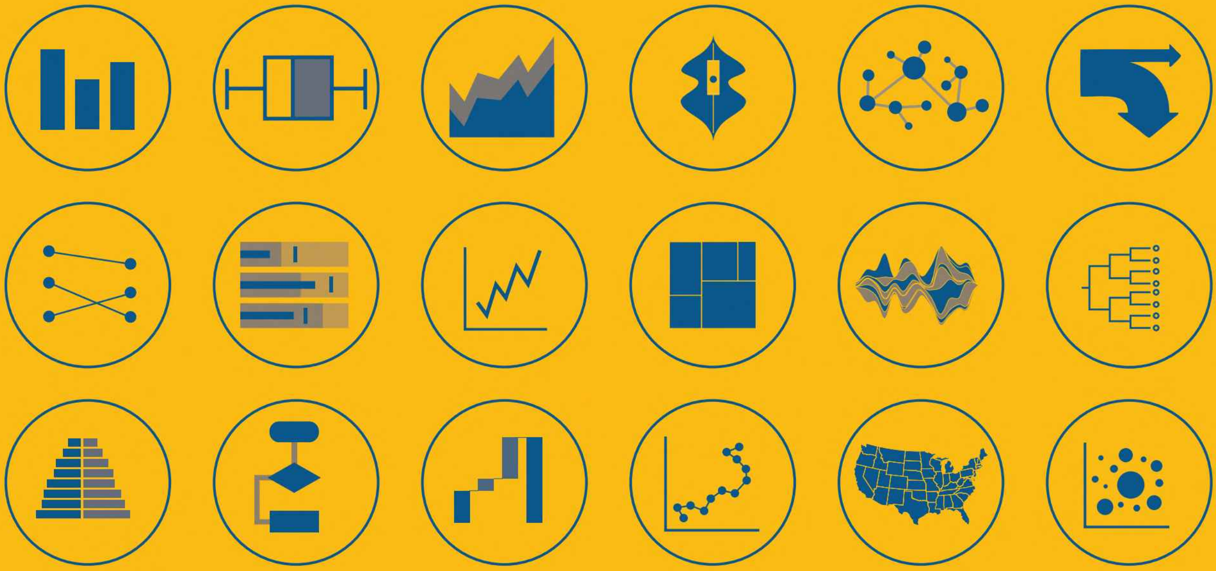
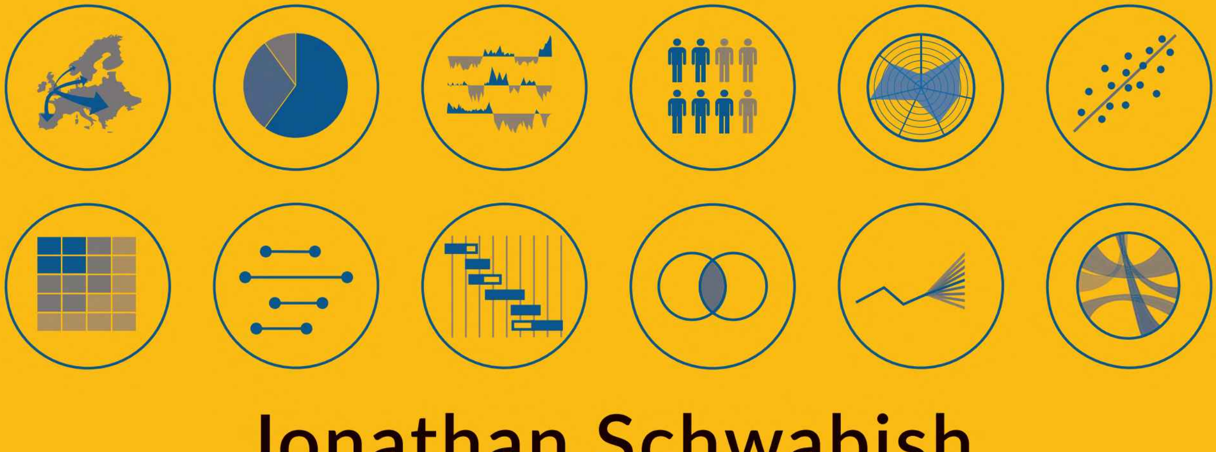


BETTER DATA VISUALIZATIONS

A Guide for Scholars, Researchers, and Wonks



Jonathan Schwabish

BETTER DATA VISUALIZATIONS

2

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.

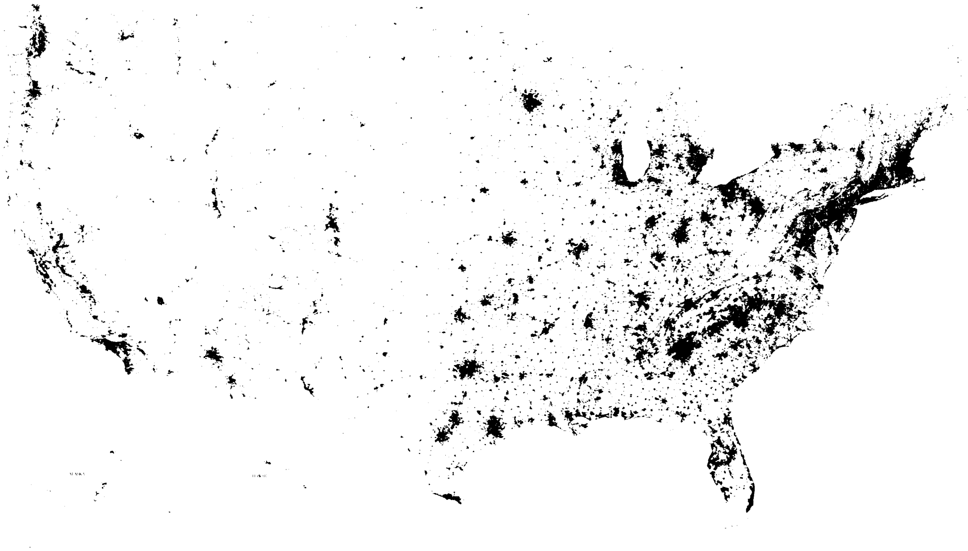
Taken together, these five guidelines remind me of the needs of my audience and how my visuals can tell them a story.

GUIDELINE 1: SHOW THE DATA

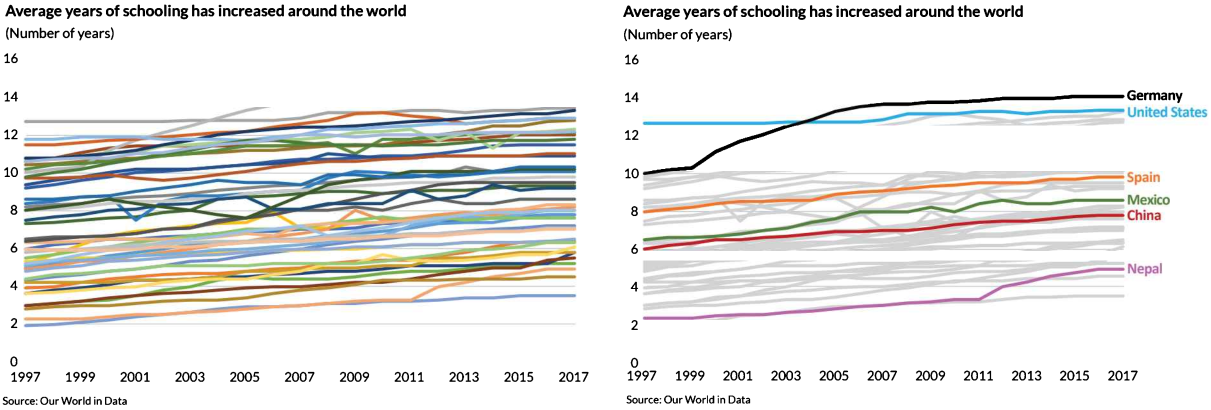
Your reader can only grasp your point, argument, or story if they see the data. This doesn't mean that *all* the data must be shown, but it does mean that you should highlight the values that are important to your argument. As chart creators, our challenge is deciding how much data to show and the best 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).



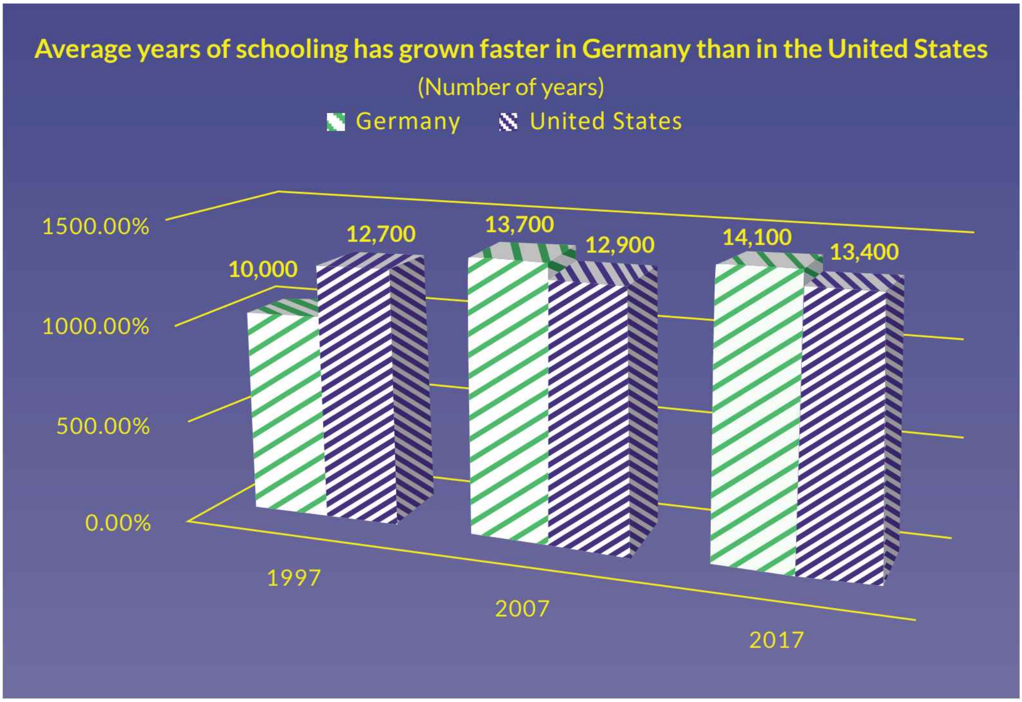
Highlighting just a few countries in the chart on the right makes it easier to read.

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.

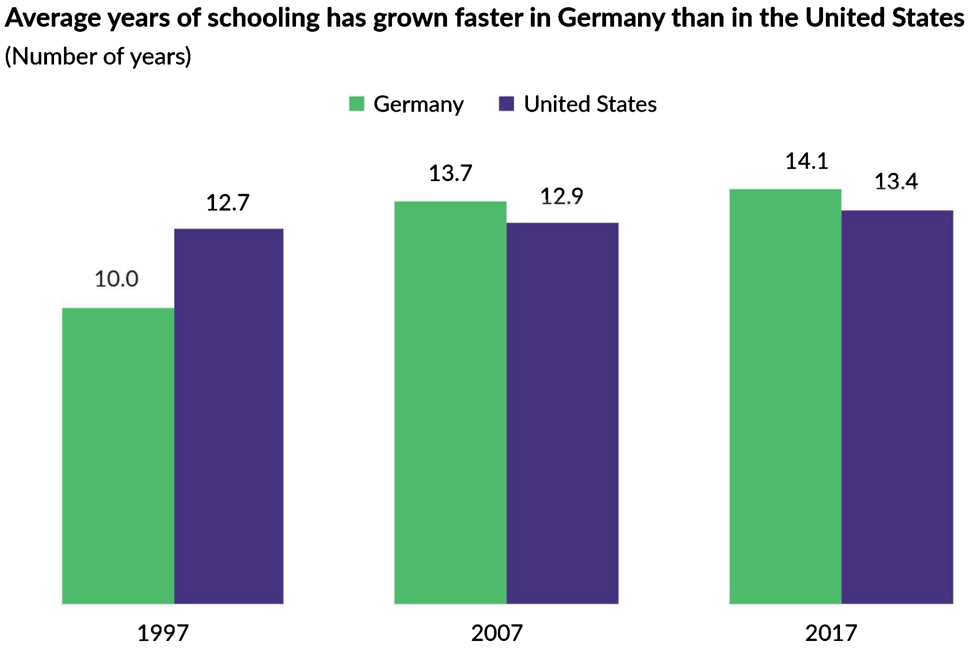
Take this three-dimensional column chart of average schooling for the United States and Germany for a few select years.



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

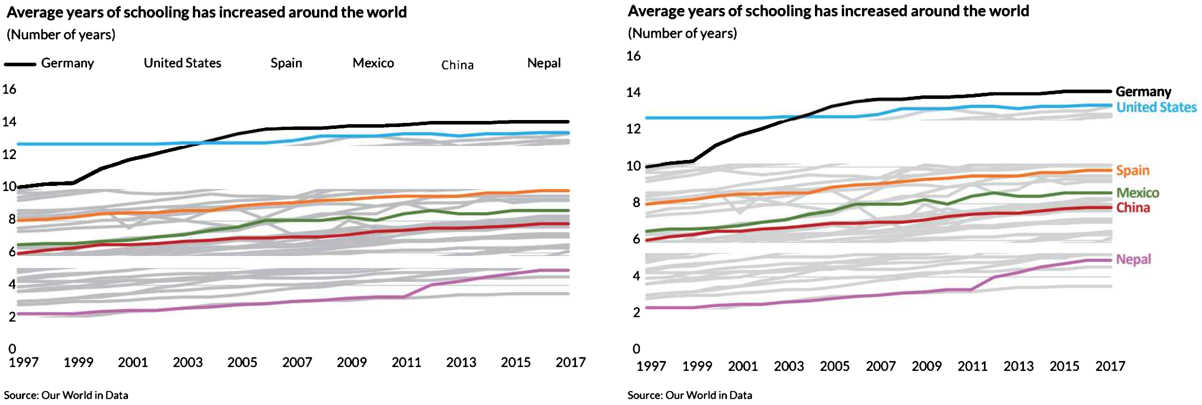
Although our primary focus on creating a visualization is the graphic elements—bars, points, or lines—the text we include in and around our graphs is just as important. Far too often, we treat the text and annotations as an afterthought, but these elements can be used to explain how to read the content 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

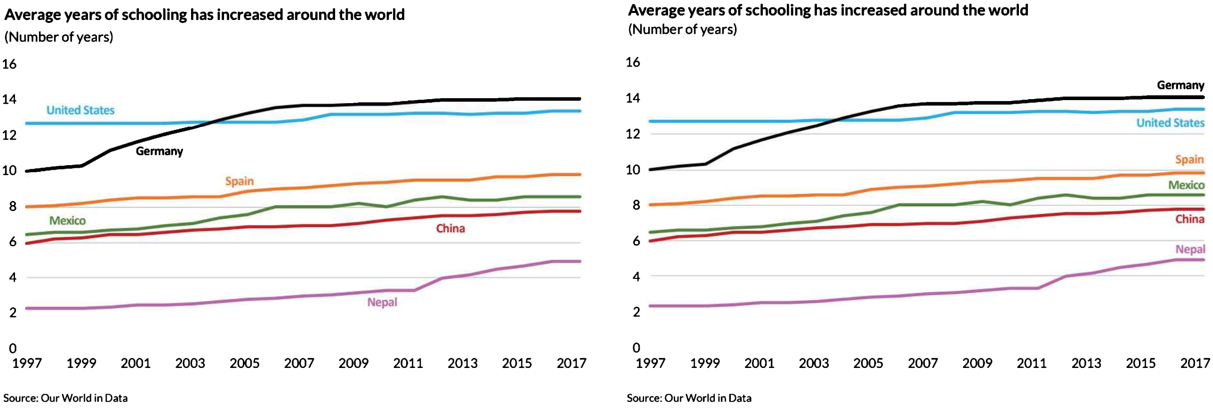
Let's start with the easiest type of annotation: Removing legends and directly labeling your data. Many software tool defaults create a data legend and place it around the chart, disconnected from the data. This forces more work upon your reader to connect each line or bar to its label. A better approach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

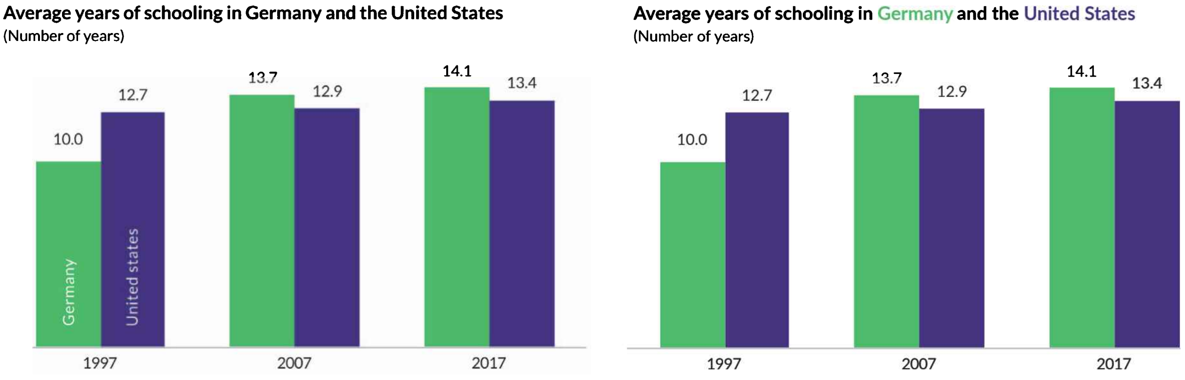
In graphs that have fewer lines, we might also be able to place the labels directly on the graph. In these cases, I try to align the labels instead of placing them in random positions. Notice how in the graph on the left of the next page your eye needs to jump around to find each label. And because we 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?

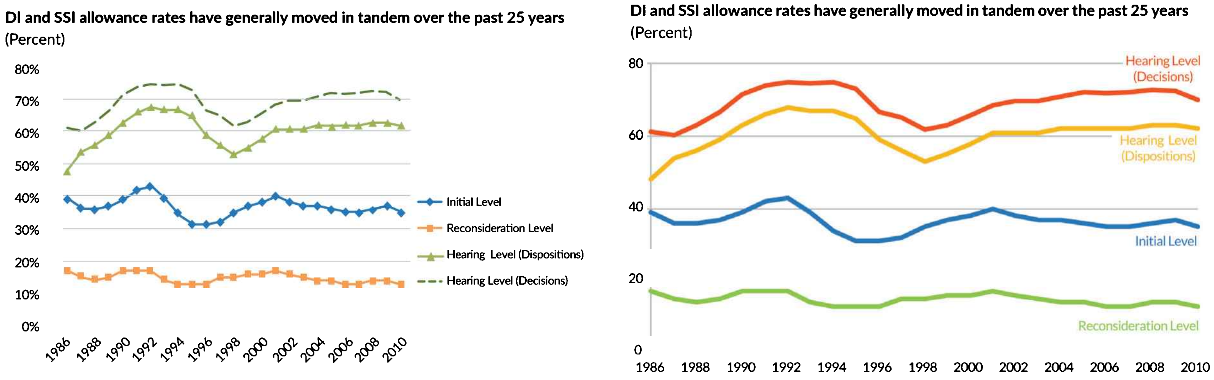
By integrating the text and the data, we're doing a better job of considering the reader's needs. Do they need to see *every* line equally, or will including all the lines clutter the graph? Is it important to label *every* point in the scatterplot, or will highlighting just a few points suffice? How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



DI and SSI allowance rates have generally moved in tandem over the past 25 years

(Percent)

DI and SSI allowance rates have generally moved in tandem over the past 25 years

(Percent)

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.

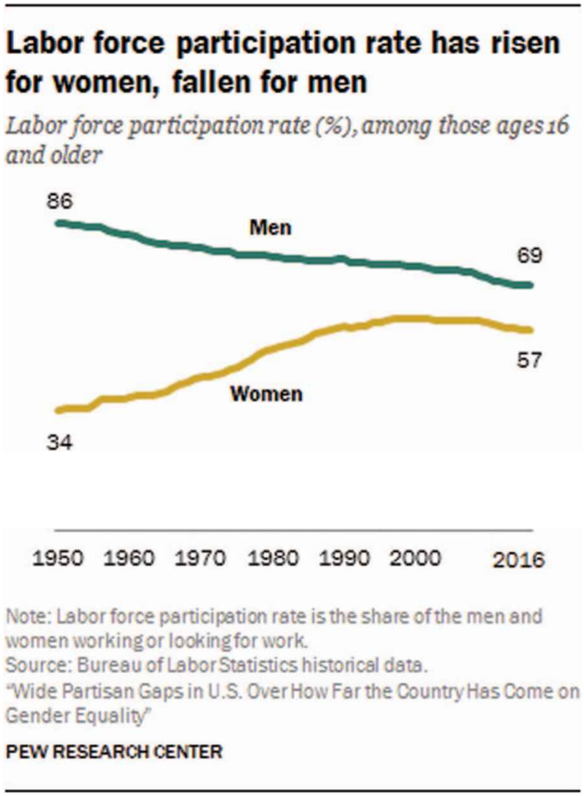
2. WRITE THE TITLE LIKE A NEWSPAPER HEADLINE

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,


Labor force participation rate has risen for women, fallen for men

Labor force participation rate (%), among those ages 16 and older

Note: Labor force participation rate is the share of the men and women working or looking for work.
Source: Bureau of Labor Statistics historical data.
"Wide Partisan Gaps in U.S. Over How Far the Country Has Come on Gender Equality"

PEW RESEARCH CENTER

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.

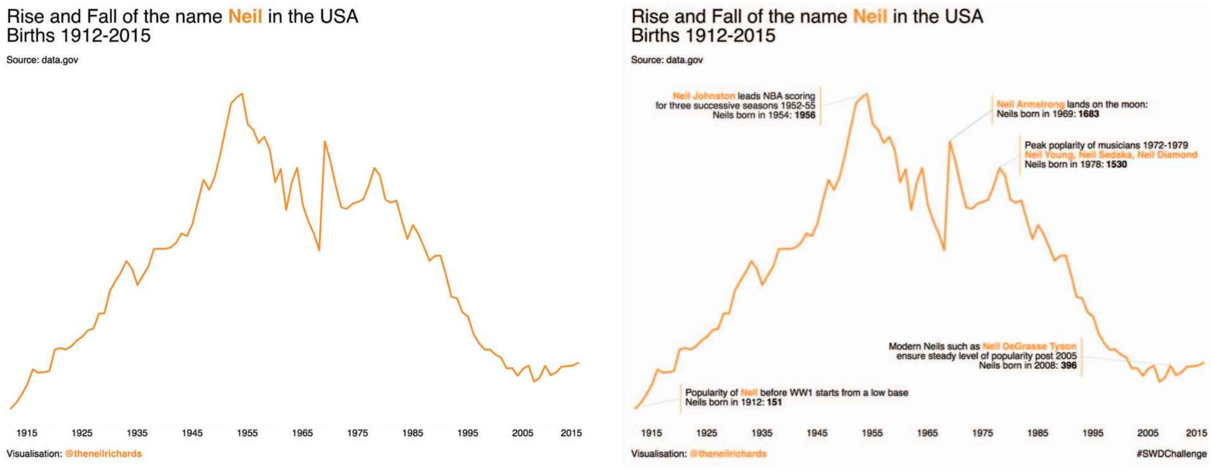
If you are having trouble coming up with a concise, active title, that may be a sign that your chart doesn't have a concise takeaway or—and maybe this is more common—you haven't thought through what you want the graph to communicate.

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?

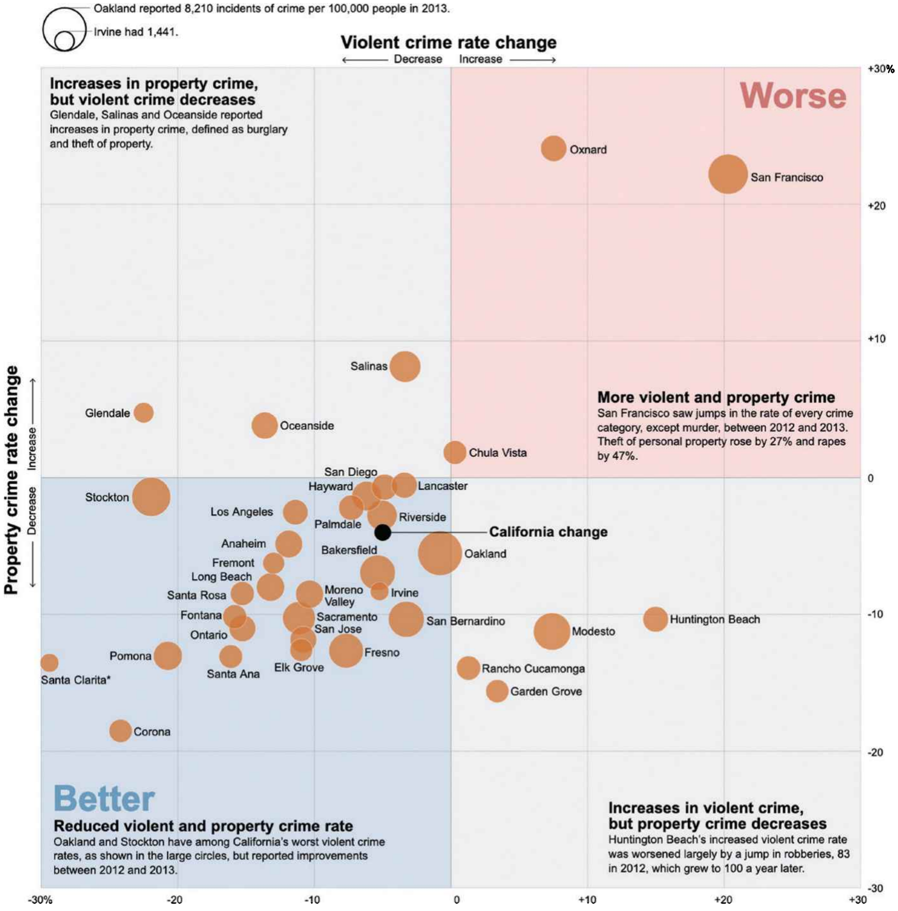


Short explainers in this graph on the right from Neil Richards explain some of the basic features of the data.

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.



Oakland reported 8,210 incidents of crime per 100,000 people in 2013.

Irvine had 1,441.

Increases in property crime, but violent crime decreases
Glendale, Salinas and Oceanside reported increases in property crime, defined as burglary and theft of property.

More violent and property crime
San Francisco saw jumps in the rate of every crime category, except murder, between 2012 and 2013. Theft of personal property rose by 27% and rapes by 47%.

Reduced violent and property crime rate
Oakland and Stockton have among California's worst violent crime rates, as shown in the large circles, but reported improvements between 2012 and 2013.

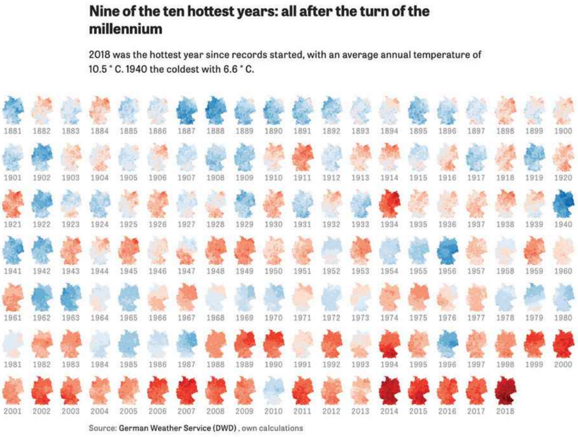
Increases in violent crime, but property crime decreases
Huntington Beach's increased violent crime rate was worsened largely by a jump in robberies, 83 in 2012, which grew to 100 a year later.

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.

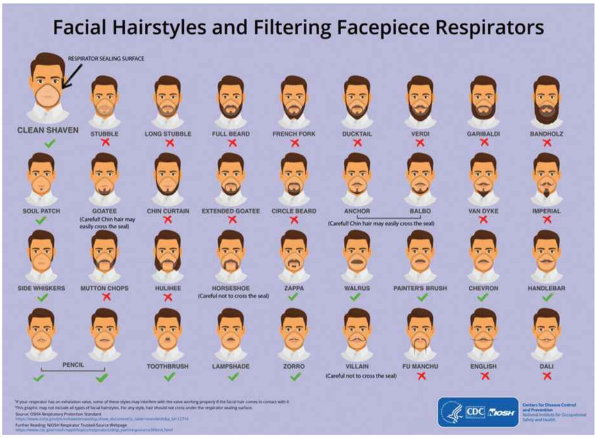
GUIDELINE 4. AVOID THE SPAGHETTI CHART

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.


Nine of the ten hottest years: all after the turn of the millennium

2018 was the hottest year since records started, with an average annual temperature of 10.5 ° C. 1940 the coldest with 6.6 ° C.

Source: German Weather Service (DWD) , own calculations


Facial Hairstyles and Filtering Facepiece Respirators

*If your respirator has an exhalation valve, some of these styles may interfere with the valve working properly if the facial hair comes in contact with it.
*This graphic may not include all types of facial hairstyles. For any style, hair should not cross under the respirator sealing surface.
Source: OSHA Respiratory Protection Standard
Further Reading: NIOSH Respirator Trusted-Source Webpage

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.

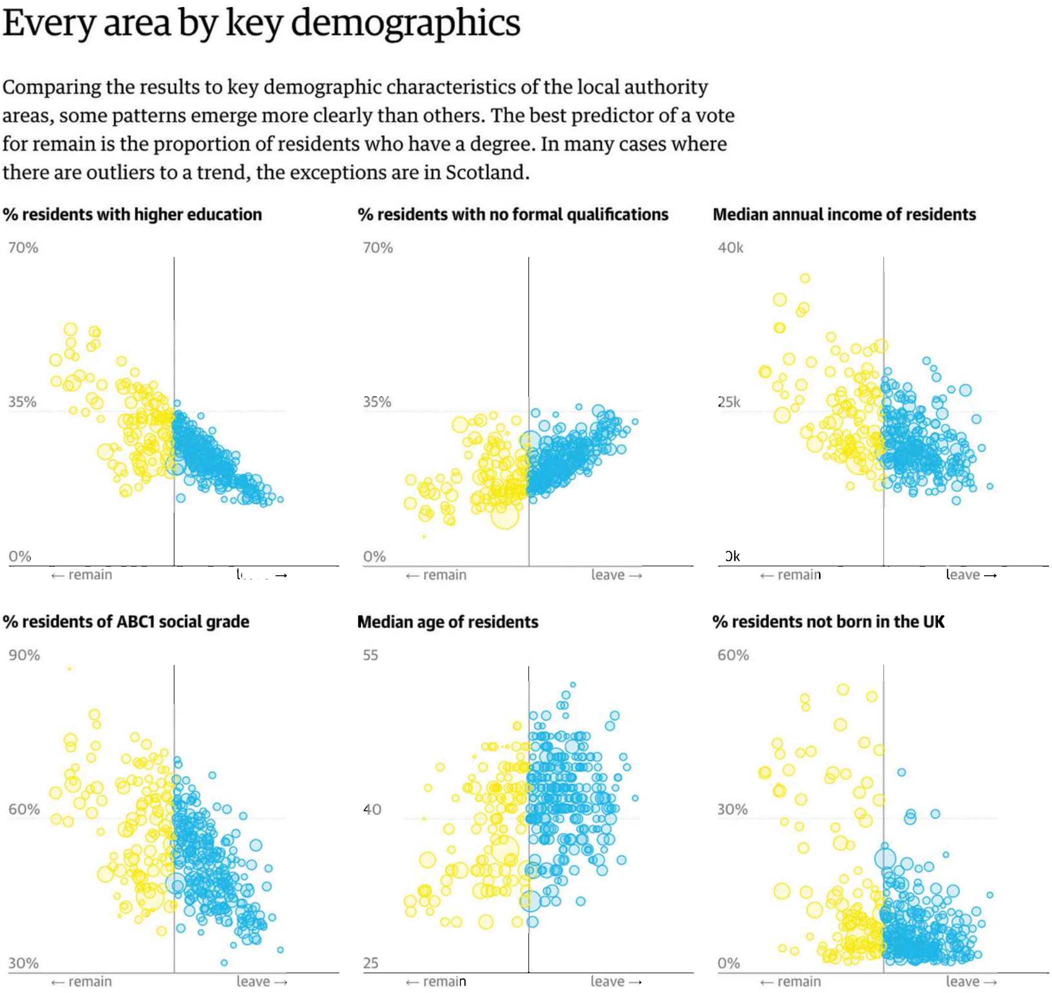
One way to address the packed, single chart is to break it into smaller parts. Known as grid charts or panel charts (also called *facets, trellis charts*, or, most commonly, *small multiples*) these are smaller charts that use the same scale, axes, and scope but spread the data across multiple visuals. 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.

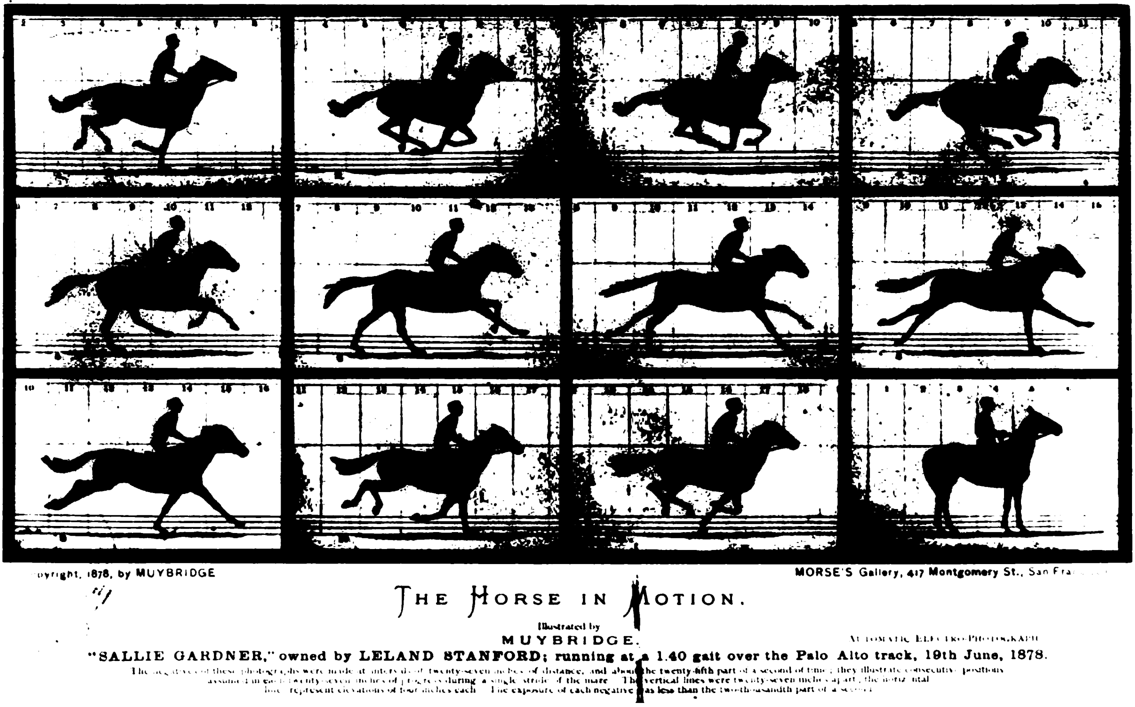
Second, the graphs should share the same layout, size, font, and color. Remember, we are breaking up one chart into many, so it should look like one chart replicated multiple times. The vertical and horizontal axes may change, but you wouldn't want to have one chart where blue dots represent "no" and 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.


opyright, 1878, by MUYBRIDGE
MORSE'S Gallery, 417 Montgomery St., San Fra
THE HORSE IN MOTION.
Illustrated by
MUYBRIDGE.
AUTOMATIC ELECTRO-PHOTOGRAPH
"SALLIE GARDNER," owned by LELAND STANFORD; running at a 1.40 gait over the Palo Alto track, 19th June, 1878.

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

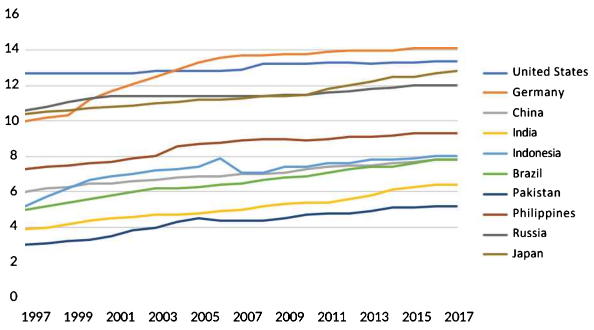
GUIDELINE 5. START WITH GRAY

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.

Average years of schooling has increased around the world

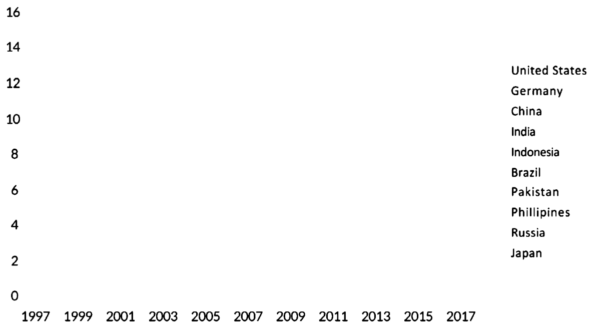
(Number of years)



Source: Our World in Data

Average years of schooling has increased around the world

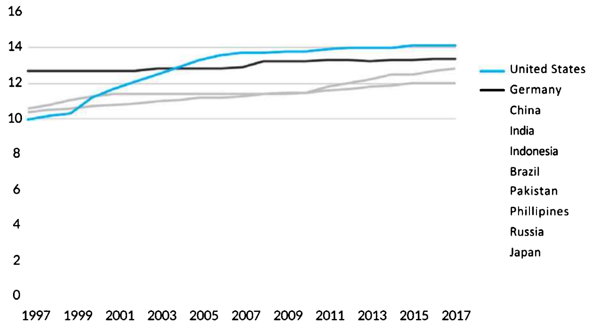
(Number of years)



Source: Our World in Data

Average years of schooling has increased around the world

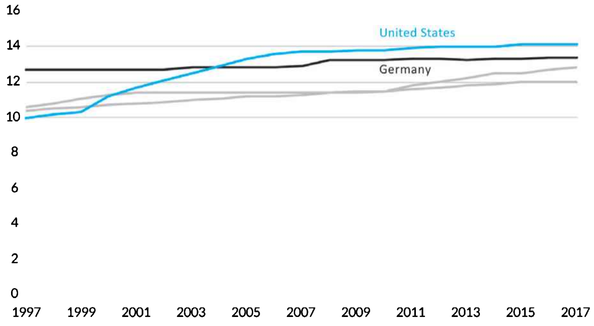
(Number of years)



Source: Our World in Data

Germany and the United States have the highest average years of completed schooling

(Number of years)



Source: Our World in Data

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.

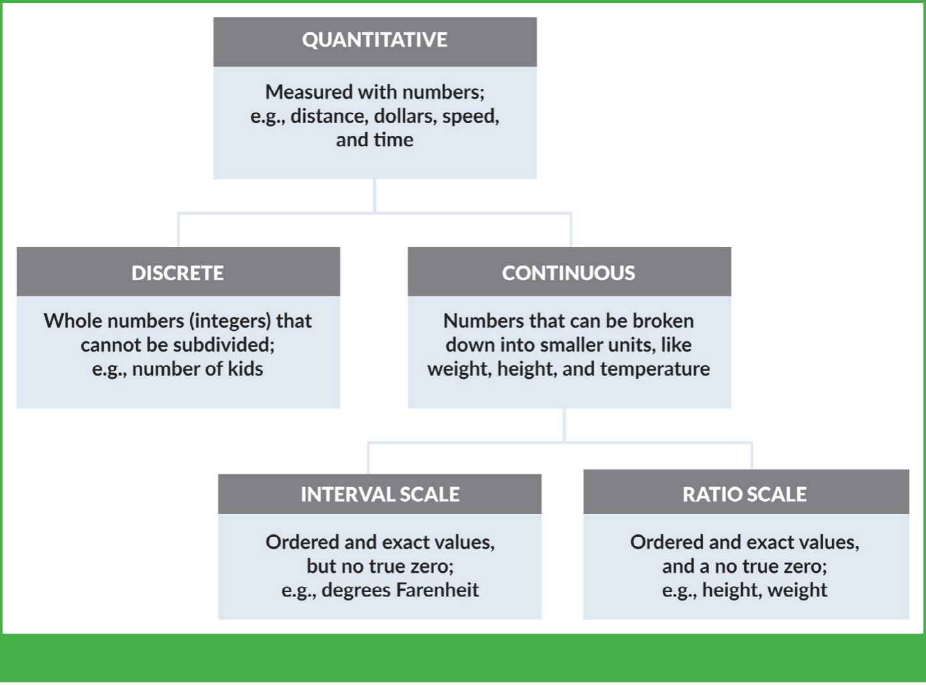
We can further break down each major data type into subcategories. On the qualitative side, we have *nominal* and *ordinal* scales. *Nominal* scales are used to label variables and don't have an order or quantitative value. In a data set of animal types, the order of lion, tiger, and bear has no meaning (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.



QUANTITATIVE

Measured with numbers; e.g., distance, dollars, speed, and time

DISCRETE

Whole numbers (integers) that cannot be subdivided; e.g., number of kids

CONTINUOUS

Numbers that can be broken down into smaller units, like weight, height, and temperature

INTERVAL SCALE

Ordered and exact values, but no true zero; e.g., degrees Farenheit

RATIO SCALE

Ordered and exact values, and a no true zero; e.g., height, weight

DATA EQUALITY & RESPONSIBILITY

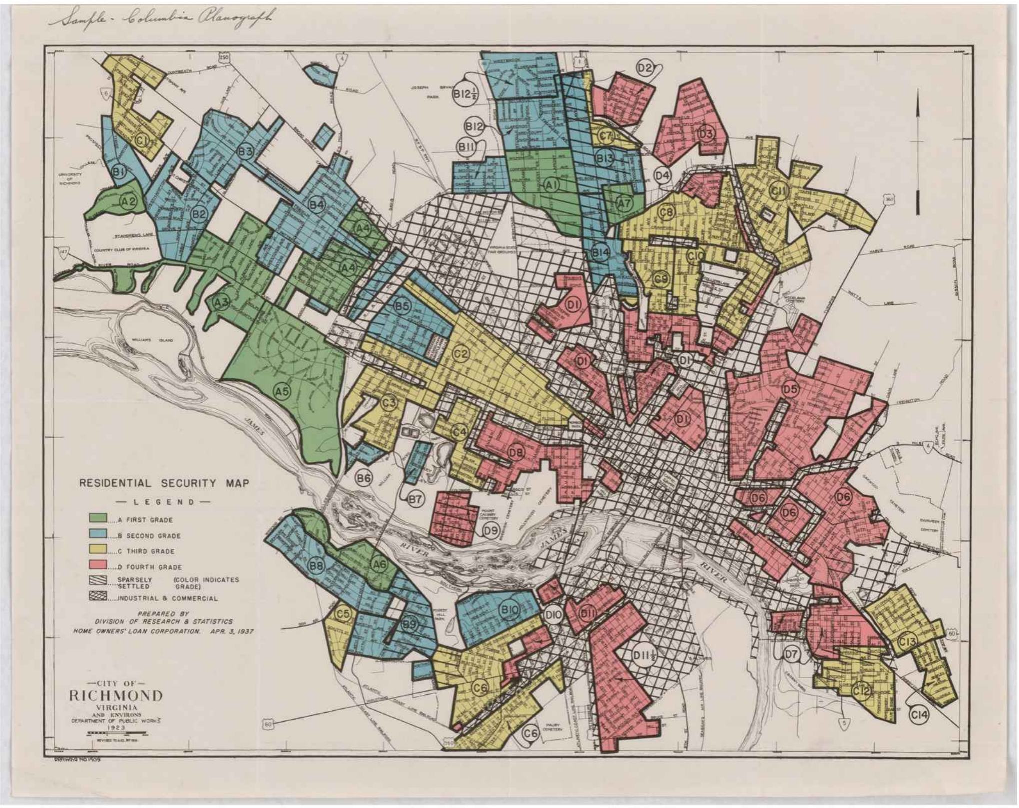
These guidelines lay out the basic approaches to effectively visualizing our data. While this is not a book about data *analysis*—how and where to get data, how to analyze underlying statistical properties, and develop statistical models—whenever we work with data it is important to recognize that visual 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.

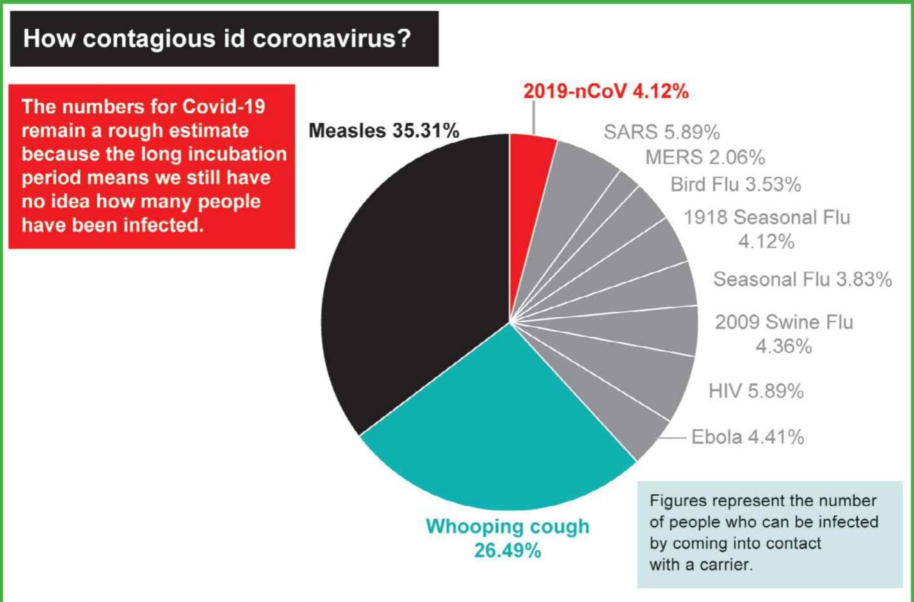
single-family homes." Systematic discrimination is and can be generated by how we use and misuse our data.

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



How contagious id coronavirus?

The numbers for Covid-19 remain a rough estimate because the long incubation period means we still have no idea how many people have been infected.

Figures represent the number of people who can be infected by coming into contact with a carrier.

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.