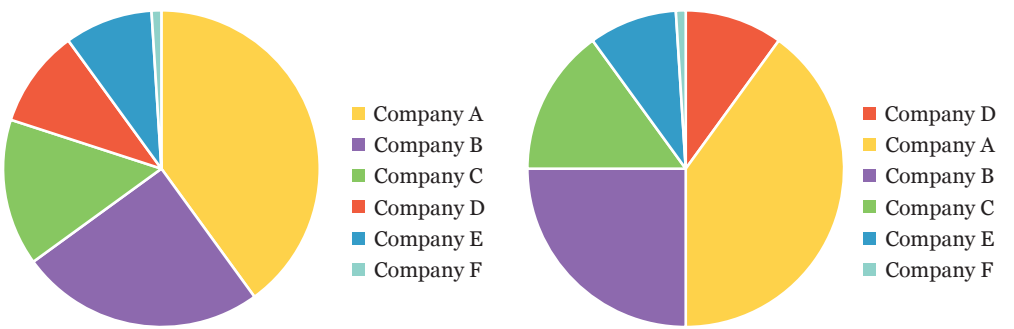
and breaks it down even further. These charts are hard to read and there are better ways to plot such data: in a bar chart or even a Sankey diagram (see page 126).

THE CASE FOR PIE CHARTS

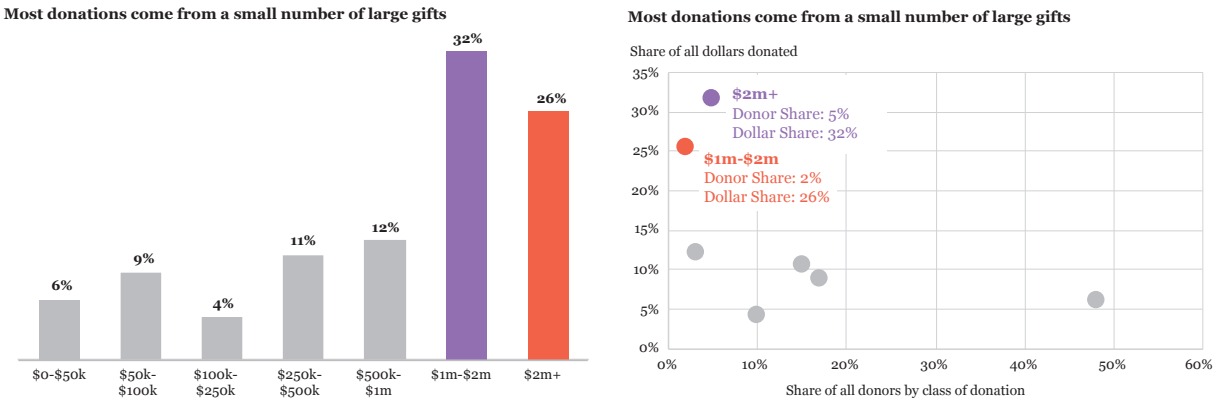
Let me now make the case *for* using pie charts. These two pie charts show the same data but arranged differently. In the version on the left, the value of Group B (the purple slice) is not immediately clear. In the version on the right, the rotation of the chart creates a familiar right angle in the center of the chart, and now Group A's value (25 percent) is clear. A pie chart is therefore perhaps best used when the value(s) of the slice or slices sums to one of these round numbers (25 percent, 50 percent, and 75 percent), for which the angle is familiar. These cases let you you easily focus your reader's attention on three or maybe four slices.

Let's look at a practical example of when a pie chart can be useful. Imagine giving a presentation to report the fundraising efforts you and your colleagues have accomplished for your nonprofit organization. Let's say you have received one hundred donations, and a bit more than half of all money you have raised has come from a few, very large gifts of $1 million or more. You break down the distribution of gifts into seven categories and break out the top categories into a group of $1 million to $2 million donations and a group of $2 million donations and above. The perceptual rankings chart may guide you to use the most perceptually accurate graphs, like a bar chart or a scatterplot.

But for your audience, a pie chart might work best. They can read and understand a pie chart almost instantaneously, and because the share you want your audience to focus on



It's easier to guess the value of the purple slice for Company B when the chart is arranged with right angles.



These two charts show that most donations to this nonprofit came from a small number of large gifts. But if the point is to show that slightly more than half came from these two groups, a pie chart is actually a good choice.

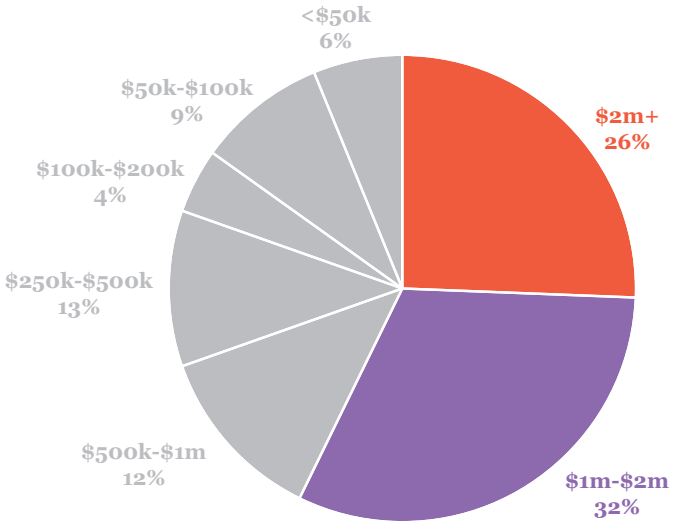
accounts for just more than half, it's easy for your audience to mentally draw the vertical line through the circle. Imagine explaining how a Marimekko chart or a scatterplot works to an audience who just wants the basic information as quickly as possible.

Notice how all seven values are plotted and labeled in the pie chart on the next page, but only the two slices of interest are colored, and together they are clearly larger than 50 percent. (Also notice how we are not following the start-at-12-o'clock rule, because arranging the slices by category value makes more sense here.) Because they are not the primary focus of the graph, the other slices are in gray and move to the background. For a presentation, we might even collapse our two groups to one and delete many of the labels to reduce the amount of information on the screen that would only distract the audience's attention from the speaker.

This example leads to another scenario in which a pie chart can be useful, and that is when you are focusing your reader's attention on a single value—the true part-to-whole relationship. In these cases, consider whether you actually need a chart at all. Do you need a visual to back up the statement that the U.S. poverty rate is 12.3 percent? Maybe on social media or on a presentation slide, but in a report or article you might do just as well leaving the number in the text.

A bad pie chart is still a bad chart. If you choose to use pie charts, do so strategically and thoughtfully. It will almost always be difficult for your reader to discern specific quantities or compare slices, but if you want them to understand a general difference in magnitude or focus on a single slice, the pie chart is appropriate.

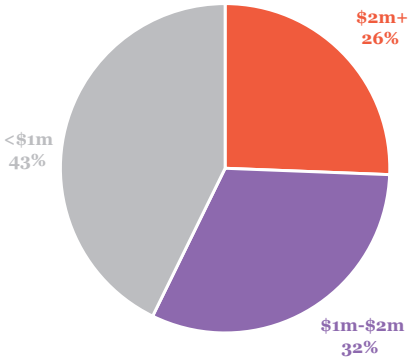
Most donations come from a small number of large gifts



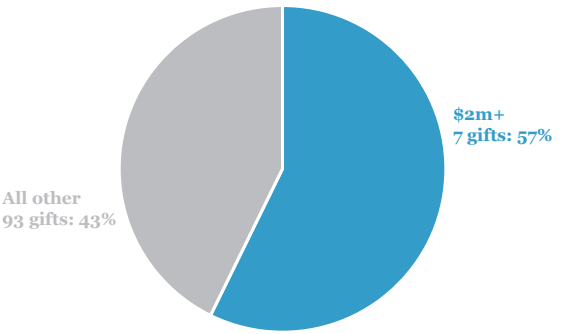
Source: The World Bank

This pie chart does a good job of highlighting the two groups of interest because (a) the groups sum to slightly more than 50 percent and (b) the other groups are gray.

Most donations come from a small number of large gifts



Most donations come from a small number of large gifts

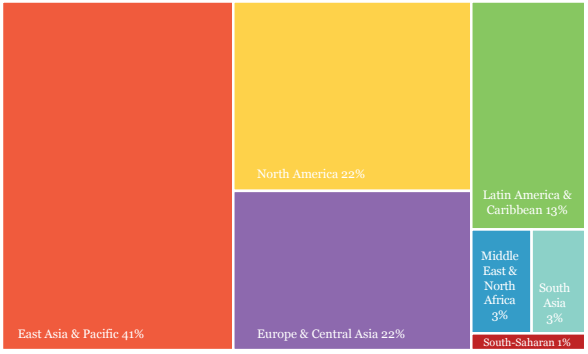


Basic simplifications to these pie charts can focus attention on the groups of interest.

TREEMAP

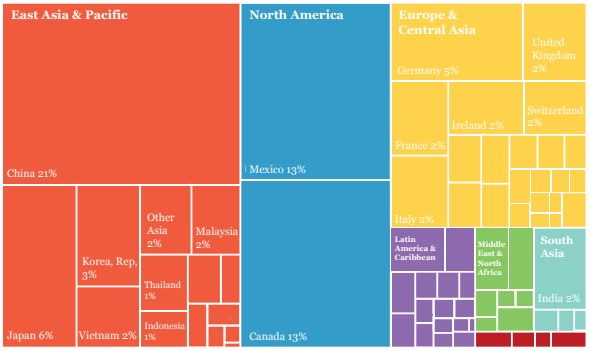
Originally developed by Ben Shneiderman at the University of Maryland, the treemap divides sections of a square or rectangles into groups to illustrate a hierarchy or part-to-whole relationship. In other words, the treemap is a squarified version of a pie chart.

Distribution of imported goods to the United States in 2016



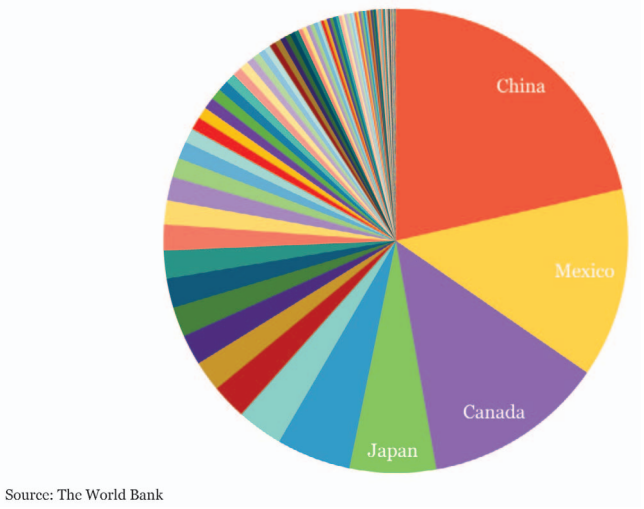
Source: The World Bank

Distribution of imported goods to the United States in 2016



Source: The World Bank

Distribution of imported goods to the United States in 2016



Source: The World Bank

Treemaps are an alternative to the pie chart and can show part-to-whole or hierarchical relationships. Notice how it is essentially impossible to read the pie chart.

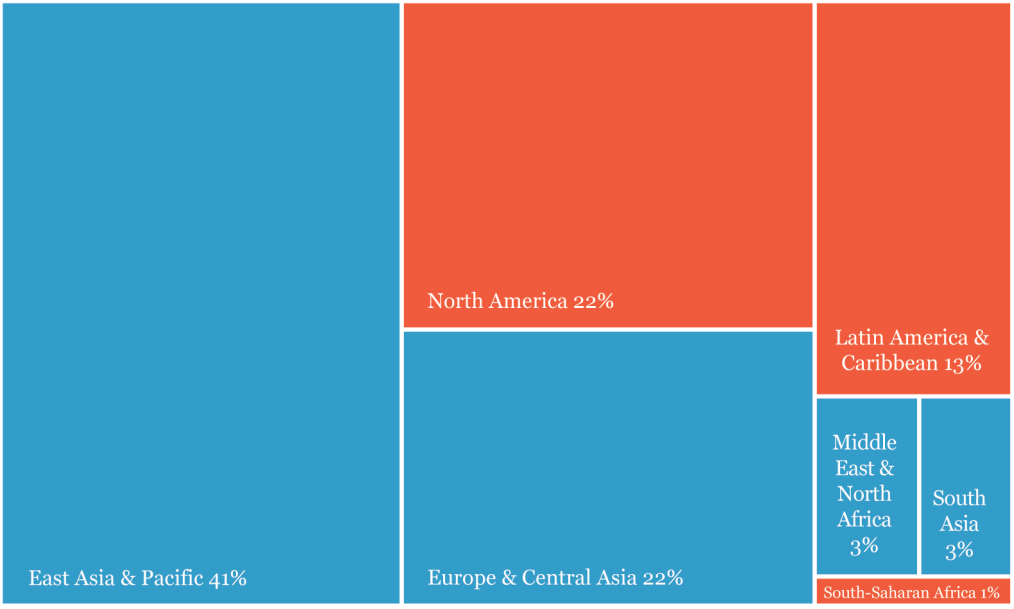
The two treemaps on the previous page show the breakdown of total imports from specific countries to the United States in 2016. You may find this easier to read than a pie chart because rectangles are more easily compared. Or, because it's an unfamiliar graph type, you might find it slower to navigate or more difficult.

The treemap on the right breaks down the regional categories further into individual countries. Not every country in the world is labeled in this chart, but you do get a clear sense of which countries send the most goods to the United States.

Try to read the pie chart that shows the same data—it's impossible.

Shneiderman originally developed the treemap to concisely view the file directory on his computer. The compactness of the treemap is one of its biggest advantages because it can include many different groups and variables. Hierarchies are easy to visualize in the treemap because subsections can be labeled and embedded within the parent group. You can also add other encodings to the treemap—for example, this treemap shows the 2016 distribution

Distribution of imported goods to the United States in 2016
(Blue denotes increases between 1996 and 2016; red denotes decreases)



Source: The World Bank

Color can add another dimension to a treemap. In this case, the change in imports between 1996 and 2016.

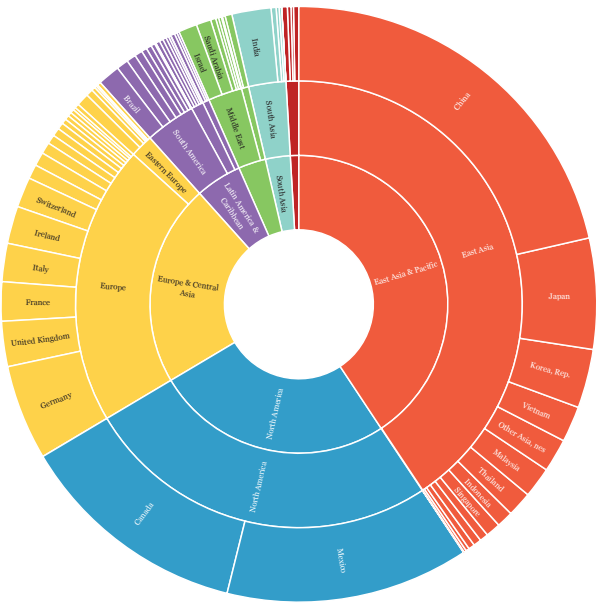
of the value of imports to the United States and the colors denote the shares that increased (blue) and decreased (red) between 1996 and 2016.

SUNBURST DIAGRAM

If you want to show the proportions of parts to a whole at several levels in a hierarchy, you might use what is called a *sunburst diagram*. Like a treemap, the sunburst can show part-to-whole relationships or hierarchical relationships. This sunburst diagram, for example, shows the same data as the detailed treemap above, with an additional ring that breaks the major regions into their components.

Each ring in a sunburst corresponds to a different level in the hierarchy, and the slices of each ring—sometimes called *nodes*—refer to the different subgroups. The central ring shows

Distribution of imported goods to the United States in 2016



Source: The World Bank

The sunburst graph shows part-to-whole or hierarchical relationships.

the top level, sometimes referred to as the *root*. Outer rings show how the groups break into subgroups. You can use color to highlight different rings, groups, or hierarchies. The sunburst graph above uses color to differentiate the seven major areas of the world.

As with all circular visualizations, it's difficult to clearly and quickly see patterns. Sunburst diagrams with too many categories are likely too cluttered for a reader to pick out patterns even after a close look. Strategic use of color, labels, and annotation can guide the reader to the most important parts of the visualization.

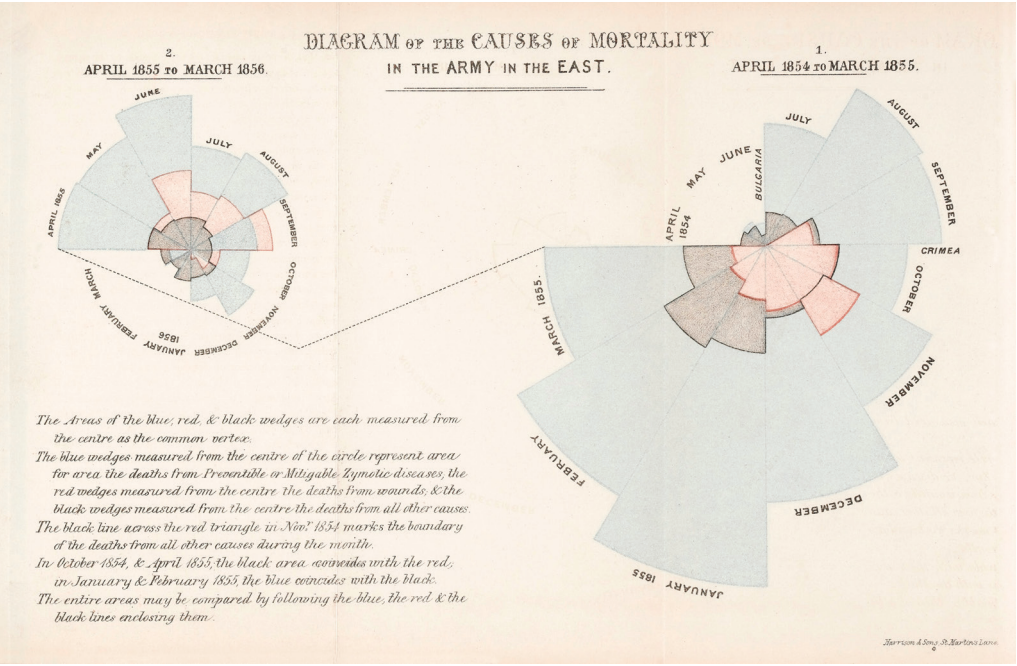
NIGHTINGALE CHART

This type of chart is sometimes called the *coxcomb* or *rose diagram* but is most often known as the *Nightingale chart*. The first and most famous of these was created by Florence Nightingale to visualize soldier casualties during the Crimean War.

Nightingale was born in the early nineteenth century to wealthy parents who provided her with a comprehensive education in liberal arts and mathematics. Deciding early on that she wanted to dedicate her life to health care and helping the poor and needy, Nightingale became a nurse in the early 1850s. With experience in organizing hospital supplies, she eventually took care of British casualties at the Barrack Hospital of Scutari in Turkey during the Crimean War. For two years, she and her team of nurses helped care for the wounded and ill, all the while keeping careful records of their patients.

Convinced that cleanliness was a major reason for the high death rates at Scutari, Nightingale wrote hundreds of publications showing her data in dozens of visualizations. Ultimately, the British government created the Sanitary Commission to investigate the poor outcomes at the hospital and soon implemented improvements to sanitation, ventilation, and cleanliness.

Her most famous chart—and the one that bears her name—is drawn around a circle with values segmented into different time periods. Each slice represents the total deaths in each month from April 1854 (in the 9 o'clock position in the right diagram) to March 1856 (in the 9 o'clock position in the left diagram) during the Crimean War. Cause of death is broken down into three categories: deaths from wounds in battle (pink); deaths from "other causes" (black); and deaths from diseases (blue). The diagram on the right shows patterns in deaths over the first twelve months of the war. The one on the left shows the next twelve months after the Sanitation Commission implemented its reforms in March 1856.

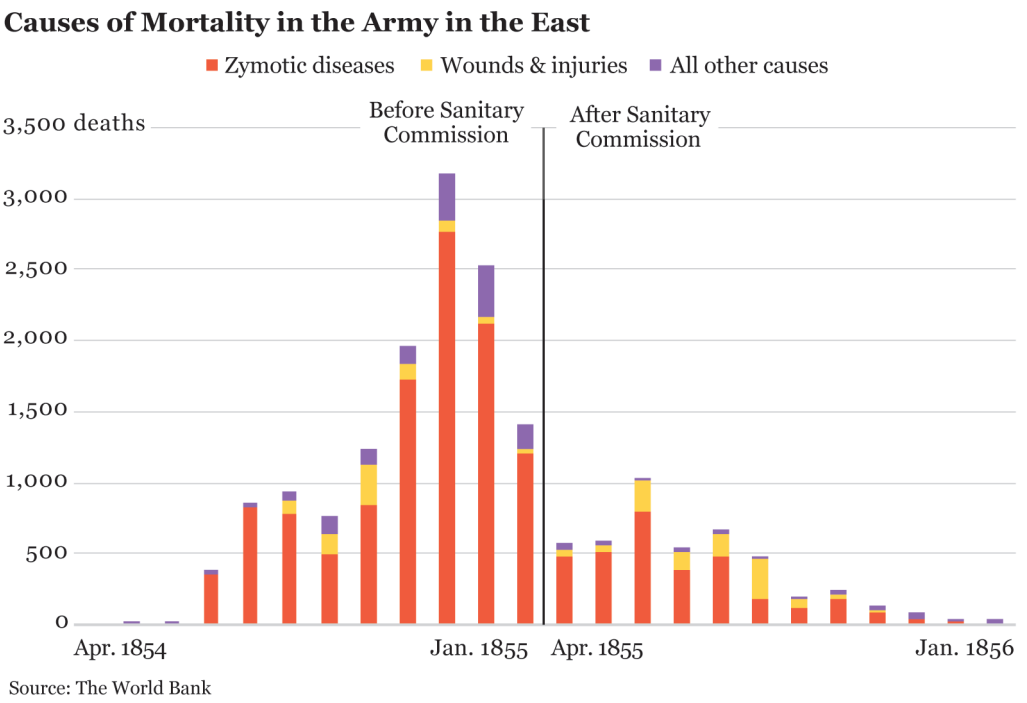


Florence Nightingale created her legendary chart to show the disproportionate share of deaths from disease like cholera, typhus, and dysentery. Image from the Wellcome Collection.

We can think of the Nightingale as a pie chart in which the slices have been expanded in different directions. The area of each slice represents its value relative to the whole, and the slices are arranged along the time dimension. The Nightingale chart therefore shows both changes over time *and* a part-to-whole relationship.

The visualization demonstrates two major points. First, deaths in battle were only a small portion of the total number deaths. Deaths from disease—in this case, cholera, typhus, and dysentery—made up a disproportionate share of deaths. Second, it illustrates how the Sanitary Commission, which started its work in the middle of the war, dramatically reduced the overall number of deaths.

Already in this book we've seen variations on the Nightingale chart. The circular column chart, for example, arranges columns around a central circle radiating outward. The radar chart also uses a circular layout to show relationships across variables aligned along the radii. As with many circular charts, the Nightingale can be difficult to read, and comparisons across the slices are hard to gauge. The major problem with the Nightingale chart is that the



This stacked bar chart shows the same data as in the original Nightingale graph—but is it as memorable?

outer segments are necessarily larger and therefore are emphasized disproportionately. This distortion increases as we move farther from the center.

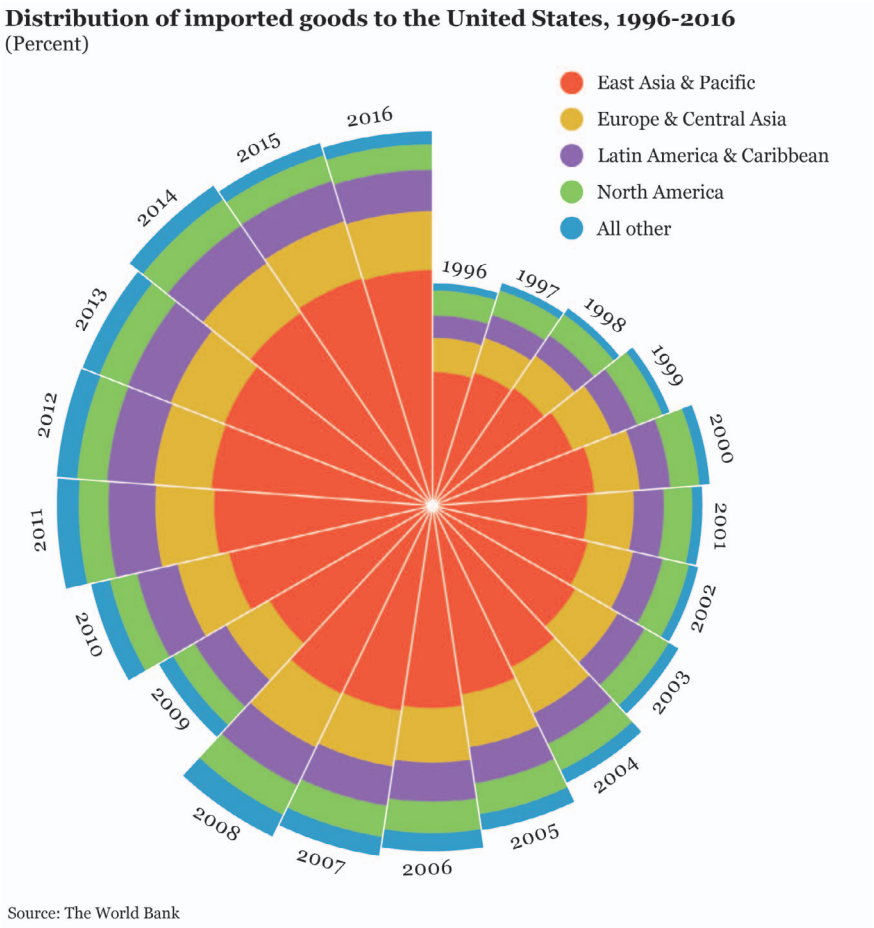
The stacked bar chart above uses the same Crimean War death data as the Nightingale chart on the previous page. I've added a vertical line between March and April 1855 to denote the break in the two circles. While the bar chart shows the totals more clearly, it does not make the break between the two years as clear as the original, nor perhaps is it necessarily as engaging.

The natural question, of course, is whether Nightingale's charts would be better as some kind of more standard chart, like this stacked bar chart. Data visualization expert R. J. Andrews addressed this debate head-on:

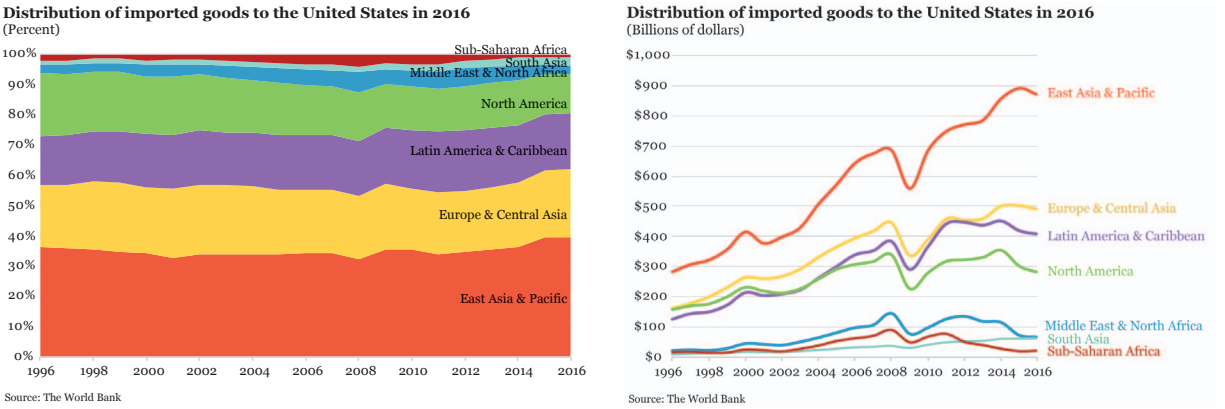
> Critics suggest that her mortality data is better shown in something more straightforward like a bar chart. But this is not true: Florence Nightingale made lots of bar charts. No one cares about them! Her roses gripped 1858 readers and they still hold our attention today.

In this case, as in many others, we must strike the balance between what is perceptually accurate and what will engage people and stick in their minds. Or, as author Alberto Cairo wrote about Nightingale, "I believe her goal wasn't just to inform but also to *persuade* with an intriguing, unusual, and beautiful picture. A bar graph conveys the same messages effectively, but it may not be as attractive to the eye."

Let's use the import data from early in this chapter and lay it out as a Nightingale chart. Again, this chart shows total imports to the United States from each of the seven regions



Nightingale charts can visualize all sorts of data, though they introduce their own perceptual hurdles.



A stacked area chart or a line chart are both simple alternatives to the Nightingale chart.

around the world from 1996 through 2016. Each region is stacked on top of each other, and the year data is arranged from clockwise from earliest to latest, beginning in 1996 at the top.

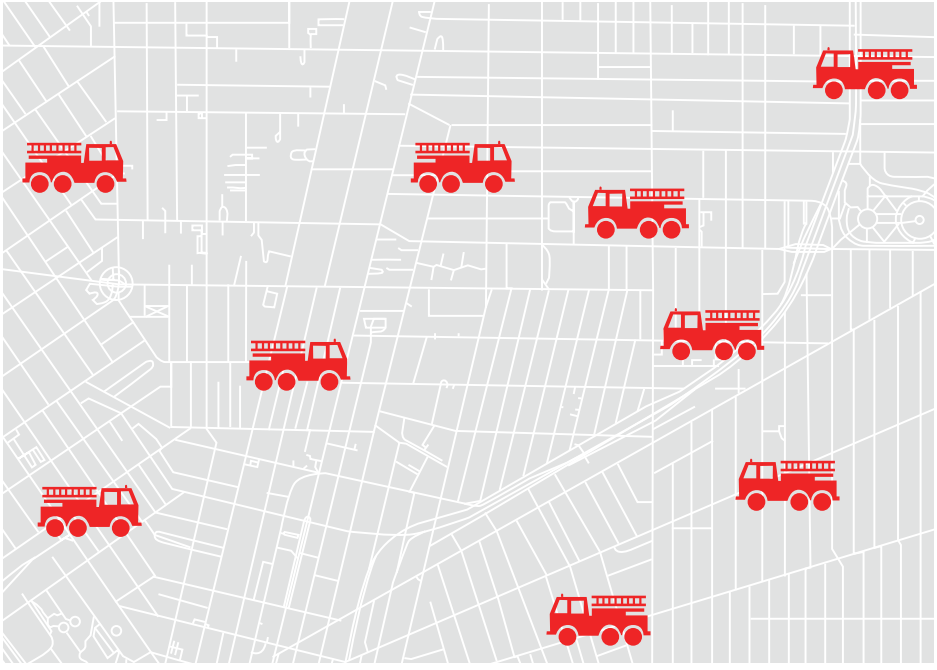
We could also use a stacked area chart or line chart to show these same data. Here, notice that the smaller values appear slightly larger in the Nightingale chart because, by definition, those areas become larger the further they are from the center of the circle. But as we've seen throughout, there might be cases where the Nightingale is useful over these standard chart types—it's more compact and looks very different from these basic chart types. And that may be a goal in itself.

VORONOI DIAGRAM

Named after Georgy Voronoi, a Russian mathematician who lived around the turn of the twentieth century, the *Voronoi diagram* divides a space of some number of points—called *sites*—into a corresponding number of shapes (polygons), which meet at boundaries that do not overlap. If we drop a new point in the space, it will be closer to the *site* in its area (polygon) than to any other in the space. Voronoi diagrams are sometimes used for geographic data, but I've placed them in this chapter because they can also show part-to-whole relationships. You can often find these diagrams in the biology literature (for cell structures), ecology (for the study of forest growth), and chemistry (for molecular positions).

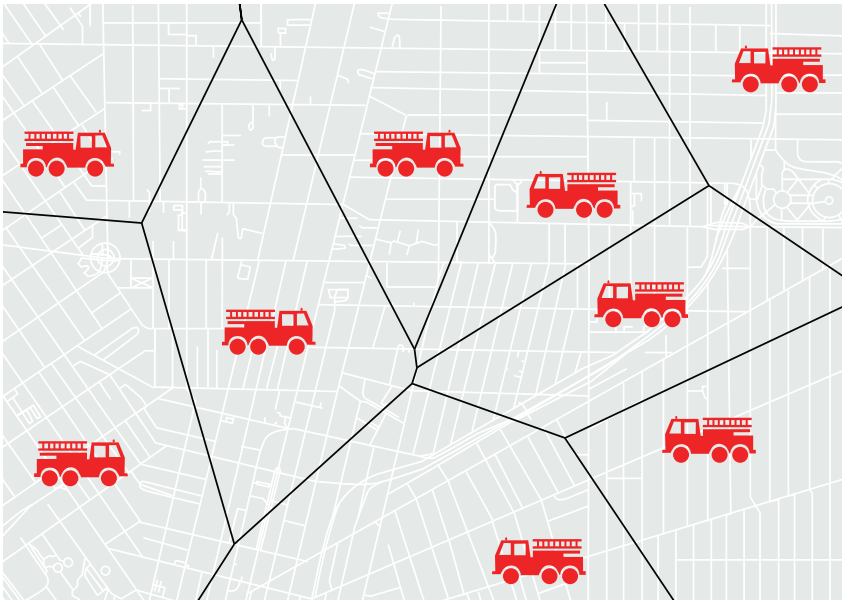
The interesting property of Voronoi diagrams is that the border of each polygon is the same distance to the two nearest sites. In other words, each polygon is defined so that the distance from the edge of the polygon to its site is the shortest it can possibly be. When three borders meet, they result in a point—called a *vertex*—which is equidistant to the three nearest sites. There are a variety of algorithms to determine the shape and position of the various polygons.

That explanation is complex, so let's take a simple example. Say you live in a city with nine fire stations, and a fire breaks out at a building somewhere in the city. Which fire station should respond? The locations of the nine fire stations are the generating points, and the polygons that surround each tells you which fire station should respond to the fire on the principle that the closest fire station should respond. (Obviously, I'm ignoring issues of roads and bridges and other obstacles to actually *get* to the fire, but more complex algorithms can factor in these kinds of obstacles.) As you can see on the next page, the Voronoi diagram splits up the city into its different components, effectively letting us visualize a part-to-whole relationship.



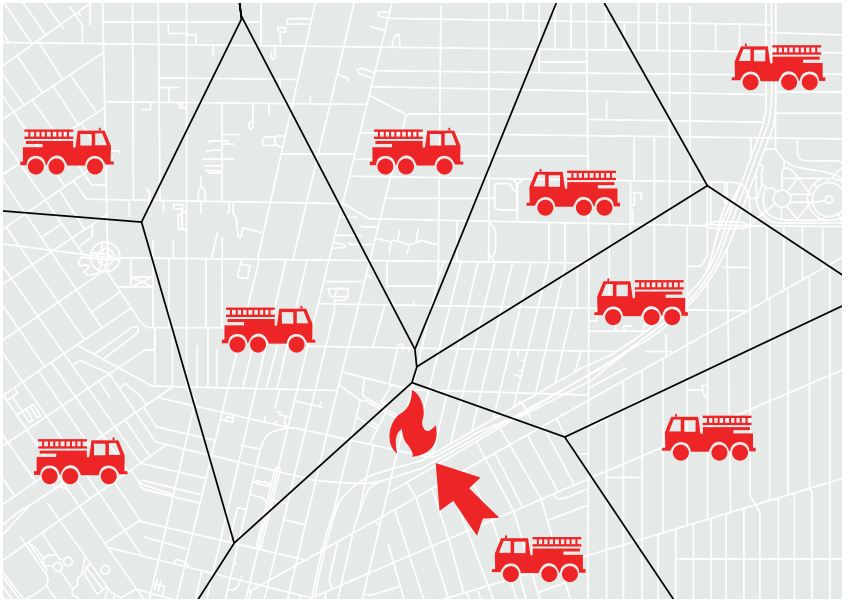
Source: Map from the City Roads project by Andrei Kashcha

To demonstrate how a Voronoi diagram works, imagine there are nine fire stations located across the city.



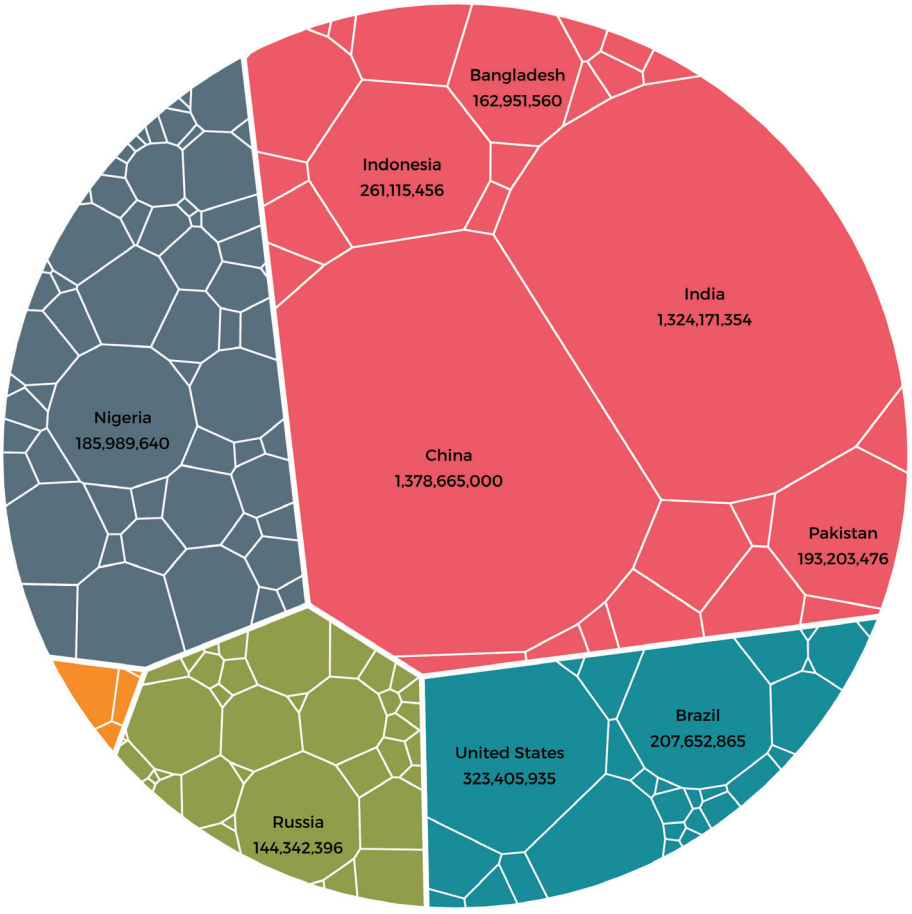
Source: Map from the City Roads project by Andrei Kashcha

Using a Voronoi map, we can divide the city into nine separate areas.



Source: Map from the City Roads project by Andrei Kashcha

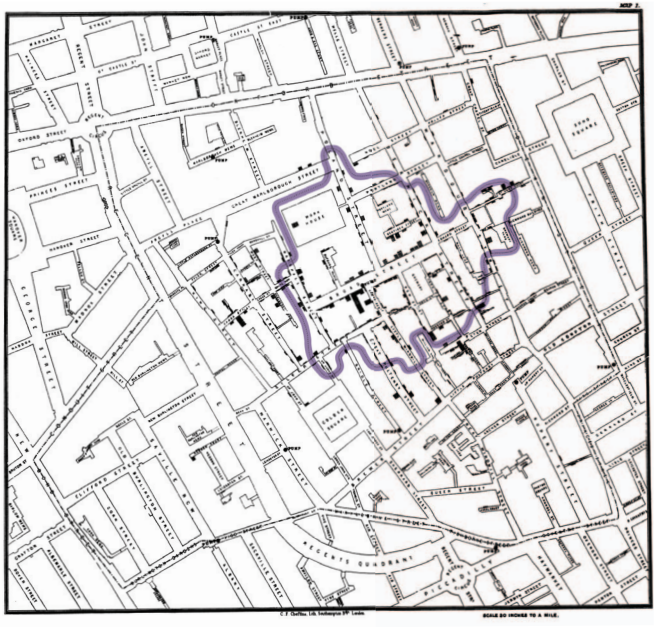
If a fire breaks out, we can identify which fire station should respond based on its proximity.



Maybe a more common use of the Voronoi chart is to show part-to-whole relationships, like this graph from Will Chase that shows the populations of countries around the world.

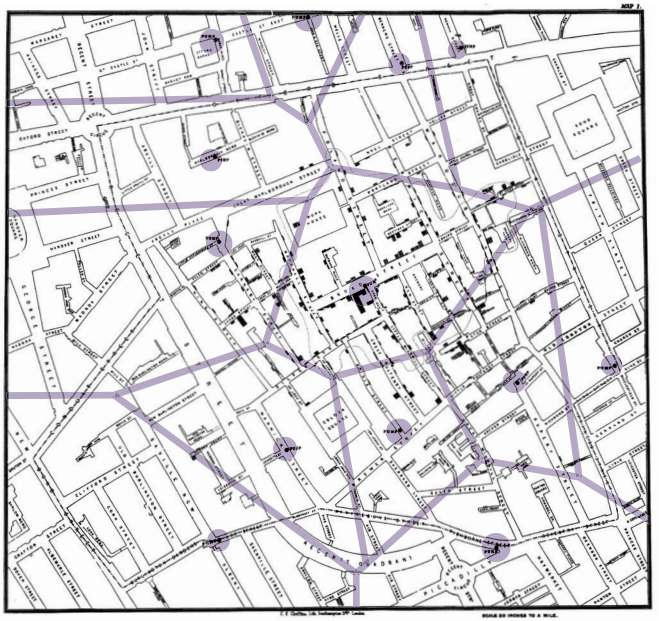
More generally, you may see Voronoi diagrams like this one, which shows the population of countries around the world categorized by region.

Many people probably don't know that one of the most famous maps in the history of data visualization is also a Voronoi diagram. In mid-1854, John Snow, an English physician and founder of the modern science of epidemiology, plotted each death from an outbreak of cholera as a small dash on a map of Soho in London. More than six hundred people died during the outbreak in a little more than a month.



Source: John Snow

John Snow's famous cholera map is actually a Voronoi diagram. There is a clear clustering of deaths around the Broad Street pump—seen as a small dot in the center of the map just to the right of the thickest black bar. I've highlighted the dashed line that Snow included to show the distance from the Broad Street pump.



Source: John Snow

If we put a purple dot at each water pump in the map, we can create a Voronoi diagram from Snow's original map.

Snow's map showed a clustering around a single water pump on Broad Street. You can see a large number of dashes just to the left of the infected pump (the dot with the "PUMP" label directly in the center of the image) as well as other deaths up and down the street. In this version of the map (published slightly later than his original), a small dashed line is drawn around the affected area (which I've highlighted in purple). According to Snow, the dotted line "shows the various points which have been found by careful measurement to be at an equal distance by the nearest road from the pump in Broad Street and the surrounding pumps."

We can also apply a direct Voronoi approach by marking each water pump on the map as a *site* (purple dots in this map) and construct the polygons around each (purple lines on the map). Thus, the "Snow Map" is actually the "Snow Voronoi."

CONCLUSION

Pie charts are the default visualization of part-to-whole relationships because they're familiar, easy to make, and easy to read. But pie charts are rife with perceptual difficulties and should be used with care. Limit the number of slices in your pie charts and remember that right angles are familiar to our eyes and brains, so 25 percent increments are useful markers. It's difficult for readers to compare values across pairs of pie charts, but if we use more than one pie chart, we should at least keep the slices in the same order.

There are alternatives to the pie chart. Treemaps are essentially squarified pie charts, which can hold more annotation and show hierarchical relationships. Sunburst diagrams also show hierarchical relationships, though a sunburst diagram with many series can look complex and cluttered. Voronoi charts show part-to-whole relationships in a different layout. They can also be used to visualize geospatial data, cell structures, or ecological data.

Some of these other graph types may be less familiar to your readers, but they also help them more accurately discern the data values. Many of the graphs we have covered in earlier chapters—like the bar chart, stacked bar chart, and slope chart—can also be used to visualize part-to-whole relationships, but they often require explanatory text to make those relationships clear.

QUALITATIVE

Until now, we've explored charts that mostly communicate quantitative data. But there are also graphs that communicate *qualitative* data, non-numerical information collected through means of observations, interviews, focus groups, surveys, and other methods. The charts in this chapter primarily communicate words and phrases.

We can build narratives and tell stories around qualitative results in ways that can be more difficult with quantitative data. Downloading a big data set, running regressions, and creating tables may provide more generalizable results, but readers don't connect to them the way they connect to a story. Qualitative data can help tell those stories.

As with quantitative data, sometimes you may want to summarize your qualitative results. Images in this chapter are drawn from a variety of news and research organizations from around the world. Unlike previous chapters, they do not use the same style guide, which helps demonstrate some of the variety in layout, design, and approach. Some of the graphs in this chapter show an overall view of your data, while others ask the reader to review the detailed results, text, or quotations. Which you use will depend on your intended purpose and where you are publishing your results.

ICONS

Data visualization design can go a long way to deliver more content and captivate your reader. We've seen in previous chapters how color, layout, and different shapes can engage

Icon examples from the Noun Project. Modern versions of Microsoft Office have a built-in icon library, or you can use websites like The Noun Project and Flaticon where you can purchase or download free icons. There are also fonts that can be used like icons—each letter in the *StateFace* font, for example, is designed as a separate icon of a US state.

readers and bring them into a graphic. Icons, images, and photographs can similarly draw your reader in and categorize your qualitative data for their ease.

Iconography, especially, can help visualize your qualitative data. Icons can be purely decorative, they can represent data (as in a unit or Isotype chart, see page 106), or they can guide the reader from one phase of a visual to the next. Icons (including emojis) are themselves a visual language, so they can simplify and communicate ideas and feelings that are otherwise difficult to express. They may also help readers with certain intellectual or cognitive disabilities engage with your work. A body of research has shown that these kinds of visual-graph forms of communication have helped advance successful language development.

As an example of how icons can be used to support research and analysis, this short graphic from the Center on Budget and Policy Priorities uses icons to offset the five ways in which the Earned Income Tax Credit and the Child Tax Credit help families. The content is inherently qualitative, and the icons organize the text to make it easier for the reader.

WORD CLOUDS AND SPECIFIC WORDS

Word clouds are perhaps the most popular and familiar way to visualize qualitative data, but they are really a way to display quantitative data: the number of times a word appears in a text. In a word cloud, the size of each word is adjusted according to its frequency in a passage.

Working-Family Tax Credits Help at Every Stage of Life

The Earned Income Tax Credit (EITC) and Child Tax Credit (CTC) not only reward work and reduce poverty for low- and moderate-income working families with children, but a growing body of research shows that they help families at virtually every stage of life:

Improved infant and maternal health: Researchers have found links between increased EITCs and improvements in infant health indicators such as birth weight and premature birth. Research also suggests receiving an expanded EITC may improve maternal health.

Better school performance: Elementary and middle-school students whose families receive larger refundable credits (such as the EITC and CTC) tend to have higher test scores in the year of receipt.

Greater college enrollment: Young children in low-income families that benefit from expanded state or federal EITCs are more likely to go to college, research finds. Researchers attribute this to lasting academic gains from higher EITCs in middle school and earlier. Increased tax refunds also boost college attendance by making college more affordable for families with high-school seniors, research finds.

Increased work and earnings in the next generation: For each $3,000 a year in added income that children in a working-poor family receive before age 6, they work an average of 135 more hours a year between ages 25 and 37 and their average annual earnings increase by 17 percent, leading researchers have found.

Social Security retirement benefits: Research suggests that by boosting the employment and earnings of working-age women, the EITC boosts their Social Security retirement benefits, which should reduce poverty in old age. (Social Security benefits are based on how much one works and earns.)

Note: For further details on the research see Chuck Marr, Chye-Ching Huang, and Arloc Sherman, "Earned Income Tax Credit Promotes Work, Encourages Children's Success at School, Research Finds," CBPP

You can use icons to support research and analysis.

Source: Center on Budget and Policy Priorities.

These graphs are probably best used to visualize *overall* patterns or where a single value is obvious and stands out. It is less appropriate in cases where finding the specific values is especially important.

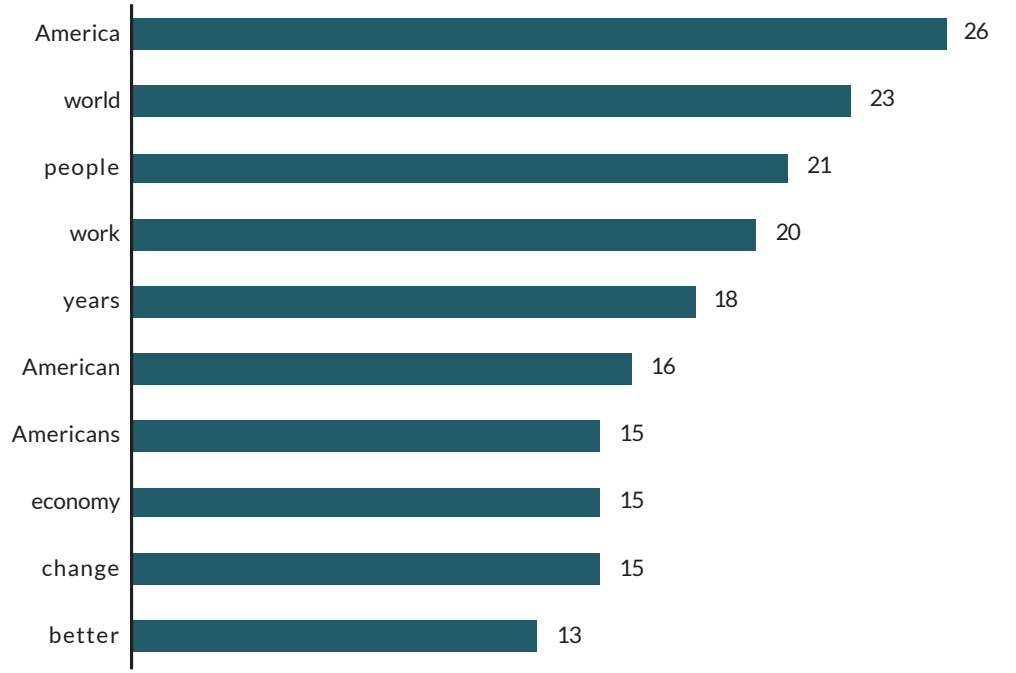
Word clouds are visually engaging, but they present two primary challenges. First, it is unclear what the *specific* frequency of each word is in the text. This word cloud uses the text from President Barack Obama's 2016 State of the Union address. You can see that he used the words *America*, *world*, and *people* frequently. But how much more frequently? It's hard to tell. If understanding the exact frequencies of words in the text is important, then this visual is insufficient.

What about the word *Americans* in this word cloud? It was also used frequently but its vertical orientation may have obscured it from your view. This is the second challenge with word clouds: some words may appear larger (and therefore more significant) than others because of their length, orientation, font, or color.

Remember, to create a word cloud you must calculate the frequency of each word. After you have quantified the text, you could use many different visuals—even a bar chart. It's much easier to see the most commonly used words in this bar chart than in the word cloud. It's worth noting that word clouds generally exclude the most common words, called *stop-words*, such as "the" and "at."

This word cloud shows the frequency of words spoken by President Barack Obama in his 2016 State of the Union address.

Data Source: The White House.



A standard bar chart that shows the ten most frequently used words in Obama's 2016 speech.
Data Source: The White House.

Another approach to creating effective word clouds may be to separate the overall body of text into *semantic* groupings—words grouped by their meaning. Following Obama's address, *USA Today* summarized his speech as follows:

> Obama defended the progress made over the last seven years and set out an agenda that will likely remain unfinished long after his presidency ends: turning back the effects of climate change, launching a "moonshot" to cure cancer, and a grassroots movement to demand changes in the political system.

Those important policy aspects of the speech don't come through in the basic word clouds, but organizing the text into semantic groups and then creating a set of smaller multiple word clouds might be a better way. In the next version, for example, you can better see the most important topics and words in the speech. This obviously requires

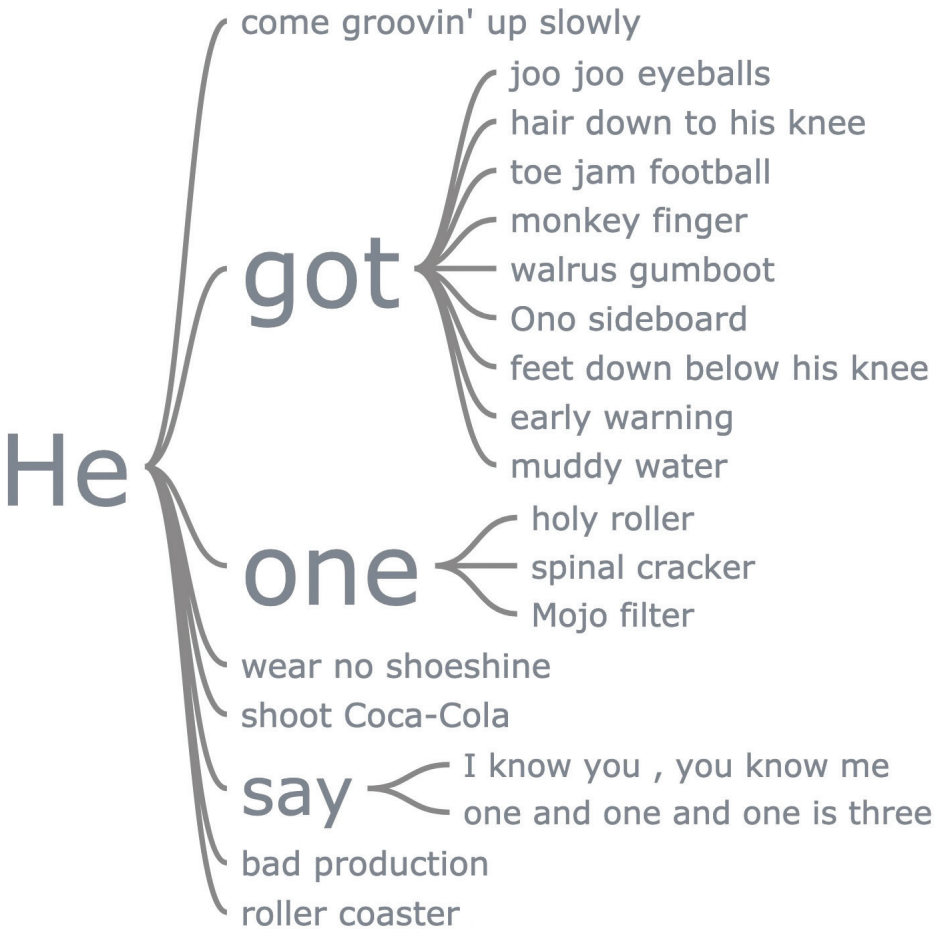
Instead of a single word cloud, Hearst et al. (2019) have suggested breaking up the text into semantic groups.

Data Source: The White House.

more work on the part of the analyst to determine the groups and clusters, but it also results in a better visualization.

WORD TREES

Another way to visualize text data is to examine the contexts in which different words appear. *Word trees*, a version of which was developed by Martin Wattenberg and Fernanda Viegas in 2007, show all of the ways in which specific words are used within a text. The tree structure shows the combinations in which each word appears within the text, and the words are sized according to their frequency.



Word trees show all of the ways in which specific words are used within a text or, in this case, the Beatles' song "Come Together."

Source: AnyChart.

We've already seen one word tree in Chapter 8 that depicted how to show hierarchical relationships. Similarly, this word tree of the lyrics from the Beatles' song "Come Together" shows the elements of the song laid out in a hierarchy.

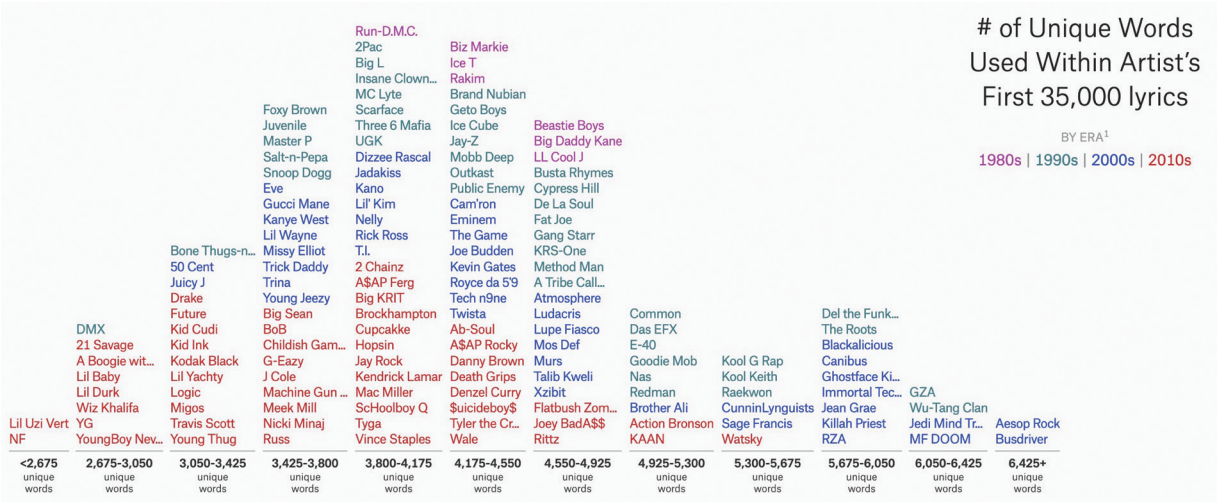
Trees can be used in other ways to show qualitative data. The next tree is based on interviews with fifty science bloggers and classifies their interview responses into their goals and

Paige Jarreau used a word tree to classify interview responses into different groups.

motivations. You can see the branching from each of nine different categories into specific comments and quotes. Word trees suffer the same pitfalls as word clouds: it is difficult to see the exact frequency of the words. But they do provide an engaging, interesting view of the text within its context.

SPECIFIC WORDS

Another way to visualize qualitative text data is to combine individual words with a quantitative metric. This 2019 visualization from Matt Daniels at the digital publication *The Pudding* is a histogram (recall page 179) of the number of unique words used by individual rap artists. Instead of a set of bars, the author plotted the artist's name, using color to distinguish the decade of album release. We see an overall view of the data, for example, that the most



Just like specific data points can be shown in a scatterplot, specific words can be included in data visualizations. Matt Daniels from *The Pudding* used such an approach in his visualization of "The Largest Vocabulary in Hip Hop."

common number of words is around 4,000. We see a shift over time toward fewer words (more names in red to the left side of the graph). And we can focus on the details of the names and eras of specific artists if we want to.

As with other graph forms we have explored—for example, the beeswarm chart (page 206) and network diagram (page 277)—there are a many ways to visualize individual words in your data. The key is to organize that information logically, so the reader can sense the pattern immediately and engage with the graph.

QUOTES

If context is important, visualizing individual words is not the most useful way to show your qualitative data. Sometimes you need to show full quotations.

Following parliamentarian elections in 2016, the *Berliner Morgenpost* published a story about how far some of the elected officials lived from their constituencies. Along with a series of maps that showed the residence of each of six officials relative to the

To communicate qualitative data, sometimes you need to just show the entire quote. The title of this story from the Berliner Morgenpost translates to "Why are you running for election from far away?" The quote from Ero Özkaraca, the image on the left, translates to, "The Social Democrat is still involved in Neukölln, although he moved to Frohnau years ago. The family had wanted a house with a garden, but that was not available in the north of Neukölln. 'I'm only there to sleep, I hardly know my way around.' But he is aware that this can certainly trigger social envy in the constituency."

district they represent, the newspaper also presented a series of direct quotes. Readers could see the crux of the story from the maps and find more detail by reading the quotes.

As an analyst working with survey data, this might be a difficult task. You may not have the exact names or pictures of your survey respondents, and even if you did, you might not be allowed to publish them because of concerns about privacy and security. But it is worth considering whether a generic quotation with a picture of a person would be sufficient in making the content more visual and personalized. As the example from an Urban Institute research project on unplanned pregnancies shows on the facing page, sometimes the quotes alone are sufficient to help communicate the content.

Again, using specific quotes or phrases can be a powerful way to communicate your message.
Source: The Urban Institute

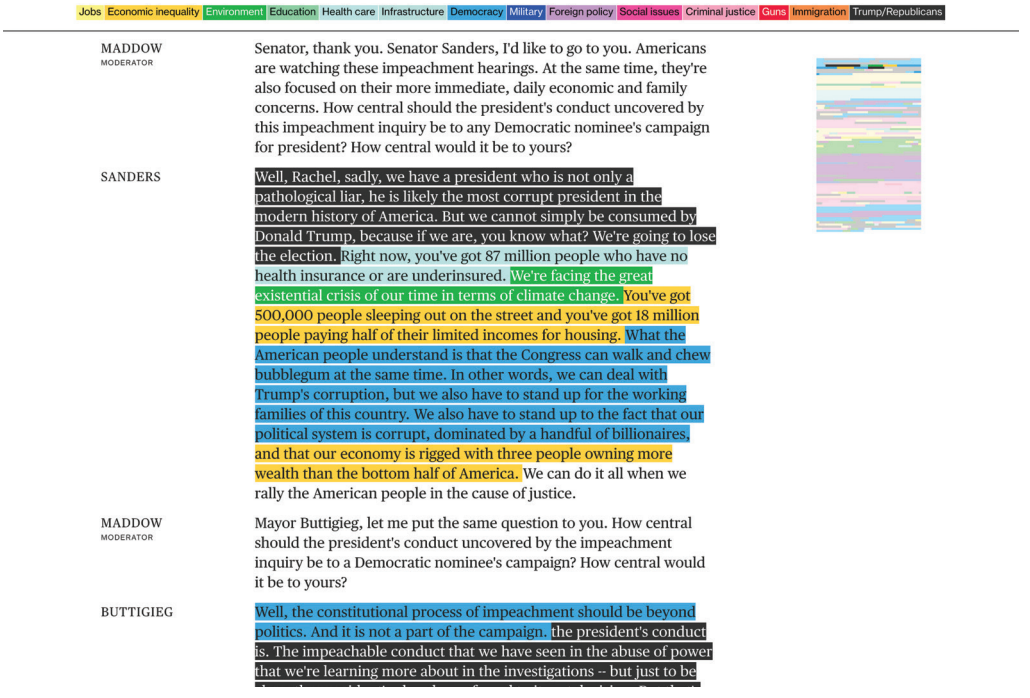
COLORING PHRASES

Passages of text build up from words to phrases to sentences and then to paragraphs. Depending on the goal of your visual, highlighting specific portions of text may be useful to summarize your analysis. You can do so by highlighting with color or boldface to make important passages visible to your reader.

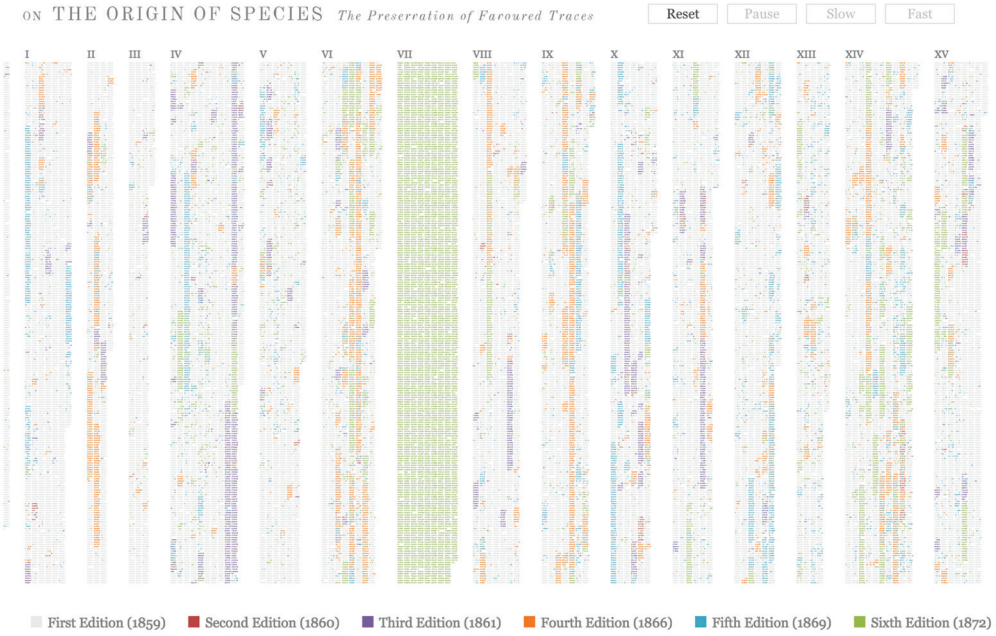
The next image is from a *Bloomberg News* story that annotated and highlighted the transcripts from the Democratic primary debates in 2019. Each color represents a different content area, which helps the reader easily navigate the piece and quickly see which topics were discussed most frequently.

The Preservation of Favoured Traces project on the next page depicts the sequence of edits in Charles Darwin's *On the Origin of Species.* Darwin's final manuscript was actually the fifth draft, so by color-coding each word in the final text by the edition in which it first appeared, we can see the evolution of Darwin's writing and thinking over several years. An interactive version enables the user to zoom in, search, and explore the text in more detail.

This qualitative data visualization approach gives us a bird's-eye view of the data, but it isn't quantified. You can also use this technique to highlight quotes or passages in text to tag them as, say, "positive" or "negative," which might be subjective but also potentially enlightening. Exciting advances in natural-language processing, text analytics, and machine learning algorithms over the past few years now let researchers more accurately measure tone and semantics.



Bloomberg News highlighted phrases in words in a news story about the Democratic primary debates in 2019.



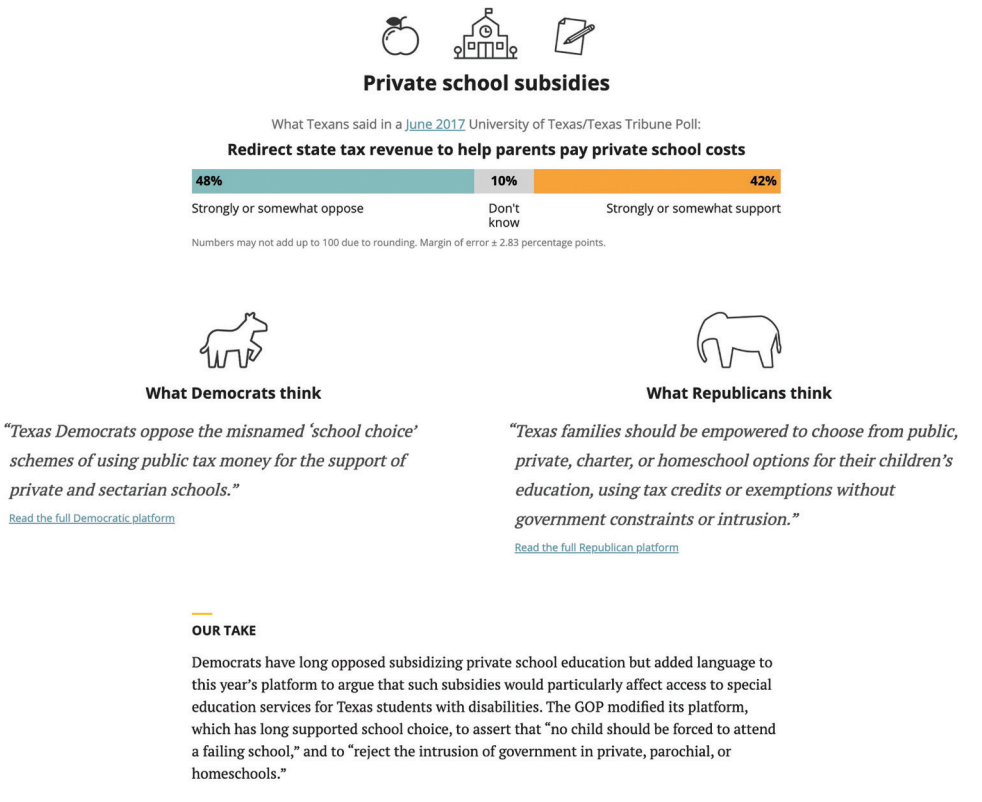
Especially in cases where there is a lot of text, highlighting specific phrases draws the reader's attention. Ben Fry used this very approach in his "Highlighting words in Darwin" project.

A closer look at "Highlighting words in Darwin."

MATRICES AND LISTS

If showing an entire qualitative data set is impossible, we can simplify and organize the data into groups or categories. Essentially, this approach creates a table of qualitative data so readers can more easily see the arguments and narratives. With this method, we can use lists to make it easier for our reader to navigate through what might be a very dense data set.

This graphic from the *Texas Tribune* organized the major issues in the platforms of the Texas Republican and Democratic parties in 2018. Combining the qualitative information



The *Texas Tribune* combined quotes, text, icons, and quantitative data about poll results.
Source: *Texas Tribune*, texastribune.org

from the published platforms with quantitative data from other surveys, the reporters provide a broader view of six major policy issues. In this example for school subsidies, we can see how each section begins with a set of three icons, a sentence describing the issue, and a bar chart of survey results from a 2017 poll. Below, separate quotes from each parties' platform documents and a summary from the reporters bring the entire issue together.

While in previous qualitative graph types, the analyst might need to make a judgment about semantics or meaning, in this case, the decision is to choose into which category the person falls. In other words, visualizing qualitative data doesn't need to be a complex process but simply a task of organizing the text into ways our reader can easily navigate and process.

CONCLUSION

Presenting qualitative data succinctly is a challenge. Taking a series of interviews and summarizing the results so that our reader can quickly and easily understand our argument is no small feat. But qualitative data can engage readers in ways that quantitative data may not. With good design and organization, we can hook our reader with qualitative data and encourage them to explore the words, quotes, and phrases.

Word clouds are one of the more common ways to visualize qualitative data, but they introduce perceptual issues depending on the length of the words, font, and layout. They are engaging, however, so they are useful in certain scenarios. Using recent research, we might consider ways to split our text into semantic groups and create a small-multiple version of the word cloud, which assuages some of these perceptual issues.

There is an array of alternatives to word clouds. Showing specific words or phrases, pairing quotes with photographs, or simply using icons in and around text are all ways to illuminate your qualitative data. As we've seen in the examples in this chapter, highlighting those specific passages and phrase often rely on specific design, color, and layout choices. But, as with many other chart types, practice and experimentation will lead to better visualizations of our qualitative data.

TABLES

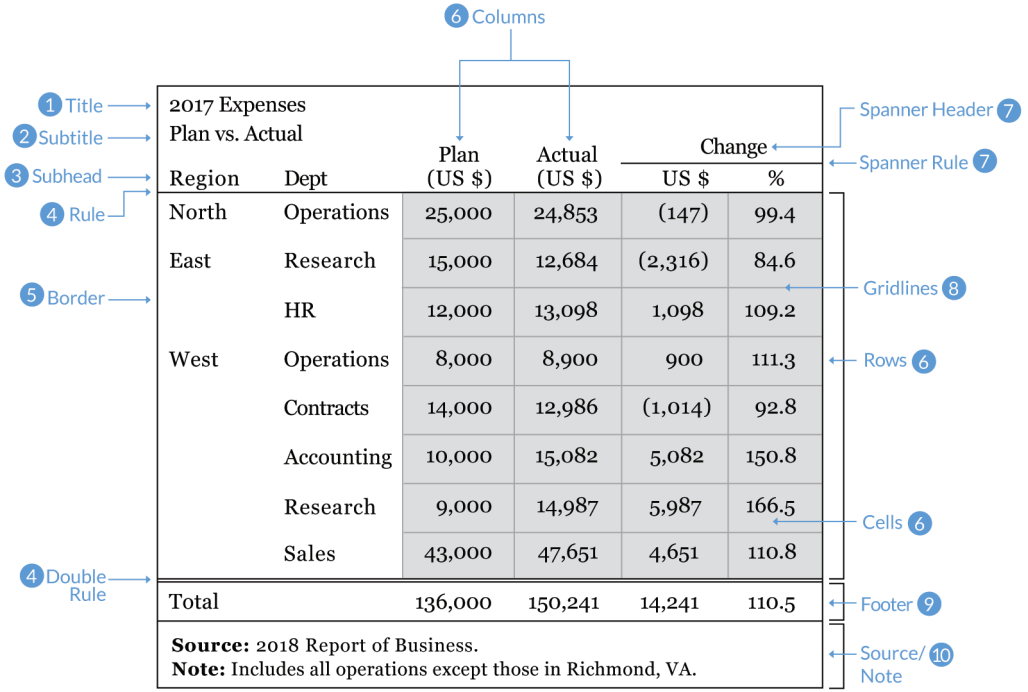
Yes, tables are a form of data visualization. If you want to show the exact amount of every value in your data, a table might be your best solution. They are not the best solution if you want to show a lot of data or if you want to show the data in a compact space—but still, a well-designed table can help your reader find specific numbers and discover patterns and outliers.

As we've seen with other graphs, gridlines, tick marks, and other clutter can crowd a visualization and obscure takeaways. Tables are especially susceptible to clutter. The same guiding principles of creating effective visualizations apply here as well—clearly *show the data* so that our reader can find the most important patterns, trends, or values; *reduce the clutter* of gridlines, extra spacing, and uneven alignment; and *integrate the table and the text* by using concise active titles and subtitles, and including unit labels like percentage signs and dollar signs with care.

In this chapter, we'll cover ten steps for making better tables.

THE PROPER ANATOMY OF A TABLE

We must first understand the components of a table before we can understand how and when to adjust them to improve our data presentation. This diagram shows the ten primary components of a table. Many of these parallel the parts of a chart that you will see in the Style Guides section in Chapter 12. As with chart style, some of the style decisions you choose for



The components of a data table.

your tables will be subjective and depend on nothing more than your preferences for shading colors, font size, and line width.

1. **Title.** Use concise, active titles. "Table 1. Regression Results" is not particularly informative. Instead, guide your reader to the conclusion with a title like "A one year increase in work experience increases annual earnings by 2.8 percent." Left-aligning the title and subtitle will line it flush with the rest of the table, creating a grid, which is easier to navigate.
2. **Subtitle.** This sits below the title, often in a smaller font size or set in a different color to differentiate it. The subtitle should specify the units of the data in the table (like "Percent" or "Thousands of dollars") or make a secondary point (such as "The experience effect is greater for men than for women").
3. **Stubheads or Column Headers.** These are the titles of your columns. Differentiate these from the rest of the table cells with boldface or separate them with a line, also called a "rule."

4. **Rules.** The lines that separate the parts of the table from one another. At minimum, place rules below the stubheads and between the bottom row and any sources or notes.
5. **Border.** The set of lines that surround the table. Including a border around the whole table depends on how the table is arranged in the rest of the document. Sometimes you need to add a visual differentiator to set the table apart, and in those cases a border is useful. If, however, too many lines and borders clutter the document, omit the border altogether.
6. **Columns, Rows, and Cells.** Columns run vertically, and rows run horizontally. The intersecting areas are called cells.
7. **Spanner Header and Spanner Rule.** The text and line that span multiple columns. Text is usually centered over the multiple columns even if the specific column headers are left- or right-aligned.
8. **Gridlines.** The intersecting lines within the table that separate the cells. As with charts, take a light touch with your gridlines—heavy gridlines clutter the table.
9. **Footer.** The bottom area of a table where you might include a row for the total or average. As with the stubhead, we should differentiate this row from the rest of the table. We can do so by bolding the numbers, separating them with a line, or color shading the cells.
10. **Sources and Notes.** The text below a table containing the citation or additional details or notes to the table. Modern Language Association (MLA) style, for example, suggests putting the sources first and the notes second.

THE TEN GUIDELINES OF BETTER TABLES

These guidelines will take us from tables that have too much color, lines, and clutter to ones in which readers can easily see the important numbers and patterns. On the next page, we can see how these guidelines move us from the table on the left to the much more clear and readable table on the right.

RULE 1. OFFSET THE HEADERS FROM BODY

Make your column titles clear. Try using boldface text or lines to offset them from the numbers and text in the body of the table. It should be clear that the headers are not data values

Role	Name	ID	Start Date	Quarterly Profit	Percent Change
Operations	Waylon Dalton	A1873	May-11	5692.88	34.1
Operations	Justine Henderson	B56	Jan-10	4905.02	43.522
Operations	Abdullah Lang	J5867	Jun-14	4919.53	38
Operations	Marcu Cruz	B395	Dec-13	9877.52	37.1
Research	Thalia Cobb	C346	Apr-13	3179.49	-9
Research	Mathias Little	D401	Mar-11	5080.26	3.2
Research	Eddie Randolph	A576	Jul-18	7218.24	43.1
Contracts	Angela Walker	B31	Feb-18	6207.53	-1.788
Contracts	Lia Shelton	C840	Jan-16	1070.61	4.31
Contracts	Hadassah Hartman	D411	Nov-15	3735.96	3.01

Role	Name	ID	Start Date	Quarterly Profit	Percent Change
Operations	Waylon Dalton	A1873	May-11	$5,693	34.1
	Justine Henderson	B56	Jan-10	4,905	43.5
	Abdullah Lang	J5867	Jun-14	4,920	38.0
	Marcu Cruz	B395	Dec-13	9,878	37.1
Research	Thalia Cobb	C346	Apr-13	3,179	-9.0
	Mathias Little	D401	Mar-11	5,080	3.2
	Eddie Randolph	A576	Jul-18	7,218	43.1
Contracts	Angela Walker	B31	Feb-18	6,208	-1.8
	Lia Shelton	C840	Jan-16	1,071	4.3
	Hadassah Hartman	D411	Nov-15	3,736	3.0

Inspired by DarkHorse Analytics

but categories or headers. In this example, which uses growth in per capita GDP, the column labels are boldface and separated from the data with a single line.

Country	2013	2014	2015	2016
China	7.23	6.76	6.36	6.12
India	5.10	6.14	6.90	5.89
United States	0.96	1.80	2.09	0.74
Indonesia	4.24	3.73	3.65	3.85
Mexico	-0.06	1.45	1.90	1.68
Pakistan	2.21	2.51	2.61	3.44

Country	**2013**	**2014**	**2015**	**2016**
China	7.23	6.76	6.36	6.12
India	5.10	6.14	6.90	5.89
United States	0.96	1.80	2.09	0.74
Indonesia	4.24	3.73	3.65	3.85
Mexico	-0.06	1.45	1.90	1.68
Pakistan	2.21	2.51	2.61	3.44

Rule 1. Offset the headers from body.

RULE 2. USE SUBTLE DIVIDERS INSTEAD OF HEAVY GRIDLINES

As with the basic principle to reduce clutter for graphs, you can lighten or even remove much of the heavy borders and dividers in your tables. There is rarely a case when every single cell border is necessary. For series that show the total, use shading, boldface, or subtle line breakers to distinguish these.

Notice in the table on the left how the two columns that show the average (between 2007–2011 and 2012–2016) blend in with the other columns. At a quick glance you don't even

Country	2007	2008	2009	2010	2011	Avg.	2012	2013	2014	2015	2016	Avg.
China	13.64	9.09	8.86	10.10	6.36	10.74	7.33	7.23	6.76	6.36	6.12	6.76
India	8.15	2.38	6.95	8.76	6.90	6.30	4.13	5.10	6.14	6.90	5.89	5.63
United States	0.82	-1.23	-3.62	1.68	2.09	-0.30	1.46	0.96	1.80	2.09	0.74	1.41
Indonesia	4.91	4.59	3.24	4.83	3.65	4.47	4.68	4.24	3.73	3.65	3.85	4.03
Mexico	0.70	-0.48	-6.80	3.49	1.90	-0.19	2.15	-0.06	1.45	1.90	1.68	1.41
Pakistan	2.72	-0.36	0.74	-0.48	2.61	0.64	1.34	2.21	2.51	2.61	3.44	2.42
Average	5.15	2.33	1.56	4.73	3.92	3.51	3.52	3.28	3.73	3.92	3.60	3.61

Country	2007	2008	2009	2010	2011	Avg.	2012	2013	2014	2015	2016	Avg.
China	13.64	9.09	8.86	10.10	6.36	10.74	7.33	7.23	6.76	6.36	6.12	6.76
India	8.15	2.38	6.95	8.76	6.90	6.30	4.13	5.10	6.14	6.90	5.89	5.63
United States	0.82	-1.23	-3.62	1.68	2.09	-0.30	1.46	0.96	1.80	2.09	0.74	1.41
Indonesia	4.91	4.59	3.24	4.83	3.65	4.47	4.68	4.24	3.73	3.65	3.85	4.03
Mexico	0.70	-0.48	-6.80	3.49	1.90	-0.19	2.15	-0.06	1.45	1.90	1.68	1.41
Pakistan	2.72	-0.36	0.74	-0.48	2.61	0.64	1.34	2.21	2.51	2.61	3.44	2.42
Average	5.15	2.33	1.56	4.73	3.92	3.51	3.52	3.28	3.73	3.92	3.60	3.61

Rule 2. Use subtle dividers instead of heavy gridlines.

notice that there is a break in the annual series. In the version on the right, a light shade in those columns sets them apart.

RULE 3. RIGHT-ALIGN NUMBERS AND HEADERS

Right-align numbers along the decimal place or comma. You might need to add zeros to maintain the alignment, but it's worth it so the numbers are easier to read and scan. Here, for example, it is much easier to compare the values in the far-right column where the numbers are right-aligned than in either of the other two columns. To maintain the grid layout, the column header is right-aligned with the numbers as well.

Along these lines, choose the fonts in your tables carefully. Some fonts use what are called "oldstyle figures," in which some numbers drop below the horizontal baseline, the same way the letters *p* or *g* or *q* do. This is fine for cases where numbers are not a matter of data—like the numbering of chapters in a novel. But in data tables, they can be distracting and more difficult to read. Always use fonts that have "lining numbers," where all the numerals hit the baseline, and none drop below it.

Notice how the commas and decimal points in the table on the next page don't line up with custom fonts like Karla and Cabin. When choosing a font, be mindful that the numerals

	2016	2016	2016
China	6,894.40	6,894.40	6,894.40
India	1,862.43	1,862.43	1,862.43
United States	52,319.10	52,319.10	52,319.10
Indonesia	3,974.73	3,974.73	3,974.73
Mexico	9,871.67	9,871.67	9,871.67
Pakistan	1,179.41	1,179.41	1,179.41
Average	**12,683.62**	**12,683.62**	**12,683.62**

Rule 3. Right-align numbers and headers.

	Calibri	Karla	Cabin	Georgia
China	6,894.40	6,894.40	6,894.40	6,894.40
India	1,862.43	1,862.43	1,862.43	1,862.43
United States	52,319.10	52,319.10	52,319.10	52,319.10
Indonesia	3,974.73	3,974.73	3,974.73	3,974.73
Mexico	9,871.67	9,871.67	9,871.67	9,871.67
Pakistan	1,179.41	1,179.41	1,179.41	1,179.41
Average	**12,683.62**	**12,683.62**	**12,683.62**	**12,683.62**

Be aware of how numbers appear in different fonts.

are not always the same size. Also be aware that oldstyle figures of Georgia drop some of the digits below the horizontal baseline (I've added an underline in each cell to make this clear).

RULE 4. LEFT-ALIGN TEXT AND HEADER

Once we've right-aligned the numbers, we should left-align the text. The English language is read from left to right, so lining up the entries in that way generates an even, vertical border and is natural for the reader. Notice how much easier it is to read the country names in the far-right column than in the other two columns.

Right-aligned and hard to read	Centered and even harder to read	Left-aligned and easiest to read
British Virgin Islands	British Virgin Islands	British Virgin Islands
Cayman Islands	Cayman Islands	Cayman Islands
Democratic Republic of Korea	Democratic Republic of Korea	Democratic Republic of Korea
Luxembourg	Luxembourg	Luxembourg
United States	United States	United States
Germany	Germany	Germany
New Zealand	New Zealand	New Zealand
Costa Rica	Costa Rica	Costa Rica
Peru	Peru	Peru

Rule 4. Left-align text and headers.

RULE 5. SELECT THE APPROPRIATE LEVEL OF PRECISION

Precision to the fifth-decimal place is almost never necessary. Strike a balance between necessary precision and a clean, spare table. The per capita GDP growth rate, for example, is never

Country	Too many decimals	Too few decimals	About right
China	6.12380	6	6.1
India	5.88984	6	5.9
United States	0.74279	1	0.7
Indonesia	3.84530	4	3.8
Mexico	1.58236	2	1.6
Pakistan	3.43865	3	3.4
Average	**2.63104**	**3**	**2.6**

Rule 5. Select the appropriate level of precision.

reported to five decimals—that would be unnecessary and suggest a level of precision that is not supported by the data. This can also go the other way: Don't report too few digits. Showing per capita GDP growth as whole numbers masks important variation across countries.

RULE 6. GUIDE YOUR READER WITH SPACE BETWEEN ROWS AND COLUMNS

Your use of space in and around the table can influence the direction in which your reader reads the data. In the table on the left, for example, there is more space between the columns than between the rows, so your eye is drawn to read the table top-to-bottom rather than left-to-right. By comparison, the table on the right has more space between the rows

Country	2014	2015	2016
China	6.76	6.36	6.12
India	6.14	6.90	5.89
United States	1.80	2.09	0.74
Indonesia	3.73	3.65	3.85
Mexico	-0.38	-4.37	-4.25
Pakistan	2.51	2.61	3.44
Average	**3.43**	**2.87**	**2.63**

Country	2014	2015	2016
China	6.76	6.36	6.12
India	6.14	6.90	5.89
United States	1.80	2.09	0.74
Indonesia	3.73	3.65	3.85
Mexico	-0.38	-4.37	-4.25
Pakistan	2.51	2.61	3.44
Average	**3.43**	**2.87**	**2.63**

Rule 6. Guide your reader with space between rows and columns.

than between the columns, so your eye is more likely to track horizontally rather than vertically. Use spacing strategically to match the order in which you want your reader to take in the table.

RULE 7. REMOVE UNIT REPETITION

Your reader knows that the values in your table are dollars because you told them in the title or subtitle. Repeating the symbol throughout the table is overkill and cluttering. Use the title or column title area to define the units, or place them in the first row only (remembering to align the numbers along the decimal). If you are mixing units within the table, be sure to make your labels clear.

Country	2014	2015	2016
China	6.76%	6.36%	6.12%
India	6.14%	6.90%	5.89%
United States	1.80%	2.09%	0.74%
Indonesia	3.73%	3.65%	3.85%
Mexico	-0.38%	-4.37%	-4.25%
Pakistan	2.51%	2.61%	3.44%
Average	**3.43%**	**2.87%**	**2.63%**

Country	2014	2015	2016
China	6.76%	6.36%	6.12%
India	6.14	6.90	5.89
United States	1.80	2.09	0.74
Indonesia	3.73	3.65	3.85
Mexico	-0.38	-4.37	-4.25
Pakistan	2.51	2.61	3.44
Average	**3.43**	**2.87**	**2.63**

Rule 7. Remove unit repetition

RULE 8. HIGHLIGHT OUTLIERS

Instead of showing just six countries and three years as in the previous example, what if we need to show twenty countries and ten years of data? In this case, we might want to highlight outlier values by making the text boldface, shading it with color, or even shading the entire cell. Some readers will wade through all of the numbers in the table because they need specific information, but many readers are more likely to look for only the most important values. Guiding them to those important numbers lets them answer their own questions about the data or better comprehend your argument.

	2010	2011	2012	2013	2014	2015	2016
China	10.10	9.01	7.33	7.23	6.76	6.36	6.12
India	8.76	5.25	4.13	5.10	6.14	6.90	5.89
United States	1.68	0.85	1.46	0.96	1.80	2.09	0.74
Indonesia	4.83	4.79	4.68	4.24	3.73	3.65	3.85
Brazil	6.50	3.00	0.98	2.07	**-0.38**	**-4.37**	**-4.25**
Pakistan	**-0.48**	0.61	1.34	2.21	2.51	2.61	3.44
Nigeria	5.00	2.12	1.52	2.61	3.52	**-0.02**	**-4.16**
Bangladesh	4.40	5.25	5.28	4.77	4.84	5.37	5.96
Russia	4.46	5.20	3.48	1.57	**-1.04**	**-3.04**	**-0.41**
Mexico	3.49	2.12	2.15	**-0.06**	1.45	1.90	1.58

	2010	2011	2012	2013	2014	2015	2016
China	10.10	9.01	7.33	7.23	6.76	6.36	6.12
India	8.76	5.25	4.13	5.10	6.14	6.90	5.89
United States	1.68	0.85	1.46	0.96	1.80	2.09	0.74
Indonesia	4.83	4.79	4.68	4.24	3.73	3.65	3.85
Brazil	6.50	3.00	0.98	2.07	**-0.38**	**-4.37**	**-4.25**
Pakistan	**-0.48**	0.61	1.34	2.21	2.51	2.61	3.44
Nigeria	5.00	2.12	1.52	2.61	3.52	**-0.02**	**-4.16**
Bangladesh	4.40	5.25	5.28	4.77	4.84	5.37	5.96
Russia	4.46	5.20	3.48	1.57	**-1.04**	**-3.04**	**-0.41**
Mexico	3.49	2.12	2.15	**-0.06**	1.45	1.90	1.58

Rule 8. Highlight outliers.

RULE 9. GROUP SIMILAR DATA AND INCREASE WHITE SPACE

Reduce repetition by grouping similar data or labels. Similar to eliminating dollars signs on every number value, we can reduce some of the clutter in our tables by grouping like terms or labels. In this next example, grouping the names of the country regions reduces the amount of repetitive information in the first column. You can also use spanner headers and rules to combine the same entry and reduce unnecessary repetition. Here, besides grouping the country names, I've also applied some of the guidelines discussed so far such as left-aligning text, right-aligning numbers, and using boldface headers and footers.

Region	Country	Per Capita GDP 2015	Per Capita GDP 2016	Percent Change
Asia	China	6496.62	6894	6.1238
Asia	India	1758.84	1862	5.8898
North America	United States	51933.40	52319	0.7428
Asia	Indonesia	3827.55	3975	3.8453
North America	Brazil	11351.57	10869	-4.2541
Asia	Pakistan	1140.21	1179	3.4387
Africa	Nigeria	2562.52	2456	-4.1601
Asia	Bangladesh	971.64	1030	5.9627
North America	Mexico	9717.90	9872	1.5824
Asia	Japan	47163.49	47661	1.0546
Africa	Ethiopia	487.29	511	4.9041
Middle East	Egypt	2665.35	2726	2.2633
Europe	Germany	45412.56	45923	1.1240
Middle East	Iran	6007.00	6734	12.1010
Middle East	Turkey	13898.75	14117	1.5734
Europe	France	41642.31	41969	0.7845
Average		15440	15631	2.6860

Region	Country	Per Capita GDP 2015	Per Capita GDP 2016	Percent Change
Africa	Ethiopia	487	511	4.90
	Nigeria	2,563	2,456	-4.16
Asia	Bangladesh	972	1,030	5.96
	China	6,497	6,894	6.12
	India	1,759	1,862	5.89
	Indonesia	3,838	3,975	3.85
	Japan	47,163	47,661	1.05
	Pakistan	1,140	1,179	3.44
Europe	France	41,642	41,969	0.78
	Germany	45,413	45,923	1.12
Middle East	Egypt	2,665	2,726	2.26
	Iran	6,007	6,734	12.10
	Turkey	13,899	14,117	1.57
North America	Mexico	9,718	9,872	1.58
	United States	51,933	52,319	0.74
South America	Brazil	11,352	10,869	-4.25
Average		**15,440**	**15,631**	**2.69**

Rule 9. Group similar data and increase white space.

While grouping like elements does help reduce the amount of clutter on the page, be aware that posting tables to the internet may require some concessions in this regard. If you post tables to websites as images, users will be unable to copy and paste the data from the table to another tool, and screen readers—which literally step through the table and read the values out loud (see Chapter 12)—will be unable to read the data values. Instead, because of current constraints in web programming languages and formats, you might need to forgo spanner headers and other special formatting decisions, depending on the tools you use to post it to the internet.

RULE 10. ADD VISUALIZATIONS WHEN APPROPRIATE

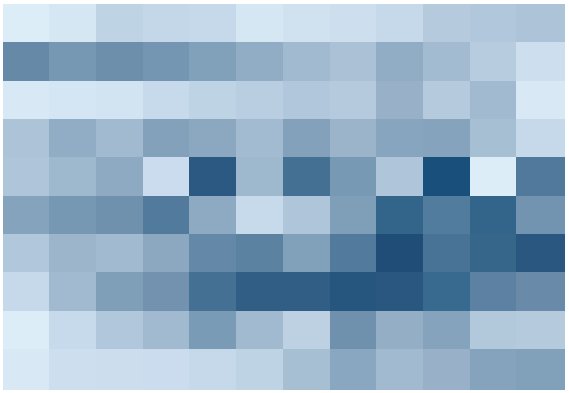
We can make larger changes to our tables by adding small visualizations. Just like highlighting outliers with color or boldface, you might add sparklines (see page 152) to visualize some data rather than showing every number. Or you can use small bar charts to visually illustrate a series of numbers. Or you could use a heatmap (see page 112) and leave the numbers in the table or hide them, which can help the reader focus on the overall patterns and ignore the details.

We can also embed a chart-type structure right into our table. If you want a full chart embedded within the table, a dot plot (see Chapter 4) is succinct and can line up well within the linear structure of a table. You can also use a modification on the standard dot plot to place the numbers in their relative positions directly in a table.

Country	2007	2016	2007-2016
China	13.64	6.12	
India	8.15	5.89	
United States	0.82	0.74	
Indonesia	4.91	3.85	
Mexico	0.70	1.58	
Pakistan	2.72	3.44	
Average	**5.15**	**3.60**	

Country	2016
China	6.12
India	5.89
United States	0.74
Indonesia	3.85
Mexico	1.58
Pakistan	3.44
Average	**3.60**

	Jan	Feb	Mar	Apr	May	June	Jul	Aug	Sept	Oct	Nov	Dec
1	10	13	24	22	21	13	16	17	21	28	30	32
2	65	57	62	58	52	45	38	33	45	37	27	17
3	12	14	15	20	24	26	30	28	42	28	38	12
4	32	45	38	51	47	37	51	41	49	50	35	21
5	31	39	46	19	92	39	80	56	31	97	10	75
6	50	57	61	74	46	20	31	53	86	73	86	59
7	30	40	38	47	66	69	52	74	98	78	85	93
8	21	38	53	60	80	90	90	94	93	83	70	64
9	10	20	30	38	55	38	25	61	44	50	29	28
10	12	17	18	19	21	24	35	48	38	42	50	52



Country	2007	2016
China	13.64	6.12
India	8.15	5.89
United States	0.82	0.74
Indonesia	4.91	3.85
Mexico	0.7	1.58
Pakistan	2.72	3.44
Average	5.86	2.63

Rule 10. Add visualizations when appropriate, for example, sparklines, bar charts, heatmaps or dot plots.

DEMONSTRATION: A BASIC DATA TABLE REDESIGN

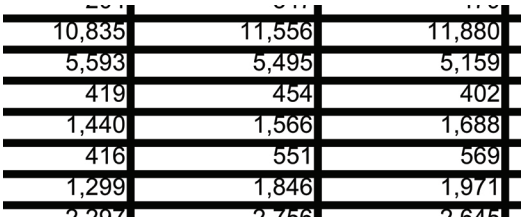
This table from the U.S. Department of Agriculture Food and Nutrition Service shows the number of people who participate in the Food Distribution Programs on Indian Reservations. The table presents participation estimates for twenty-four states over fiscal years 2013 through 2016, plus preliminary estimates for fiscal year 2017. Note the very dark, thick

FOOD DISTRIBUTION PROGRAM ON INDIAN RESERVATIONS: PERSONS PARTICIPATING
(Data as of March 9, 2018)

State	FY 2013	FY 2014	FY 2015	FY 2016	FY 2017
					Preliminary
Alaska	204	347	479	650	724
Arizona	10,835	11,556	11,880	11,887	11,235
California	5,593	5,495	5,159	4,795	4,463
Colorado	419	454	402	442	353
Idaho	1,440	1,566	1,688	1,706	1,530
Kansas	416	551	569	592	613
Michigan	1,299	1,846	1,971	2,061	1,960
Minnesota	2,297	2,756	2,645	2,600	2,487
Mississippi	701	863	958	1,056	1,169
Montana	2,375	3,144	3,149	3,313	3,271
Nebraska	1,010	1,229	1,339	1,396	1,267
Nevada	1,373	1,611	1,508	1,468	1,328
New Mexico	2,533	2,853	2,966	2,890	2,809
New York	380	384	369	452	350
North Carolina	584	736	743	700	671
North Dakota	3,840	4,800	4,976	5,661	5,569
Oklahoma	25,678	29,012	31,042	33,588	32,795
Oregon	678	871	800	785	687
South Dakota	7,457	8,123	8,208	8,505	8,525
Texas	117	131	142	124	114
Utah	117	167	217	421	384
Washington	3,164	3,185	3,284	3,410	3,221
Wisconsin	2,441	2,978	3,240	3,442	3,367
Wyoming	657	742	881	1,096	1,190
TOTAL	**75,608**	**85,397**	**88,615**	**93,038**	**90,083**

FDPIR is an alternative to the Supplemental Nutrition Assistance Program for Indian tribal organizations which prefer food distribution. Participation numbers are 12-month averages. Data are subject to revision.

This table from the U.S. Department of Agriculture Food and Nutrition Service is cluttered and difficult to read.



10,835	11,556	11,880
5,593	5,495	5,159
419	454	402
1,440	1,566	1,688
416	551	569
1,299	1,846	1,971

Notice those heavy gridlines in the USDA table.
Source: US Department of Agriculture

gridlines, which make the table cluttered and difficult to read. As we zoom in, we can see that the numbers are top-aligned in each cell, which cuts them off ever-so-slightly.

There is a better way. Instead of including all of the gridlines—and making them dark and thick—we can remove them and keep only the line below the column header row. The

Number of People Participating in Food Distribution Programs on Indian Reservations
(Data as of March 9, 2018)

State	FY 2013	FY 2014	FY 2015	FY 2016	Preliminary, FY 2017
Alaska	204	347	479	650	724
Arizona	10,835	11,556	11,880	11,887	11,235
California	5,593	5,495	5,159	4,795	4,463
Colorado	419	454	402	442	353
Idaho	1,440	1,566	1,688	1,706	1,530
Kansas	416	551	569	592	613
Michigan	1,299	1,846	1,971	2,061	1,960
Minnesota	2,297	2,756	2,645	2,600	2,487
Mississippi	701	863	958	1,056	1,169
Montana	2,375	3,144	3,149	3,313	3,271
Nebraska	1,010	1,229	1,339	1,396	1,267
Nevada	1,373	1,611	1,508	1,468	1,328
New Mexico	2,533	2,853	2,966	2,890	2,809
New York	380	384	369	452	350
North Carolina	584	736	743	700	671
North Dakota	3,840	4,800	4,976	5,661	5,569
Oklahoma	25,678	29,012	31,042	33,588	32,795
Oregon	678	871	800	785	687
South Dakota	7,457	8,123	8,208	8,505	8,525
Texas	117	131	142	124	114
Utah	117	167	217	421	384
Washington	3,164	3,185	3,284	3,410	3,221
Wisconsin	2,441	2,978	3,240	3,442	3,367
Wyoming	657	742	881	1,096	1,190
Total	**75,608**	**85,397**	**88,615**	**93,038**	**90,083**

Note: FDPIR is an alternative to the Supplemental Nutrition Assistance Program for Indian tribal organizations which prefer food distribution. Participation numbers are 12-month averages. Data are subject to revision.

A simple redesign of the USDA table removes the clutter and lightens the view.

column header text is now bold to distinguish it from the numbers in the table. A line at the bottom of the table separates it from the note, and the *Total* row is bolded to set it apart from the body of the table.

Let's take this a step further by adding some visuals and color to the table.

Number of People Participating in Food Distribution Programs on Indian Reservations
(Data as of March 9, 2018)

State	FY 2013	FY 2014	FY 2015	FY 2016	Preliminary, FY 2017
Alaska	204	347	479	650	724
Arizona	10,835	11,556	11,880	11,887	11,235
California	5,593	5,495	5,159	4,795	4,463
Colorado	419	454	402	442	353
Idaho	1,440	1,566	1,688	1,706	1,530
Kansas	416	551	569	592	613
Michigan	1,299	1,846	1,971	2,061	1,960
Minnesota	2,297	2,756	2,645	2,600	2,487
Mississippi	701	863	958	1,056	1,169
Montana	2,375	3,144	3,149	3,313	3,271
Nebraska	1,010	1,229	1,339	1,396	1,267
Nevada	1,373	1,611	1,508	1,468	1,328
New Mexico	2,533	2,853	2,966	2,890	2,809
New York	380	384	369	452	350
North Carolina	584	736	743	700	671
North Dakota	3,840	4,800	4,976	5,661	5,569
Oklahoma	25,678	29,012	31,042	33,588	32,795
Oregon	678	871	800	785	687
South Dakota	7,457	8,123	8,208	8,505	8,525
Texas	117	131	142	124	114
Utah	117	167	217	421	384
Washington	3,164	3,185	3,284	3,410	3,221
Wisconsin	2,441	2,978	3,240	3,442	3,367
Wyoming	657	742	881	1,096	1,190
Total	**75,608**	**85,397**	**88,615**	**93,038**	**90,083**

Note: FDPIR is an alternative to the Supplemental Nutrition Assistance Program for Indian tribal organizations which prefer food distribution. Participation numbers are 12-month average. Data are subject to revision.

Adding a little color to the USDA table—such as a simple heatmap—makes it easier and faster for our reader to pick out specific values or patterns.

Number of People Participating in Food Distribution Programs on Indian Reservations
(Data as of March 9, 2018)

State	FY 2013	FY 2014	FY 2015	FY 2016	FY 2017	Average FY 2013-FY 2017
Alaska	204	347	479	650	724	481
Arizona	10,835	11,556	11,880	11,887	11,235	11,479
California	5,593	5,495	5,159	4,795	4,463	5,101
Colorado	419	454	402	442	353	414
Idaho	1,440	1,566	1,688	1,706	1,530	1,586
Kansas	416	551	569	592	613	548
Michigan	1,299	1,846	1,971	2,061	1,960	1,827
Minnesota	2,297	2,756	2,645	2,600	2,487	2,557
Mississippi	701	863	958	1,056	1,169	949
Montana	2,375	3,144	3,149	3,313	3,271	3,050
Nebraska	1,010	1,229	1,339	1,396	1,267	1,248
Nevada	1,373	1,611	1,508	1,468	1,328	1,458
New Mexico	2,533	2,853	2,966	2,890	2,809	2,810
New York	380	384	369	452	350	387
North Carolina	584	736	743	700	671	687
North Dakota	3,840	4,800	4,976	5,661	5,569	4,969
Oklahoma	25,678	29,012	31,042	33,588	32,795	30,423
Oregon	678	871	800	785	687	764
South Dakota	7,457	8,123	8,208	8,505	8,525	8,164
Texas	117	131	142	124	114	126
Utah	117	167	217	421	384	261
Washington	3,164	3,185	3,284	3,410	3,221	3,253
Wisconsin	2,441	2,978	3,240	3,442	3,367	3,094
Wyoming	657	742	881	1,096	1,190	913
Total	**75,608**	**85,397**	**88,615**	**93,038**	**90,083**	**86,548**

Note: FDPIR is an alternative to the Supplemental Nutrition Assistance Program for Indian tribal organizations which prefer food distribution. Participation numbers are 12-month average. Data are subject to revision.

Number of People Participating in Food Distribution Programs on Indian Reservations
(Data as of March 9, 2018)

State	FY 2013	FY 2014	FY 2015	FY 2016	FY 2017	Percent Change FY 2013-FY 2017
Alaska	204	347	479	650	724	254.9 ▲
Arizona	10,835	11,556	11,880	11,887	11,235	3.7 ▲
California	5,593	5,495	5,159	4,795	4,463	-20.2 ▼
Colorado	419	454	402	442	353	-15.8 ▼
Idaho	1,440	1,566	1,688	1,706	1,530	6.3 ▲
Kansas	416	551	569	592	613	47.4 ▲
Michigan	1,299	1,846	1,971	2,061	1,960	50.9 ▲
Minnesota	2,297	2,756	2,645	2,600	2,487	8.3 ▲
Mississippi	701	863	958	1,056	1,169	66.8 ▲
Montana	2,375	3,144	3,149	3,313	3,271	37.7 ▲
Nebraska	1,010	1,229	1,339	1,396	1,267	25.4 ▲
Nevada	1,373	1,611	1,508	1,468	1,328	-3.3 ▼
New Mexico	2,533	2,853	2,966	2,890	2,809	10.9 ▲
New York	380	384	369	452	350	-7.9 ▼
North Carolina	584	736	743	700	671	14.9 ▲
North Dakota	3,840	4,800	4,976	5,661	5,569	45.0 ▲
Oklahoma	25,678	29,012	31,042	33,588	32,795	27.7 ▲
Oregon	678	871	800	785	687	1.3 ▲
South Dakota	7,457	8,123	8,208	8,505	8,525	14.3 ▲
Texas	117	131	142	124	114	-2.6 ▼
Utah	117	167	217	421	384	228.2 ▲
Washington	3,164	3,185	3,284	3,410	3,221	1.8 ▲
Wisconsin	2,441	2,978	3,240	3,442	3,367	37.9 ▲
Wyoming	657	742	881	1,096	1,190	81.1 ▲
Total	**75,608**	**85,397**	**88,615**	**93,038**	**90,083**	**19.1**

Note: FDPIR is an alternative to the Supplemental Nutrition Assistance Program for Indian tribal organizations which prefer food distribution. Participation numbers are 12-month average. Data are subject to revision.

Adding other visualizations—bar charts or icons denoting change—are other ways to add visual elements to your tables.

The first example is a heatmap. Until I made this, I didn't realize by how much program participation in Oklahoma exceeded the rest of the states. It was only after the row appeared in dark blue that the magnitude became clear.

Another approach is to maintain the core look of the original table but add additional visual elements. In the tables above, the version on the left in the above pair adds a new data point—the average between fiscal year 2013 and 2017—and a bar chart to its right. This small graphic element gives the table a visual anchor and directs the eye to the states with more participation. The chart on the right adds the percentage change between 2013 and 2017 and a small up- or down-arrow to signal the change.

DEMONSTRATION: A REGRESSION TABLE REDESIGN

A typical regression table contains point estimates, standard errors, and some symbol (usually asterisks) to denote the level of statistical significance, such as 1 percent, 5 percent, and 10 percent. Such basic tables are especially useful when readers need the detailed numbers.

	Model 1	Model 2	Model 3
r_age	0.0509***	0.0119***	0.0207***
	(0.0062)	(0.0044)	(0.0026)
gndr	0.0442***	0.0616***	0.0630***
	(0.0057)	(0.0037)	(0.0043)
_educ	0.0027***	0.0052***	0.0157***
	(0.0087)	(0.0050)	(0.0072)
hrswkd	0.0397***	0.0075***	0.0211***
	(0.0053)	(0.0025)	(0.0029)
expr	0.0003***	0.0043***	0.0030***
	(0.0051)	(0.0026)	(0.0024)
marstat	0.0191***	0.0066***	0.0069***
	(0.0053)	(0.0025)	(0.0027)

* p < 0.05, ** p < 0.01, *** p < 0.001

	Model 1	Model 2	Model 3
Age	0.0509***	0.0119***	0.0207***
	(0.0062)	(0.0044)	(0.0026)
Gender	0.0442***	0.0616***	-0.0630***
	(0.0057)	(0.0037)	(0.0043)
Education	0.0027	0.0052	0.0157**
	(0.0087)	(0.0050)	(0.0072)
Hours Worked	0.0397***	0.0075*	0.0211***
	(0.0053)	(0.0044)	(0.0029)
Experience	0.0003	0.0043*	0.0030
	(0.0051)	(0.0026)	(0.0024)
Married	0.0191***	0.0066***	0.0069*
	(0.0053)	(0.0025)	(0.0041)

* p < 0.05, ** p < 0.01, *** p < 0.001

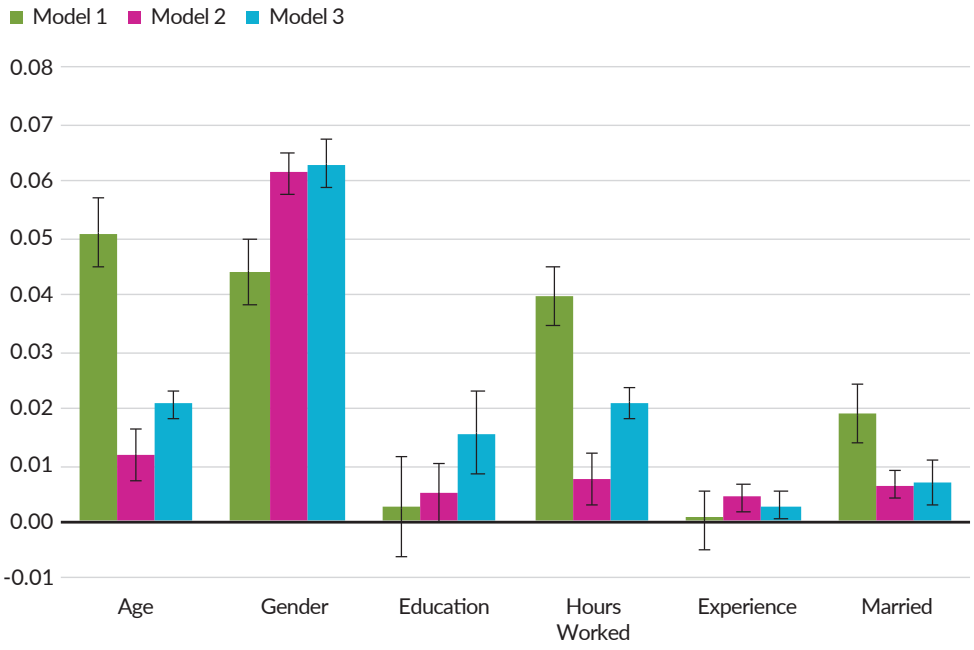
A table of basic regression results can be improved following the ten rules shown earlier.

We can make a table of regression estimates clearer and more visually engaging by following the ten table rules and the visualization strategies from earlier in this chapter. We might also consider putting the dense table in an appendix (in the paper itself or maybe online) and using a graph in the main body of the paper instead.

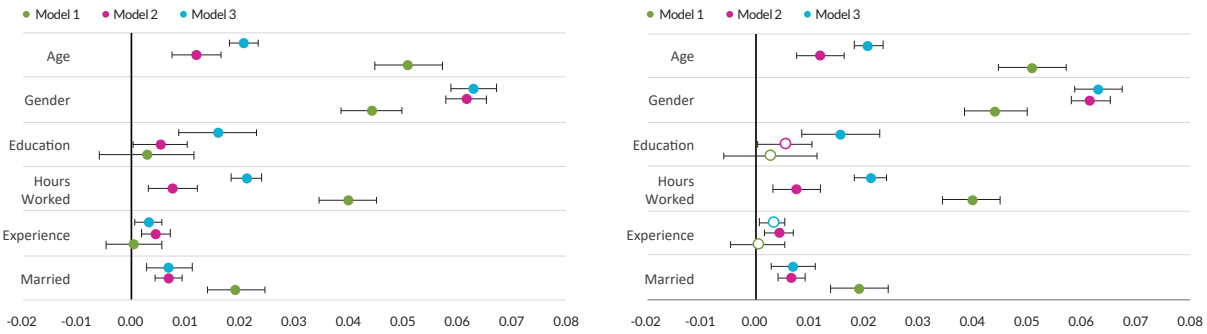
Consider this relatively simple regression table that includes the coefficient estimates with asterisks, standard errors with parentheses, and unreadable variable names in the first column. Don't use variable names to list your results! Your reader—even a reader of an academic journal article—does not know what "_educ" or "expr" means. For our final table, let's use real words like "Education" and "Experience," and use the rules above to make the table cleaner and easier to read.

We can also convert these kinds of tables into data visualizations. A standard way is to use a bar chart with error bars, though, as noted in Chapter 6, some research has shown that we tend to discount the end of the error bar that sits within the bar itself.

Or we could try a dot plot approach (with or without error bars) and maybe use color to further signify statistical significance. In the graph on the right, solid circles contain estimates that are statistically significant and empty circles are estimates that are not statistically significant.



Regression results can be shown as a bar chart instead of a table.



A dot plot is another way to visualize regression results.

Ultimately, if you decide such visual elements are unnecessary or insufficient, stick with the ten rules to make the table clearer and easier to read. Remember, the goal of our tables is to let the reader to more easily find the important numbers and patterns in the data and not ask them to wade through clutter.

CONCLUSION

Tables are themselves a form of data visualization, and the same rules apply. Many researchers and scholars rely heavily on tables, likely because they don't require much creative thinking—filled with text and numbers, columns and rows intersect, and details are left for the reader to navigate and decipher. And while tables are valuable and have their place, we can use these ten strategies to elevate our tables and make them clearer and easier to read.

Rule 1. Offset the headers from body
Rule 2. Use subtle dividers instead of heavy gridlines
Rule 3. Right-align numbers and headers
Rule 4. Left-align text and header
Rule 5. Select the appropriate level of precision
Rule 6. Guide your reader with space between rows and columns
Rule 7. Remove unit repetition
Rule 8. Highlight outliers
Rule 9. Group similar data and increase white space
Rule 10. Add visualizations when appropriate

PART THREE

DESIGNING AND REDESIGNING YOUR VISUAL

DEVELOPING A DATA VISUALIZATION STYLE GUIDE

A data visualization style guide does for graphs what the *Chicago Manual of Style* does for English grammar. It defines the components of a graph and their proper, consistent use. Like a writing style guide, a comprehensive data visualization style guide breaks down the parts of graphs, charts, and tables to demonstrate best practices and strategies to design and style your charts. Elements like font and color, the widths of lines and style gridlines, and the use of tick marks are all choices that determine whether a graph is clear, engaging, and consistent—or whether it isn't.

The difference between a grammar guide and a data visualization guide is that many of our data style decisions are subjective. While the word *their* is objectively different than *they're*, and the use of one in a particular case is either correct or incorrect, there is no objectively correct or incorrect line thickness for a chart. There are, however, certain principles to consider, many of which we have covered so far. But for the most part, the styles you choose will reflect you and your organization's preferences.

THE ELEMENTS OF A DATA VISUALIZATION GUIDE

In organizations, a data visualization style guide serves three purposes.

First, it provides team members with the detailed styles and expectations about what should and should not be included in a visualization. Where should the title go? How large should it be? What font? What color?

Second, it guides those who may not be familiar with (or care about) all the styling and branding guidelines the organization may value. Instead of asking researchers and analysts to compile the data, create the graph, and then worry about which colors and fonts to use, a style guide makes those decisions easier. Building these styles into software tools streamlines the process and automates the application of graph styles.

Finally, a style guide sets the tone and expectations for people in the organization that the style, look, and details about data visualization are as important as other branding materials.

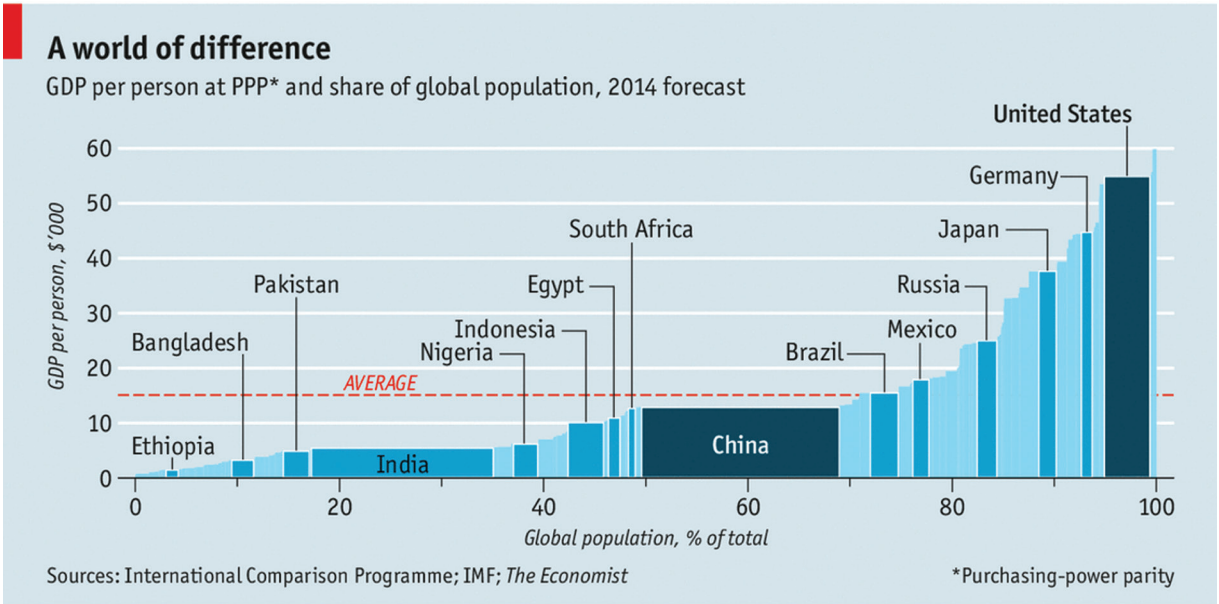
Even if you're an individual working with data, a style guide can be worthwhile. A custom style guide will make your work more consistent and efficient, and it will build your individual brand so your work stands out. A good style guide handles the basic style decisions for you, so you can focus on more important aspects of creating data visualizations.

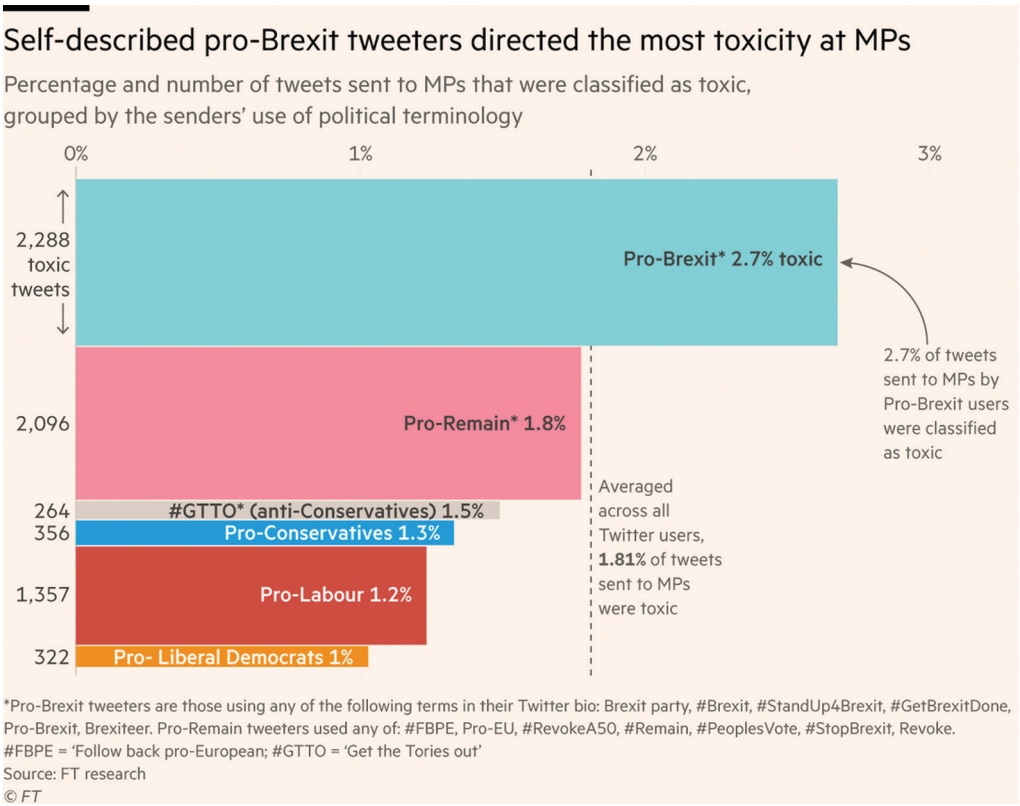
As you build your style guide, test the components to make sure you or your team members can use and implement them. The style needs of your charts may differ from those of other branding materials. Colors that might look great in a logo may not work in a line chart or bar chart. Also remember to treat your data visualization style guide as a living document, just as you would a style guide for text or design. The guide should change as your personal or organizational aesthetic changes and evolve alongside changes in publication types and software tools.

Consider these Marimekko charts from the *Economist* and the *Financial Times*. Both publications have a distinct look and feel, and even those who are not regular readers may recognize the style. This branding is an important aspect of the organization's identity.

There can be many sections to a detailed style guide, which we'll cover in detail in this chapter, but here are the basics that any data visualization style guide should cover.

1. **Graph Anatomy.** Where should labels, titles, and other elements be placed? What is the proper size of charts and should this size differ for different types of output?
2. **Color Palette.** What colors should be used across graph and data types? Does the color palette vary across graph types? Does it vary for print and digital products?
3. **Font.** What font should be used and how should its size, boldness, and position vary? Should there be one font style for the title and another for the text in the body of the graph?





The color, fonts, and overall style of these Marimekko charts from the *Economist* (top) and the *Financial Times* (bottom) make them easily identifiable.

4. **Graph Types.** Are there special considerations for specific graphs? For example, are pie charts forbidden in all situations? Is there a maximum number of series allowed in a line chart?
5. **Exporting Images.** How should team members move graphs from their software tool to the final report or website? Should they use PNG, JPEG, or some other image format? How should people create those image formats if they are not native to their software tool?
6. **Accessibility, Diversity, and Inclusion.** What steps do you and your organization need to take to make your graphs accessible to people with vision impairments or intellectual or other disabilities? Are you being mindful of how you're presenting results for different races, genders, and other groups?

THE ANATOMY OF A GRAPH

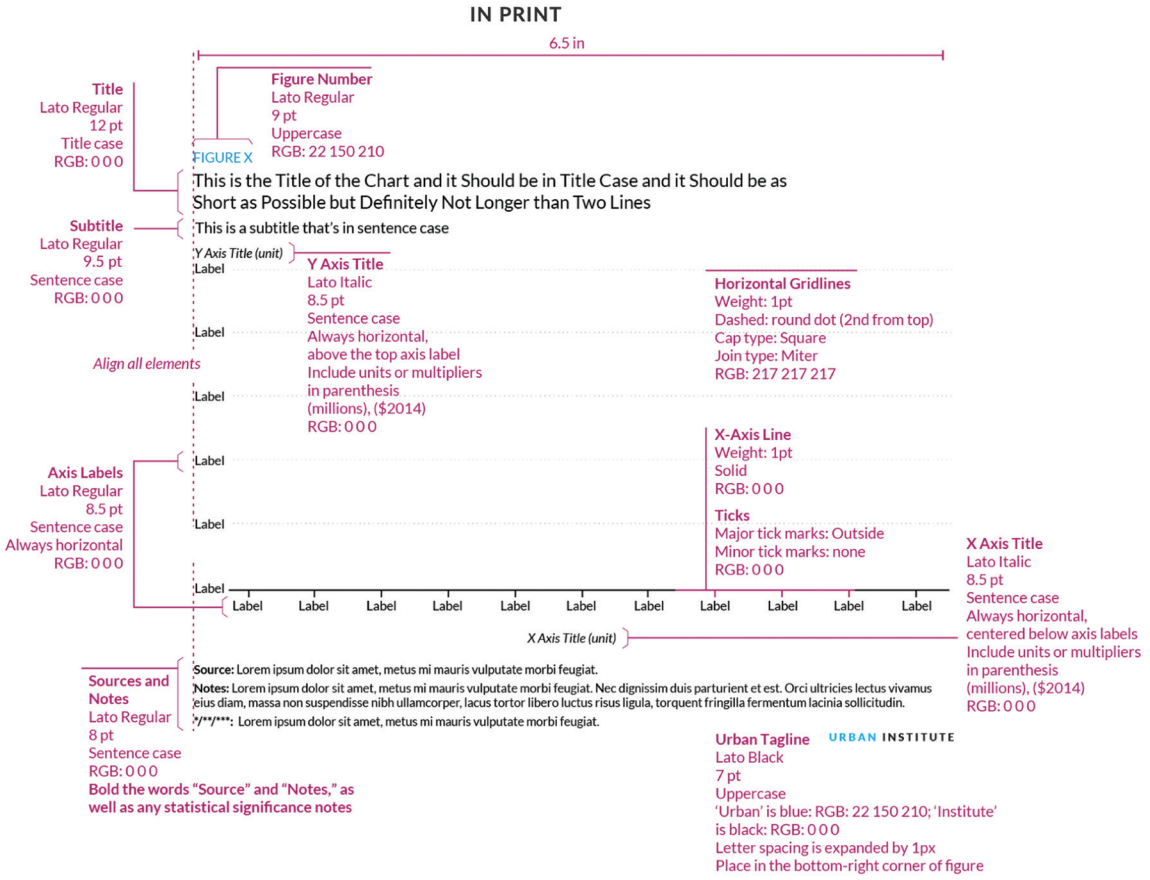
To set graph styles, we should first define each part of a graph. To illustrate how this can be done in practice, we can use the basic template in the style guide published by the Urban Institute, a nonprofit research institution in Washington, DC.

1. OVERALL DIMENSIONS

Specify the overall size of the chart. This may differ for different product types—an online graph, for example, is often measured in pixels, while a chart for a print document is typically measured in inches or centimeters. The dimensions may depend on the tools and workflow your organization uses. In the Urban Institute guide, the horizontal dimensions for a print graphic are specified at the top (6.5 inches) to fit on an 8.5 × 11 inch page.

2. FIGURE NUMBER

Figures and tables can either be numbered, lettered, or left alone. You can place it above the chart title, centered or left-aligned, in a different font size and color. You could also leave it in-line with the chart title like, "Figure 1. Chart Title." As you can see in the image, the Urban style is to put the figure number above the chart title in all capital letters in its standard blue color. The note includes details for font (Lato Regular), font size (9 pt), capitalization (Uppercase), and color (RGB: 22 150 210).



The Urban Institute style guide defines each part of the chart and style choices that apply to them.

3. TITLE

Will the title be aligned with the left side of the chart or centered over the plot space? Left-aligning the title (and other text objects) has the advantage of creating a vertical grid along the left edge, which makes it appear more organized and therefore easier to read. You should also lay out a basic style for the tone of the titles. Will they be purely descriptive or more active, summarizing the important points of the chart? Will you use sentence case or title case? The Urban style is to use title case, as detailed in the sample title. Notice also how the

use of the title is a direction to the user—it should be title case, short as possible, and no more than two lines.

4. SUBTITLE

If there are going to be subtitles, how will they be used in the chart? Is this a place to insert a more active statement or is it a place to list the units in the chart? The subtitle is a good place to include the vertical axis title because, when left-aligned, it's located close to the top of the axis. To offset it from the title, you might place it in parentheses, make the size smaller, or even change the color. In the Urban example, the subtitle is written out in sentence case with a smaller size and black color.

5. AXIS TITLES

Where will the vertical and horizontal axis titles go? In many software tools, the vertical axis title is rotated and placed alongside the vertical axis. A better position is to have it horizontally oriented and positioned above the vertical axis, aligned with the title and subtitle (or, as mentioned, it might be the subtitle). For the horizontal axis title, you might need to decide how far below the axis labels it will sit. There are cases—such as months or years—where the units are obvious and a horizontal axis title can be omitted. Axis titles can be differentiated by using smaller text or different colors. You should also decide whether to spell out and capitalize units like "dollar" or "percent," or to use a symbol. The Urban style is to place the vertical (y-) axis title above the axis and to have units in parentheses; the horizontal axis label sits below the axis in 8.5 pt Lato Italic font, horizontal, and centered.

6. AXIS LABELS

How these should be formatted? Boldface, italics, different font size? The vertical axis labels (as distinct from the title) typically sit to the left of the chart, though they can also be added to the right side if the chart is very wide. For the horizontal axis labels, are there specific formats for certain units? For example, when using years along the axis, would a series like 2000, '01, '02 . . . be acceptable or should each number be written out in full?

7. AXIS LINES AND TICK MARKS

What color and thickness will you use for the axis lines? Will the tick marks be inside or outside the chart? Some organizations leave out the vertical axis line altogether, but the horizontal axis line is typically included to give the chart a consistent anchor. I prefer to make the zero-axis line slightly darker than the other gridlines because it acts as a baseline. This is especially true in cases with negative values: We want to make it clear that the zero-axis line is not at the bottom of the chart. Tick marks are likely not needed in the space between the bars in a bar chart, but may be necessary in a line chart. In the Urban example, there is no vertical axis line, but the horizontal axis line is a 1 pt black line with major tick marks that are outside the chart.

8. GRIDLINES

Many charts include horizontal gridlines, though the exact formatting varies. Will they be solid, dashed, or dotted? How thick will they be? And what color? At what increments will they be added? Many charts do not include vertical gridlines, though the occasional scatterplot will include them to create a visible grid.

9. SOURCES AND NOTES

Data sources should be documented and note any important modeling or modifications. A box for sources and notes is typically found at the bottom of the chart, left-aligned with the vertical axis labels, title, and subtitle. In many cases, the word *Source* and *Note* are boldfaced. The Chicago Manual of Style (section 3.20), for example, suggests placing the source line above the note line. In the Urban style, the words *Source* and *Notes* are in bold face and ordered in that way.

10. LOGO

If you want to include a logo on the graph, decide where it will go and what size it will be (and be sure to use a high-resolution image). Logos are often placed in the bottom-right corner, but sometimes in other places. The advantage of placing it in the bottom-right corner is that

it is out of the way of the title/subtitle and sources/notes areas. Urban adds one of its logo formats to the bottom-right area of the graph, with specific instructions for color and spacing.

11. LEGEND

Will a legend be used and if so, where will it go, what size will it be, and what markers will be used? It is not labeled on this image, but the Urban style guide includes a separate section that specifies font sizes for other elements of graphs, including the legend.

12. DATA MARKERS

Will graphs, especially line graphs, include data markers, like circles or squares? Will the markers be filled or hollow? When will data values be labeled? You may want to set rules about using data markers for graphs with some number of values.

13. DATA LABELS

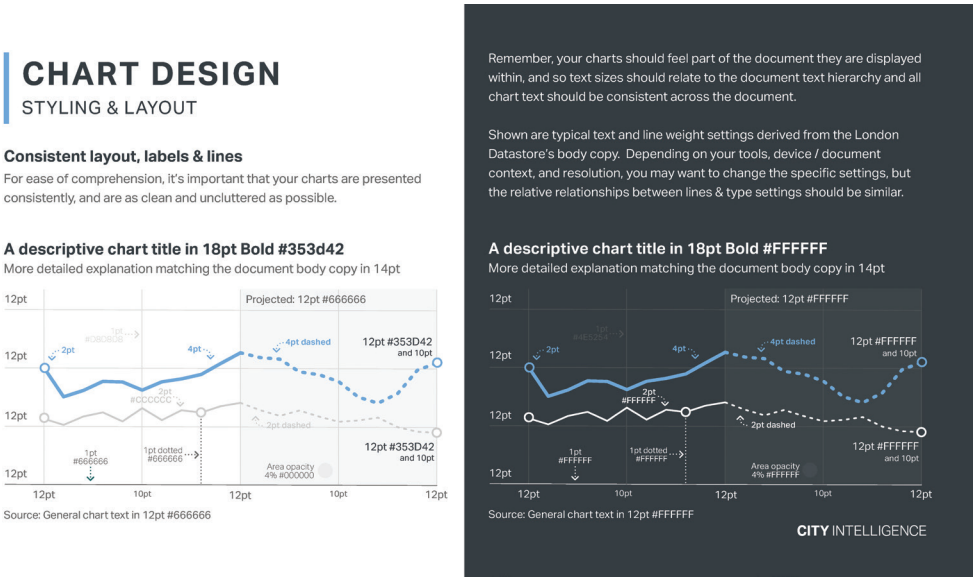
Determine when data points should be labeled and how they should be placed and formatted. The Urban guide has a separate table of font sizes that describes how these labels should appear.

14. DATA SERIES

This will vary by chart type—thickness of lines, space between bars and columns, colors for each element. You may need a separate section of the style guide to address issues of specific chart types, depending on the complexity of the charts your organization uses.

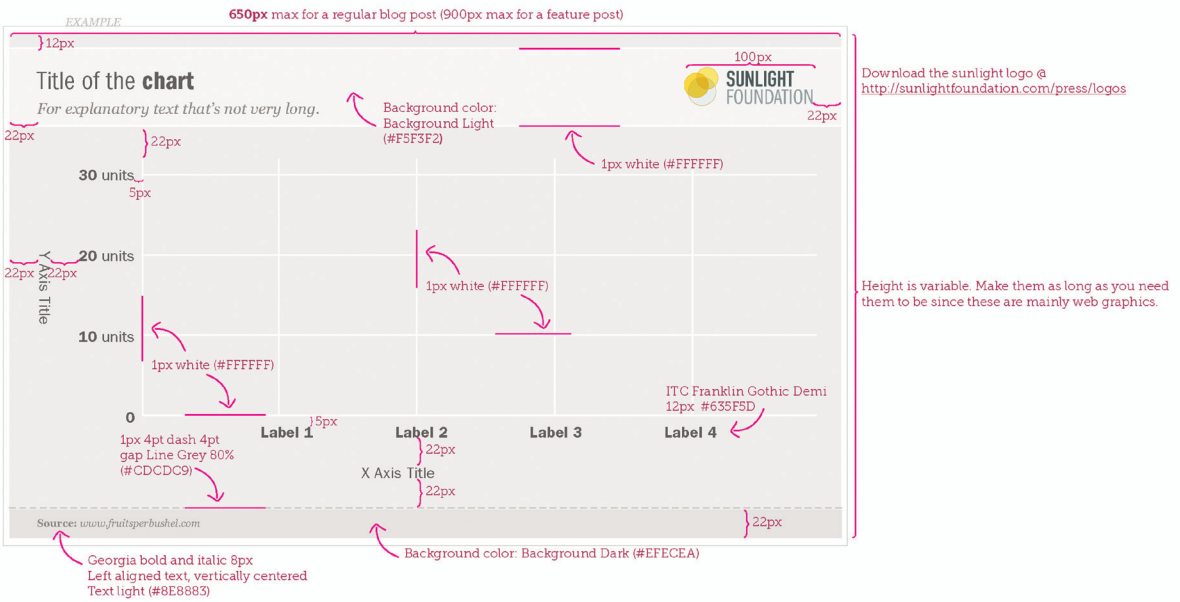
▸ ▸ ▸ ▸ ▸

As just two other examples of how published style guides define different parts of a chart, the first image is from the data visualization style guide published by the London Datastore, an effort by the city of London to open and share its data and resources. Similarly, the Sunlight Foundation, a nonpartisan organization that advocates for open government, has a style guide that defines each part of their charts to reflect their styling preferences.



Source: Mike Brondbjerg of the Greater London Authority, reproduced under the Open Government License

Basic Structure



Source: Sunlight Foundation

A data visualization style guide should lay out specific chart fonts, styles, colors, and sizes.

COLOR PALETTES

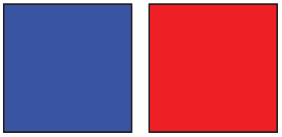
Color has unmistakable power in our visualizations. It may be the first thing people notice about our graphs. Color can evoke emotions and draw attention. As Vincent van Gogh wrote to his brother in 1885, "Color expresses something in itself. One can't do without it; one must make use of it."

Successful brands have recognizable color palettes for everything from their logo and letterhead to their data visualizations. But a palette that works for company letterhead or website may not necessarily work for a line chart with five lines. There are a number of free online color tools to develop color palettes: Adobe Color, Color Brewer, Colour Lovers, and Design Seeds are a few examples, and the Appendix contains a longer list. Besides the basic colors, we will also need different shades and tints for each color in the palette.

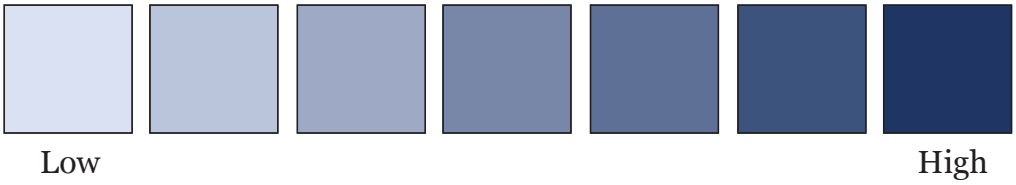
A style guide should contain different schemes to guide chart creators in their color choices. The easier you make it for an analyst to apply the branding and design elements, the more time they have to work with the data and develop the best graph for their purpose.

There are five primary color schemes you can apply to your data visualizations.

Binary. Nominal differences divided into two (binary) categories: urban-rural, Democrat-Republican, agree-disagree.

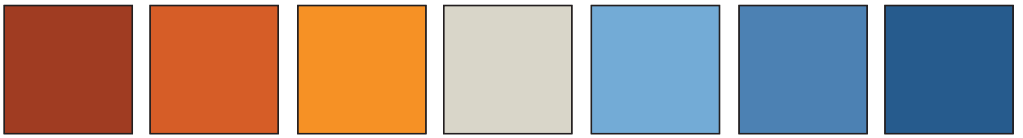


Sequential. Data values that are logically arranged from low to high should use sequential color schemes. Low values are usually represented by light colors, and high values by

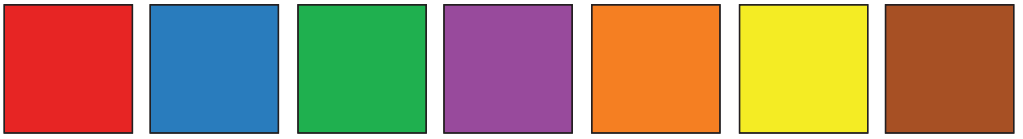


dark colors. Choropleth maps that show poverty rates or population, for example, would use sequential color palettes.

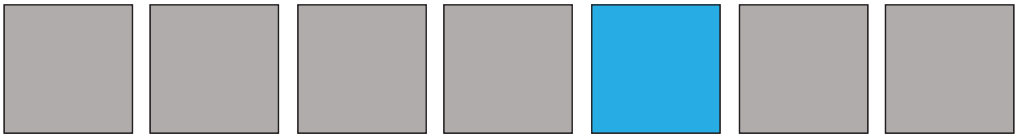
Diverging. In this scheme, the colors progress outward, growing darker from a central midpoint. A diverging color palette will share sequential schemes on two different colors and diverge from a shared, lighter color, for example, deviations from zero or a central number.



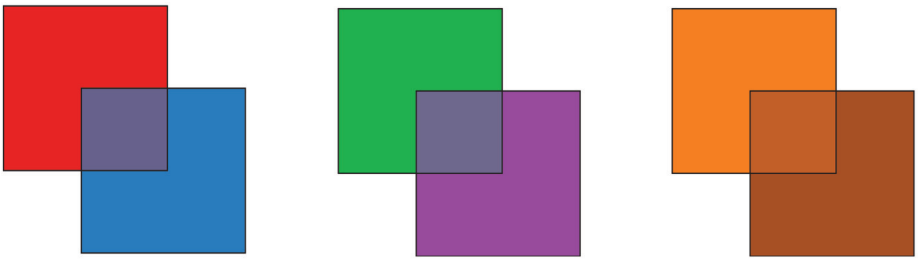
Categorical. Color schemes that use separate colors to represent nominal differences, for example, different race or gender groups.



Highlighting. This is a special case of the categorical color scheme. These color schemes highlight a certain value or group within the visualization. For example, we could use this palette to emphasize a single or small group of points in a scatterplot.

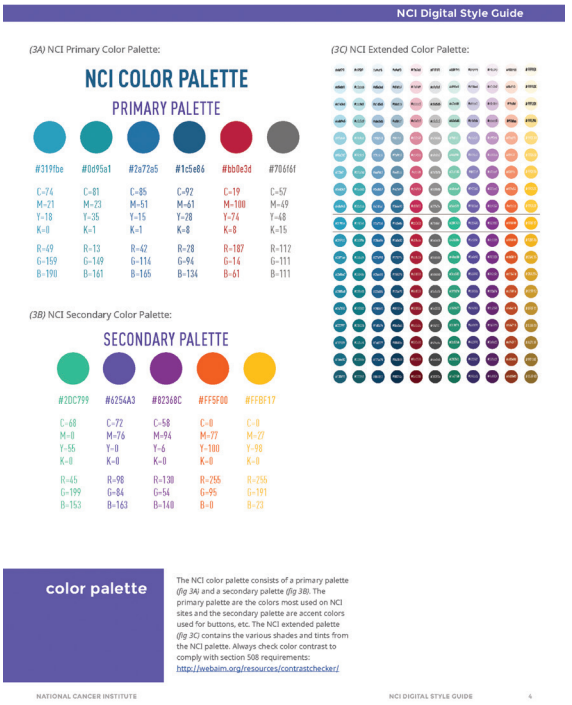


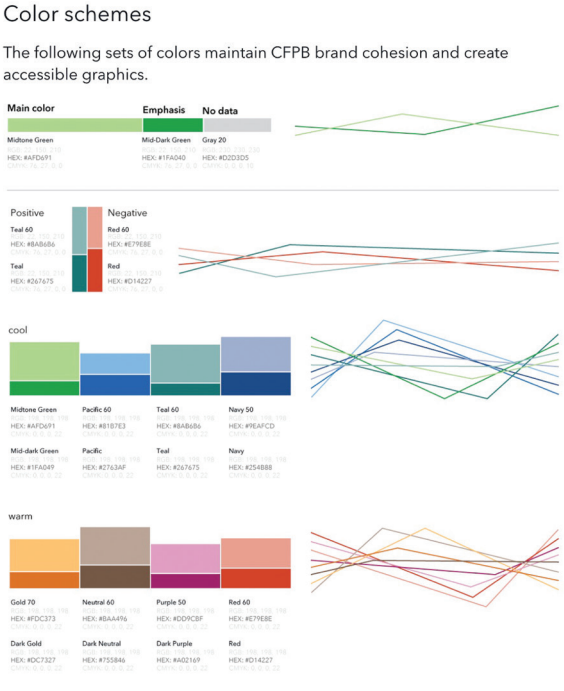
Transparency. Not so much a color scheme but a technique for using color, transparency in graph objects lets us (and our reader) see the object behind it. We've seen this technique a few times already (see the area chart section on page 157 as an example). You can use transparent colors—with or without a solid border—to make overlapping objects visible.



As an example of how color palettes are defined in practice, this section of the National Cancer Institute style guide shows primary and secondary color palettes along with a full list of tints and shades. The Consumer Finance Protection Bureau (CFPB) style guide includes sets of colors to "maintain CFPB brand cohesion."

We should also be mindful of readers with color vision deficiency (CVD) or color blindness. About 300 million people around the world have some form of CVD, most of them





There are many ways to define branded colors and styles. The National Cancer Institute (left) and the Consumer Finance Protection Bureau (right) are demonstrations of two such ways.

men, and most of them have difficulty discerning between reds and greens, though there are other forms as well. There are a variety of online color and color contrast checking tools such as Vischeck.com and WebAim that can be used to test colors.

AVOID THE RAINBOW

When choosing a color palette, avoid the rainbow color palette. In most cases, the rainbow palette is a poor choice for visualizing your data for at least three reasons. First, while a color ramp from light blue (small data values) to dark blue (large data values) makes logical sense, it isn't really logical to say that "purple" means more than "orange." Second, and more importantly, the rainbow color palette does not map to our number system. Notice how wide the green area is in the rainbow palette below compared with the thin light blue area. If we were to show a unit change from, say, 1 to 2, we might not see a change in greens; but the same unit change from, say, 9 to 10, in the blue spectrum might shift all the way from teal to

Avoid the rainbow color palette. It doesn't map to our number system, isn't logical when mapped to data, is not comprehensible for people with color vision deficiency, and does not translate to grayscale.

navy. Finally, the rainbow palette is not consistent for people with CVD (the middle image) or when printed in black and white (the last image).

COLORS AND CULTURE

Finally, be mindful that colors can reinforce stereotypes or hold different meanings in different cultures. For many years, pink and blue colors were used to differentiate data values for women and men. But in modern-day western cultures, these colors come with gendered stereotype baggage: pink suggests weakness and blue suggests strength. Interestingly, this was not always the case—up until about the mid-twentieth century, it was the opposite. In her book, *The Secret Lives of Color*, Kassia St. Clair writes, "Pink is, after all, just faded red, which in the era of scarlet-jacketed soldiers and red-robed cardinals was the most masculine color, while blue was the signature hue of the Virgin Mary." Instead of the standard pink-blue pairing, consider using other color combinations such as purples and greens (as in the *Telegraph* newspaper) or blues and oranges (as in the *Guardian*).

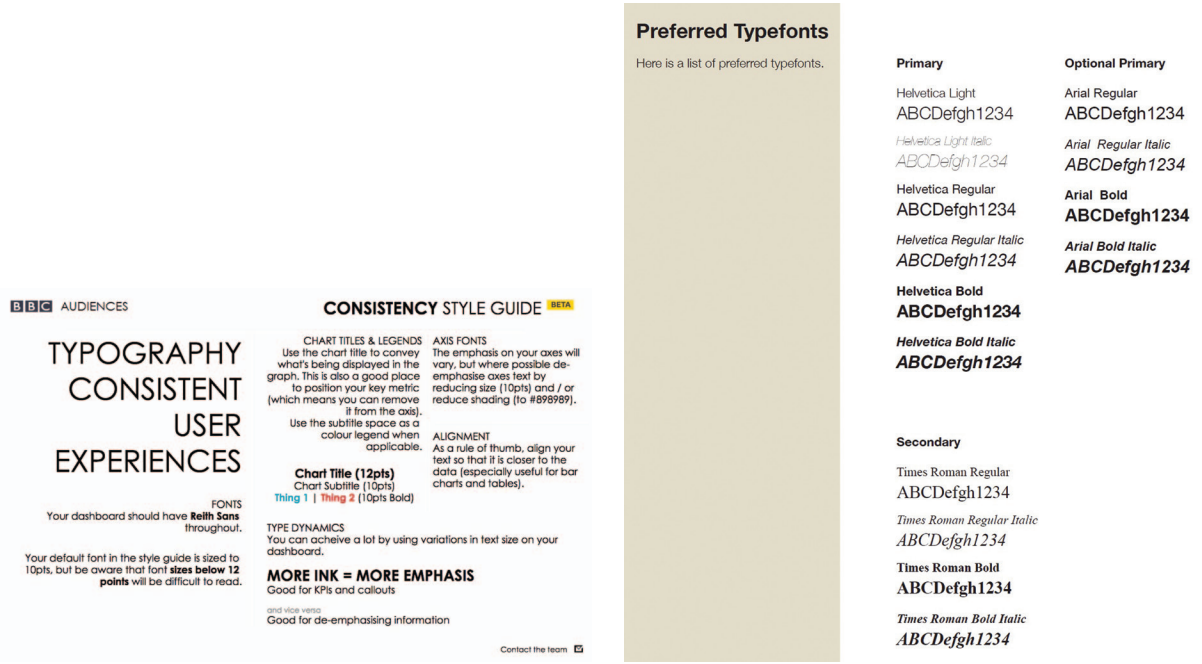
More generally, also consider how different cultures use and perceive different colors. In western cultures, for example, the color red may evoke emotions of passion and excitement and has both positive and negative associations. In eastern cultures, however, the color red represents happiness, joy, and celebration. In India, red relates to purity, and in Japan it is associated with life, anger, and danger.

DEFINING FONTS FOR THE STYLE GUIDE

A data visualization style guide should define typefaces (or fonts) for each part of the chart. You probably don't need more than two different fonts, and one will usually suffice. Also remember that you can vary the look of a single font by changing its thickness (thin, bold), angle (italics), and color.

BE WARY OF CUSTOM FONTS

A custom font can differentiate your style from the standard fonts found in, for instance, the Microsoft Office package. But beware: using custom fonts requires that they be installed on any machine that shows the graph. While custom fonts make your graphs stand out,



A data visualization style guide should also define appropriate fonts to use and when to use them. The BBC (left) and US Department of Agriculture (right) are just two examples of how to provide this guidance.

they can also raise trouble when sharing files or presenting from a laptop different from your own.

Default fonts like Century Gothic, Tahoma, Trebuchet MS, and Verdana are examples of effective fonts for data visualization that are available on most operating systems but are less commonly used and therefore appear more novel.

The BBC style guide for data visualization in Tableau (on the left) includes a typography section that demonstrates which fonts to use, where, and how to align them to the broader chart space. Their Reith Sans font is not a default font type, and they must therefore make sure that everyone in the organization has that font installed on their computers. The *Visuals Standards Guide* from the U.S. Department of Agriculture (on the right) features a broad set of fonts that are used across their publication types and at least two of the three (Arial and Times New Roman) are typically default fonts.

GUIDANCE FOR SPECIFIC GRAPH TYPES

Another section you might want to include in your style guide is a set of examples or instructions about specific chart types. Your organization may want to specify certain styling or data visualization best practices that differ chart to chart. You might also include examples of less common chart types to broaden your organization's data visualization toolbox, just like you did while reading this book.

Start by constructing guidelines for the most common chart types your organization uses. For example, you might specify that dual-axis line charts should never be used (see page 143) or that pie charts should have some upper limit of series (see page 289). There are also more granular specifications, such as where exactly labels should sit in a stacked bar chart or whether data points on a line chart sit on or between the tick marks. Or you might specify never to include tick marks on bar charts, or that whenever data labels are included, gridlines and tick marks must be omitted.

Another issue you might address in the specific chart area is how to manage the color palette among different data series. If the main colors in your palette are blue, red, and orange, the order of those colors may change if you have two or three series or may vary for, say, a paired bar chart versus a stacked bar chart.

TIPS FROM THE URBAN INSTITUTE STYLE GUIDE

- All of Urban's charts will be full-width (685px), so it is important to keep the data density as high as possible. Always include a text reference to your figure to give the data context to the content of the report/brief/blog post. If your chart has only two or three values, consider a couple sentences of text to explain the figure.
- If you find your explanatory sentences do a better job of distilling the information, you might want to consider going without a chart.
- Title: Keep it short and simple. Try to explain the chart in a few words. If you need to add qualifiers (e.g., years, dollars) or further clarification, use a subtitle
- Source and Notes: This is where the technical information about methodology can go. Try to avoid putting this information in the title, labels, or on the chart.
- Legends: Stretch legends across the top of the chart, or to the right. Order them in a logical way, mirroring the order of the data in the charts.

Source: The Urban Institute Style Guide, accessed January 2020.

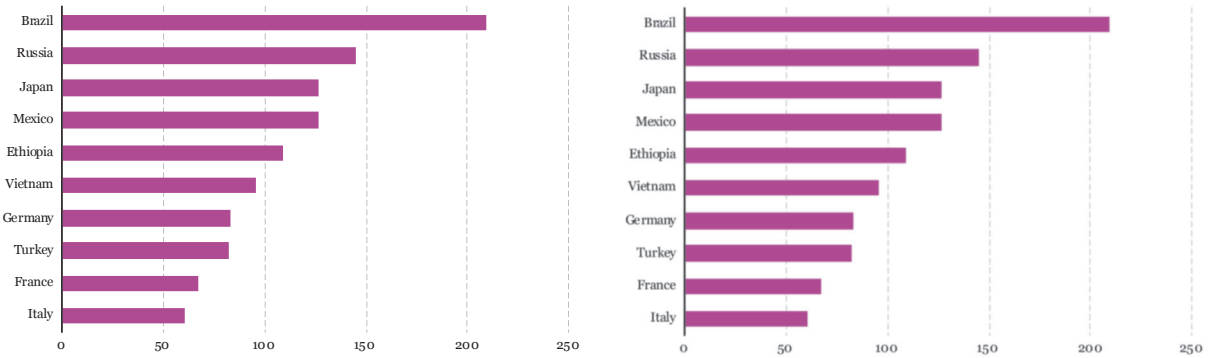
You might also include data visualization tips and tricks. These five tips are listed at the very top of the Urban Institute's guide. You can develop your own rules and tips or borrow those published by other groups and organizations.

EXPORTING IMAGES

Once a visualization is ready for external consumption, the chart creator must export it to a usable file format. This is another opportunity for things to go wrong: improper exporting might compress the resolution and pixelate the whole image. You can see the difference in resolution between these two versions of the same chart from Chapter 4. You can spend all the time you want creating a great, effective graph with clear colors and fonts, so don't waste all that effort with a blurry, hard-to-read final image.

Choosing the right file format for your visualizations is key. There are many file formats from which to choose, each with its own advantages and disadvantages. The biggest difference between image file formats is whether they are bitmap or vector. Images in bitmap format (also called raster) are stored as a series of squares (called pixels), each assigned a specific color. When you take a bitmap image and stretch it out, the pixels get larger and the resolution falls. You may have seen something like this if you put a photograph in a document and then tried to make it bigger—each pixel is now larger, and the crispness of the image deteriorates.

The other image format is vector. As opposed to bitmap images, vector images contain information about the actual shape in the image. A vector image is recreated when you



When moving images from your data visualization tool to your final product—either a report or web image—be sure the image has a sufficiently high resolution. The key is to test the image before you publish or post it.

Type	Acronym	Name	Application
vector	pdf	Portable Document Format	general purpose
	eps	Encapsulated PostScript	general purpose
	svg	Scalable Vector Graphics	online
bitmap	png	Portable Network Graphics	optimized for line drawings
	jpeg	Joint Photographic Experts Group	optimized for photographic images
	tiff	Tagged Image File Format	print production; better color reproduction
	gif	Graphics Interchange Format	typically used for animations

Source: Adapted from Claus O. Wilke, *Fundamentals of Data Visualization*

stretch it, so it won't lose resolution the way bitmap images do. Vector images are also called "resolution-independent" because they can be stretched forever without ever losing sharpness or detail. It might not surprise you to learn, then, that one of the biggest downsides of vector images is that the file size can be surprisingly large.

How you and your organization guide the export of graphs from software tool to the final product depends on a variety of factors, including the primary data visualization software tool, operating system, and where the final output will appear: Will it be in a PDF report? A standalone image on a website? Embedded in a tweet? The best strategy is to try a variety of approaches, but double-check the final product to make sure you have the sharpest, clearest image possible.

ACCESSIBILITY, DIVERSITY, AND INCLUSION

Many people with vision impairments rely on screen readers to navigate the internet. A screen reader reads the content on a screen aloud to the user, so if you post a graph with the filename "Image1.png," that is what the user will hear. People with other physical, cognitive, or intellectual disabilities may have difficulty reading your work or using your website if you have not taken into account their accessibility needs. Accessibility also extends to whether people can access the internet and the speed of their Internet connection. It is worth considering how your content (and website more generally) can be made more accessible by users who may require different levels of assistance.

To create accessible content, you might follow the guidelines laid out in Section 508 of the Rehabilitation Act of 1973. Section 508 requires U.S. federal government agencies to

develop, procure, maintain, and use information and communications technology (ICT) that is accessible to people with disabilities. This means that federal agencies that fall under Section 508 compliance rules must make their ICT—such as online training and websites—accessible for everyone.

One 508 standard for images that we can all apply is to use "alternative text" (commonly called "alt text") in our images. Alt text succinctly describes the content in the image. For data visualizations, this might be text communicating the general conclusion or message of the chart. In other words, what is the single concise sentence that summarizes your chart?

You can find some basic issues to consider by following the recommendations of the Web Content Accessibility Guidelines (WCAG), an international group charged with leading "the Web to its full potential." With respect to accessibility, WCAG defines four main areas:

1. **Perceivable.** Information must be presented in ways that users can perceive them. This might mean making non-text content available to other forms people need such as speech, symbols, or large print. Text should have sufficient color contrast with the background and images should have information ("alt tags") that will make them readable by screen readers and other assistive technology.
2. **Operable.** Make all functionality available from a keyboard. This means, for example, that using the tab, enter, and space bar keys enables the user to navigate the page and each interaction can be triggered.
3. **Understandable.** Text content should be readable and understandable, and web pages should operate in predictable ways. For example, significantly rearranging the content on the page can make content more difficult to read and understand.
4. **Robust.** Online content should be robust enough to be compatible with current and future users as well as assistive technologies. This might mean, for example, developing the website in such ways that screen readers and other assistive technologies can accurately interpret the content.

To date, there are no concrete rules about how to make a website completely accessible, though there are existing threads of research exploring how to use different assistive technologies to make visual content more accessible. Computer operating systems, browsers, and programming languages change and evolve, and any accessibility guide would be trying to hit a moving target. What we *can* do, however, is to consider how people with different abilities can or cannot access our content. A lot of these strategies are just good practices we

can use to more effectively communicate our work with text and explanations. Considering better accessibility then leads to better usability for everyone.

Another issue to keep in mind in data visualization is how you refer to different groups. You may have considered this when using terms like "Black," "African American," or "Hispanic" in your writing, tables, or graphs. Use the phrasing accepted and recognized by your audience and the communities you are studying. Consider the lived experiences of the people and groups you study and write about. Also consider using "people-first" language, such as "people with disabilities" instead of "disabled people." It is important to remember that data are a reflection of the lives of real people.

This also applies to the layout of your graphs and the language you use. How do you order the bars or lines in your tables and graphs? Is it alphabetical, based on sample size, or is it based on some unknown, arbitrary decision? Again, there are not many answers to these questions, but it is worth taking some time to consider approaches and strategies to make your work more accessible and inclusive of different groups.

PUTTING IT ALL TOGETHER

There are not necessarily right or wrong answers to some of these questions and style decisions. Whether the thickness of your gridlines is 1 pt or 2 pt, one shade of gray or another—these are primarily style decisions, but they are also functional decisions. As you saw in the first chapter, the goal is to emphasize the data over the gridlines, tick marks, and markers.

An effective, comprehensive data visualization style guide is best developed at the organizational level. If possible, bring your design and data teams together to determine branding guidelines that meet the needs of your organization, including data visualization. If your organization does not have these divisions, or if you are working to develop your own individual style guide, you might reach out to experts or refer to other published style guides to develop branding guidelines and styles.

Remember to treat your data visualization style guide as a living document. Revisit the guide as technologies and trends change. And remember to be flexible to the different needs, tools, and skills in your organization. Creating an instructive and clear guide that can be accessed and implemented by everyone can serve you, your organization, and your reader.

REDESIGNS

By this chapter, your data visualization toolbox contains much more than it did when you began this book. We've seen dozens of graphs, many of which may have been new to you. As you develop your own eye for data visualization, you'll find places where these new graph types may be especially useful.

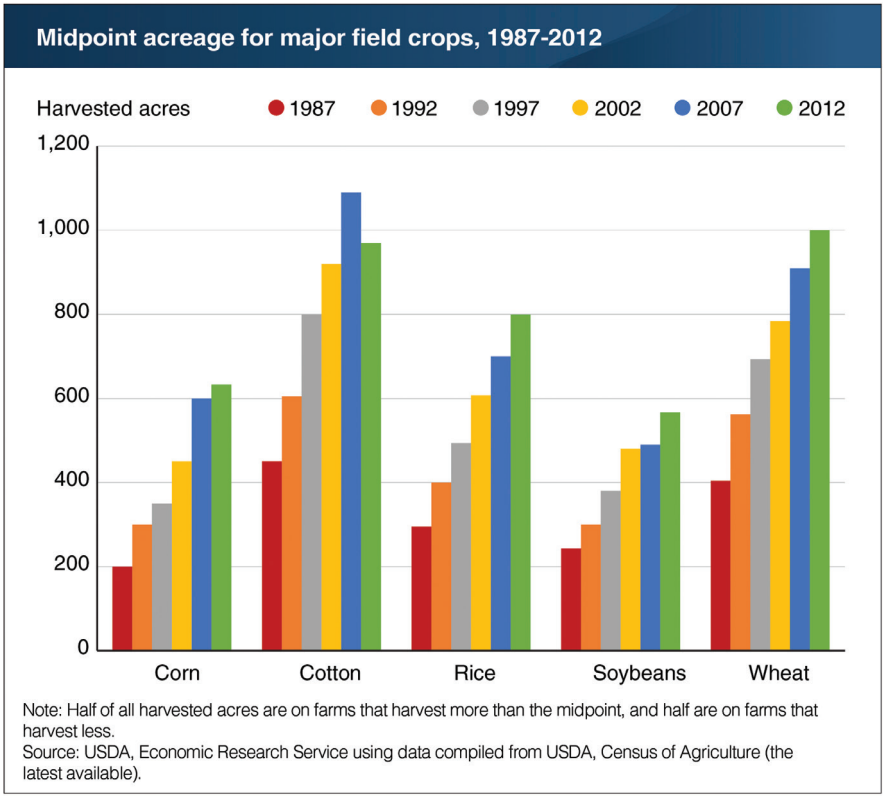
In this chapter, we'll cover a handful of data visualization redesigns. The graphs I choose to redesign here are not all especially *bad* graphs. Some are simply chosen because I believe there are more effective ways to plot the data. My goal is not to criticize these chart creators or their efforts but to demonstrate how the lessons we have learned can be applied to making data visualizations cleaner, clearer, and more effective.

The changes made here are by no means the only ways to modify these graphs, but each redesign follows the guidelines discussed throughout this book. In general, there is no "right" or "wrong" approach, just different ways of making improvements. As you develop an eye for better data visualization design, you will develop your own aesthetic and preferences.

PAIRED BAR CHART: ACREAGE FOR MAJOR FIELD CROPS

Take a moment and examine this bar chart from the U.S. Department of Agriculture that shows the number of harvested acres for five major crops in the United States for six different years. What do you see first?

My guess is you saw what I first saw: The acreage for all five crops increased over time. Your second observation, which quickly follows the first, is that cotton acreage (the second



Basic bar chart from the U.S. Department of Agriculture.

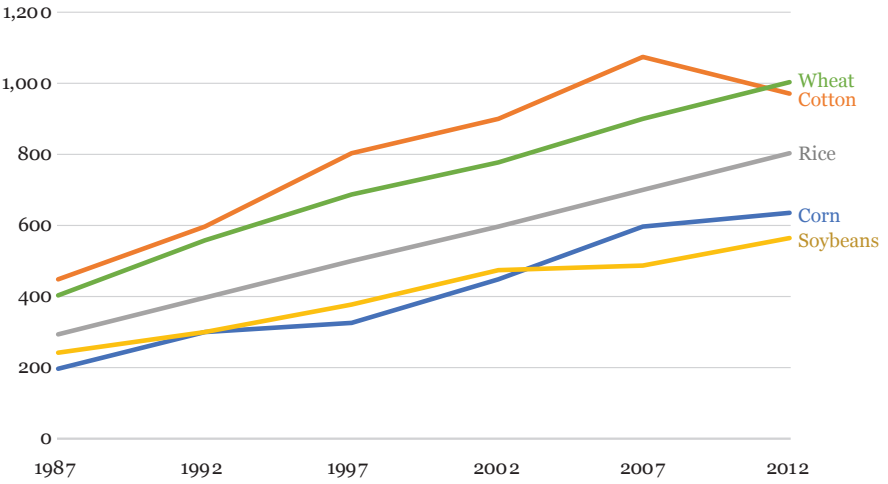
group) fell in the last year. Unlike the other groups, the last bar for cotton (the green bar) is shorter than the bar for the preceding year. But it doesn't jump out at you because there is so much ink and color in the graph.

If the goal with this chart is to show relative trends in acreage among five crops, a bar chart is a poor choice. The paired bar chart is good at showing exact values, but the relative trends are not clear or immediately evident.

We could redesign this as a simple line chart.

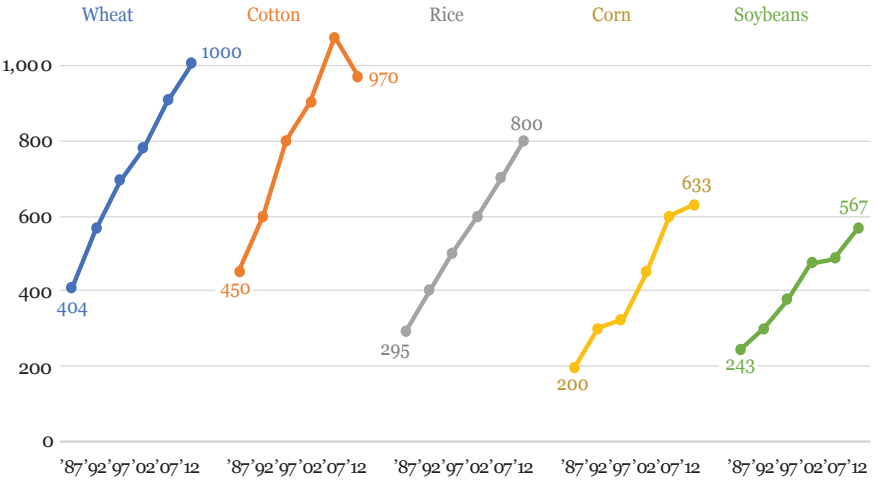
Here, the drop in acreage for cotton is very clear, as are the relative sizes of the five crops. In the bar chart, I couldn't see immediately that rice acreage sits right in the middle of the five crops, but here I can see that right away. I didn't use a legend here, as might be the default approach, but instead added the labels at the end of each line, using color to link them with the lines.

Midpoint acreage for major field crops, 1987-2012
(Midpoint acreages more than doubled for all five major field crops)



Source: U.S. Department of Agriculture

Midpoint acreage for major field crops, 1987-2012
(Midpoint acreages more than doubled for all five major field crops)



Source: U.S. Department of Agriculture

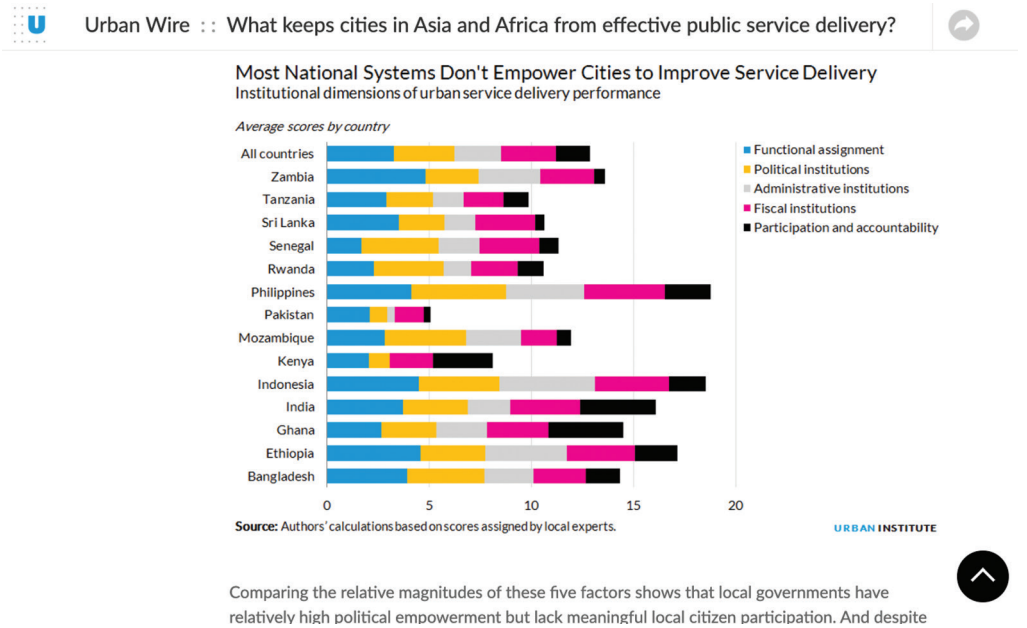
Two ways to redesign the USDA bar chart: A line chart or a cycle chart.

Another approach is a cycle chart. Instead of putting the lines together, this cycle chart is essentially a small multiples line chart where each crop gets its own panel. The advantage is that there's more space for the graph and it's perhaps a little more engaging because it's different. The disadvantage is that relative patterns are slightly less clear than in the line chart.

STACKED BAR CHART: SERVICE DELIVERY

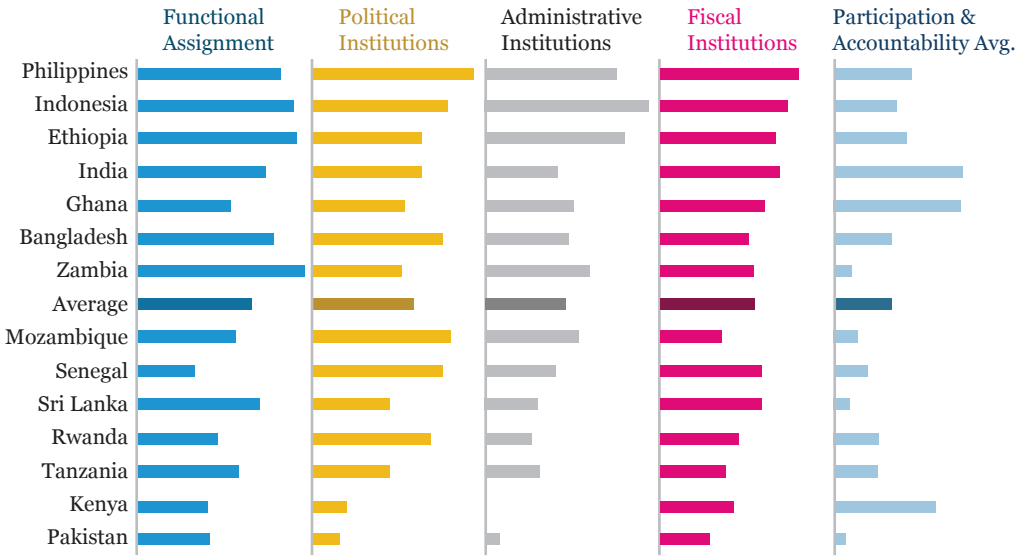
Let's go back to page 14 in Chapter 1 and consider the perceptual rankings diagram. At the very top are graphs positioned along common scales—the bar chart or line chart with a single horizontal axis, for example. One step below are those graphs that are not positioned along common scales graphs. It is slightly harder to accurately assess the values in these.

This graph contains data from both sections of the ranking diagram. We can clearly discern the differences between the values of the blue series (Functional assignment) because they all sit on the same vertical baseline. We are not as well equipped, however, to similarly



We can barely see that the value for Political Institutions is larger for *All countries* than *Zambia*.
Source: Roth and Malik, 2016

Most national systems don't empower cities to improve service delivery
(Institutional dimensions of urban service delivery performance, Average scores by country)



Source: Roth and Malik, 2016

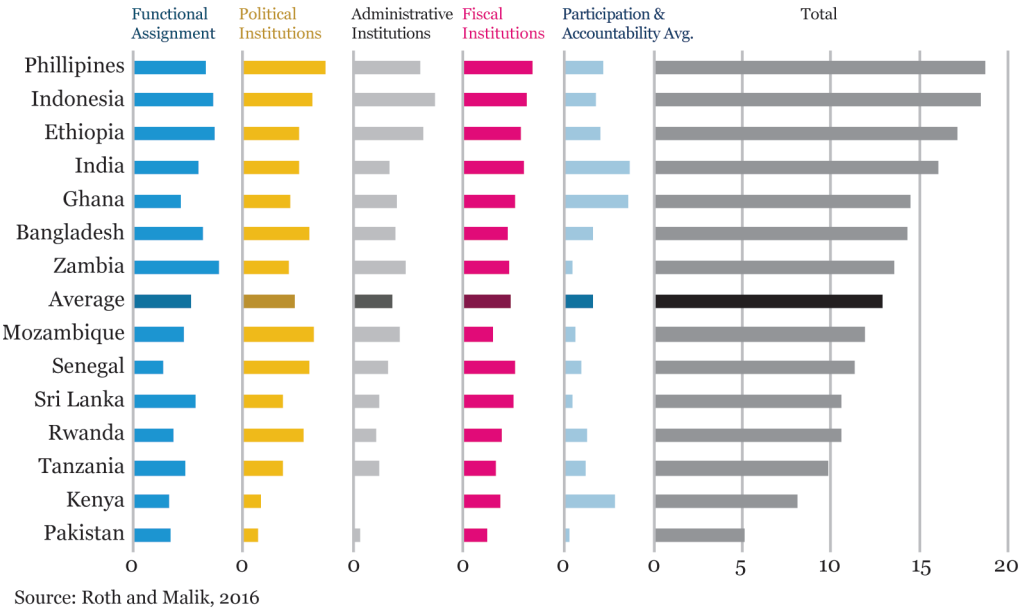
One way to redesign the stacked bar chart is to break them up and use a small multiples approach.

assess the values for the other series, because they don't share the same baseline. You can test this yourself: Is the value for Political Institutions (the yellow series) larger for "All Countries" or Zambia (the first two series)?

Instead of packing all of the data onto a single chart, we can break it into five separate charts. In this case, each series is given its own vertical baseline, so it's easier to make comparisons across countries within each series. The important point with making a graph like this is that the horizontal space for each series is the same. If we shrank the space for "Fiscal Institutions," for example, it might look like those values are larger than others.

This approach, however, doesn't tell you much about the *overall* values between countries. There's nothing wrong with adding a series for the overall *Total*, again, as long as we use the same horizontal spacing. In other words, the space between each gridline is the same. This approach works even better in cases where the values sum to the same total or to 100 percent, because the total length will be the same for all of the bars and thus a *Total* segment is unnecessary.

Most national systems don't empower cities to improve service delivery
(Institutional dimensions of urban service delivery performance, Average scores by country)



Source: Roth and Malik, 2016

When breaking up stacked bar charts, it is sometimes important to include the totals.

LINE CHART: THE SOCIAL SECURITY TRUSTEES

Each year, the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds report on the current and projected status of the U.S. Social Security program. The Trustees are responsible for estimating the current and future financial picture of the program to communicate to the public and policymakers the challenges the program faces. The Social Security Technical Panel is an independent expert panel responsible for reviewing the work of the Trustees, including the methodological details, economic and demographic assumptions, and the Trustees communication efforts.

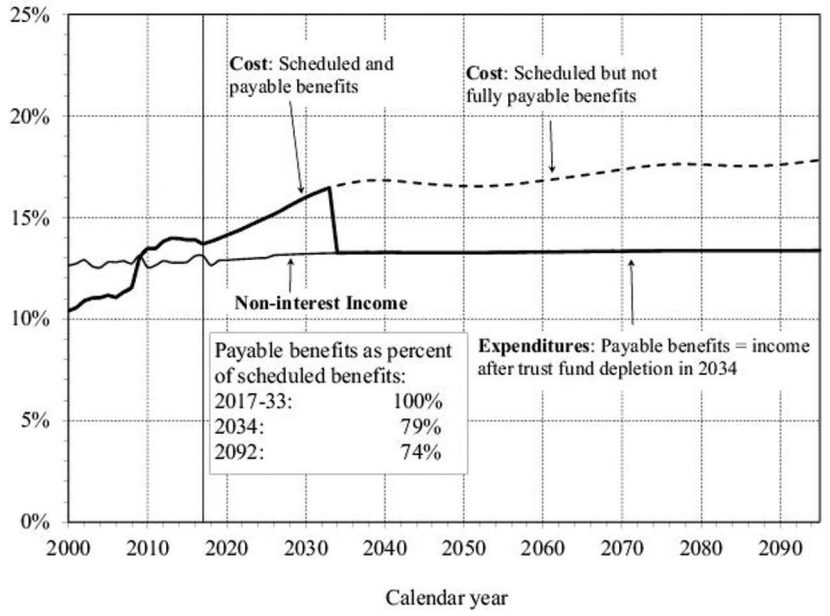
The 2019 Technical Panel placed an emphasis on this latter category: "The Panel believes that trust in public institutions is enhanced by greater understanding . . . In this context, we believe it is paramount for the Trustees to communicate clearly and effectively with the general public about its finances." The panel emphasized clear, plain language, a focus on the core message, and better data visualizations in the Trustee's work.

Let's look, then, at two of their data visualizations.

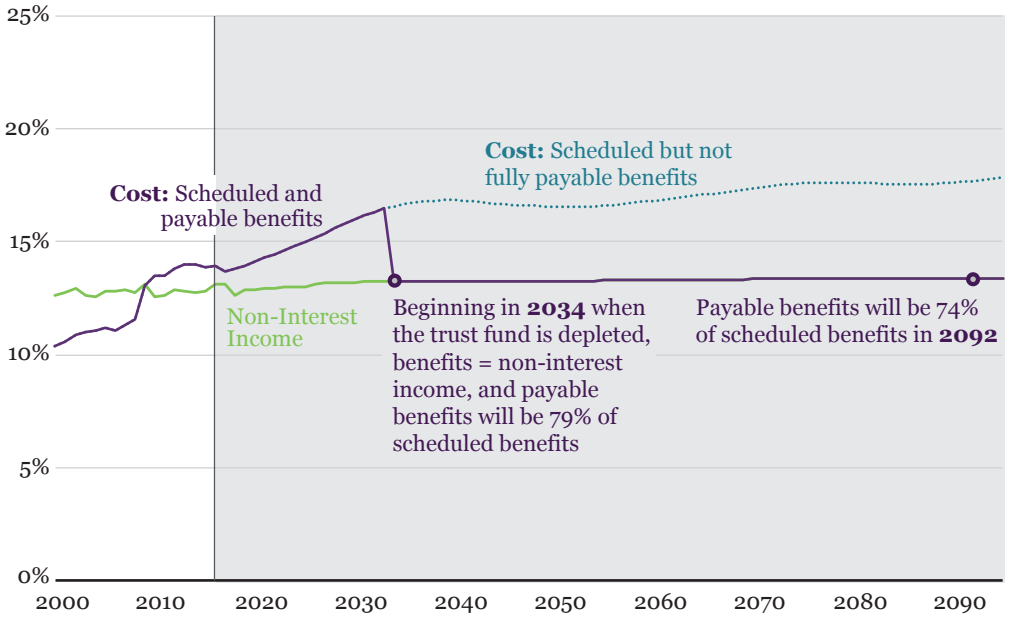
A CLEANUP

The first example is a relatively simple clean-up rather than a wholesale redesign. This line chart—which has appeared in virtually every Trustees Report—shows the time series of the basic finances of the Social Security system. System income (taxes paid into the system) are set next to system costs (benefits paid to beneficiaries) for a short historical period (here from 2000 to 2018) and out in the longer projection period (here from 2018 through 2092). Two sets of costs are shown: one that shows how many benefits are *scheduled* to be paid (the dashed line) and one that shows how many benefits can *actually* be paid (the solid bold line).

The existing graph has annotation and labels to help the reader better understand the content and the concepts. A small table near the bottom of the graph lists benefit shares in



The Social Security Administration (2019) has published this graph showing the basic finances of the Social Security system for many years.



Source: Social Security Administration, 2019

Some basic cleanup and annotation improves the clarity of the Social Security finances chart.

specific years. But there is also a lot of ink used on extraneous details: horizontal and vertical gridlines and tick marks for every percentage point and year.

Let's take a simple approach to redesigning this graph by removing some of these extraneous details and markers. Here, I've removed the vertical gridlines and all of the tick marks. I deleted the small table and instead directly labeled the years those numbers referenced. I used some slight color here—which is consistent with black-and-white printing—and added a gray box to the projection period (after 2018) to draw attention to the imbalance.

A BETTER DOT PLOT

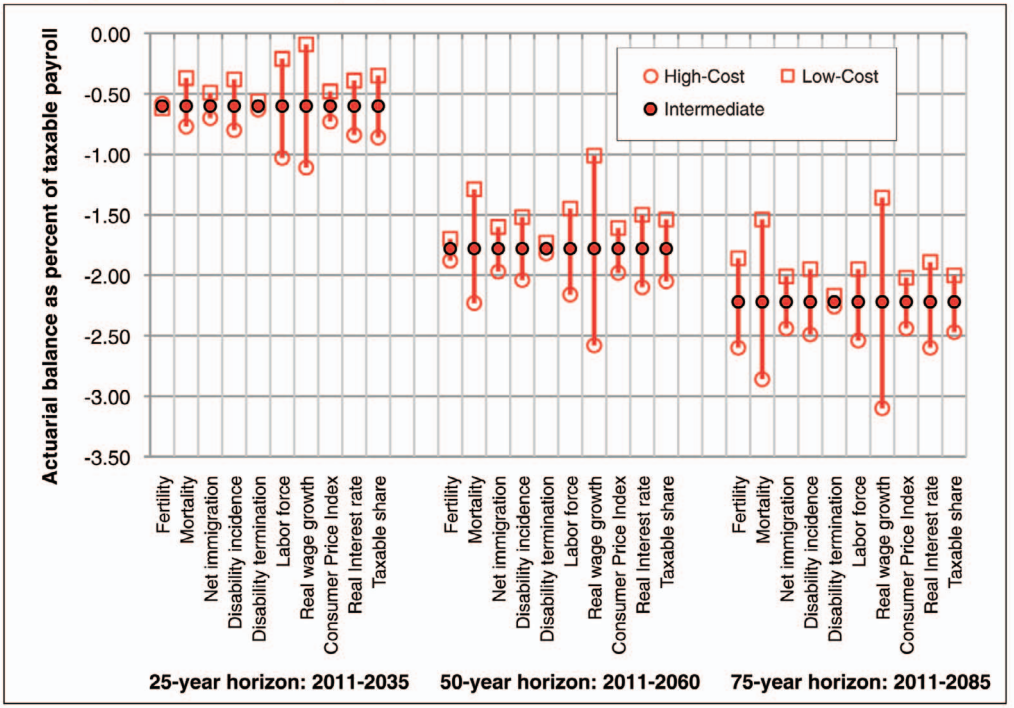
The next graph appeared in the 2011 Technical Panel Report. It shows the sensitivity of different assumptions of the Social Security model. Most of the six Technical Panel Reports published since 1999 include this information as a series of tables. But in 2011, the Panel

presented these estimates in what amounts to a dot plot or a simplified box-and-whisker plot. Here, instead of two dots with a connecting line, the chart has an "intermediate" estimate in the middle and a "low-cost" and "high-cost" options sit on either side.

Notice the different shapes, vertical and horizontal gridlines, and rotated axis labels. Using our basic guidelines from Chapter 2—showing the data, reducing the clutter, integrating the text and the graph, use more graphs, and start with gray—we can improve this visualization to make it clearer and easier to read.

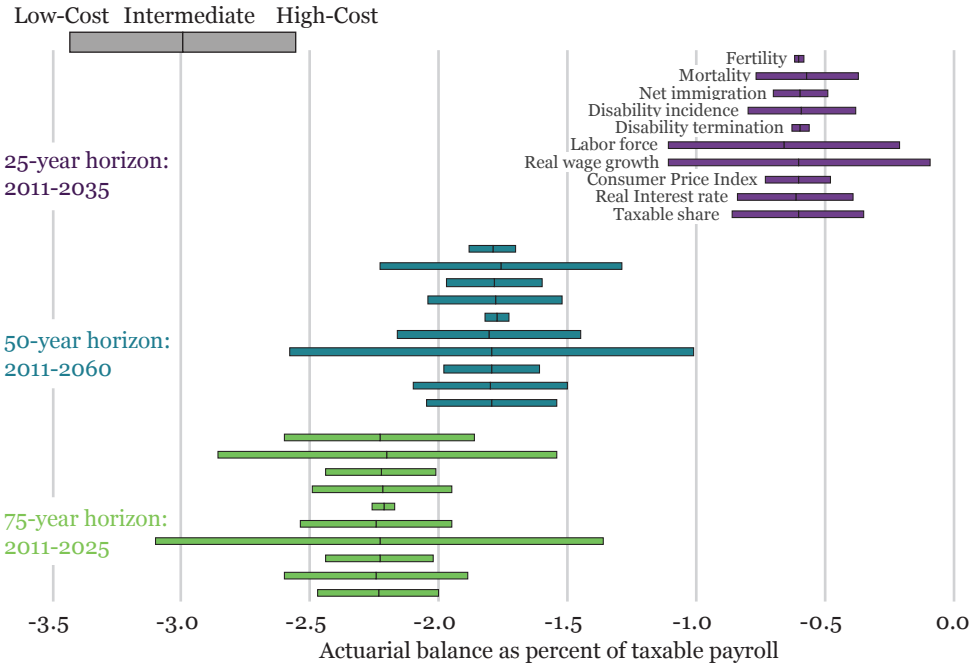
A simple start is to rotate the entire visualization. Now we don't need to turn our heads to read the labels, and we can place what was formerly along the vertical axis along the horizontal axis. Instead of dots—which works just fine—I converted them to boxes, which

Figure 4. Sensitivity of Summarized Actuarial Balance to Range of Assumptions: 25-, 50-, and 75-Year Horizons (as a Percent of Taxable Payroll)[4]



Source: 2011 Trustees Report, Appendix D; additional estimates provided by Office of the Chief Actuary, Social Security Administration.

Gridlines, rotated text, and general clutter make this graph hard to read.
Source: 2011 Technical Panel Report on Assumptions and Methods



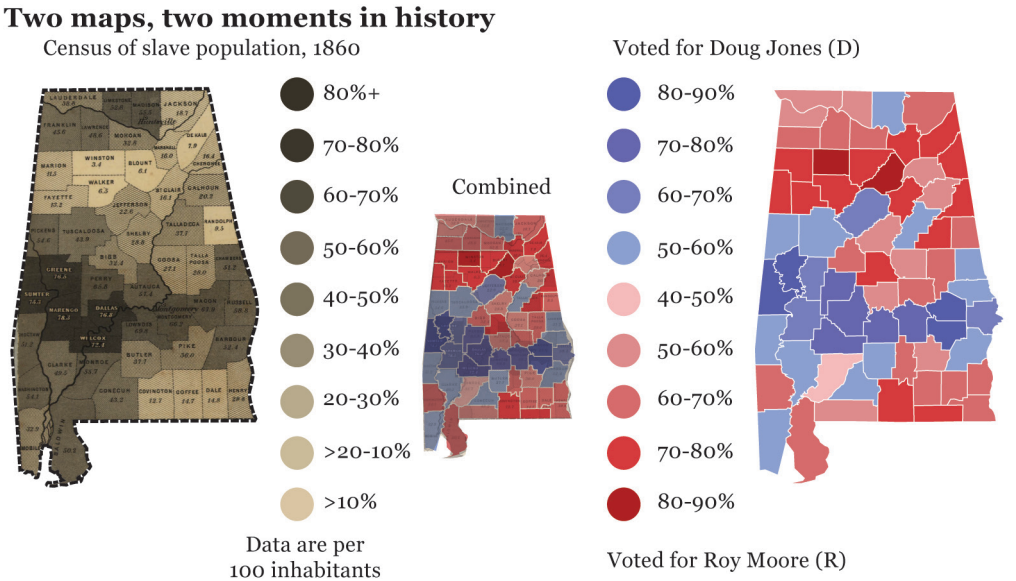
Source: Social Security Administration, 2011

Rearranging the plot and removing some of the clutter makes the graph easier to read.

reduces the extra lines and dots in the graph. By using a box, one end can encode the low-cost value; one end, the high-cost value; and a middle marker, the intermediate value. We can also place the labels for each metric right next to the first set of boxes though we could also repeat them if we thought it was necessary.

CHOROPLETH MAP: ALABAMA SLAVERY AND SENATE ELECTIONS

In late 2017, there was a runoff election for a U.S. Senate seat in Alabama. In a tense and competitive election, journalist Sarah Slobin (then at Quartz) wrote a story about the relationship between voting behavior and the distribution of enslaved people in 1860. Slobin wrote that, "[W]hile correlation is not causation, there is a startling visual parallel when you zoom in to Alabama . . . and compare it to how Alabama just voted this week."



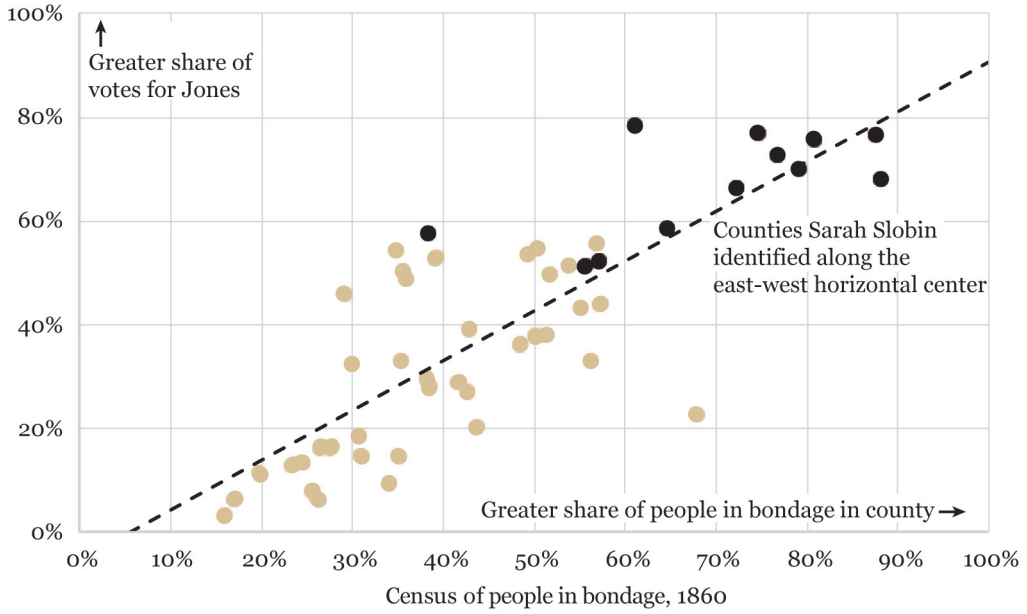
Using an 1860 map from the US Bureau of the Census and voting results in the 2017 Alabama Senate election, we can see similar bands of darker colors running through the middle of the state.

Source: Author's rendering based on original chart from Quartz. Map from the US Bureau of the Census; voting data provided courtesy of Sarah Slobin

Using voting data from that election, there is a clear horizontal band of dark blue (representing more Democratic votes) in the map on the right. The map on the left, from an 1860 map published by the U.S. Bureau of the Census, shows a streak of darker colors representing counties with a larger proportion of people held in bondage. Slobin writes: "If you focus on the 'black belt' moving horizontally across both maps, you can see that in areas with a history of slavery, the vote went to [the Democratic candidate] Jones."

Though visually striking, this pairing forces the reader to jump between the two maps to see the similarity in the bands. Can we take a different approach?

The share of the vote going to the Democrat in Alabama's sixty-seven counties ranged from 16.1 percent to 88.1 percent. These roughly (though not perfectly) overlap with fifty-one counties I could identify in the Census Bureau map, which range from 3.1 percent to 78.3 percent. (There are some data issues we will ignore here as counties have changed, merged, or broken up between 1860 and 2017).



Source: Voting data provided courtesy of Sarah Slobin

An alternative—or addition—to two maps is to use a scatterplot.

Plotting the two variables in a scatterplot lets us more easily see the positive relationship with the (darker) circles in the top-right part of the graph marking those twelve counties along the east-west corridor.

Slobin's original maps are visually striking. For a news story, they may well be the best way to show the data. It's easy to see the basic band pattern running through each map. By comparison, the scatterplot may take some additional explanation for a casual reader who may be less familiar with this graph type. We could publish *both* graphs to pair the visually-striking maps—to draw readers in—with the more technical scatterplot for those who want to see the detailed comparison. If I were publishing this in a peer-reviewed academic journal, I would lean toward the scatterplot because it clearly shows the association between the two series.

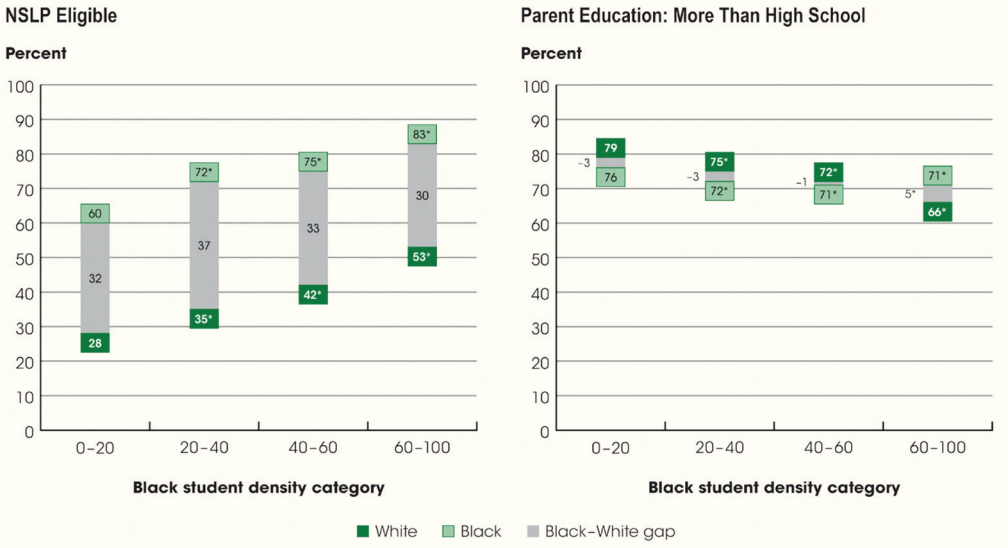
DOT PLOT: THE NATIONAL SCHOOL LUNCH PROGRAM

In reviewing a report from the National Assessment of Educational Progress (NAEP) on the achievement gaps between Black and white students, I came across this chart, which shows

differences in school achievement scores for Black and white students, arranged by scores of Black students.

Let's focus on the bars on the far-left side of the graph. There are three numbers here: 28 percent, 32 percent, and 60 percent. The numbers in the green boxes show test scores for white (60 percent) and Black (28 percent) students, and the middle number shows the gap between the two groups (32 percent). But the green boxes make it appear as if the 28 percent represents a range of numbers, from, say, 22 percent to 28 percent. By using rectangles instead of points or markers, it resembles a stacked chart rather than the dot plot, which was likely the intention.

Figure 8. Percentage of Black and White students who were National School Lunch Program (NSLP) eligible and percentage who had a parent with more than a high school education, by Black student density category: 2011

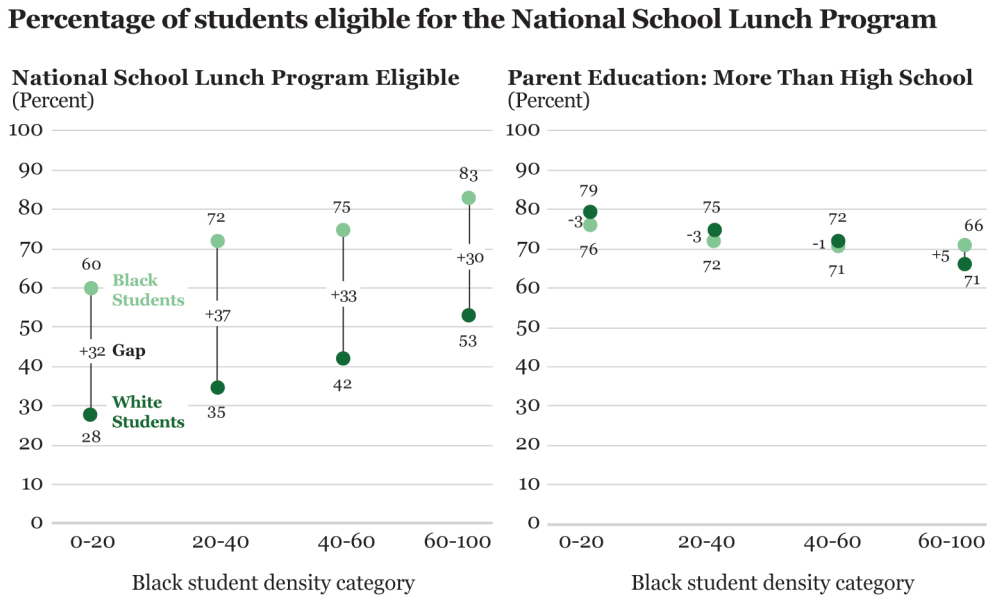


* Significantly different ($p < .05$) from the 0 percent to 20 percent density category.

NOTE: The measures displayed in this figure are percentages of students within each Black student density category.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics Grade 8 Assessment.

This graph from the National Center for Education Statistics shows the percentage of students eligible for the National School Lunch Program.



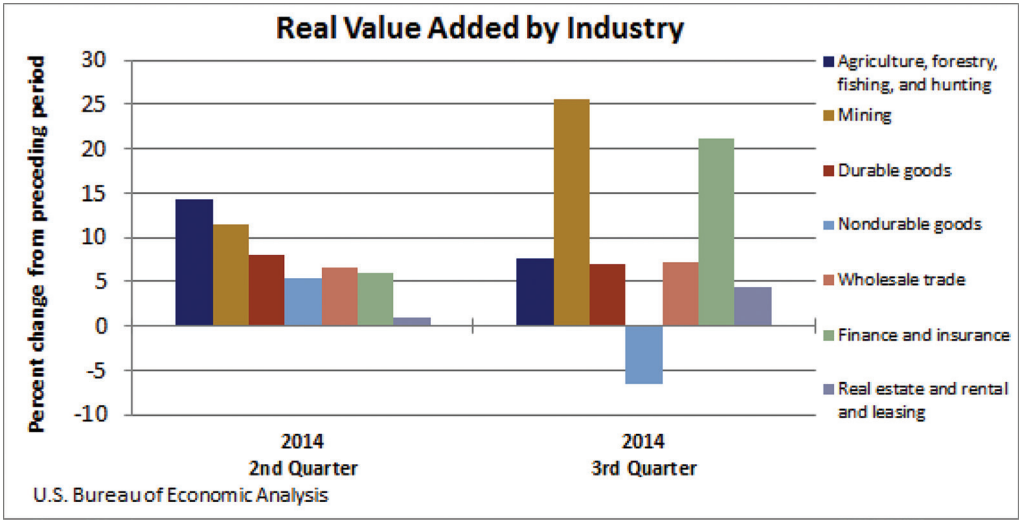
Changing shapes and removing some of the clutter makes the graph from the National Center for Education Statistics easier to read.

As an alternative, let's make it a true (vertical) dot plot. We can replace the green boxes with green circles and connect them with a gray, vertical line. Now we perceive the green circles as specific points rather than ranges or a stacked set of values.

You might also notice that I deleted the legend and labeled the three series directly on the left chart. I didn't repeat the labeling in the chart on the right for two reasons: First, because the gaps are smaller, there is less space for the labels. And second, the reader doesn't need to be reminded of the definition of each dot and line at every single occurrence on the page.

DOT PLOT: GDP GROWTH IN THE UNITED STATES

Every quarter, the U.S. Bureau of Economic Analysis (BEA)—the federal agency responsible for producing some of the most important measures of the U.S. economy—releases their



This bar chart from the Bureau of Economic Analysis's quarterly report does not match what's written in the text.

report about changes in gross domestic product (GDP). And each quarter, they publish a press release on changes over time and in specific industries.

The graph above is from the third quarter of 2014 press release. It shows the "Real Value Added"—a measure of each industry's contribution to GDP—for major industries in the country. Given what you've learned so far, there are a variety of things you might change to make this graph more effective. You might directly label the bars, rotate the vertical axis legend and place it near the title, and lighten some of the gridlines.

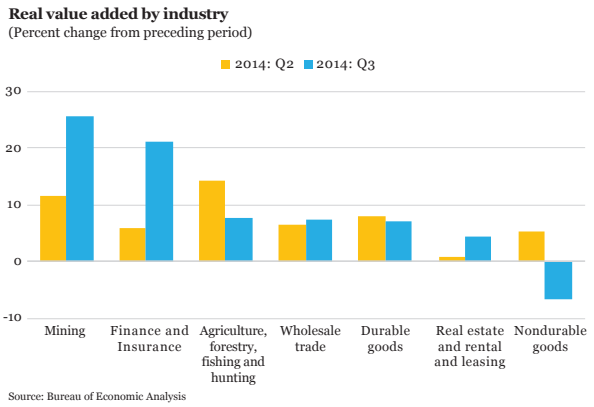
More importantly, let's take a look at what this graph is *supposed* to show. Here are the six bullet points that surround the Real Value Added (RVA) by Industry graph in the BEA document:

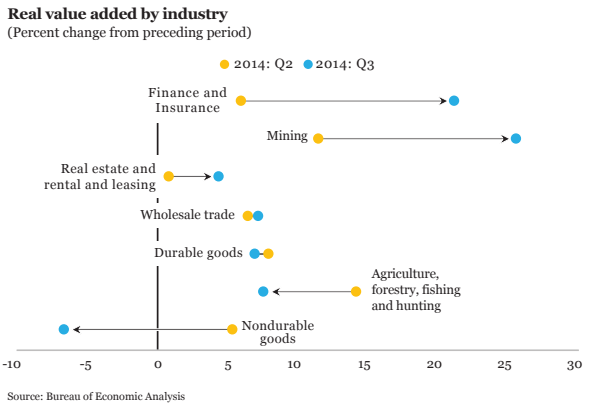
- Finance and insurance real value added—a measure of an industry's contribution to GDP—increased 21.2 percent in the third quarter, after increasing 6.0 percent in the second quarter.

- Mining increased 25.6 percent, after increasing 11.5 percent. This was the largest increase since the fourth quarter of 2008.
- Real estate and rental and leasing increased 4.4 percent, after increasing 0.9 percent.
- Real value added for manufacturing increased 0.5 percent, after increasing 6.8 percent. Durable-goods increased 7.0 percent following an increase of 8.0 percent, while nondurable-goods decreased 6.6 percent, after increasing 5.4 percent.
- Agriculture, forestry, fishing, and hunting increased 7.6 percent after increasing 14.2 percent.
- Wholesale trade continued to show strong growth, increasing 7.3 percent, after increasing 6.5 percent.

What do you notice about how these points are arranged? Each one details how RVA changed between the first and last period for each industry. The structure of the graph, however, is formatted to compare *across* industries *within* each period.

A better approach for the graph would match the text and show the inverse: the change within each industry across periods. A paired bar chart and dot plot are two effective ways to do this. In the paired bar chart, I've sorted the data according to the most recent period (2014:Q3) to subtly guide the reader to the best-performing industries. In the dot plot, the data are sorted based on the *change* between the two periods—the largest positive changes are at the top of the graph and the largest declines at the bottom.





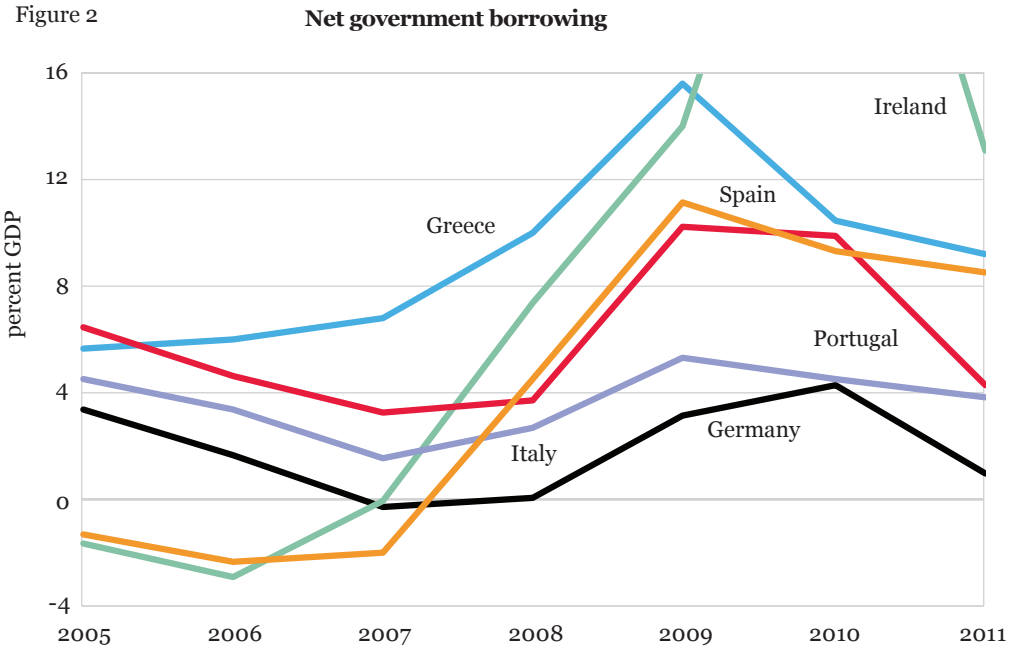
Two options to redesign the original BEA graph to match the organization of the press release.

In either case, these two graphs better illustrate the takeaways from the text. They are sorted by industry so that now, for example, you can more easily see that, "Mining increased 25.6 percent, after increasing 11.5 percent." Your data visualizations should not be intended to break up long sections of text or to provide a "visual break." They are there to support your argument. Integrate them with your writing for a seamless reading experience.

LINE CHART: NET GOVERNMENT BORROWING

As I mentioned in Chapter 5, I'm a big fan of line charts. They clearly show changes over time, and everyone knows how to read them. Consider, however, this line chart in a 2012 Economic Policy Paper from the Federal Reserve Bank of Minneapolis.

Figure 2



This line chart, originally published by the Federal Reserve Bank of Minneapolis, simply cuts off the data for Ireland.

Source: Author's rendering based on original chart from Arellano, Conesa, and Kehoe (2012).

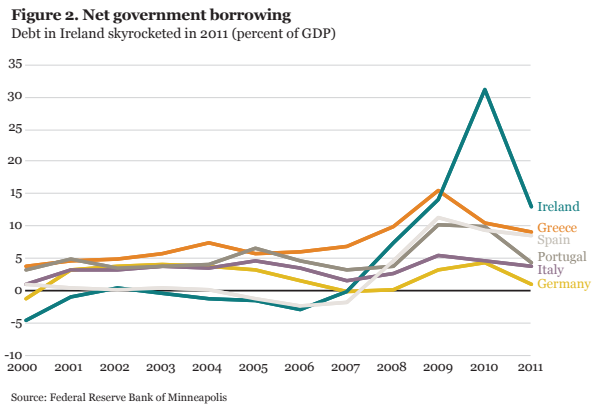
Notice anything strange about this graph? Anything aside from the title split into three parts ("Figure 2" in the top-left; "Net government borrowing" centered over the graph; and "percent GDP" along the vertical axis)? Anything besides the equal-weighted gridlines even though zero is not at the bottom? Or the mix of pastel and bright colors?

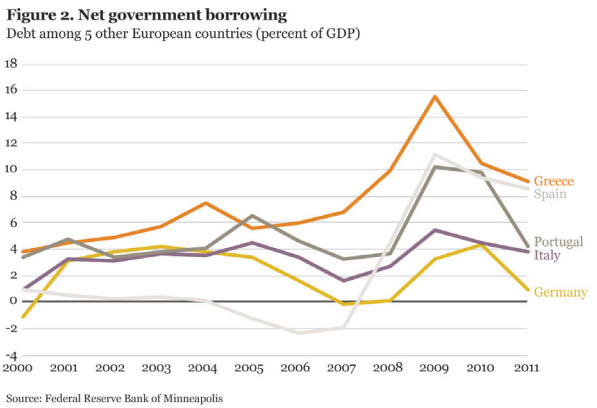
How about how the line for Ireland shoots off the top of the graph? You must have a very good reason not to show all of the data on the graph, and putting the value in the footnote does not count!

The chart creator here faced a problem: Ireland had a debt spike in 2011, far more than the other countries. To show all the data in one chart would scrunch up and lose the detail among the other countries.

But there is a better solution: Use two graphs. One that includes Ireland with a vertical axis that ranges from −10 percent to 35 percent, and another with a vertical axis from −4 percent to 18 percent that shows the detail among the other countries. I could leave these as equal-sized charts, or even make the second one smaller in a sort of zoom-out/zoom-in comparison. In either case, I use the subtitle to explain that one graph includes Ireland and the other does not.

Do not be afraid to use more than one graph, if that's what it takes to clearly communicate your argument. We are now a digital-first society—using more space only requires more computer memory, not more paper.





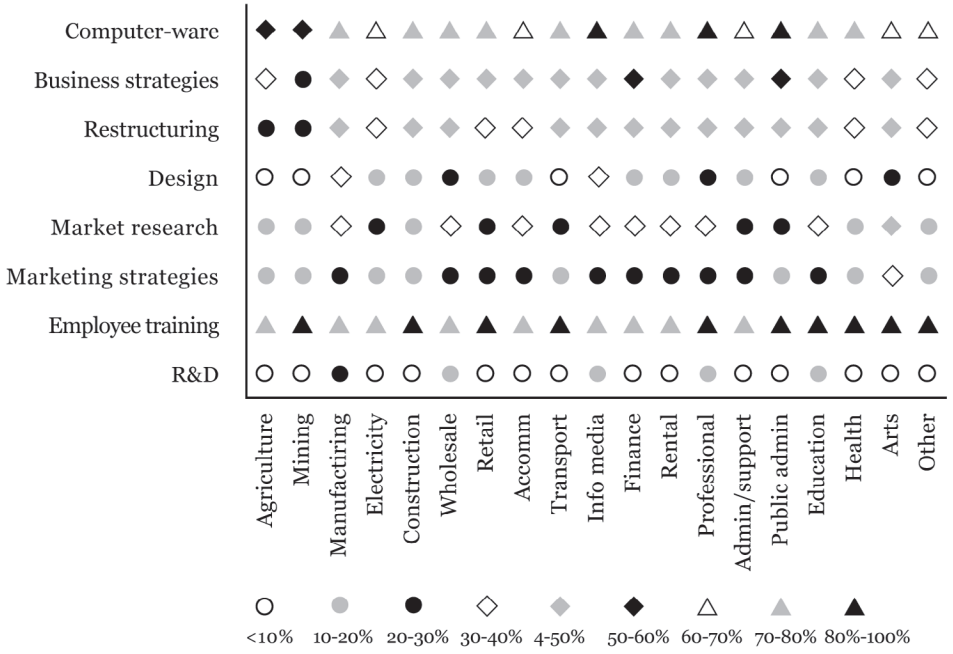
Instead of packing all of the information on a single chart, try breaking them up and create a "zoom in" view.

TABLE: FIRM ENGAGEMENT

As we saw in Chapter 11, there are many ways to make our tables more visual. We can add color, icons, bars, or other elements to highlight the important values for our reader instead of asking them to sift through all the data values.

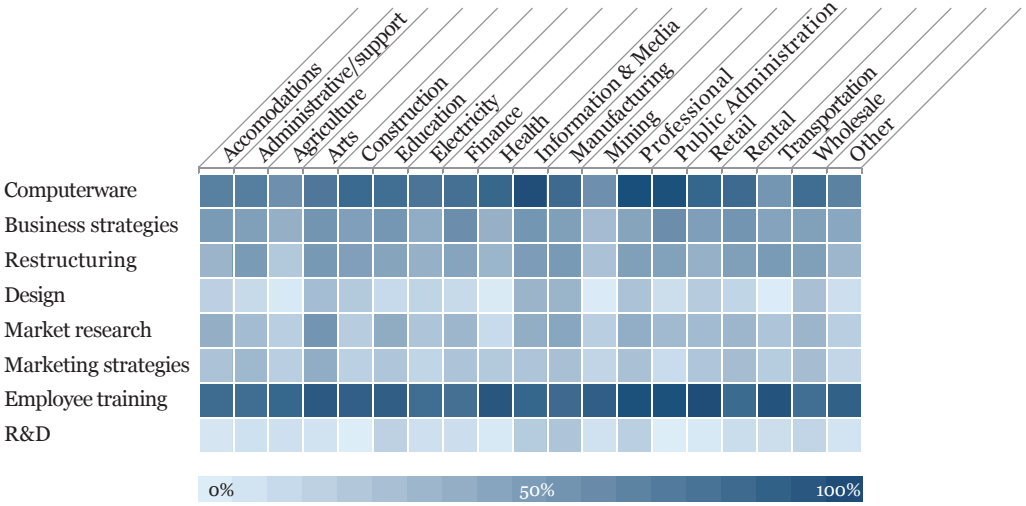
This table uses different shapes and shades of gray to show the share of firms that engage in different business activities like design and market research. As the reader, we must understand which shapes correspond to which percentages and then figure out the different

Figure 1: Proportion of firm-years engaging in each intangible acitivty, by industry



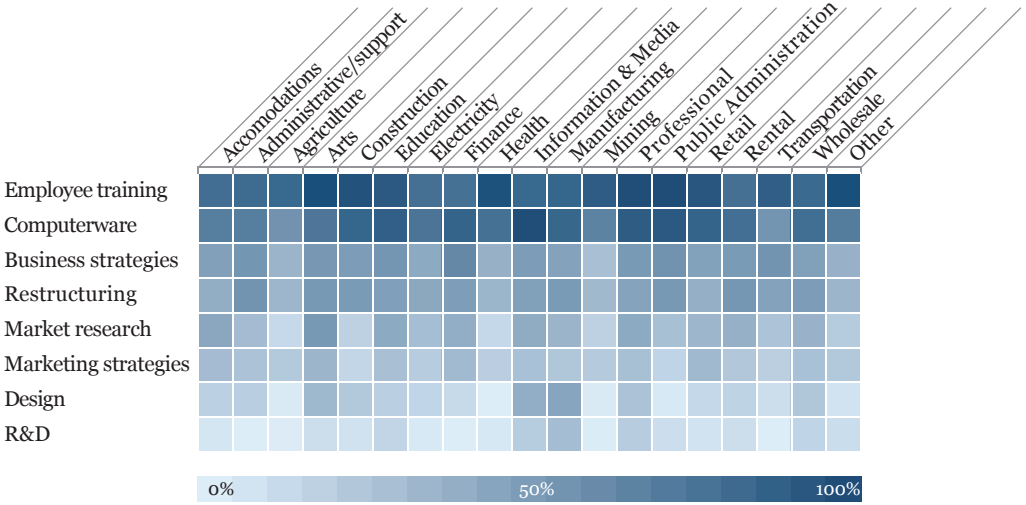
Source: Author's rendering of original chart by Chappell and Jaffe, 2018
Note: Data based on a visual inspection of the original graphic.

Author's rendering of an original chart by Chappell and Jaffe (2018), which could be improved by changing how the values are displayed.



Source: Chappell and Jaffe, 2018
Note: Data based on a visual inspection of the original graphic.

A heatmap is one alternative to the Chappel and Jaffe (2018) chart.



Source: Chappell and Jaffe, 2018
Note: Data based on a visual inspection of the original graphic.

This heatmap alternative to the Chappel and Jaffe (2018) chart sorts the data.

shading styles. Of course, triangles don't necessarily mean "more" of something than circles, so the rank-ordering of the values is hard to interpret.

Instead, what if we use a monochromatic color ramp moving from a light blue for the lower percentages to darker blues for the higher values? In this heatmap approach, it's much easier to see that there is a lot of time spent on *Employee training*, the dark blue row towards the bottom of the table.

We can take this a small step forward and sort the data, which will naturally focus the reader's visual attention. The longer labels either require us to use rotated labels, as I've done here, or perhaps rotate the entire graph and change the spacing or cell size so all of the text can fit.

CONCLUSION

With more graphs in your data visualization toolbox to choose from, and having seen more graphs and best practices, I'm confident that you're ready to improve upon your own graphs. Finding and redesigning even the simplest graph—I find that mining the academic peer-reviewed literature a good place to start—can help you refine your skills and develop your own data visualization aesthetic. Like any other skill, practice makes better.

Two important caveats. First, if you critique a graph publicly, keep in mind that someone made that graph and that even your well-intentioned efforts to redesign it may not be appreciated. The chart creator may have had time pressures, software limitations, or organizational demands of which you are not aware. Reaching out to the person who created the original graph may be worth your effort. Second, try to identify the central goal of the chart and the possible challenges of the data series. This will help lead you to the best chart type for the task at hand.

CONCLUSION

Now, at the end of this book, your data visualization toolbox has expanded considerably. Instead of using whatever the default graphs are available in your favorite software tool, you can now draw upon more examples to visualize your data in the ways that best serve your reader, user, and audience.

The graphs presented in these chapters have been tested over and over again. Analysts, researchers, reporters, and scholars have used them with different data sets and layers of text, annotation, color, font, purpose, and platform. But the set of graphs available to you is infinite. The bar chart did not exist before someone invented it. Maybe it's you who will invent the next great graph type.

Sources of inspiration for new and innovative visualizations are all around us, from public and private organizations, the media, data scientists, designers, and artists. The images on the next few pages are from some popular data visualization projects and stories. They use different shapes, layouts, approaches, and techniques to visualize data in ways that are unique to the field and form.

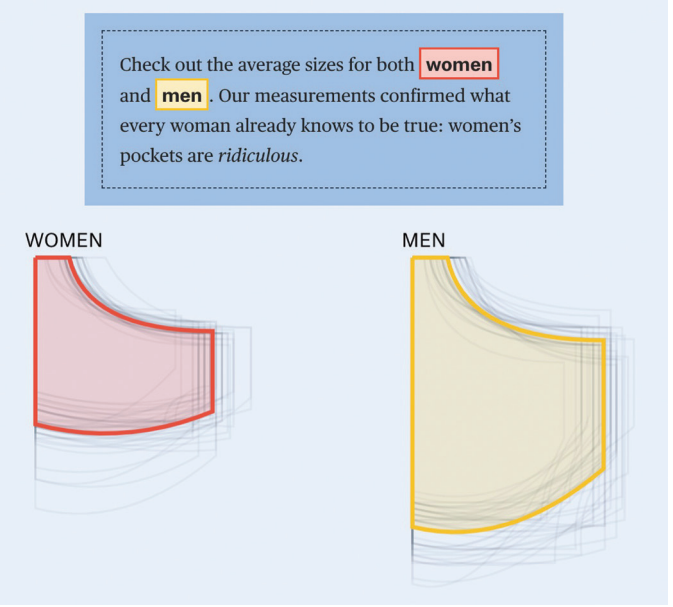
Freelance data designer Maarten Lambrechts maintains a special project website, Xenographics, with the tagline, "Weird but (sometimes) useful charts." It is a repository of "novel, innovative and experimental visualizations to inspire you, to fight xenographphobia and popularize new chart types." There, you can find the different, unique, and strange graphs that can be used to show data in different ways. By experimenting with such forms, we can move the field forward and communicate data and information in new and better ways.

Source: Nadieh Bremer, Visual Cinnamon

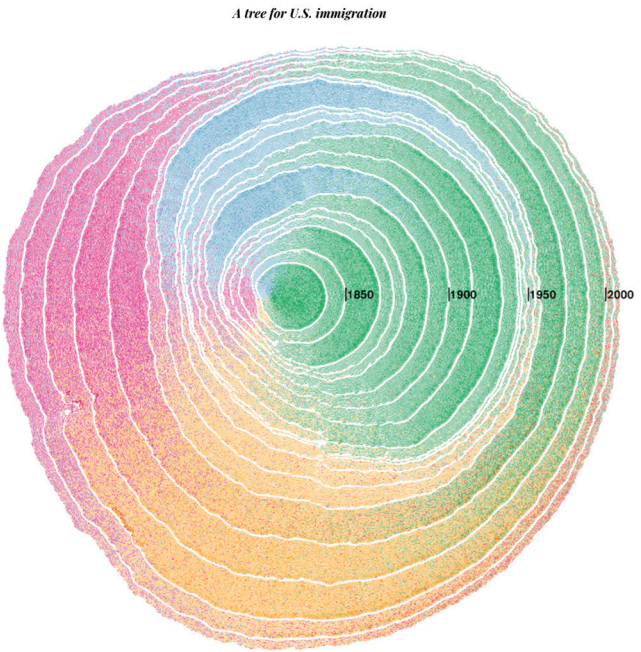
As you go forth and create your visualizations, continue to explore. There are yet undiscovered options, forms, shapes, and graphic types that can best show your data. Even if you come up with a great idea but don't know how to create it on the computer, you may find colleagues or partners who can put your ideas in practice.

LESSONS LEARNED

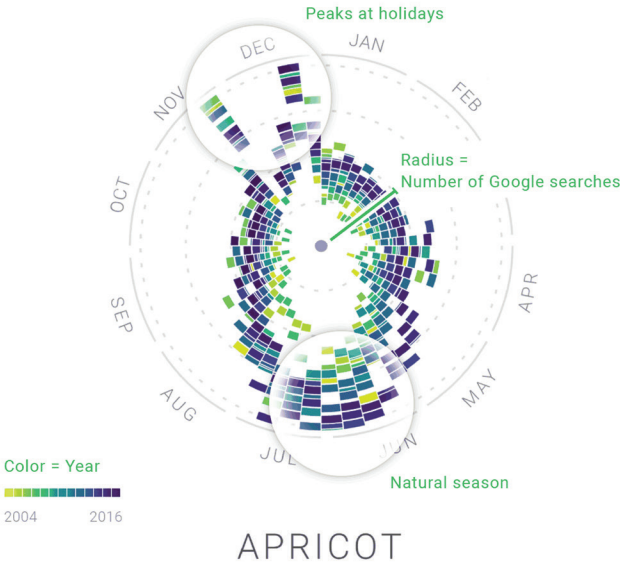
Over the course of this book, we have covered lots of rules, principles, and guidelines. If I were to sum these up into the absolute universal principles of data visualization, I would leave you with just these four.



Source: Jan Diehm and Amber Thomas, "Women's Pockets are Inferior," The Pudding



Source: Cruz, Wihbey, Ghael, and Shibuya, 2018



Source: Moritz Stefaner

SHOW YOUR DATA

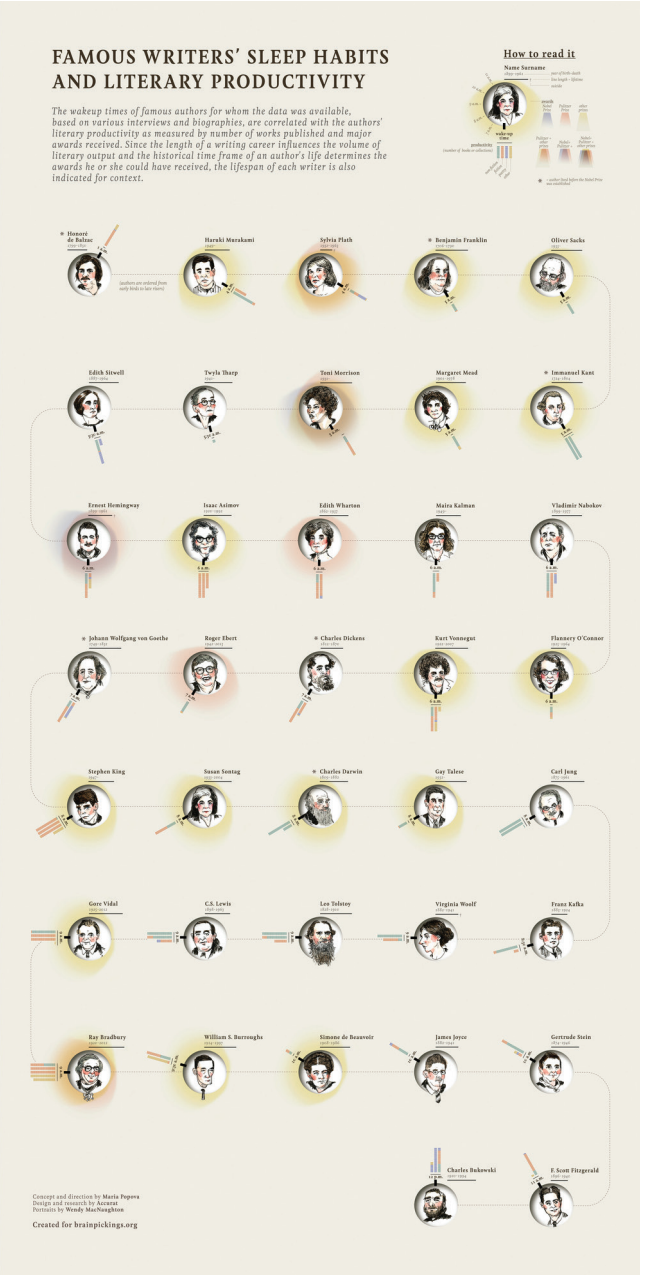
People are reading your graph to learn something, and they do that best by seeing the data. This doesn't mean you need to show them *all* of the data, but that you should always highlight the most important data.

REDUCE CLUTTER

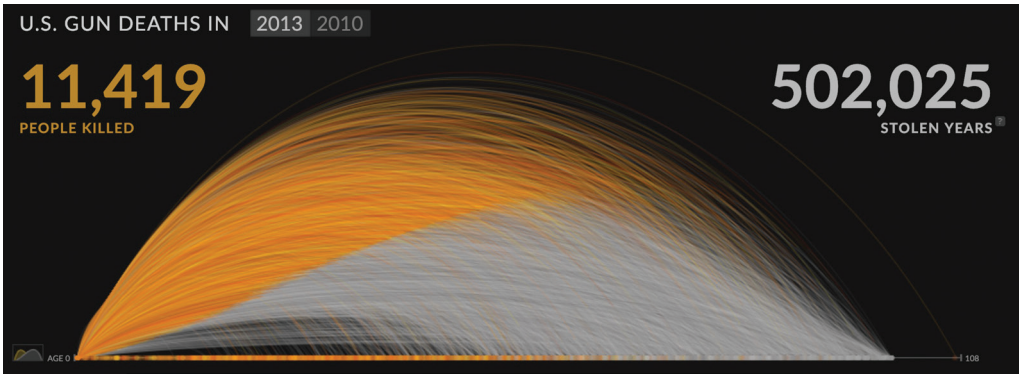
Reduce and remove all of the clutter that distracts your audience from the data or distorts the representation. Make it as easy as possible for your reader to see the most important points in your graph.

INTEGRATE THE GRAPHICS WITH THE TEXT

Directly label your data, remove legends, use active titles, and employ good labels and annotation. You may need to guide your reader through the graph before you help them



Source: Accurat. Portraits by Wendy MacNaughton; Design and research by Giorgia Lupi, Simone Quadri, Gabriele Rossi, Davide Ciuffi, Federica Fragapane, Francesco Majno. 2013.



Source: Periscopic: Do good with data

understand the content, but even that lesson can be accomplished with simple, smart labels and callouts.

CONSIDER YOUR AUDIENCE

Always remember with whom you are communicating. Academic researchers are looking for different things than practitioners, and practitioners are looking for different things than managers and policymakers. Try to identify your likely audience—try to talk to them if possible—and design your graphs to meet their needs. That way you can help them find insights, make discoveries, and do their jobs better.

FINAL THOUGHTS

I first became interested in data visualization after seeing much of my and my colleagues' work go unnoticed and unused. I did not come to the field with a degree in design or computer science or data science. And because I did it, I believe you can too. In fact, anyone can effectively communicate their data by thinking critically about their own work and the needs of their audience, readers, and users.

APPENDIX 1

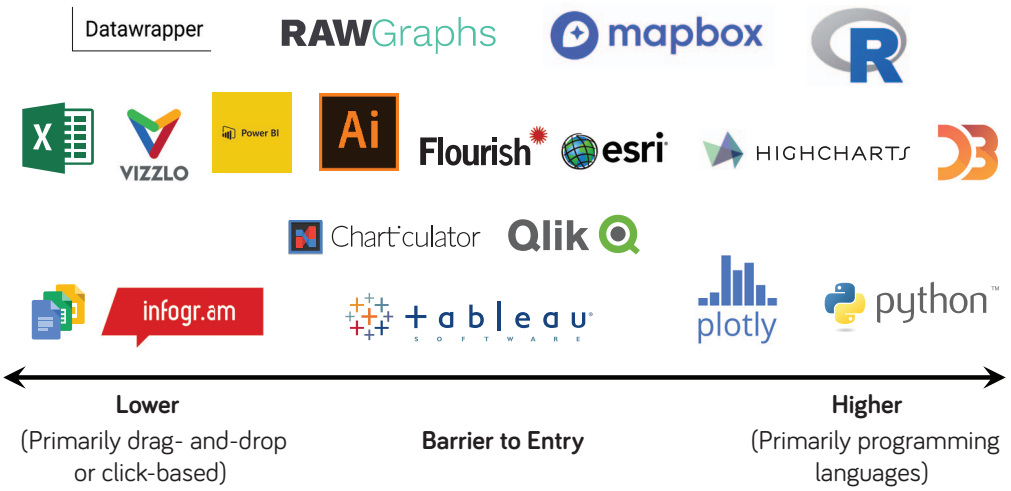
DATA VISUALIZATION TOOLS

This book is tool agnostic. My goal is not to show you how to create each of the eighty-plus graphs we explored. There are far too many tools you can use and far too many approaches within each tool. What matters is not which tool you use, but that it helps you create the graphs you need to best serve the needs of your audience.

There are many, many data visualization tools available, for use on different platforms and with different purchasing and subscription options. The number, types, and capabilities of these tools is constantly changing to reflect updated underlying technologies and coding languages. Which tool you use will depend on your personal preferences and skills, as well as those of your colleagues and support within your organization.

Data visualization tools live along a spectrum. On one end are click or drag-and-drop tools like Excel, which allow the user to click and insert a chart. On the other end are programming languages like R and JavaScript that require written code to create data visuals. The barrier to entry is much lower in Excel—virtually anyone can create a line chart or bar chart in seconds. It's much different on the other end, where programming languages require an understanding of how to write code and the different syntax across the different languages. Yet programming languages give you substantially more flexibility, while tools like Excel can box you in within a small subset of graphs.

How "difficult" these tools are depends on the person and sometimes the organization. You may have an affinity for computer programming languages, in which case JavaScript, Python, or R may be appropriate, but for a variety of reasons, your organization may not allow you to use these open source tools. Many social science researchers use



statistical languages like SAS, SPSS, and Stata, but, in my opinion, the graphing capabilities of these tools lag behind those of other languages. Some of the drag-and-drop tools like Excel and Tableau are easier to start with but designing more bespoke visualizations is either impossible or may require some coding (for example, Calculated Fields in Tableau).

I used a variety of tools to create the graphs in this book. Many, if not all, could probably be created in programming language tools like R or JavaScript; a sizable proportion could be created in drag-and-drop tools like Excel and Tableau. By utilizing a variety of tools, I found that some were easier to create the graph, but harder to style the way I liked and some were more difficult to learn while others were more intuitive. This list of tools is not exhaustive (by far) and I based my decision on which tools to include on my experience in the field, not on any formal survey or data set. Before you invest substantial time or money to learn any of these, you should explore the landscape of available tools and products.

PRIMARILY DRAG-AND-DROP OR CLICK-BASED

Adobe Illustrator. This is primarily a design tool. Adobe Illustrator, and the rest of the Adobe Creative Suite like Photoshop and InDesign, are the workhorse tools for designers. The graphing library is actually pretty poor in Illustrator, but you can insert graphs made in

other tools to add more styling, labels, and annotation. The Adobe Creative Cloud is now primarily subscription-based, but is not a particularly cheap tool to purchase.

Charticulator. Launched in 2018, Microsoft's Charticulator is an online tool that enables you to create a custom chart layout. It differs from some other tools in that you don't select from predefined charts but instead the creators have transformed chart specifications into mathematical layout parameters, such as marks (e.g., rectangle, line, or text) and axes (e.g., property of one direction in a plot). Charticulator charts are primarily static, but can be integrated with Microsoft's PowerBI tool to create interactive visualizations. At the time of this writing, Charticulator is free to use.

Datawrapper. An online tool from a team based in Germany where you can upload your data, select a graph template, refine and style, and publish or download. Created charts can be embedded in websites and can be made interactive. Datawrapper is free for most purposes, and paid versions allow for custom themes and additional exporting options. There are many tools in this same vein (Flourish and RAW below are two other examples)—some are better than others with different default options and user experience.

Excel. Likely the primary data and data visualization tool for many people around the world. As part of the Microsoft Office suite, Excel is not free, but it is your basic click-based tool. At the time of this writing, you can create more than 16 basic chart types in Excel, with variations within most of them. There is a basic library of charts that you can expand with clever "hacking" or additional coding using the Visual Basic for Applications (VBA) programming language that sits behind all Microsoft Office tools.

Flourish. Launched in 2016, Flourish is an online tool primarily aimed at newsrooms to help journalists create both static and interactive data visualizations in a drag-and-drop framework. There are options to customize and further develop Flourish graphs using the underlying JavaScript framework. Pricing options range from the free Public version to the paid Personal version (which gives you additional features) and the paid Business version (targeted towards large teams and organizations). Through a partnership with the Google News Lab, Flourish provides newsrooms free premium accounts.

Google Sheets. Akin to Excel, Google Sheets is part of the Google suite of tools. It works very similarly to Excel, though without some of the more sophisticated options. Because it is based online, the sharing capabilities are somewhat better than Excel (with the side effect being you need to have internet access to use it).

PowerBI. Microsoft's business intelligence tool that allows you to create interactive dashboards and visualizations. It directly links with the rest of the Microsoft Office suite

(especially Excel) and can be modified and customized in ways similar to Tableau. There is the free Power BI Desktop version, the paid Power BI Pro, and the Power BI Premium package for organizations.

RAW. Created by DensityDesign Research Lab in Italy in 2013, RAW was an early project designed to help creators link spreadsheet tools like Excel to graphic editing tools like Adobe Illustrator. It is an open source tool, which means you can download the code to further customize the visualization options. There is also an online platform in which, like other tools, you upload your data, select the graph, and customize. RAW has a variety of options for certain non-standard graphs (like streamgraphs and bump charts) that are not typically available in other tools. RAW is free to use.

Tableau. Perhaps the most popular business intelligence dashboarding tool, Tableau's drag-and-drop interface enables you to create interactive dashboards and visualizations. Like Excel, users have customized their Tableau work to create an array of visualizations outside the basic graph menu. There are a number of versions of Tableau, from the free Tableau Public (but which means you save your work to the Tableau website) to paid versions like Tableau Desktop and Tableau Server (for large organizations).

ONLINE TOOLS (CLICK-BASED)

Infogram, Venngage, and Vizzlo. These are just three of the many click-based online tools that are aimed more for people who want to quickly create infographics and reports. In my experience, these tools sometimes have more graph options than other online tools, but they are not always based on best practices. Pricing varies from free packages that usually mean your data and visualizations can be viewed by anyone, to enterprise packages for large teams and organizations.

PROGRAMMING LANGUAGES

D3. We first need to define JavaScript. JavaScript is programming language that allows you to implement information onto a webpage. Every time a web page does something, like display updates, animate graphics, or play videos, JavaScript is probably involved. D3 is a JavaScript library for manipulating objects based on data and was developed by Mike Bostock along with Jeff Heer and Vadim Ogievetsky at Stanford University in the early 2010s. Most of the interactive data visualizations we currently see on the web are run on D3—virtually

every interactive graph you play with on the *New York Times*, *Washington Post*, and *Guardian* websites is built with D3. Like other programming languages, there is a steep learning curve to using D3, but you are basically unlimited in the kinds of visualizations you can create. D3 is an open source language, which means it is free to use.

Highcharts. Launched in 2009 by a team in Norway, Highcharts—and its cousin tools Highstock, Highmaps, Highcharts Cloud, and Highslide—is a suite of interactive data visualization tools rooted in JavaScript. You do need to know a bit of coding to use Highcharts, but templates and libraries help you create the basics of a graph and then add additional styling and formats. Highcharts is free for personal use and nonprofit organizations; pricing then varies by number of licenses and package.

Python. Python is used in a wide number of fields and industries from basic and complex data analysis to web application to artificial intelligence. Like D3 and R, the language is open source, which means there are a lot of open, free libraries to help you create data visualizations including, for example, Matplotlib, Seaborn, Bokeh, and ggplot.

R. Conceived in 1992 and initially released in 1995, R is a free, open source programming language for statistical computing and graphics. R is becoming more and more widely used for data visualization, especially since the launch of the "ggplot2" package by Hadley Wickham in 2005 (based on the "Grammar of Graphics" by Leland Wilkinson). R allows you to conduct statistical analyses and create customizable data visualizations; additional tools and packages can enable you to create interactive visualizations.

APPENDIX 2

FURTHER READINGS AND RESOURCES

GENERAL DATA VISUALIZATION BOOKS

SCOTT BERINATO. Berinato's book *Good Charts: The HBR Guide to Making Smarter, More Persuasive Data Visualizations* focuses on graphs for the business community. The book is wide-ranging and discusses more of the differences between explanatory and exploratory charts than some of the other leading data visualization books. He has a follow-up workbook that provides hands-on examples and tutorials.

ALBERTO CAIRO. Author of several books on data visualization including *The Functional Art*, *The Truthful Art*, and *How Charts Lie*. Cairo is a journalism professor, so his books focus primarily on creating data visualizations for telling stories to a wide audience. His books provide fundamental overviews of data, data visualization, introductory statistics, and how to create visualizations. His most recent book *How Charts Lie*, helps readers spot lies in deceptive graphs and how to become better consumers of data visualization.

JORGE CAMÕES. His book *Data at Work* covers a wide range of data visualization principles and strategies, ranging from rules of visual perception to design considerations to data preparation and visualization.

STEPHEN FEW. Author of several books on data visualization, his *Show Me the Numbers* and *Now You See It: Simple Visualization Techniques for Quantitative Analysis* are comprehensive overviews of how to present data effectively and strategically.

ANDY KIRK. Author of two books on data visualization, his *Data Visualisation: A Handbook for Data Driven Design* provides readers with a system to conceptualize and develop data visualizations, and a process to help readers make design choices that result in clear and effective visualizations.

JUSSO KOPONEN AND JONATAN HILDÉN. Their *Data Visualization Handbook* (translated from Finnish) is a practical guide to data visualization and contains lots of examples of data visualizations and information graphics.

COLE NUSSBAUMER KNAFLIC. Knaflic's book *Storytelling with Data*, and blog of the same name, provides an introductory treatment of data visualization, and how to pair text with graphs to tell effective, compelling stories. Her follow-up book, *Storytelling with Data: Let's Practice!*, takes the reader through a series of hands-on exercises to practice their data visualization skills.

ISABEL MEIRELLES. A professor of design, Meirelles' *Design for Information* surveys current examples of data visualizations for both elements of content and design. Numerous examples provide a library of visualization types and approaches.

TAMARA MUNZNER. Her *Visualization Analysis and Design* book approaches data visualization from a more systematic academic-based perspective. It features a unified approach to reading and creating visualizations, all rooted in the academic literature. Munzner's book is more along the lines of a true data visualization textbook.

CLAUS O. WILKE. Wilke's book *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*, takes a fundamental and practical approach to data visualization. Wilke presents basic principles of good data visualization strategies and practices and also shows how to create the visuals in his book using the R programming language. Like this book, it is one of the few to include a wider array of graphs than the typical line, bar, pie, and map.

STEVE WEXLER, JEFF SHAFFER, AND ANDY COTGREAVE. Their *Big Book of Dashboards* is one of the more comprehensive treatments of creating interactive dashboards. Mostly rooted in the Tableau software tool, the authors review nearly thirty examples to help the reader determine good design and interactive experiences.

DONA WONG. Wong's *Guide to Information Graphics* dedicates individual pages to specific graph types, how and why to choose the best chart to fit the data, the most effective way to communicate data, and what to include and not include in different graphs.

HISTORICAL DATA VISUALIZATION BOOKS

R. J. ANDREWS. Andrews' *Info We Trust: How to Inspire the World with Data* takes a different perspective on data visualization techniques, drawing heavily on historical visualizations and design approaches by luminaries in the field over the past century.

WHITNEY BATTLE-BAPTISTE AND BRITT RUSERT. The first full exploration of W.E.B. DuBois's data visualizations from the 1900 Paris Exposition, *W.E.B. Du Bois's Data Portraits: Visualizing Black America*, is an image-by-image description of these early data visualizations.

BRUCE BERKOWITZ. His biography of William Playfair shows the real life of the man who invented "statistical graphics." Berkowitz goes beyond the statistics and graphics that Playfair is best known for to tell a detailed story of his life.

MANUEL LIMA. In his two books, *The Book of Trees: Visualizing Branches of Knowledge* and *The Book of Circles: Visualizing Spheres of Knowledge*, Lima explores the long history of the tree diagram and circular information design. Both take you on a tour of the long history of both kinds of visualization methods.

SANDRA RENDGEN. Similar to the DuBois book, in Rendgen's *The Minard System: The Complete Statistical Graphics of Charles-Joseph Minard*, she explores the career and story behind Minard's graphs, maps, and tables.

BOOKS ON DATA VISUALIZATION TOOLS

The list in Appendix 1 includes just a selection of tools I've used for this book. Which tool is best for you depends on your existing expertise, your needs, and the needs of your audience. In my experience, these books have the best tool-specific data visualization tutorials. There are countless online blogs and resources you can also explore.

GARRETT GROLEMUND AND HADLEY WICKHAM. Probably my go-to book to learn R, *R for Data Science*, takes a comprehensive view at how to work with and visualize data in R. It is especially strong on the side of wrangling and visualizing data.

KEIRAN HEALY. Maybe more of an introduction to basic core principles of data visualization, Healy's *Data Visualization: A Practical Introduction* mixes in a how-to in R. There is an online companion with code snippets.

NORMAN MATLOFF. At more than 400 pages, Matloff's *The Art of R Programming* is a beast—and that's because it covers everything in R, an essential reference book in your library.

ERIC MATTHES. Every Python programmer I know loves Matthes' *Python Crash Course: A Hands-On, Project-Based Introduction to Programming*. Beginners and more experienced coders are really fond of how he leads the reader through the language and the useful exercises in each chapter.

RYAN SLEEPER. One of the more recent (and better) books on creating calculations and custom visualizations in the Tableau software tool, Sleeper's *Practical Tableau* is a great start to using Tableau.

AMELIA WATTENBERGER. One of the few books about the D3 JavaScript library, likely because coding examples and snippets are really important, *Fullstack D3 and Data Visualization* (and the companion website) is a comprehensive introduction to D3. This is the most recent book about D3 (and perhaps the best) because it walks the reader through the entire process and includes a large amount of code snippets and examples on the book's website.

DATA VISUALIZATION LIBRARIES

There is no one-to-one mapping between data types and graph types. A column chart, for example, can be used to show changes over time or to compare differences between categories. There are a variety of resources you can use to help you select a graph, but the ultimate decision will rest with you, your data, your audience, and your creativity.

THE CHARTMAKER DIRECTORY. A crowd-sourced interactive matrix that consists of 50 graph types and 40 different tools. Users can post links to tutorials on how to make the graphs in the different tools. http://chartmaker.visualisingdata.com/

CHART SUGGESTIONS—A THOUGHT STARTER. A relatively smaller, static project with about 20 graphs broken into four categories. One of the first of these charting libraries. http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

THE DATA VISUALISATION CATALOGUE. Online graph resource with descriptions, anatomy, and some video guides to each chart type. https://datavizcatalogue.com/

THE DATA VIZ PROJECT. An interactive website with more than 100 graph types including select examples of each. http://datavizproject.com/

THE GRAPHIC CONTINUUM. A poster, smaller sheet, flash cards, and card game I produced with a friend, *The Graphic Continuum* shows more than 90 different graph types grouped into six categories. https://policyviz.com/product/graphic-continuum-poster/

INTERACTIVE CHART CHOOSER. An interactive chart chooser that lets you sort and filter more than 30 graph types based on the data you are trying to visualize. Links to real examples help you see these chart types in action. https://depictdatastudio.com/charts/

THE R GRAPH GALLERY AND PYTHON GRAPH GALLERY. A pair of webpages that contain hundreds of charts made with the R and Python programming languages. Each chart contains reproducible code. https://www.r-graph-gallery.com/ and https://python-graph-gallery.com/

TEXT VISUALIZATION BROWSER. Nearly 450 examples of ways to visualize text data. https://textvis.lnu.se/

THE VISUAL VOCABULARY. A poster project from the graphics desk at the *Financial Times*, the Visual Vocabulary shows more than 70 different graph types in 9 categories. https://github.com/ft-interactive/chart-doctor/blob/master/visual-vocabulary/Visual-vocabulary.pdf

XENO.GRAPHICS. A collection of unusual charts and maps to help you expand your visualization repertoire even further. https://xeno.graphics/

WHERE TO PRACTICE

Maybe the best way to improve and refine your data visualization technique is to create visualizations. Exploring different data sets, visualization types, and tools can help you refine your aesthetic, and play with different techniques and forms. There are a few community projects that you may want to explore to help you on your way.

MAKEOVER MONDAY. A weekly learning project in which participants work with a sample data set to try to create better, more effective visualizations. It tends to be focused in the

Tableau community, but visualizations can be created in any tool. Visualizations are created and posted publicly with review and feedback on social media and a mid-week webinar to enable people to give feedback and refine their ideas.

OBSERVE, COLLECT, DRAW. Primarily shown on Instagram, *ObserveCollectDraw* is based on the book by designers Giorgia Lupi and Stefanie Posavec that encourages people to collect their own personal data and draw their visualizations. This project—and their original book *Dear Data*—shows people how to collect and record their personal experiences in analog ways.

STORYTELLING WITH DATA. The Storytelling with Data Community is a place to practice your data visualization skills, get and give feedback, and discuss topics related to effectively communicating data. There is a monthly "SWDchallenge," which asks participants to address a specific data visualization challenge and a periodic "SWDexercise" that is based on real-world scenarios with actual data.

TIDY TUESDAY. A weekly project aimed at R users, project managers post a dataset and a published chart or article related to the dataset. Participants are asked to explore the data and publish their results (and code) for others to use and adapt.

OTHERS. There are other similar projects, but tend to have dedicated audiences in different ways. The Data Visualization Society, a community of data visualization practitioners, hosts Slack channels revolving around practice and critique. There are also various forums on HelpMeViz, Reddit, Stack Exchange, Twitter, and other social media platforms where you can practice and publish your work.

ACKNOWLEDGMENTS

In 2014, I wrote an introductory article on data visualization for the *Journal of Economic Perspectives (JEP)*. I had always wanted to publish in the *JEP*, but never imagined that data visualization—rather than public policy or economics—would be the topic of interest. Shortly after the article was published, I received a call from Bridget Flannery-McCoy at Columbia University Press asking to see if I was interested in writing a full book on data visualization. At that time, I wasn't ready to write this book—there were a lot of books just out or coming out and to do it right, I would need to dedicate a lot more time and resources than I had available. Bridget's call provide fruitful nonetheless, ultimately resulting in my first book, *Better Presentations.*

Fast-forward five years and Bridget called again to see if I was ready for a follow-up book. Now, having changed jobs and focusing more on data communication, I found myself with more to say and an idea for a book that would place it in a unique space in the field.

I wrote the first draft of the entire manuscript on two train rides between Washington, DC and New York City. But it would take another two-and-a-half years to fully develop it, reorganize it, expand it, and create the graphs. The result is the book you are now holding.

For getting me to this point, my gratitude belongs to several important people: Brittany Fong, Ajjit Narayanan, Jon Peltier, Anthea Piong, and Aaron Williams helped me with various Tableau, Excel, R, and JavaScript challenges and tasks. RJ Andrews, John Burn-Murdoch, Alberto Cairo, Jennifer Christiansen, Alice Feng, John Grimwade, Steve Haroz, Robert Kosara, and Severino Ribecca were generous in their time to have discussions, provide

feedback, and assist with a number of requests to find research and images, both modern and historical.

More thanks are due to the people and organizations who have granted me permission to include their work in this book. Their work is truly special and I am grateful they allowed me to include it in these pages.

Additional thanks to Ken Skaggs for helping to manage the PolicyViz Podcast and to the more than 200 guests who have appeared on the show.

A special thanks to Alberto Cairo, Nigel Holmes, Jessica Hullman, David Napoli, and Chad Skelton for reviewing parts or all of the manuscript and providing invaluable feedback.

Very special thanks are due to Kenneth Field for creating many of the maps and for keeping up with my "feature creep." Special thanks are also due to Hiram Henriquez for creating original illustrations and making all of the graphs ready for print.

This book could not have come together without the help of my Columbia University Press editor, Stephen Wesley, who kept up with my incessant emails and questions. Additional thanks to Christian Winting, Ben Kolstad, and the rest of the Columbia University Press team for helping to bring this project to life.

I am grateful for my Urban Institute colleagues whose dedication to fact-based research has the power to improve public policy and practice, strengthen communities, and transform people's lives for the better. They have helped create a special place that not only values in-depth scholarship, but also innovative ways to communicate that work.

I am also indebted to the many, many friends and strangers with whom I have discussed and debated aspects of data visualization and data communication over the years. The data visualization community is a special place—the warm welcome it gave me about 10 years ago changed my career in unimaginable ways, and I hope to pass it forward to another crop of data communicators. I am also especially thankful to people in the field whose creativity continues to inspire.

Writing a third book is simultaneously easier and harder than writing the first and second. I could not have summoned the courage to go through another project without the love and support of friends and family who have cheered me on and taken interest in even the most mundane details of my work. Special thanks are due to the ongoing support and love of my entire family.

Finally, my most special thanks are due to the three most important people in my life: my wife and kids. My kids, Ellie and Jack, are my favorite sources of happiness, pride, and fun. They have watched their dad struggle with code, complain about color palettes and fonts,

and watched me stay up late into the night, reading and writing. Through it all, they supplied me with steady encouragement, love, and support. As a dad, you know you're doing something right when they lean over your shoulder and say, "How's it coming with that heatmap?"

My deepest thanks goes to my wife, Lauren, who has edited every page of this book, cutting out repetitive text and keeping the language clear and descriptive, always working with my reader in mind. She keeps all of us moving forward, especially when I'm writing, speaking, or traveling demands take me far and wide. She routinely reminds me she's the best thing that's ever happened to me. A truer statement has never been spoken.

REFERENCES

Ahmed, Naema, Cassi Pollock, and Alex Samuels. "How the Texas Democratic and Republic Party Platforms Compare." *Texas Tribune*, July 5, 2018, https://apps.texastribune.org/features/2018/party-platforms/?_ga=2.129478090.770685496.1576106215–1729798919.1576004948.

Alda, Alan. *If I Understood You, Would I Have This Look on My Face?: My Adventures in the Art and Science of Relating and Communicating*. New York: Random House, 2018.

Andrews, R. J. "Florence Nightingale Is a Design Hero." Medium, July 15, 2019, https://medium.com/nightingale/florence-nightingale-is-a-design-hero-8bf6e5f2147.

Andrews, R. J. *Info We Trust: How to Inspire the World with Data*. New York: Wiley, 2019.

Andris, Clio, David Lee, Marcus J. Hamilton, Mauro Martino, Christian E. Gunning, and John Armistead Selden. "The Rise of Partisanship and Super-Cooperators in the US House of Representatives." *PloS one* 10, no. 4 (2015): e0123507.

Anscombe, Francis J. "Graphs in Statistical Analysis." *The American Statistician* 27, no. 1 (1973): 17–21.

Arellano, Cristina, Juan Carlos Conesa, and Timothy J. Kehoe. "Chronic Sovereign Debt Crises in the Eurozone, 2010–2012." Federal Reserve Bank of Minneapolis Economic Policy paper 12–4, May 2012, https://www.minneapolisfed.org/~/media/files/pubs/eppapers/12-4/epp_12-4_chronic_sovereign_debt_crisis_eurozone.pdf.

Avery, Beth. "Ban the Box: U.S. Cities, Counties, and States Adopt Fair Hiring Policies." National Employment Law Project. July 1, 2019. https://www.nelp.org/publication/ban-the-box-fair-chance-hiring-state-and-local-guide/.

AxisMaps. "CartographyGuide." https://www.axismaps.com/guide/general/map-projections/, accessed November 2019.

Bard Graduate Gallery. "Marimekko: Fabrics, Fashion, Architecture." https://www.bgc.bard.edu/gallery/exhibitions/43/marimekko.

Battle-Baptiste, Whitney, and Britt Rusert. "WEB Du Bois's Data Portraits: Visualizing Black America: The Color Line at the Turn of the Twentieth Century." Amherst, Massachusetts: WEB Du Bois Center at the University of Massachusetts Amherst (2018).

BBC. "BBC Audiences Tableau Style Guide." https://public.tableau.com/profile/bbc.audiences#!/vizhome/BBCAudiencesTableauStyleGuide/Hello

Béland, Antoine, and Thomas Hurtut. "Unit Visualizations for Visual Storytelling." OSF Preprints, https://osf.io/bshpc/.

Bennet, Tim. "How 52 Ninth-Graders Spell 'Camouflage', Sankey Diagram." https://www.reddit.com/r/dataisbeautiful/comments/6a4pb8/how_52_ninthgraders_spell_camouflage_sankey/?st=J2NBTEoQ&sh=ddd5c5e1

Berinato, Scott. "The Power of Visualizations' 'Aha!' Moments." *Harvard Business Review*, March 19, 2013, https://hbr.org/2013/03/power-of-visualizations-aha-moment.

Berman, Jacob. "Washington Metro Map Print Original Art." Etsy listing, 2019, https://www.etsy.com/listing/647623499/washington-metro-map-print-original-art

Bontemps, Xtophe. "Why You Should Never Use Radar Plots."data.visualisation.free.fr, March 2017, https://rpubs.com/Xtophe/268920.

Borkin, Michelle A., Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. "Beyond Memorability: Visualization Recognition and Recall." *IEEE Transactions on Visualization and Computer Graphics* 22, no. 1 (2015): 519–528.

Börner, Katy, Andreas Bueckle, and Michael Ginda. "Data Visualization Literacy: Definitions, Conceptual Frameworks, Exercises, and Assessments." *Proceedings of the National Academy of Sciences* 116, no. 6 (2019): 1857–1864.

Bortins, I., Demers S., & Clarke, K. (2002). Cartogram Types. http://www.ncgia.ucsb.edu/projects/Cartogram_Central/types.html

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. "D^3 Data-Driven Documents." *IEEE transactions on visualization and computer graphics* 17, no. 12 (2011): 2301–2309.

Brehmer, Matt. "Visualizing Ranges Over Time on Mobile Phones." Medium, November 3, 2018, https://medium.com/multiple-views-visualization-research-explained/ranges-900a04d7d32a.

Brehmer, Matthew, Bongshin Lee, Petra Isenberg, and Eun Kyoung Choe. "Visualizing ranges over time on mobile phones: a task-based crowdsourced evaluation." *IEEE transactions on visualization and computer graphics* 25, no. 1 (2018): 619–629.

Bremer, Nadieh. "More Fun Data Visualizations with the Gooey Effect." Visual Cinnamon (blog). June 20, 2016. https://www.visualcinnamon.com/2016/06/fun-data-visualizations-svg-gooey-effect.html.

Bremer, Nadieh. "Techniques for Data Visualization on Both Mobile & Desktop." Visual Cinnamon (blog). April 17, 2019. https://www.visualcinnamon.com/2019/04/mobile-vs-desktop-dataviz#recap.

Brinton, Willard Cope. *Graphic Methods for Presenting Facts*. Engineering magazine company, 1919.

Broderick, Ryan. "Who Watches More Porn: Republicans or Democrats?" BuzzFeed News. April 2014. https://www.buzzfeednews.com/article/ryanhatesthis/who-watches-more-porn-republicans-or-democrats.

Bureau of Labor Statistics. Unemployment rate. https://www.bls.gov/cps/

Burn-Murdoch, John. "Episode #155: John Burn-Murdoch." The PolicyViz Podcast, June 18, 2019, https://policyviz.com/podcast/episode-155-john-burn-murdoch/.

Burn-Murdoch, John. Twitter post. October 28, 2019, 4:04 a.m., https://twitter.com/jburnmurdoch/status/1188728193945100288.

Bush, George W. "Address Before a Joint Session of the Congress on the State of the Union." The American Presidency Project, February 2, 2005, https://www.presidency.ucsb.edu/documents/address-before-joint-session-the-congress-the-state-the-union-14.

Byron, Lee, and Martin Wattenberg. "Stacked Graphs–Geometry and Aesthetics." *IEEE Transactions on Visualization and Computer Graphics* 14, no. 6 (2008): 1245–1252.

Cairo, Alberto. "Annotation, Narrative, and Storytelling in Infographics and Visualization." The Functional Art (blog), April 16, 2014, http://www.thefunctionalart.com/2014/04/annotation-narrative-and-storytelling.html.

Cairo, Alberto. "Download the Datasaurus: Never Trust Summary Statistics Alone; Always Visualize Your Data." The Functional Art (blog), August 29, 2016, http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html.

Cairo, Alberto. *The Functional Art: An Introduction to Information Graphics and Visualization*. Berkeley, CA: New Riders, 2012.

Cairo, Alberto. "How Charts Lie." (2019).

Cairo, Alberto. "Visual Storytelling w/ Alberto Cairo and Robert Kosara." The Data Stories Podcast, Episode #35, April 16, 2014, https://datastori.es/data-stories-35-visual-storytelling-w-alberto-cairo-and-robert-kosara/.

Camões, Jorge. "Perception: Gestalt Laws." excelcharts.com (blog), January 10, 2012, https://excelcharts.com/data-visualization-excel-users/gestalt-laws/.

Carpenter, Patricia A., and Priti Shah. "A Model of the Perceptual and Conceptual Processes in Graph Comprehension." *Journal of Experimental Psychology: Applied* 4, no. 2 (1998): 75.

Center on Budget and Policy Priorities. "Working-Family Tax Credits Help at Every Stage of Life." Graphic, https://www.cbpp.org/working-family-tax-credits-help-at-every-stage-of-life-0.

Centers for Disease Control and Prevention. "CDC Wonder." https://wonder.cdc.gov/.

Centers for Disease Control and Prevention. "Community Mitigation Guidelines to Prevent Pandemic Influenza—United States, 2017." *Morbidity and Mortality Weekly Report* 66, No. 1, April 21, 2017. https://stacks.cdc.gov/view/cdc/45220.

Centers for Disease Control and Prevention. "National, Regional, and State Level Outpatient Illness and Viral Surveillance." https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.

Cesal, Amy. "The Sunlight Foundation's Data Visualization Style Guidelines." Sunlight Foundation, March 12, 2014, https://sunlightfoundation.com/2014/03/12/datavizguide/.

Chalabi, Mona. Presentation, Tapestry Conference, University of Miami, November 29, 2018.

Chang, Kenneth. "A Different Way to Chart the Spread of Coronavirus." *New York Times*, March 20, 2020. https://www.nytimes.com/2020/03/20/health/coronavirus-data-logarithm-chart.html.

Chappell, Nathan and Adam B. Jaffe. "Intangible Investment and Firm Performance." *National Bureau of Economic Research Working Paper No. 24363*, March 2018, https://www.nber.org/papers/w24363.

Chase, Will. "Voronoi Treemap" (website). May 2, 2019, https://observablehq.com/@will-r-chase/voronoi-treemap.

Chase, Will. "Why I'm Not Making COVID19 Visualizations, and Why You (Probably) Shouldn't Either" (blog post). March 31, 2020, https://www.williamrchase.com/post/why-i-m-not-making-covid19-visualizations-and-why-you-probably-shouldn-t-either/.

Cherdarchuk, Joey. "Clear off the Table." Dark Horse Analytics (blog). March 27, 2014, https://www.darkhorseanalytics.com/blog/clear-off-the-table.

Choi, Jinho, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmqvist. "Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization." *Computer Graphics Forum* 38, no. 3 (2019): 249–260.

Christiansen, Jen. "Pop Culture Pulsar: Origin Story of Joy Division's Unknown Pleasures Album Cover." *Scientific American*, February 18, 2015, https://blogs.scientificamerican.com/sa-visual/pop-culture-pulsar-origin-story-of-joy-division-s-unknown-pleasures-album-cover-video/.

Cleveland, William S. *The Elements of Graphing Data*. Belmont, CA: Wadsworth, 1985.

Cleveland, William S. *Visualizing data*. Hobart Press, 1993.

Cleveland, William S., and Robert McGill. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79, no. 387 (1984): 531–554.

Colour Blindness Awareness. "Colour Blindness." http://www.colourblindawareness.org/colour-blindness/, accessed November 2019.

Congressional Budget Office. "CBO's August 2018 report An Update to the Economic Outlook: 2018 to 2028." www.cbo.gov/publication/54318.

Congressional Budget Office. "Changes in the Distribution of Workers' Annual Earnings Between 1979 and 2007." Congressional Budget Office, October 2009, https://www.cbo.gov/sites/default/files/111th-congress-2009-2010/reports/10-02-workers.pdf.

Congressional Budget Office. "Social Security Policy Options." Congressional Budget Office, July 2010, https://www.cbo.gov/sites/default/files/111th-congress-2009-2010/reports/07-01-ssoptions_forweb.pdf

Congressional Budget Office. "The 2012 Long-Term Budget Outlook | House Budget Committee." Congressional Budget Office, September 12, 2014, https://www.youtube.com/watch?v=dqhqZYYGNnA

Congressional Budget Office. "The Budget and Economic Outlook: Fiscal Years 2012 to 2022." Congressional Budget Office." January 31, 2012, https://www.cbo.gov/publication/42905?index=12699

Congressional Budget Office. "The Budget and Economic Outlook: Infographic." Congressional Budget Office." June 5, 2012, https://www.cbo.gov/publication/43289

Congressional Budget Office. "The Social Security Disability Insurance Program." Infographic. https://www.cbo.gov/publication/43432

Congressional Budget Office. "The Supplemental Nutrition Assistance Program." Infographic. https://www.cbo.gov/publication/43174.

Cook, Albert M., and Janice Miller Polgar. *Assistive Technologies-E-Book: Principles and Practice*. St. Louis, MO: Elsevier, 2014.

Correll, Michael, and Michael Gleicher. "Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error." *IEEE Transactions on Visualization and Computer Graphics* 20, no. 12 (2014): 2142–2151.

Correll, Michael, Enrico Bertini, and Steven Franconeri. "Truncating the Y-Axis: Threat or Menace?" In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ed. Regina Bernhaupt et al., 1–12. New York: Association for Computing Machinery. 2020.

Correll, Michael. "Ethical Dimensions of Visualization Research." Tableau Research, CHI 2019, 2019.

Correll, Michael. "Visualization Design Principles for the Pandemic." Medium, April 3, 2020, https://medium.com/@mcorrell/visualization-design-principles-for-the-pandemic-e65388280d16.

Cousins, Carrie. "Color and Cultural Design Considerations." WebDesignerDepot (blog), June 11, 2012, https://www.webdesignerdepot.com/2012/06/color-and-cultural-design-considerations/.

Cox, Amanda. "Amanda Cox—Eyeo Festival 2011." https://vimeo.com/29391942

Cruz, Pedro de, John Wihbey, and Felipe Shibuya. *Simulated Dendrochronology of the United States*. 2018.

D'Ignazio, Catherine and Lauren F. Klein. *Data Feminism.* Boston, MA: MIT Press. 2020.

Dallas Morning News. *Dallas Morning News Graphics Stylebook*. https://knightcenter.utexas.edu/mooc/file/tdmn_graphics.pdf.

Daniels, Matt. "The Largest Vocabulary In Hip Hop." The Pudding, January 21, 2019, https://pudding.cool/projects/vocabulary/index.html.

DeBelius, Danny. "Let's Tesselate: Hexagons For Tile Grid Maps." NPR visuals team, May 11, 2015, https://blog.apps.npr.org/2015/05/11/hex-tile-maps.html.

DeBold, Tynan and Dov Friedman. "Battling Infectious Diseases in the 20th Century: The Impact of Vaccines." *Wall Street Journal*, February 11, 2015, http://graphics.wsj.com/infectious-diseases-and-vaccines/.

Desilver, Drew. "A Record Number of Women Will Be Serving in the New Congress." Pew Research Center, December 18, 2018. https://www.pewresearch.org/fact-tank/2018/12/18/record-number-women-in-congress/

Dorling, Daniel. "Area cartograms: their use and creation." *The map reader: Theories of mapping practice and cartographic representation* (2011): 252–260.

DuBois, William Edward Burghardt, ed. *The College-bred Negro: Report of a Social Study Made Under the Direction of Atlanta University; Together with the Proceedings of the Fifth Conference for the Study of the Negro Problems, Held at Atlanta University, May 29–30, 1900*. No. 5. Atlanta University Press, 1900.

Economist. "The Climate Issue." https://www.economist.com/leaders/2019/09/19/the-climate-issue.

Economist. "The Dragon Takes Wing." May 1, 2014, https://www.economist.com/news/finance-and-economics/21601568-new-data-suggest-chinese-economy-bigger-previously-thought-dragon.

Economist. "Flattening the Curve: Covid-19 Is Now in 50 Countries, and Things Will Get Worse." *The Economist*, February 29, 2020. https://www.economist.com/briefing/2020/02/29/covid-19-is-now-in-50-countries-and-things-will-get-worse.

Eurostat, "Graphical Style Guide—A Practical Layout Guide for Eurostat Publications—2016 Edition." *Eurostat News*, February 25, 2016, https://ec.europa.eu/eurostat/web/products-eurostat-news/-/STYLE-GUIDE_2016.

Federal Home Loan Bank Board. Home Owners' Loan Corporation. 1933-7/1/1939. "City of Richmond, Virginia and Environs." U.S. National Archives and Records Administration. https://catalog.archives.gov/id/85713737

Federal Reserve Board of Governors. "Household Debt Services and Financial Obligations Ratios." Accessed January 2020. https://www.federalreserve.gov/releases/housedebt/

Few, Stephen. "Bullet Graph Design Specification." Perceptual Edge, Visual Business Intelligence Newsletter, October 10, 2013, https://www.perceptualedge.com/articles/misc/Bullet_Graph_Design_Spec.pdf.

Few, Stephen. "The DataVis Jitterbug: Let's Improve an Old Dance." Perceptual Edge, Visual Business Intelligence Newsletter, April/May/June 2017, https://www.perceptualedge.com/articles/visual_business_intelligence/the_datavis_jitterbug.pdf.

Few, Stephen. *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Oakland, CA: Analytics Press, 2004.

Few, Stephen. "Unit Charts Are For Kids." Perceptual Edge, Visual Business Intelligence Newsletter, October, November, December 2010, https://www.perceptualedge.com/articles/visual_business_intelligence/unit_charts_are_for_kids.pdf.

Field, Kenneth. *Cartography*. Redlands, CA: Esri, 2018.

FiveThirtyEight. "50 Years Of World Cup Doppelgangers." July 15, 2018, https://projects.fivethirtyeight.com/world-cup-comparisons/romelu-lukaku-2018/

Florence, Philip Sargant. *Only an Ocean Between*. New York: Essential, 1946.

Fontenot, Kayla, Jessica Semega, and Melissa Kollar. "Income and Poverty in the United States: 2017." U.S. Bureau of the Census, September 2008, https://www.census.gov/content/dam/Census/library/publications/2018/demo/p60-263.pdf.

Friendly, Michael. "Playfair Balance of Trade Data." September 13, 2018, http://www.datavis.ca/courses/RGraphics/R/playfair-east-indies.html.

Fruchterman, Thomas MJ, and Edward M. Reingold. "Graph Drawing by Force-Directed Placement." *Software: Practice and experience* 21, no. 11 (1991): 1129–1164.

Fry, Ben. "Tracing the Origin of Species." https://fathom.info/traces/.

Gallo, Carmine. *The Presentation Secrets of Steve Jobs: How to Be Insanely Great in Front of Any Audience*. Upper Saddle River, NJ: Prentice Hall, 2010.

Gambino, Megan. "Do Our Brains Find Certain Shapes More Attractive Than Others?" Smithsonian Magazine. November 14, 2013. https://www.smithsonianmag.com/science-nature/do-our-brains-find-certain-shapes-more-attractive-than-others-180947692/.

Gantt, Henry Laurence. "A Graphical Daily Balance in Manufacture." No. 1002 in American Society of Mechanical Engineers. *Transactions of the American Society of Mechanical Engineers*. New York City: The Society, 1880. https://babel.hathitrust.org/cgi/pt?id=mdp.39015023119541&view=1up&seq=1363

Gee, Alastair, Julia Carrie Wong, Paul Lewis, Adithya Sambamurthy, Charlotte Simmonds, Nadieh Bremer, and Shirley Wu. "Bussed Out: How America Moves Its Homeless." *The Guardian*, December 20, 2017, https://www.theguardian.com/us-news/ng-interactive/2017/dec/20/bussed-out-america-moves-homeless-people-country-study

Google Finance. http://finance.google.com. Accessed January 2020.

Gramacki, Artur. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Cham, Switzerland: Springer International, 2018.

Groeger, Lena V. "A Big Article About Wee Things." ProPublica (blog), September 25, 2014, https://www.propublica.org/nerds/a-big-article-about-wee-things.

Guardian. "EU Referendum: Full Results and Analysis." *The Guardian*, https://www.theguardian.com/politics/ng-interactive/2016/jun/23/eu-referendum-live-results-and-analysis.

Guardian. "Full US 2012 Election County-Level Results to Download." https://www.theguardian.com/news/datablog/2012/nov/07/us-2012-election-county-results-download#data.

Hansen, Wallace R. Suggestions to Authors of the Reports of the United States Geological Survey, *US GPO, January 1991.*

Haroz, Steve, Robert Kosara, and Steven L. Franconeri. "Isotype Visualization: Working Memory, Performance, and Engagement with Pictographs." In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 1191–1200. ACM, 2015.

Hawkins, Ed. "147 | Iconic Climate Visuals with Ed Hawkins." Data Stories podcast, September 10, 2019, https://datastori.es/147-iconic-climate-visuals-with-ed-hawkins/.

Hawkins, Ed. ShowYourStripes.com. https://showyourstripes.info/

Healey, Christopher, and James Enns. "Attention and Visual Memory in Visualization and Computer Graphics." *IEEE transactions on visualization and computer graphics* 18, no. 7 (2011): 1170–1188.

Hearst, Marti, Emily Pedersen, Lekha Priya Patil, Elsie Lee, Paul Laskowski, and Steven Franconeri. "An Evaluation of Semantically Grouped Word Cloud Designs." *IEEE transactions on visualization and computer graphics* (2019).

Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky. "A Tour Through the Visualization Zoo." *Commun. Acm* 53, no. 6 (2010): 59–67.

Hermann, E.P. "Maps and Sales Visualization" *Personal Efficiency: The How and Why Magazine*, Published by the LaSalle Extension University, US Department of Education, July 1922, pages 6–7.

Hofman, Jake M., Daniel G. Goldstein, and Jessica Hullman. "How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Hofmann, Heike, and Marie Vendettuoli. "Common Angle Plots as Perception-True Visualizations of Categorical Associations." *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013): 2297–2305.

Horowitz, Juliana Mesace, Kim Parker, and Renee Stepler. "Wide Partisan Gaps in U.S. Over How Far the Country Has Come on Gender Equality." Pew Research Center, October 18, 2017, https://www.pewsocialtrends.org/2017/10/18/wide-partisan-gaps-in-u-s-over-how-far-the-country-has-come-on-gender-equality/.

Hullman, Jessica and Matthew Kay. "Uncertainty + Visualization, Explained." Medium, Visualization Research Explained (blog), https://medium.com/multiple-views-visualization-research-explained/uncertainty-visualization-explained-67e7a73f031b.

Hullman, Jessica, Eytan Adar, and Priti Shah. "Benefitting Infovis with Visual Difficulties." *IEEE Transactions on Visualization and Computer Graphics* 17, no. 12 (2011): 2213–2222.

Hullman, Jessica, Paul Resnick, and Eytan Adar. "Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences About Reliability of Variable Ordering." *PloS one* 10, no. 11 (2015).

Hullman, Jessica. "Leading with the Unknowns in COVID-19 Models." *Scientific American*, April 11, 2020. https://blogs.scientificamerican.com/observations/leading-with-the-unknowns-in-covid-19-models/.

Ingraham, Christopher. "Kansas Is the Nation's Porn Capital, According to Pornhub." Wonkviz, no date, https://wonkviz.tumblr.com/post/82488570278/kansas-is-the-nations-porn-capital-according-to

Institute for Health Metrics and Evaluation. "Causes of Death (COD) Visualization | Viz Hub." Accessed January 2020. https://vizhub.healthdata.org/cod/.

International Social Survey Programme (ISSP). "ISSP 2009 'Social Inequality IV'—ZA No. 5400." Accessed January 2020. https://www.gesis.org/issp/modules/issp-modules-by-topic/social-inequality/2009/

Irvin-Erickson, Yasemin, Jonathan Schwabish, and Nicole Weissman. "What We Know About Gun Violence in the United States: Who's Affected?" *Urban Wire Urban Institute*, October 4, 2016, https://www.urban.org/urban-wire/what-we-know-about-gun-violence-united-states-whos-affected.

Japanese: https://seibu.ekitan.com/pdf/20180310/235-16-1-0.pdf

Jarreau, Paige. "#MySciBlog Interviewee Motivations to Blog about Science." Figshare.com (blog), March 21, 2015, https://figshare.com/articles/_MySciBlog_Interviewee_Motivations_to_Blog_about_Science/1345026/2.

Johnson, Brian, and Ben Shneiderman. "Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures." *VIS '91: Proceedings of the 2nd conference on Visualization '91*. IEEE, 1991.

Johnson, Matt. "Which Metro Stations Are the Most Balanced?" Greater Greater Washington, November 29, 2012, https://ggwash.org/view/29468/which-metro-stations-are-the-most-balanced.

Joint Committee on Taxation. "A Distribution of Returns by the Size of the Tax Change for the 'Tax Cuts and Jobs Act,' As Ordered Reported by the Committee on Finance on November 16, 2017." November 27, 2017.

Joint Committee on Taxation. "Distributional Effects of Public Law 115–97." Scheduled for a Public Hearing Before the House Committee on Ways and Means on March 27, 2019, March 25, 2019, JCX-10-19.

Kaiser Family Foundation. "Status of State Action on the Medicaid Expansion Decision." Accessed January 2020. https://www.kff.org/health-reform/state-indicator/state-activity-around-expanding-medicaid-under-the-affordable-care-act/.

Kastellec, Jonathan P., and Eduardo L. Leoni. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5, no. 4 (2007): 755–771.

Katz, Josh. "Who Will Be President?" *New York Times*, November 8, 2016, https://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html

Kiefer, Len. Twitter post. September 26, 2019, 4:44 p.m., https://twitter.com/lenkiefer/status/1177323018143580160.

Kight, Stef W. and Harry Stevens. "By the Numbers: How Trump Properties Profited from His Presidency." *Axios*, June 28, 2018, https://www.axios.com/donald-trump-properties-taxpayer-campaigns-presidency-91e3755d-23cd-42d1-897d-c0c81ac509dd.html.

Kirk, Andy. "The Problems with B'Arc Charts." Visualisingdata.com, September 1, 2017, https://www.visualisingdata.com/2017/09/problems-barc-charts/.

Klein, Scott. "Infographics in the Time of Cholera." *Pro Publica*, March 16, 2016, https://www.propublica.org/nerds/infographics-in-the-time-of-cholera.

Koblin, Aaron. "Flight Patterns." http://www.aaronkoblin.com/work/flightpatterns/.

Koblin, Aaron. "Visualizing Ourselves . . . with Crowd-Sourced data." TED2011, March 2011, https://www.ted.com/talks/aaron_koblin_visualizing_ourselves_with_crowd_sourced_data.

Kommenda, Niko, Caelainn Barr, and Josh Holder. "Gender Pay Gap: What We Learned and How to Fix It." *The Guardian*, April , 2018, https://www.theguardian.com/news/ng-interactive/2018/apr/05/women-are-paid-less-than-men-heres-how-to-fix-it.

Krzywinski, Martin, Inanc Birol, Steven JM Jones, and Marco A. Marra. "Hive plots—rational approach to visualizing networks." *Briefings in bioinformatics* 13, no. 5 (2011): 627–644.

Krzywinski, Martin. "Hive plots—rational approach to visualizing networks." http://www.hiveplot.com/.

Lekovic, Jovan. "How We Made the BBC Audiences Tableau Style Guide." Medium, August 29, 2018, https://medium.com/bbc-data-science/how-we-made-the-bbc-audiences-tableau-style-guide-4f0a6b7525ce.

Likert, Rensis. "A Technique for the Measurement of Attitudes." *Archives of psychology* (1932).

Lima, Manuel. "Why Humans Love Pie Charts." Medium (blog). July 23, 2018. https://blog.usejournal.com/why-humans-love-pie-charts-9cd346000bdc?.

Lima, Manuel. *The Book of Circles: Visualizing Spheres of Knowledge*. New York: Princeton Architectural Press, 2017.

Liu, Yang, and Jeffrey Heer. "Somewhere Over the Rainbow: An Empirical Assessment of Quantitative Colormaps." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. 2018.

Luxembourg Income Study. https://www.lisdatacenter.org/.

Malkulec, Amanda. "Leading with the Unknowns in COVID-19 Models." Medium, March 11, 2020, https://medium.com/nightingale/ten-considerations-before-you-create-another-chart-about-covid-19-27d3bd691be8.

Manski, Charles F., and Francesca Molinari. "Rounding Probabilistic Expectations in Surveys." *Journal of Business & Economic Statistics* 28, no. 2 (2010): 219–231.

Matejka, Justin, and George Fitzmaurice. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1290–1294. ACM, 2017.

Mayr, Eva and Günther Schreder. "Isotype Visualizations: A Chance for Participation and Civic Education." *Journal of Democracy* 6(2): 136–150, 2014.

McCann, Adam. "Arc Chart in Tableau." August 27, 2015, http://duelingdata.blogspot.com/2015/08/arc-chart-in-tableau.html.

McKee, Robert. *Substance, Structure, Style, and the Principles of Screenwriting.* New York: HarperCollins, 1997.

Medina, John. *Brain Rules: 12 Principles for Surviving and Thriving at Work, Home, and School. ReadHowYouWant.com, 2011.*

Meeks, Elijah. "Exploratory Design in Data Visualization: Understanding and leveraging chart similarity." *Medium*, January 15, 2019, https://medium.com/@Elijah_Meeks/exploratory-design-in-data-visualization-87bc60ce7f04.

Meeks, Elijah. "Sketchy Data Visualization in Semiotic." Medium (blog), September 11, 2017, https://medium.com/@Elijah_Meeks/sketchy-data-visualization-in-semiotic-5811a52f59bc.

Meyer, Bruce D., Nikolas Mittag, and Robert M. Goerge. *Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation*. No. w25143. National Bureau of Economic Research, 2018.

Meyer, Bruce D., Wallace KC Mok, and James X. Sullivan. "Household Surveys in Crisis." *Journal of Economic Perspectives* 29, no. 4 (2015): 199–226.

Miller, Matthew, Steven J. Lippmann, Deborah Azrael, and David Hemenway. "Household firearm ownership and rates of suicide across the 50 United States." *Journal of Trauma and Acute Care Surgery* 62, no. 4 (2007): 1029–1035.

Minard, Charles-Joseph. "*Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813*." 1861, *Photo courtesy of Ecole nationale des ponts et chaussées*

Monmonier, Mark. *How to Lie with Maps*. Chicago: University of Chicago Press, 2018.

Munzner, Tamara. *Visualization Analysis and Design*. AK Peters, 2014.

National Conference of State Legislatures. "State Family and Medical Leave Laws." July 2016. http://www.ncsl.org/research/labor-and-employment/state-family-and-medical-leave-laws.aspx

National Conference of State Legislatures. "State Minimum Wages | 2020 Minimum Wage by State." January 2020. http://www.ncsl.org/research/labor-and-employment/state-minimum-wage-chart.aspx#Table.

National Highway Traffic Safety Administration. Fatality Analysis Reporting System (FARS). https://one.nhtsa.gov/Data/Fatality-Analysis-Reporting-System-(FARS).

National Institute on Drug Abuse. "Overdose Death Rates." January 2019. https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates.

NBC News. "Presidential Election Results." http://elections.nbcnews.com/ns/politics/2012/all/president/#.XiZR4lNKhTZ.

Nediger, Midori. "How to Use an Icon Story to Take Your Infographic to the Next Level." Venngage (blog), April 27, 2018, https://venngage.com/blog/icon-story/.

Neurath, O. *International Picture Language. The First Rules of Isotype*. London: Kegan Paul, 1936.

Neurath, O. *Modern Man in the Making*. New York: Knopf, 1939.

New York Magazine. "Approval Matrix." provided via email.

New York Times. Paths to the White House.

New York Times. Twitter post. August 7, 2016, 10:52 p.m. https://twitter.com/nytgraphics/status/762481520565030919?ref_src=twsrc%5Etfw

Newman, George E., and Brian J. Scholl. "Bar Graphs Depicting Averages are Perceptually Misinterpreted: the Within-the-Bar Bias." *Psychonomic Bulletin and Review* 19, no. 4 (2012): 601–607.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: New York University Press. 2018.

O'Brady, Kevin. "Fight to Repeal." Joint Economic Committee Republicans, 2010, https://kevinbrady.house.gov/obamacare/fighting-to-repeal-obamacare.htm.

O'Neill, Catherine. *Weapons of Math Destruction*: *How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway, 2016.

Obama, Barack. "Remarks of President Barack Obama—State of the Union Address As Delivered." State of the Union Address, Washington, DC, January 13, 2016, https://obamawhitehouse.archives.gov/the-press-office/2016/01/12/remarks-president-barack-obama-%E2%80%93-prepared-delivery-state-union-address.

Olson, J. M. "Noncontiguous Area Cartograms." *The Professional Geographer*, 28, no. 4 (1976): 371–380.

Ondov, Brian, Nicole Jardine, Niklas Elmqvist, and Steven Franconeri. "Face to Face: Evaluating Visual Comparison." *IEEE Transactions on Visualization and Computer Graphics* 25, no. 1 (2018): 861–871.

Organization for Economic Cooperation and Development (OECD). "Better Life Index." Accessed December 2019. http://www.oecdbetterlifeindex.org/.

Organization for Economic Cooperation and Development (OECD). Program for International Student Assessment (PISA), 2015 Reading, Mathematics and Science Assessment.

Organization for Economic Cooperation and Development (OECD). Social Expenditure Database. Accessed January 2019. https://www.oecd.org/social/expenditure.htm.

Padilla, Lace, Matthew Kay, and Jessica Hullman. "Uncertainty Visualization." PsyArXiv. April 27, 2020. doi:10.31234/osf.io/ebd6r.

Pätzold, André, Julius Tröger, Joachim Fahrun, Christopher Möller, David Wendler and Marie-Louise Timcke. "These Candidates Live the Furthest Away from Their Voters." *Berliner Morgenpost*, September 8, 2016, https://interaktiv.morgenpost.de/waehlernaehe-berlin/.

Perez, Caroline Criado. *Invisible Women: Exposing Data Bias in a World Designed for Men*. New York: Random House, 2019.

Playfair, William. *Playfair's Commercial and Political Atlas and Statistical Breviary*. Cambridge: Cambridge University Press, 2005.

Porter, Mark and Rhys Blakely. 2020. "Coronavirus: How Safe Are You? All Your Health Questions Answered." *Australian*, March 31, 2020, https://www.theaustralian.com.au/world/the-times/coronavirus-how-safe-are-you-all-your-health-questions-answered/news-story/c40ec5841763c6d57f986c3ea3e19e4f

Presentitude. "A Guide to 44 Safe Fonts & 74 Safe Combos for PowerPoint." Presentitude (blog) no date, access January 2020, http://presentitude.com/fonts/.

Priestley, Joseph. "A Specimen of a Chart of Biography." 1765.

Project Tycho. https://www.tycho.pitt.edu/data.

Rendgen, Sandra. *The Minard System: The Complete Statistical Graphics of Charles-Joseph Minard*. New York: Princeton Architectural Press, 2018.

R Graph Gallery. "Network Graph." https://www.r-graph-gallery.com/network.html.

Richards, Neil. "Is White Space Always Your Friend?" blog post, January 6, 2018, https://questionsindataviz.com/2018/01/06/is-white-space-always-your-friend/.

Robbins, Naomi B. *Creating More Effective Graphs*. New York: Wiley, 2012.

Rogers, Simon. "Jon Snow's Data Journalism: The Cholera Map That Changed the World." *Guardian Datablog*, March 15, 2013, https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map.

Rosenberg, Daniel. "Joseph Priestley and the Graphic Invention of Modern Time." *Studies in Eighteenth-Century Culture* 36, no. 1 (2007): 55–103.

Roser, Max and Esteban Ortiz-Ospina. "Global Education." Our World in Data. Accessed January 2020. https://ourworldindata.org/global-education. 2020.

Roser, Max and Esteban Ortiz-Ospina. "World Development Indicators." Our World in Data. Accessed January 2020. https://ourworldindata.org/global-education.

Roser, Max, Hannah Ritchie, and Esteban Ortiz-Ospina. "Coronavirus Disease (COVID-19)—Statistics and Research." Our World in Data, https://ourworldindata.org/coronavirus#the-growth-rate-of-covid-19-deaths, accessed April 2020.

Rosling, Ted. "The Best Stats You've Ever Seen." TED2006, TED Talk, February 2006, https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen.

Rost, Lisa Charlotte. "An Alternative to Pink and Blue: Colors for Gender Data." Chartable, a blog by Datawrapper, July 10, 2018, https://blog.datawrapper.de/gendercolor/.

Rost, Lisa Charlotte. "How to Read a Log Scale."Chartable, (a blog by Datawrapper). May 31, 2018. https://blog.datawrapper.de/weeklychart-logscale/

Rost, Lisa Charlotte. "What I Learned Recreating One Chart Using 24 Tools." Source (blog), December 8, 2016, https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/.

Roth, Madeline and Amma Malik. "What Keeps Cities in Asia and Africa from Effective Public Service Delivery?" Urban Wire blog, Urban Institute, August 2, 2016, https://www.urban.org/urban-wire/what-keeps-cities-asia-and-africa-effective-public-service-delivery.

Rothstein, Richard. *The Color of Law: A Forgotten History of How Our Government Segregated America*. New York: Liveright, 2017.

Rousseeuw, Peter J., Ida Ruts, and John W. Tukey. "The Bagplot: a Bivariate Boxplot." *American Statistician* 53, no. 4 (1999): 382–387.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 9.0 [American Community Survey 2017]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D010.V9.0

Sankey, M.H. "Minutes of Proceedings of The Institution of Civil Engineers." Vol. CXXXIV, Session 1897–98. Part IV.

Schleuss, Jon and Rong-Gong Lin II. "California Crime 2013." *Los Angeles Times*, https://graphics.latimes.com/california-crime-2013/

Schuetz, Jenny. "Who Is the New Face of American Homeownership?" Brookings Institution, October 9, 2017, https://www.brookings.edu/blog/the-avenue/2017/10/09/who-is-the-new-face-of-american-homeownership/.

Schwabish, Jonathan A. "Take a Penny, Leave a Penny: The Propensity to Round Earnings in Survey Data." *Journal of Economic and Social Measurement* 32, no. 2–3 (2007): 93–111.

Schwabish, Jonathan, and Alice Feng. "Applying Racial Equity Awareness in Data Visualization." (2020). http://osf.io/x8tbw/

Schwabish, Jonathan. "How a Map's 'Bins' Can Influence Your Perception of an Important Policy Issue." *Urban Wire Urban Institute*, November 14, 2017, https://www.urban.org/urban-wire/how-maps-bins-can-influence-your-perception-important-policy-issue.

Shneiderman, Ben. "A Grander Goal: A Thousand-Fold Increase in Human Capabilities." Educom Review, 32, 6, Nov/Dec 1997, http://www.ifp.illinois.edu/nabhcs/abstracts/shneiderman.html

Shneiderman, Ben. "Tree Visualization with Tree-Maps: 2-d Space-Filling Approach." *ACM Transactions on Graphics (TOG)* 11, no. 1 (1992): 92–99.

Skau, Drew, and Robert Kosara. "Arcs, angles, or areas: Individual data encodings in pie and donut charts." *Computer Graphics Forum* 35, no. 3. 2016.

Skau, Drew, and Robert Kosara. "Judgment Error in Pie Chart Variations." In *Short Paper Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization (EuroVis)*, 91–95. 2016.

Slobin, Sarah. "Two Maps Explain the Racial History Behind Alabama's Senate Vote." Quartz, December 13, 2017, https://qz.com/1155837/how-doug-jones-beat-roy-moore-maps-of-alabama-in-1860-and-2017-offer-a-startling-comparison/.

Social Security Administration. "The 2019 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds." 2019, https://www.ssa.gov/OACT/TR/2019/.

Social Security Advisory Board, Aspects of Disability Decision Making: Data and Materials, February 2012. https://www.ssab.gov/Portals/0/OUR_WORK/REPORTS/Chartbook_Aspects%20of%20Disability%20Decision%20Making_2012.pdf

Social Security Advisory Board. "2011 Technical Panel Report on Assumptions and Methods." 2011, https://www.ssab.gov/Portals/0/OUR_WORK/REPORTS%20TO%20THE%20BOARD/TPAM_Report_2011.pdf.

Soffen, Kim. "To Reduce Suicides, Look at Guns." *Washington Post*, July 13, 2016, https://www.washingtonpost.com/graphics/business/wonkblog/suicide-rates/?tid=a_inl

Spear, Mary Eleanor. *Charting Statistics*. New York: McGraw-Hill, 1952.

Spence, Ian, and Howard Wainer. "William Playfair: A Daring Worthless Fellow." *Chance* 10, no. 1 (1997): 31–34.

St. Clair, Kassia. *The Secret Lives of Color*. New York: Penguin, 2017.

Strunk Jr, William, and E. B. White. *The Elements of Style*. New York: Macmillan, 1959.

Szűcs, Krisztina. "Spotlight on Profitability." http://krisztinaszucs.com/my-product/hollywood/

Texas Tribune News Apps Style Guide. https://apps.texastribune.org/styles/

Texas Tribune. "What the Texas delegation says about impeachment." *Texas Tribune*, accessed September 26, 2019, https://graphics.texastribune.org/graphics/impeachment-delegation-tracker-2019-09/.

Torres, Nicole. "Why It's So hard for Us to Visualize Uncertainty." *Harvard Business Review*, November 11, 2016, https://hbr.org/2016/11/why-its-so-hard-for-us-to-visualize-uncertainty.

Treisman, Anne. "Preattentive Processing in Vision." *Computer Vision, Graphics, and Image Processing* 31, no. 2 (1985): 156–177.

Tse, Archie. "Why We Are Doing Fewer Interactives." presentation at the 2016 Malofiej Conference, https://github.com/archietse/malofiej-2016/blob/master/tse-malofiej-2016-slides.pdf

Tufte, Edward R. *Beautiful Evidence*, vol. 1. Cheshire, CT: Graphics Press, 2006.

Tukey, John W. *Exploratory Data Analysis*, vol. 2. Reading, MA: Addison-Wesley, 1977.

Turner, Cory. "Is There A Better Way To Pay For America's Schools?" *NPR*, May 1, 2016, https://www.npr.org/2016/05/01/476224759/is-there-a-better-way-to-pay-for-americas-schools.

Uncertainty Toolkit for Analysts in Government. Website. Accessed February 2020. https://analysts uncertaintytoolkit.github.io/UncertaintyWeb/index.html.

United Nations. "Population by Age, Sex, and Urban/Rural Residence." http://data.un.org/Data.aspx?d=POP&f=tableCode%3A22.

Urban Institute. "Urban Institute Data Visualization Style Guide." No date, http://urbaninstitute.github.io/graphics-styleguide/, accessed January 2020.

US Bureau of Economic Analysis. "Gross Domestic Product by Industry and Input-Output Statistics." January 22, 2015, https://apps.bea.gov/histdata/fileStructDisplay.cfm?HMI=8&DY=2014&DQ=Q3&DV=Quarter&dNRD=January-22-2015.

US Bureau of the Census. "Map Showing the Distribution of the Slave Population of the United States." https://www.census.gov/history/pdf/1860_slave_distribution.pdf

US Department of Agriculture, Economic Research Service. "Midpoint acreage more than doubled for all five major field crops." May 10, 2018, https://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=88868.

US Department of Agriculture. "Food Distribution Program on Indian Reservations." https://www.fns.usda.gov/fdpir/food-distribution-program-indian-reservations.

US Department of Agriculture. "Visual Standards Guide." January 2013, https://www.usda.gov/sites/default/files/documents/visual-standards-guide-january-2013.pdf.

US Department of Education. "School Composition and the Black-White Achievement Gap." National Center for Education Statistics, June 2015, https://nces.ed.gov/nationsreportcard/subject/studies/pdf/school_composition_and_the_bw_achievement_gap_2015.pdf.

USA Today. "All of Obama's State of the Union Speeches in Word Clouds." *USA Today*, January 12, 2016, https://www.usatoday.com/story/news/politics/onpolitics/2016/01/12/obama-state-of-the-union-word-clouds/78712780/.

van der Bles, Anne Marthe, Sander van der Linden, Alexandra LJ Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval, and David J. Spiegelhalter. "Communicating Uncertainty About Facts, Numbers and Science." *Royal Society open science* 6, no. 5 (2019): 181870.

van Gogh, Vincent. Vincent van Gogh to Theo van Gogh, Amsterdam, October 28, 1885, http://www.vangoghletters.org/vg/letters/let537/letter.html

Vessey, Iris. "Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature." *Decision Sciences* 22, no. 2 (1991): 219–240.

Vigen, Tyler. Spurious Correlations. Website. http://www.tylervigen.com/spurious-correlations.

Wagemans, Johan, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. "A Century of Gestalt Psychology in Visual Perception: I. Perceptual grouping and figure–ground organization." *Psychological Bulletin* 138, no. 6 (2012): 1172.

Wagemans, Johan, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R. Pomerantz, Peter A. Van der Helm, and Cees Van Leeuwen. "A Century of Gestalt Psychology in Visual Perception: II. Conceptual and Theoretical Foundations." *Psychological Bulletin* 138, no. 6 (2012): 1218.

Wainer, Howard, and Ian Spence. *Playfair's Commercial and Political Atlas and Statistical Breviary.* Cambridge: Cambridge University Pres, 2005.

Walker, Mason. "Americans Favor Mobile Devices Over Desktops and Laptops for Getting News." Pew Research Center, November 19, 2019, https://www.pewresearch.org/fact-tank/2019/11/19/americans-favor-mobile-devices-over-desktops-and-laptops-for-getting-news/.

Ware, Colin. *Information Visualization: Perception for Design*, 3rd ed. Waltham, MA: Morgan Kaufmann, 2020.

Washington Metropolitan Area Transit Authority. Public Access to Records Policy request. October 2019.

Watson, Cecelia. *Semicolon: How a Misunderstood Punctuation mark Can Improve Your writing, Enrich Your Reading and Even Change Your Life*. HarperCollins UK, 2019.

Wattenberg, Martin, and Fernanda B. Viégas. "The Word Tree, an Interactive Visual Concordance." *IEEE Transactions on Visualization and Computer Graphics* 14, no. 6 (2008): 1221–1228.

Wattenberg, Martin. "Arc Diagrams: Visualizing Structure in Strings." In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, pp. 110–116. IEEE, 2002.

Web Content Accessibility Guidelines (WCAG) 2.0. https://www.w3.org/TR/WCAG20/#understandable, accessed November 2019.

Weissgerber, Tracey L., Natasa M. Milic, Stacey J. Winham, and Vesna D. Garovic. "Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm." *PLoS biology* 13, no. 4 (2015): e1002128.

Wellcome Collection. "Diagram of the Causes of Mortality in the Army." https://wellcomecollection.org/works/sz9sms2m

Wexler, Steve, Jeffrey Shaffer, and Andy Cotgreave. *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*. New York: Wiley, 2017.

The White House. "President Obama's Final State of the Union." February 12, 2016. https://obamawhitehouse.archives.gov/sotu.

Wickham, Hadley, and Lisa Stryjewski. "40 Years of Boxplots." *American Statistician* (2011).

Wikipedia. "Right-to-Work Law." Accessed January 2020. https://en.wikipedia.org/wiki/Right-to-work_law.

Wikipedia. "State Income Tax." Accessed January 2020. https://en.wikipedia.org/wiki/State_income_tax.

Wilke, Claus O. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. Sebastopol, CA: O'Reilly, 2019.

Wilkinson, Krista M., and William J. McIlvane. "Perceptual Factors Influence Visual Search for Meaningful Symbols in Individuals with Intellectual Disabilities and Down Syndrome or Autism Spectrum Disorders." *American Journal on Intellectual and Developmental Disabilities* 118, no. 5 (2013): 353–364.

Wilkinson, Leland. "Dot Plots." *American Statistician* 53, no. 3 (1999): 276–281.

Wilkinson, Leland. "The Grammar of Graphics." In *Handbook of Computational Statistics*, 375–414. Berlin: Springer, 2012.

Wilkinson, Leland. *The Grammar of Graphics*. Chicago: Springer Science & Business Media, 2013.

Witt, Jessica. "Graph Construction: An Empirical Investigation on Setting the Range of the Y-Axis." http://amplab.colostate.edu/reprints/Witt_Graphs_YaxisRange.pdf.

Wood, Jo, Petra Isenberg, Tobias Isenberg, Jason Dykes, Nadia Boukhelifa, and Aidan Slingsby. "Sketchy rendering for information visualization." *IEEE Transactions on Visualization and Computer Graphics* 18, no. 12 (2012): 2749–2758.

World Bank. "World Development Indicators." The World Bank Group. Accessed January 2020. https://datacatalog.worldbank.org/dataset/world-development-indicators.

XKCD. "Correlation." https://xkcd.com/552/.

Yau, Nathan. "Vehicles Involved in Fatal Crashes." FlowingData blog, January 11, 2012, https://flowingdata.com/2012/01/11/vehicles-involved-in-fatal-crashes/.

INDEX