5

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.

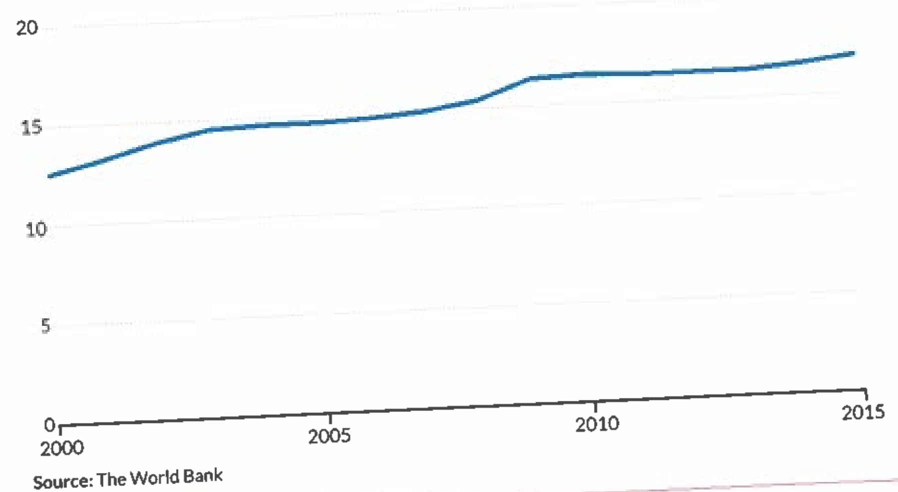
Graphs in this chapter are styled following the guidelines published by the Urban Institute, a nonprofit research institution based in Washington, DC. Urban's style guide outlines their color palette, fonts, and guidance for different chart types.

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.

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

(Percent of GDP)



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.

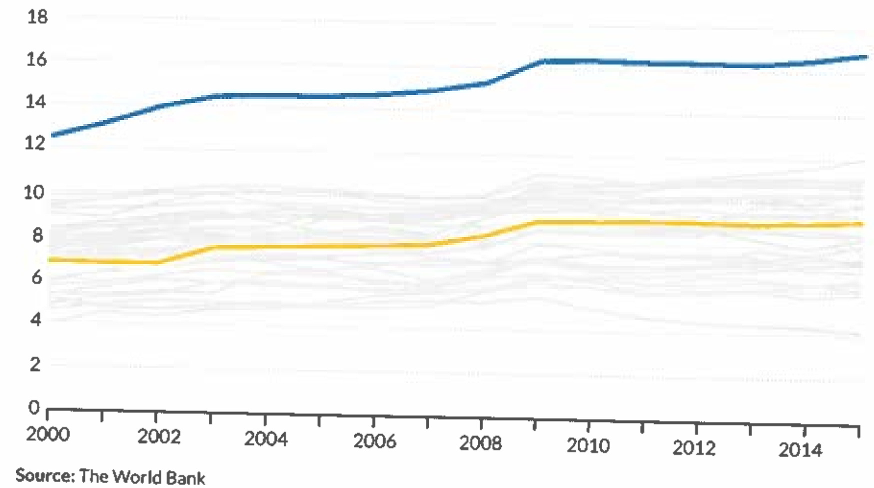
THERE IS NO LIMIT TO THE NUMBER OF LINES YOU PLOT

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

Total health care spending in the United States and Germany increased between 2000 and 2015

(Percent of GDP)



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

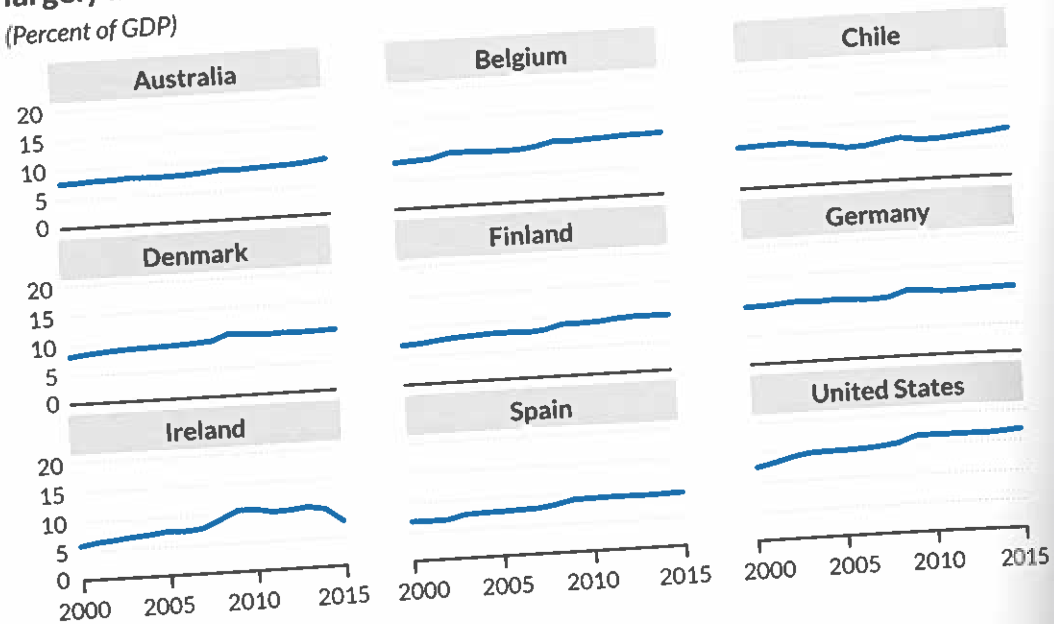
thirty-four countries that make up the Organisation of Economic Co-operation and Development (OECD). To do that, instead of giving each line the same color saturation and thickness, we might only color and thicken the lines for the United States and Germany and leave the lines for the other thirty-two 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

Health care spending across major countries has largely increased since 2000

(Percent of GDP)



Source: The World Bank

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.

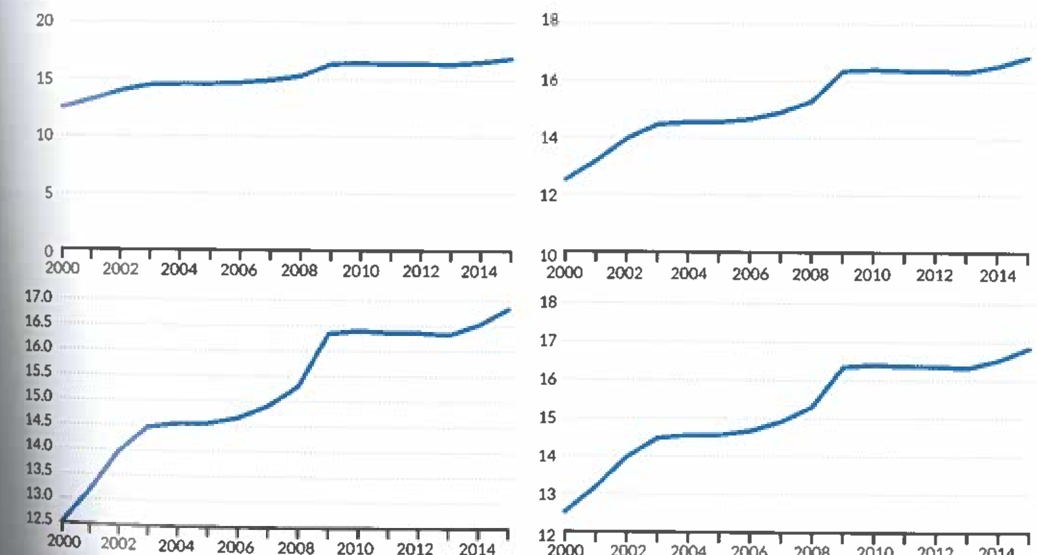
This does not hold true for line charts. The axis of a line chart does not necessarily need to start at zero. As with many aspects of visualizating data, there are complications and different perspectives. If we say the axis does not need to start at zero, what is an appropriate range? Where should we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



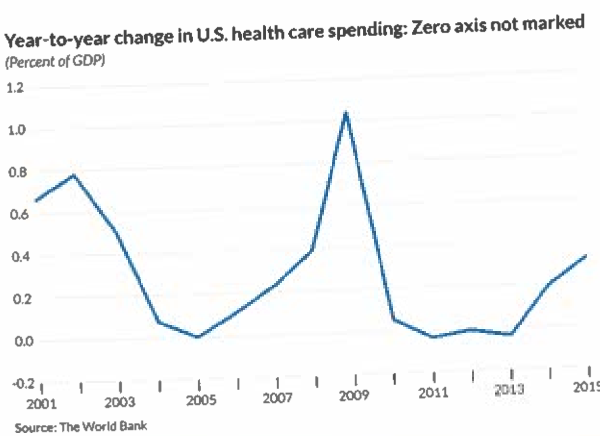
Source: The World Bank

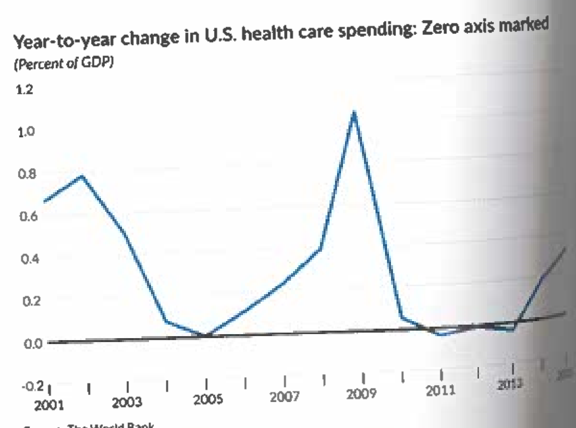
There is no right answer to the choice of the vertical axis. The answer depends on the data and your goal.

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.



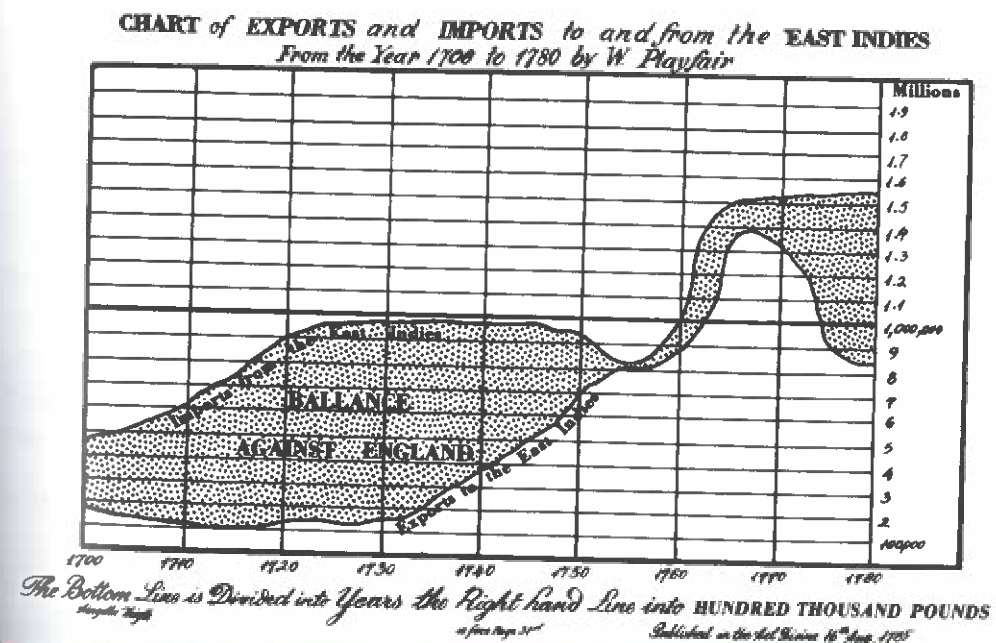


We are inclined to think the bottom of the vertical axis is zero, so using a different color or thickness for the axis makes it clear where zero lies.

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.

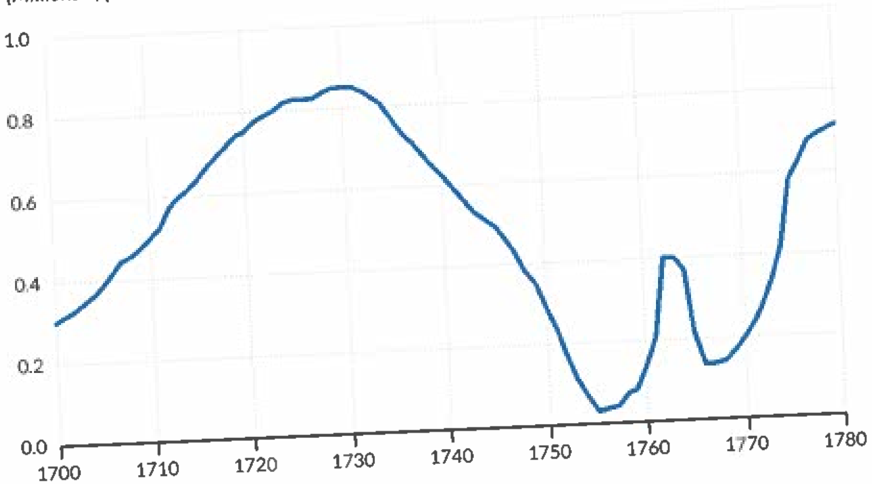


In his chart from 1785, William Playfair, a Scottish engineer and political scientist who is often credited as the founder of graphic methods of statistics plots, shows exports and imports between England and the East Indies.

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.

Gap between exports and imports between the UK and the East Indies
(Millions of pounds)

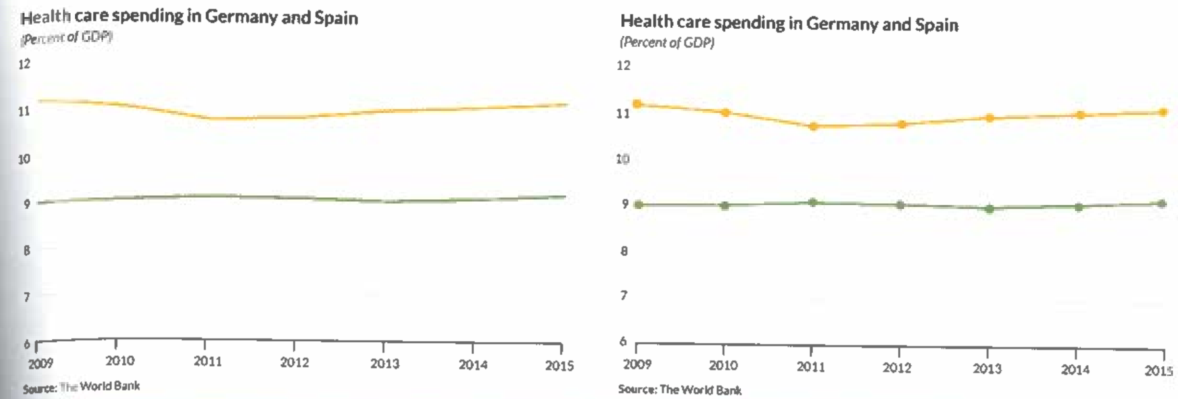


Source: Based on Cleveland and McGill, 1984; data from Michael Friendly

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.

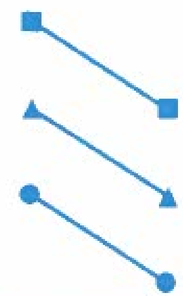


I like to add data markers to my line charts when I don't have a lot of data or when I want to highlight or label specific values.

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.

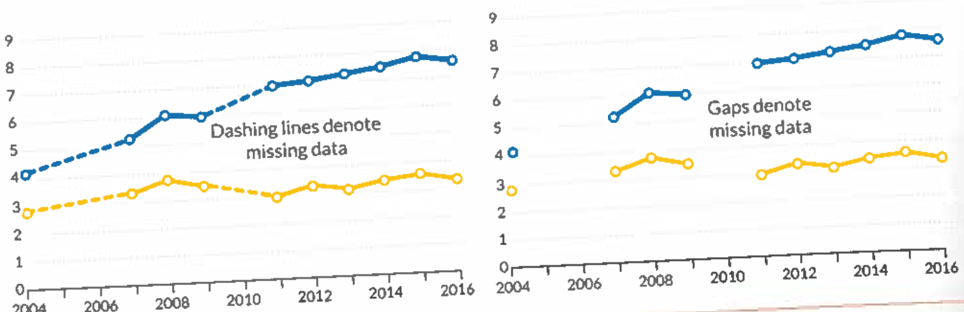


Circles are a symmetrical shape, which is why I prefer to use them as data markers.

USE VISUAL SIGNALS FOR MISSING DATA

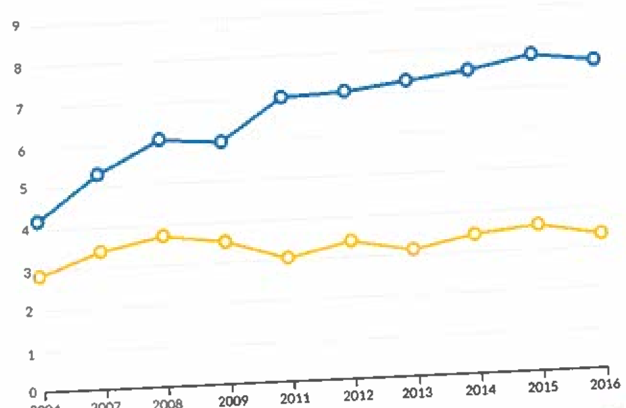
At some level, there is *always* missing data. People change jobs every day, not just when unemployment numbers are published. Lots of things happen in the ten years between the publication of each U.S. Census. Most data are a snapshot in time, but we often treat them as continuous.

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

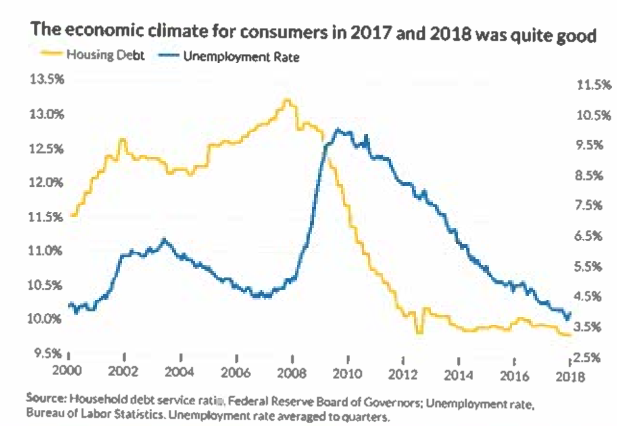
You might be inclined to add another vertical axis to your line chart when comparing changes in two or more series of different units. Resist that urge. Consider this dual-axis line chart that shows the share of income devoted to paying for housing on the left axis and the quarterly unemployment rate 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.

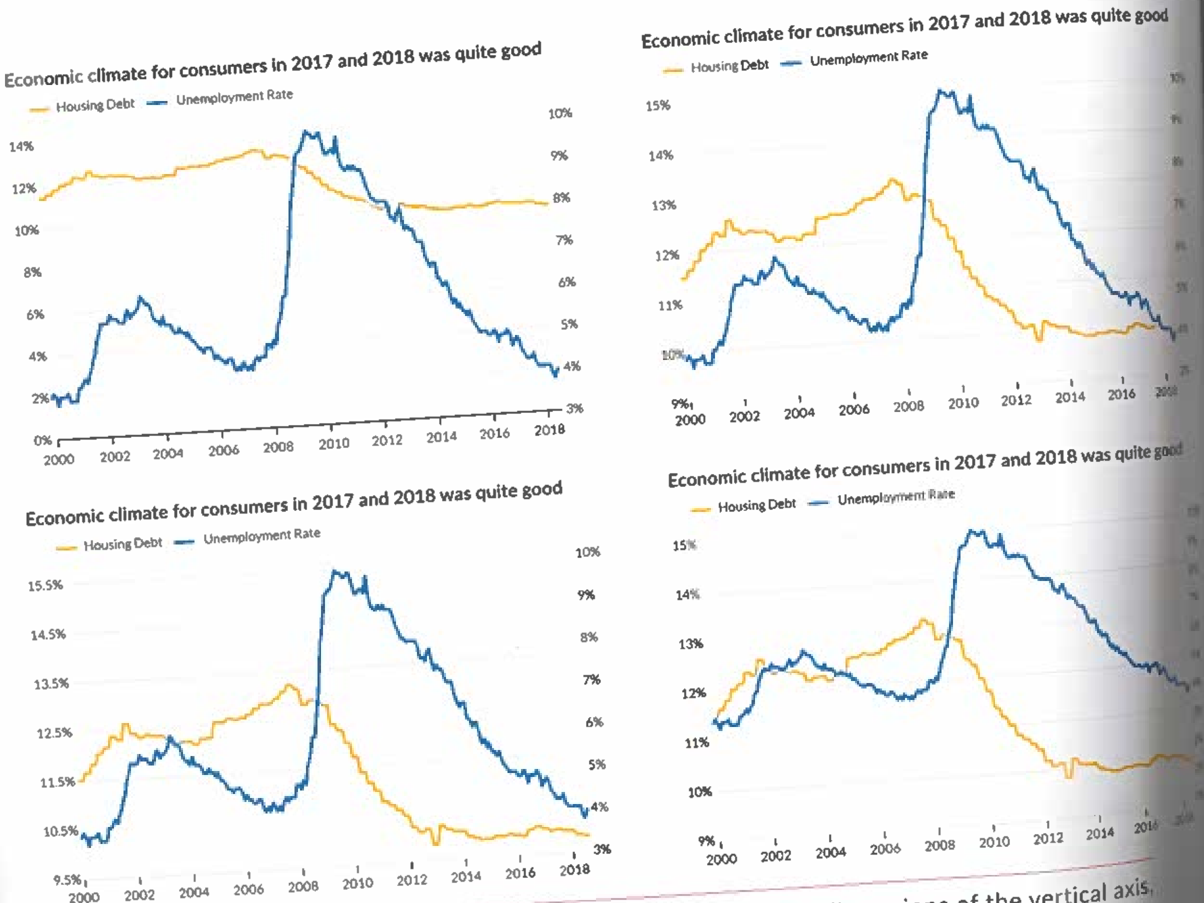
Third, and most importantly, the point where the lines cross becomes a focal point, even though it may have no real meaning. In this graph, the eye is drawn to the middle of



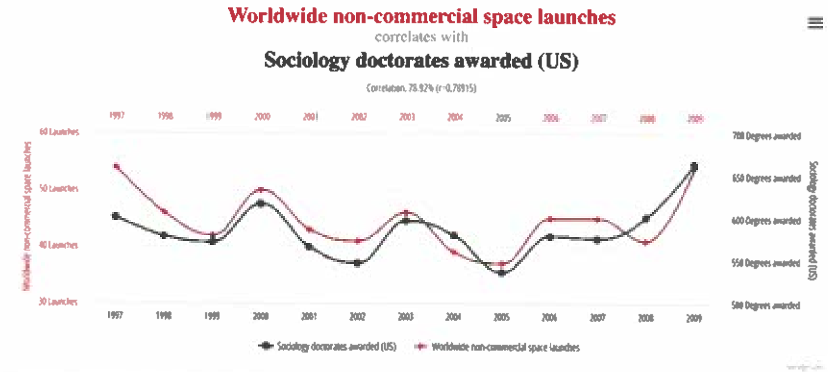
The dual-axis chart introduces a series of perceptual issues, maybe the most important of which is that the eye is drawn to where the lines intersect, even though that may be meaningless.

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.

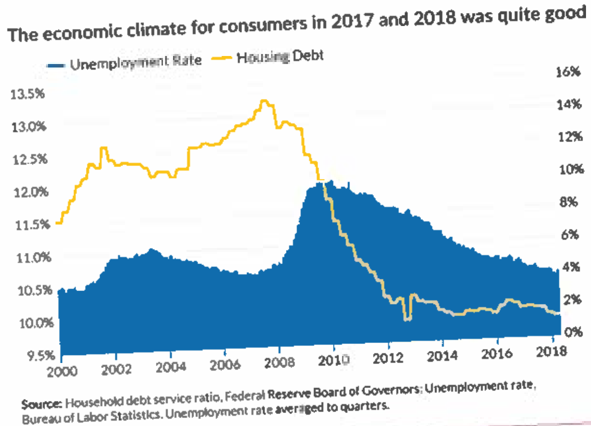
Each of these four graphs are reasonable ways to set the vertical axes, and by manipulating those ranges, I can make the series look like they (a) are closely matched for a few years around 2010 and 2012; (b) cross in the middle and the end; (c) intersect around 2003 and then again a few years later;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

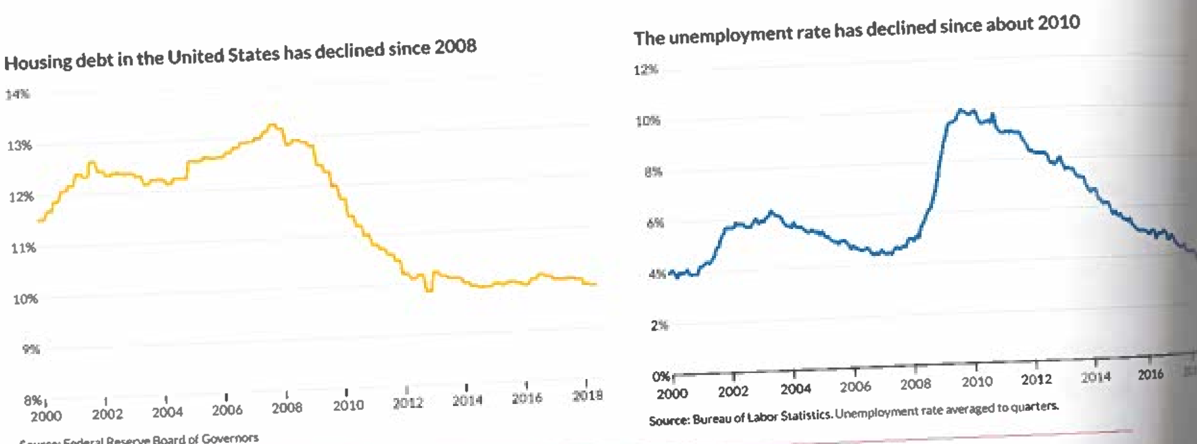
Similar difficulties, though to a slightly different extent, exist in dual-axis charts that combine different graph types. On the next page you can see the same graph of the unemployment rate and housing debt, now with the unemployment rate plotted as an area chart. The right axis starts at zero so the 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.



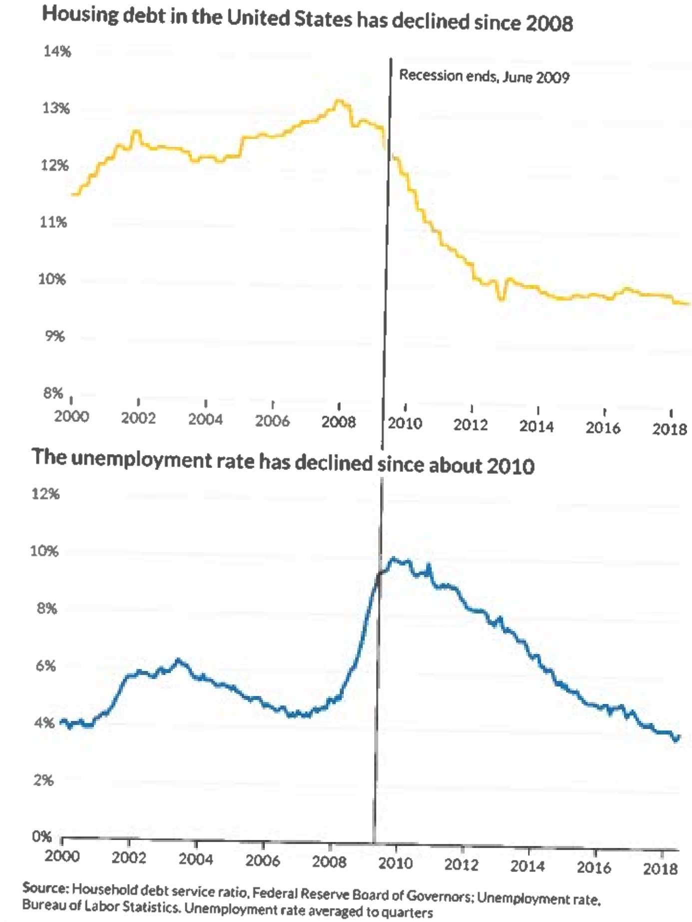
The issue with dual-axis charts is not resolved by combining area charts and line charts.



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

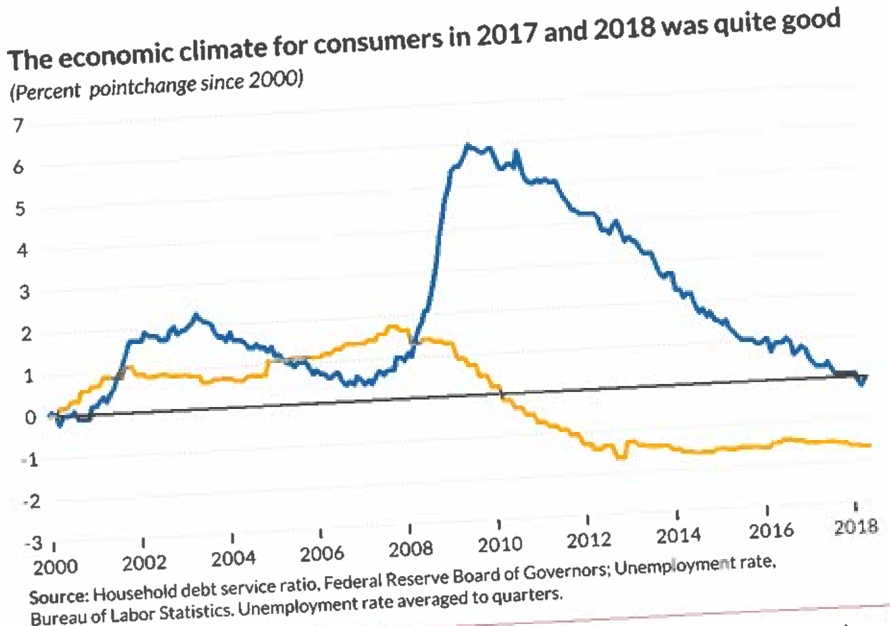
If it's important to annotate a specific point on the horizontal axis, you could also vertically arrange the two and draw a line across both. This will change the rotation of the final graphic, but is an easier way to label a specific value or year.

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*.

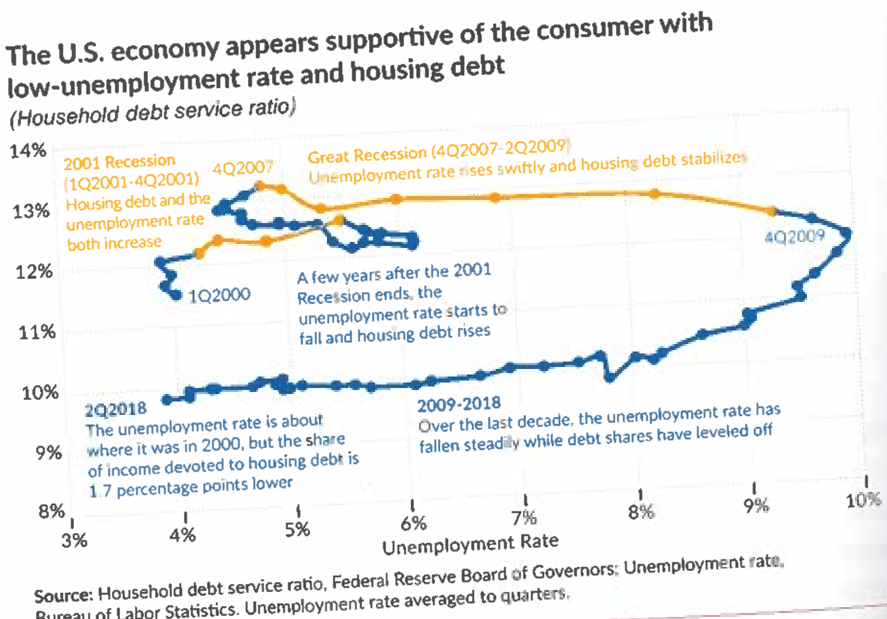


One alternative to the dual axis line chart is to use two charts aligned vertically, which makes it a little easier to mark a specific data point on both charts.

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.



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



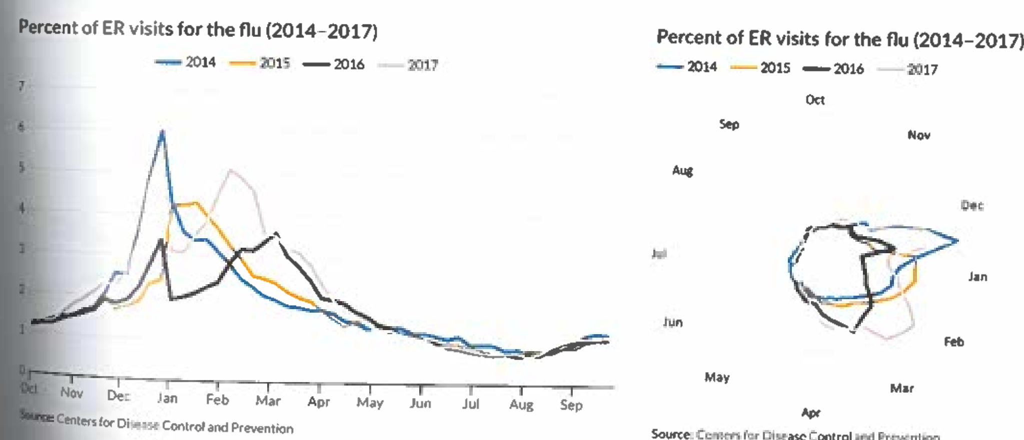
Yet another alternative to the dual-axis chart is to use the connected scatterplot in which one data series corresponds to the horizontal axis and another to the vertical axis.

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.

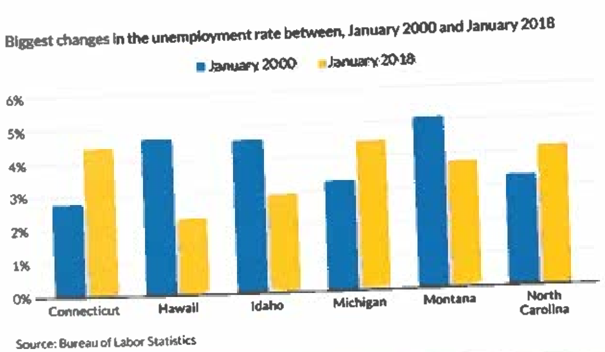
SLOPE CHART

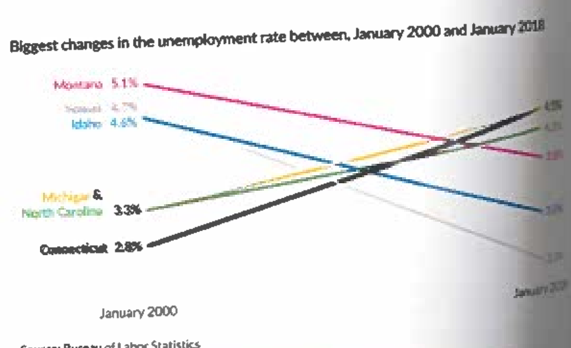
In some cases, it may not be necessary to show *all* of the data in your time series. In these cases, a slope chart—which is really just a simplified line chart—is a useful alternative.

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

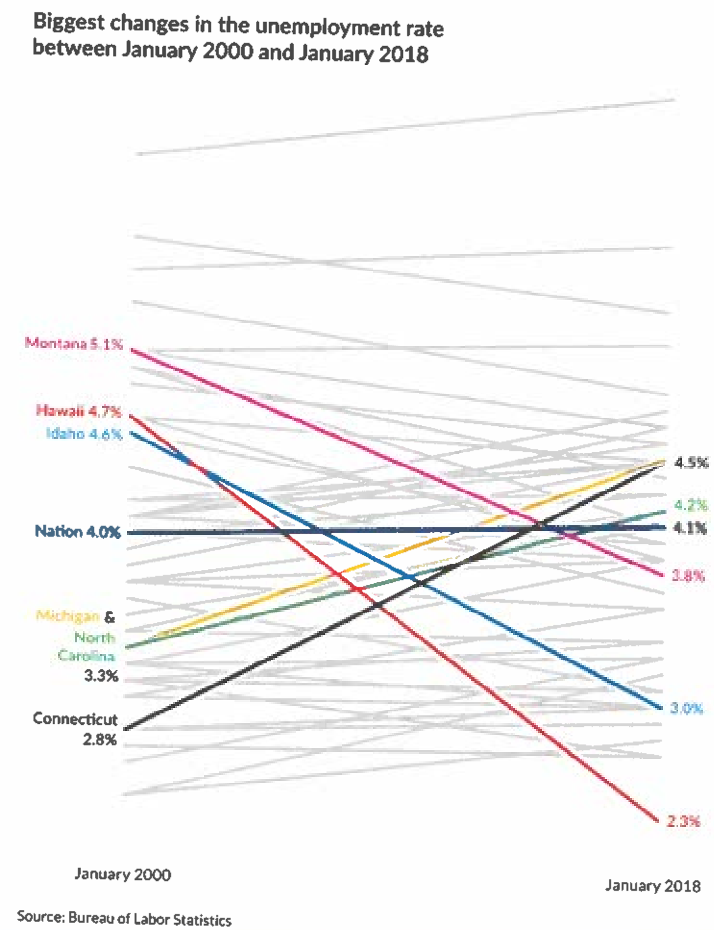




The paired bar chart asks the reader to make several comparisons on their own simultaneously, while the slope chart visualizes these comparisons for the reader.

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

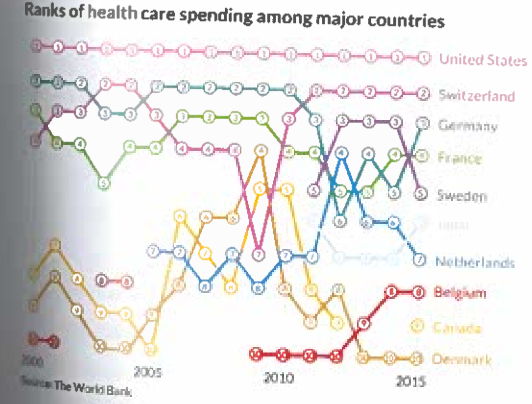
Sparklines are a type of small multiples line chart typically used within data-rich tables.

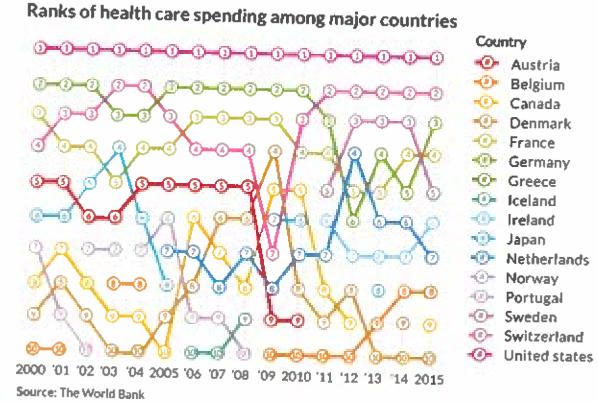
BUMP CHART

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.

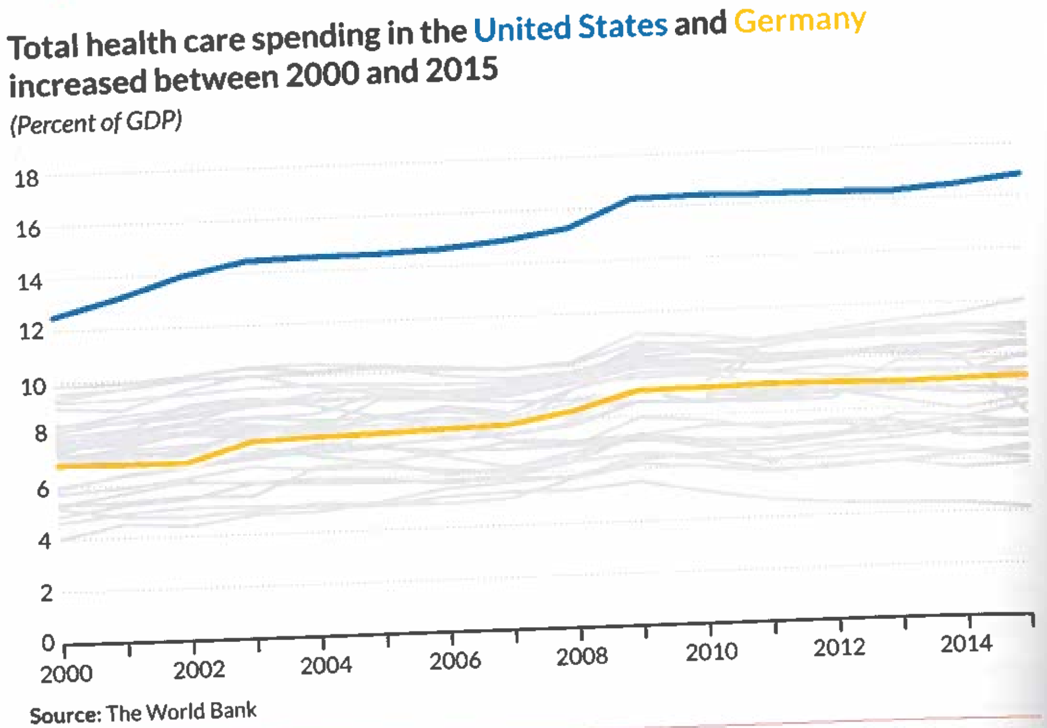
A bump chart is, of course, a compromise. It does not show the raw values, which are often preferred, but it can be especially useful if your data have outliers. By plotting the ranks, we abstract from the large differences in magnitude.

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.





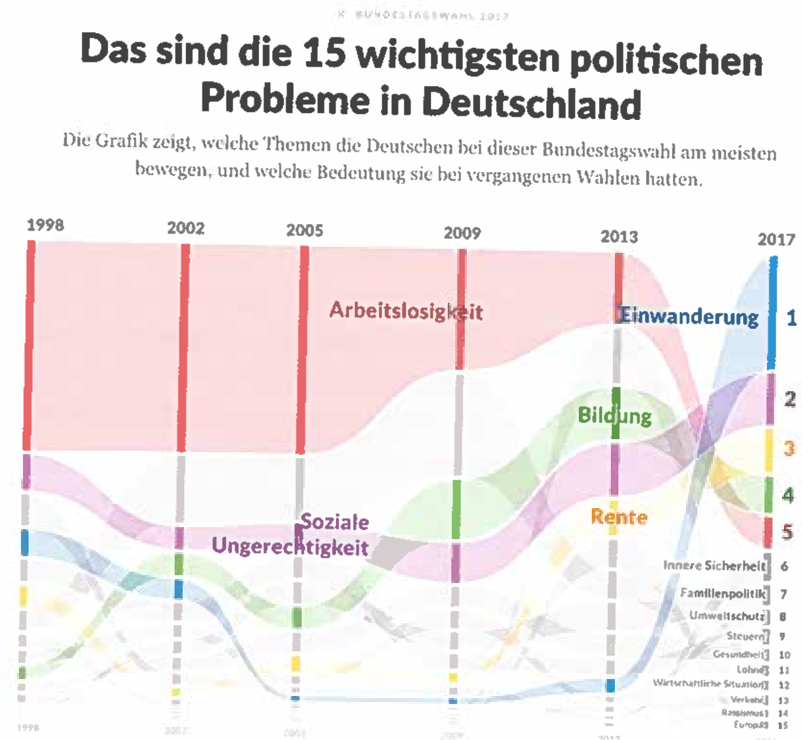
Bump charts plot changes in ranks over time.



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.



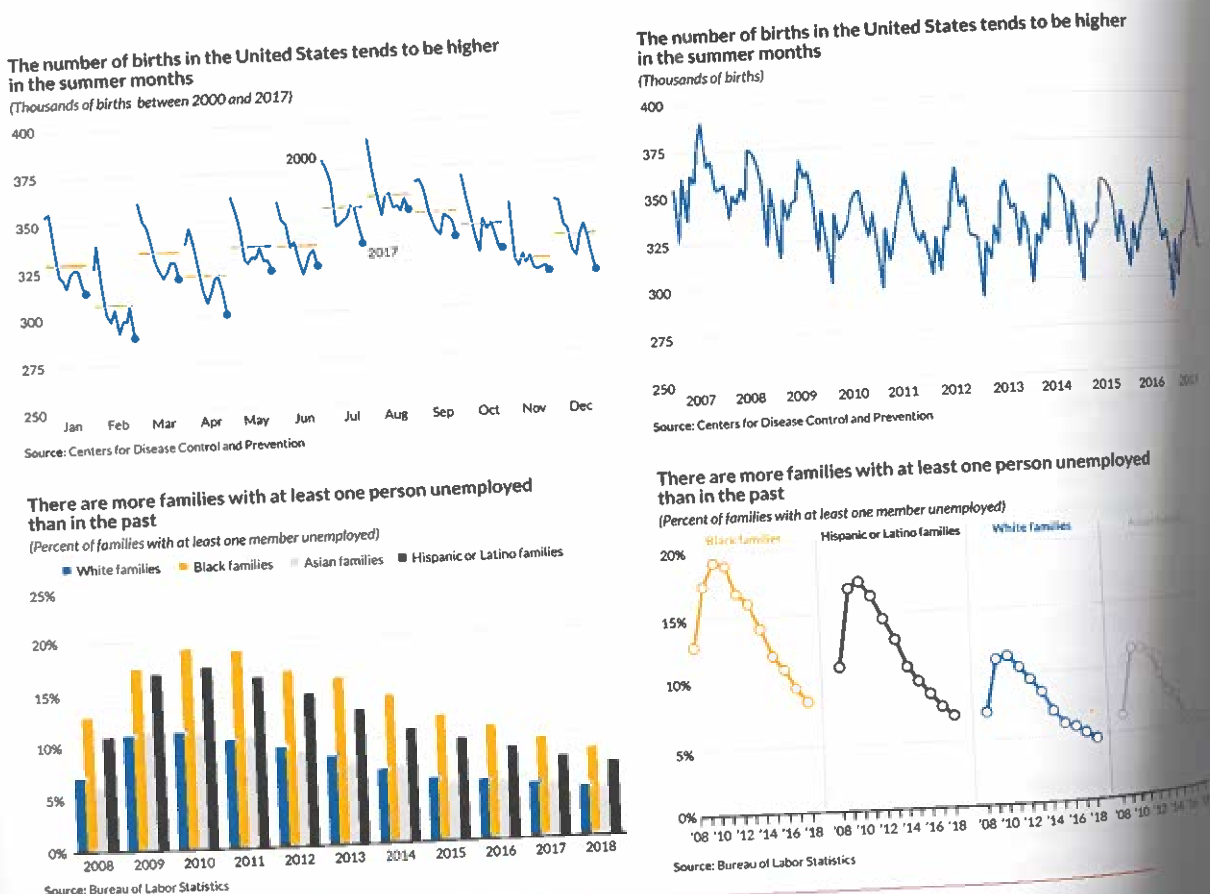
The ribbon effect is a modification on the standard bump chart. The title in this chart from the Berliner Morgenpost translates to "These are the 15 most important political problems in Germany". Arbeitslosigkeit translates to Unemployment; Einwanderung to Immigration; Bildung to education; and so on. 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



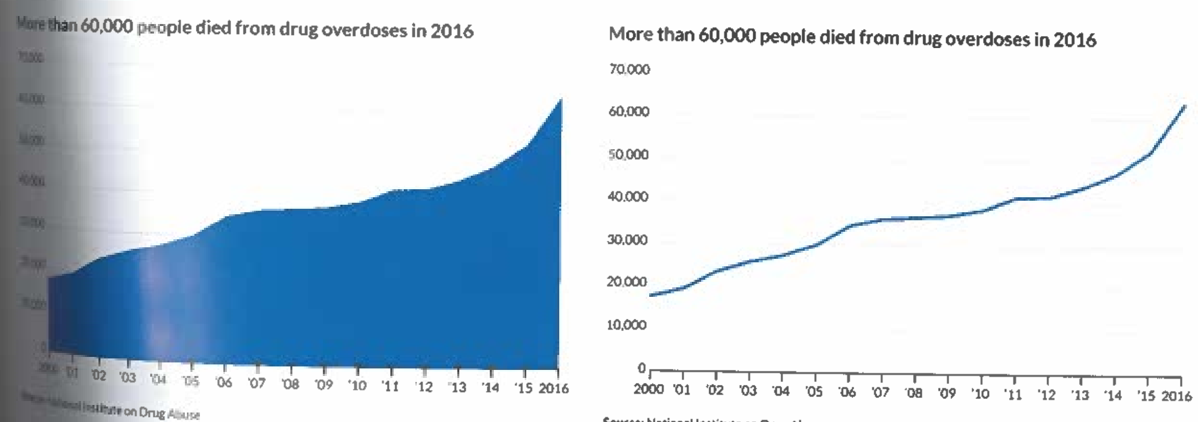
Cycle graphs compare small units of time, such as weeks or months, across a multiyear time frame.

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.

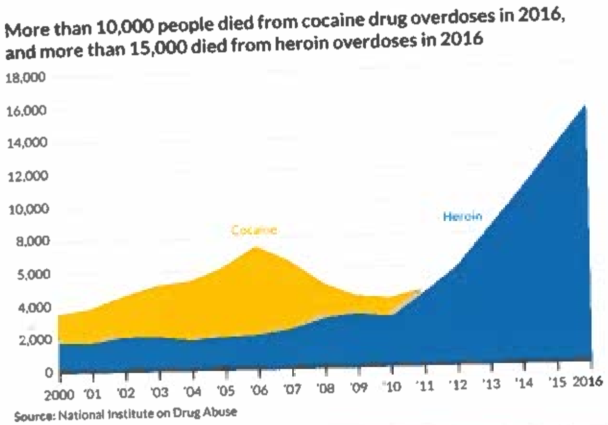
AREA CHART

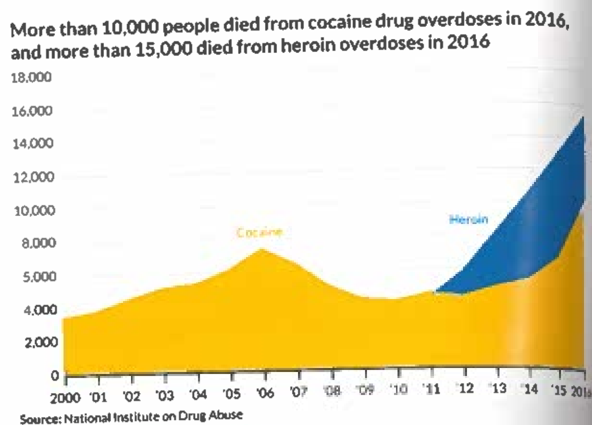
Area charts are line graphs with the area below the line filled in, giving the series more visual weight. The area chart on the left and the line chart on the right both show the number of people who died from prescription opioid overdoses in the United States between 2000 and 2016. You might think of 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.

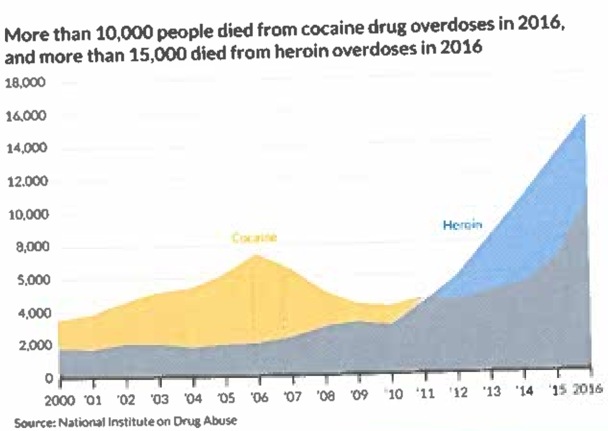


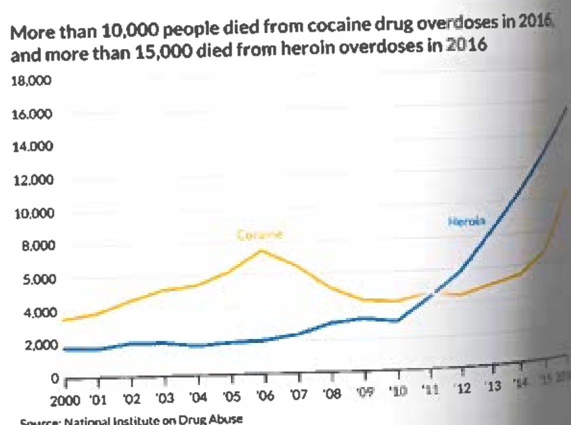


Placing two or more series in an area chart can spell trouble, because one series can hide (or "occlude") the other.

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.

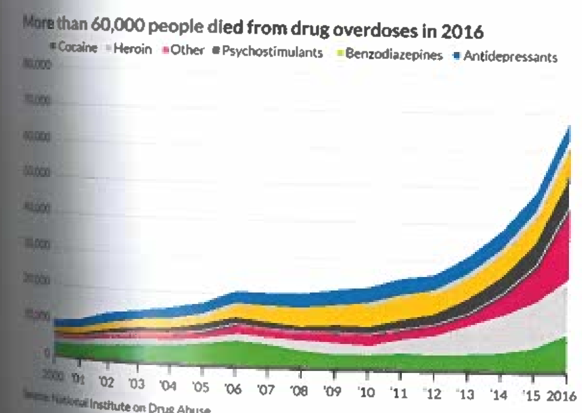
STACKED AREA CHART

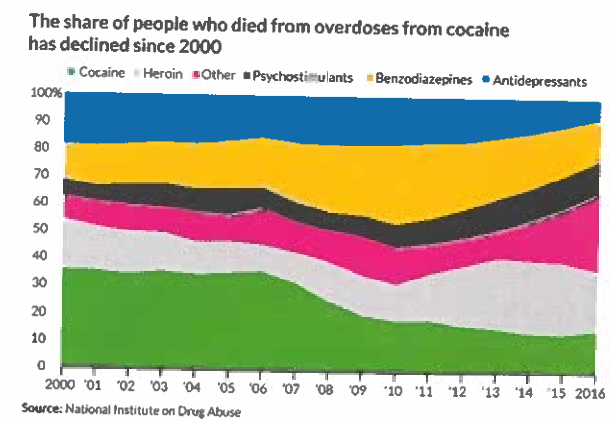
Stacked area charts build on the typical area chart by showing multiple data series simultaneously. Instead of sitting independently of one another as in the previous chart, the data in a stacked area chart sum to a total or a percentage.

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.

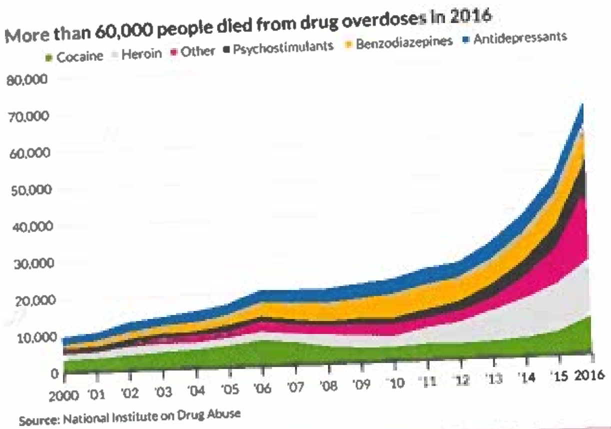
The reader will take away different conclusions from these two representations. In the graph on the left, the eye is drawn to the large increase in overall deaths over the period. In the version on the right, it is drawn to the changes in the distribution of deaths—a decline in overdoses from cocaine,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

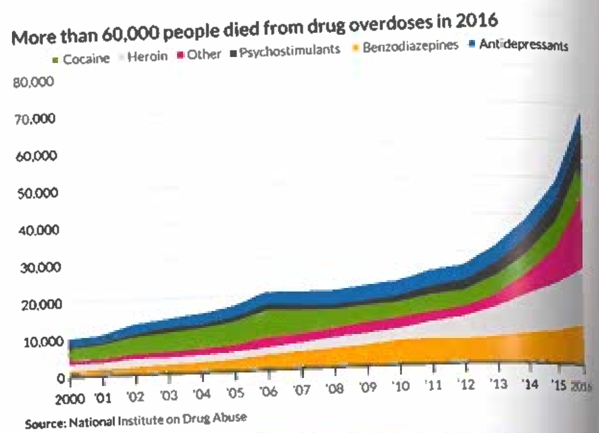
There are three disadvantages with the stacked area chart on the left. First, as earlier, we again see the line-width illusion—we tend to view steep changes as bigger than they actually are. Second, only the bottom series sits on a horizontal axis, so it is hard to accurately compare the changes over 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.

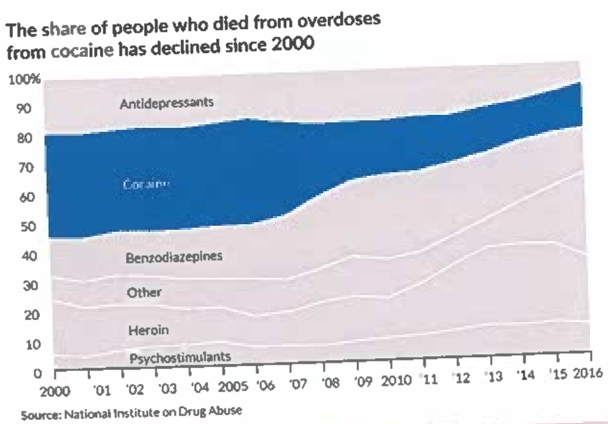


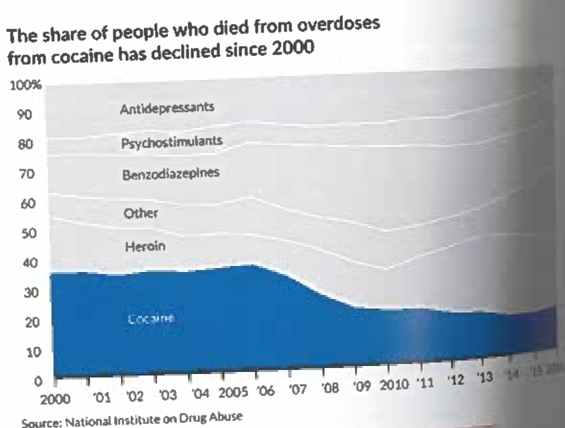


Recall that the perceptual ranking list suggests we can better compare values when they sit on the same axis. That's why it's easiest to compare values for the bottom series.

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.



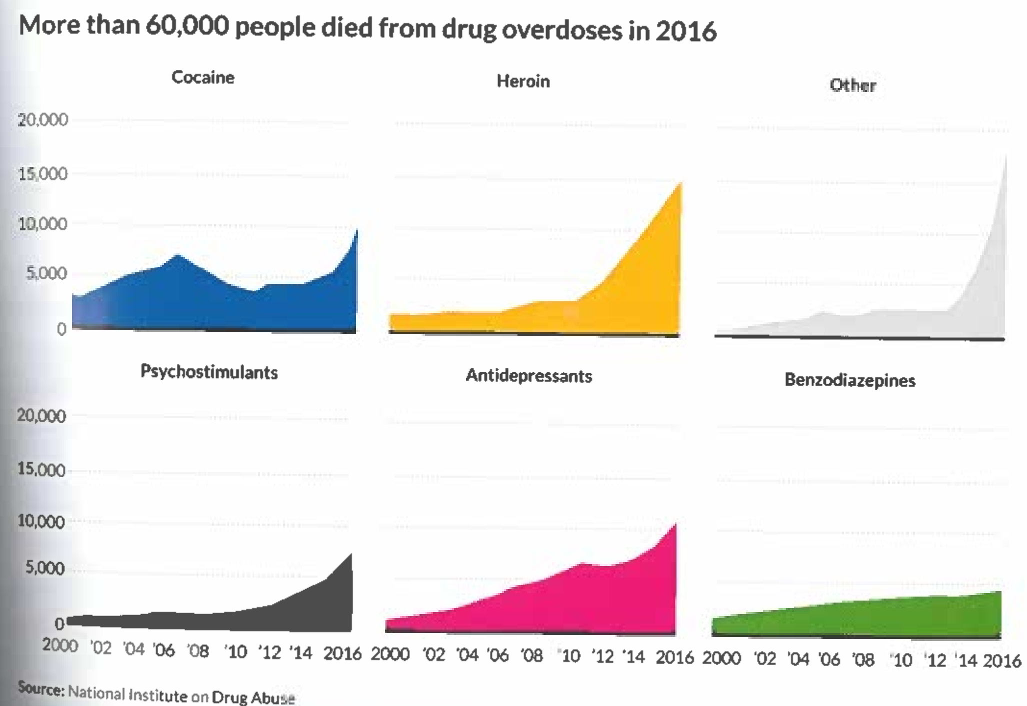


There isn't necessarily a right way to stack the series in a stacked area chart, but how you decide to arrange them will influence how your reader perceives the data.

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

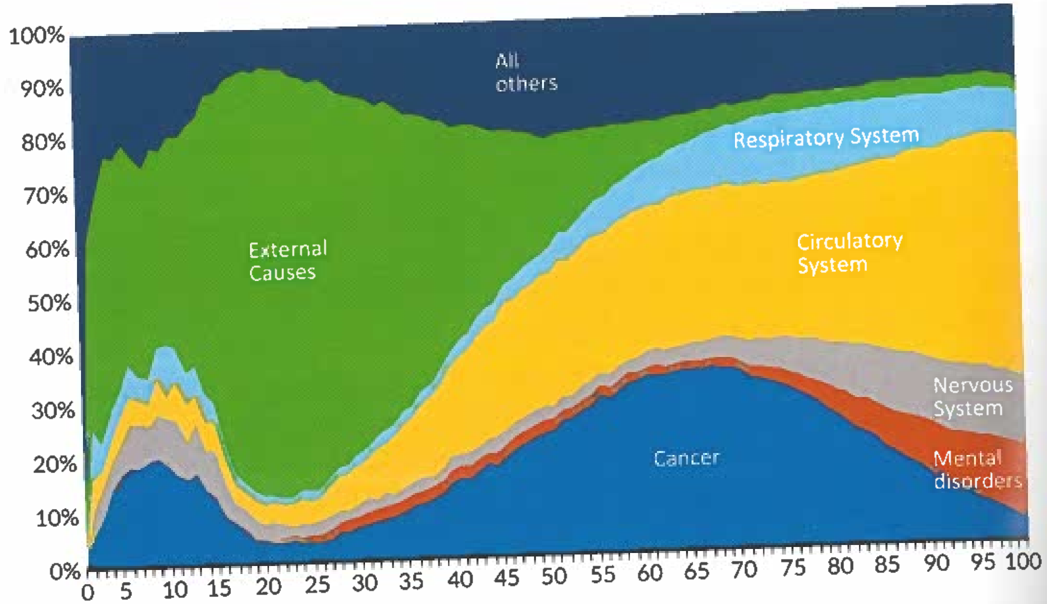
share of overdose cocaine deaths over this period, you could keep it in the same position as above, but use the "start with gray" strategy and use color in just the cocaine series. Even with the line-width illusion, you can still see the share of deaths has declined. If it's important for your reader 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.

Causes of death by age in the United States in 2017



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

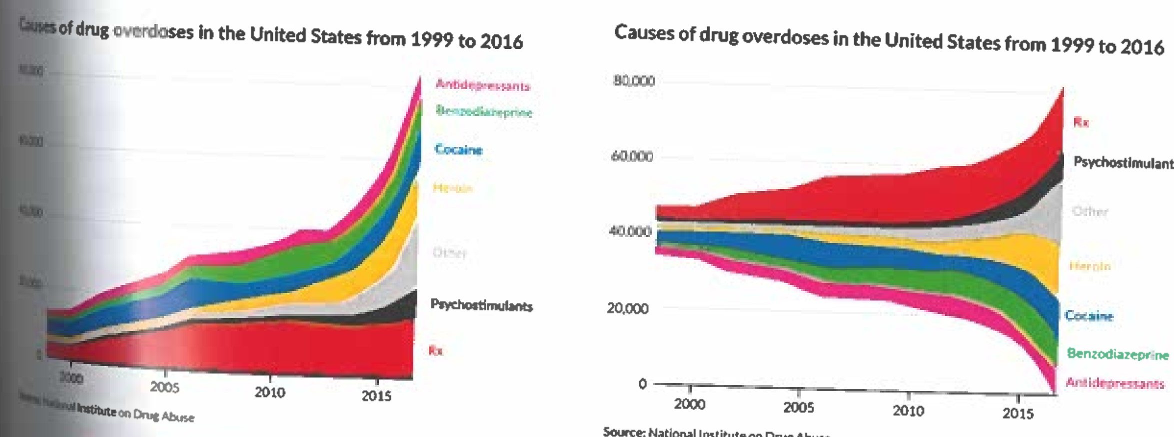
Like the stacked area chart, a streamgraph also stacks the data series, but the central horizontal axis does not necessarily signal a zero value. Instead, data can be positive on both sides of the axis. Together, the streamgraph illustrates fluctuations in data over time in a

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.

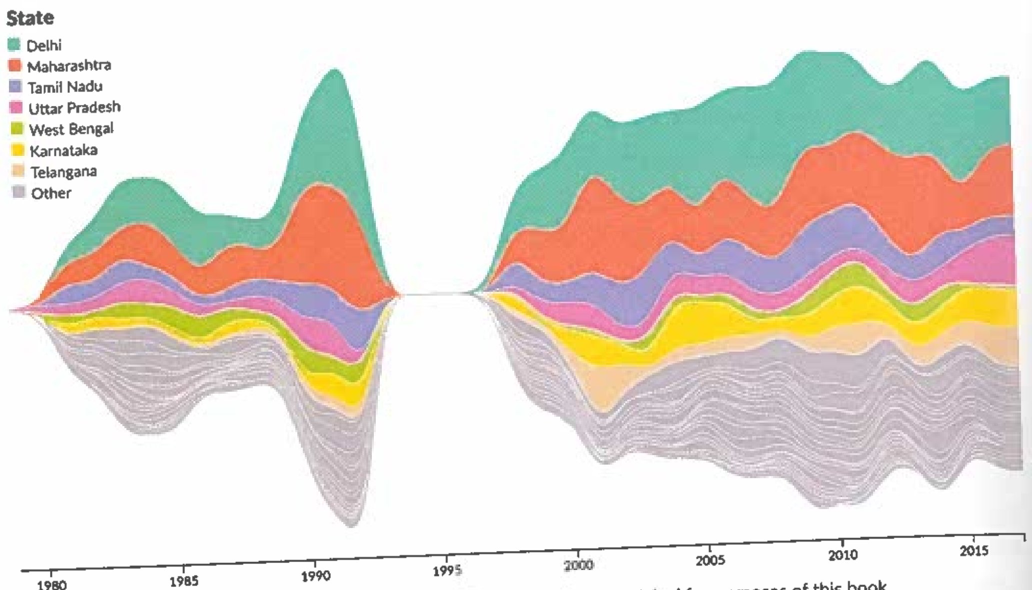
Researchers are aware of how unusual the streamgraph looks, and how it may be more difficult for readers to understand. In a review of a streamgraph published by the *New York Times* in 2008, researchers noted that they "suspect that some of the aesthetically pleasing—or at least engaging—qualities may 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.

Delhiites have received the most awards, followed by Maharashtrians



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.

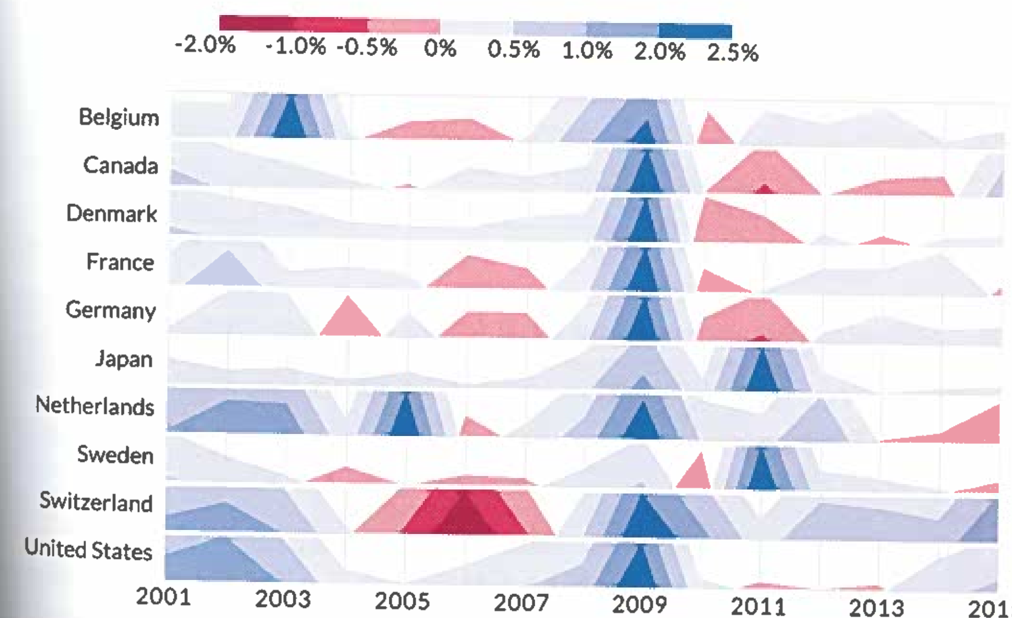
Color is the most important attribute in a horizon chart. Darker colors represent larger values and lighter colors smaller values. Like sparklines and to some extent heatmaps, the purpose of the horizon chart is not necessarily to enable readers to pick out specific values, but instead to easily spot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

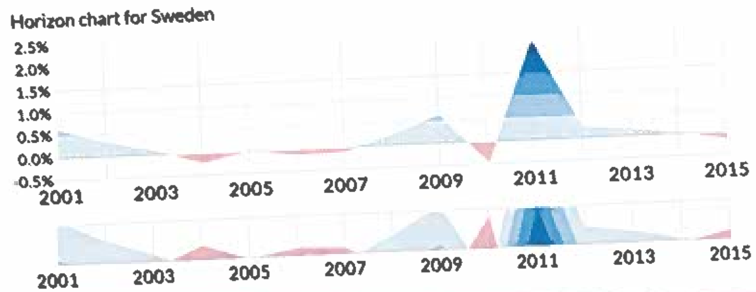
Color is the key here. The same visualization as a series of line charts does not have the same punch. The eye scans the entire visualization for important trends, but no particular region of the visual draws our attention. That could be modified by adding color to the

Change in health spending as a % of GDP

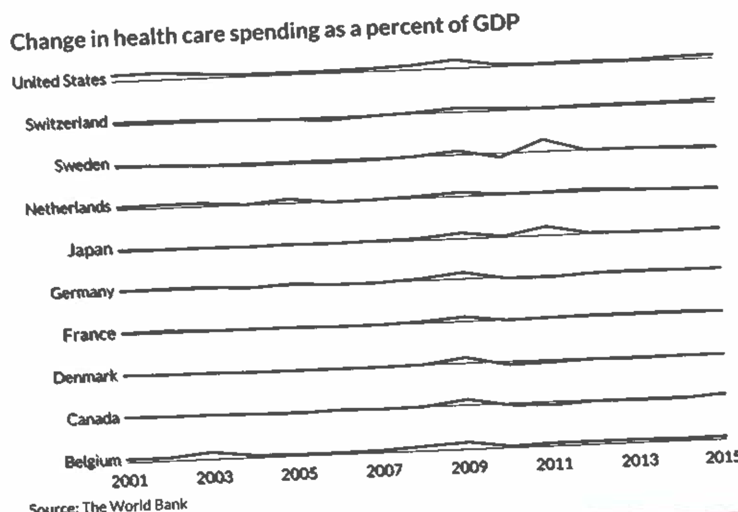


Source: The World Bank

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.

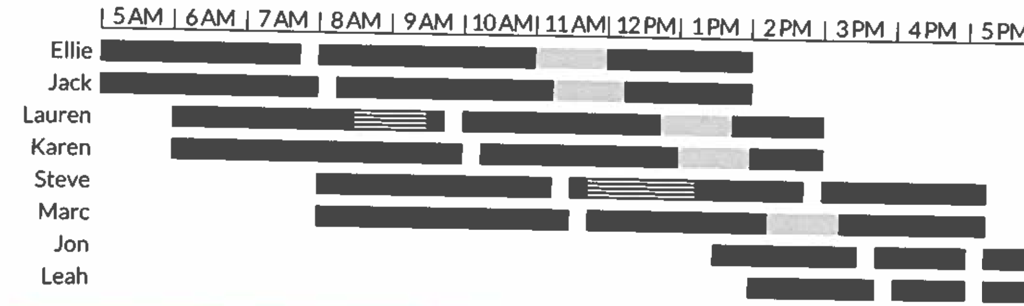


A line chart is sufficient, but the use of color in the horizon chart attracts and directs the reader's attention.

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



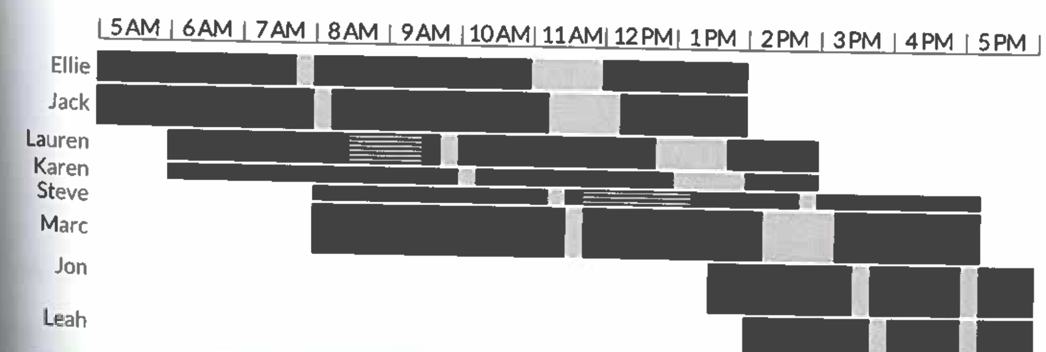
Gantt charts often show processes or schedules.

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.

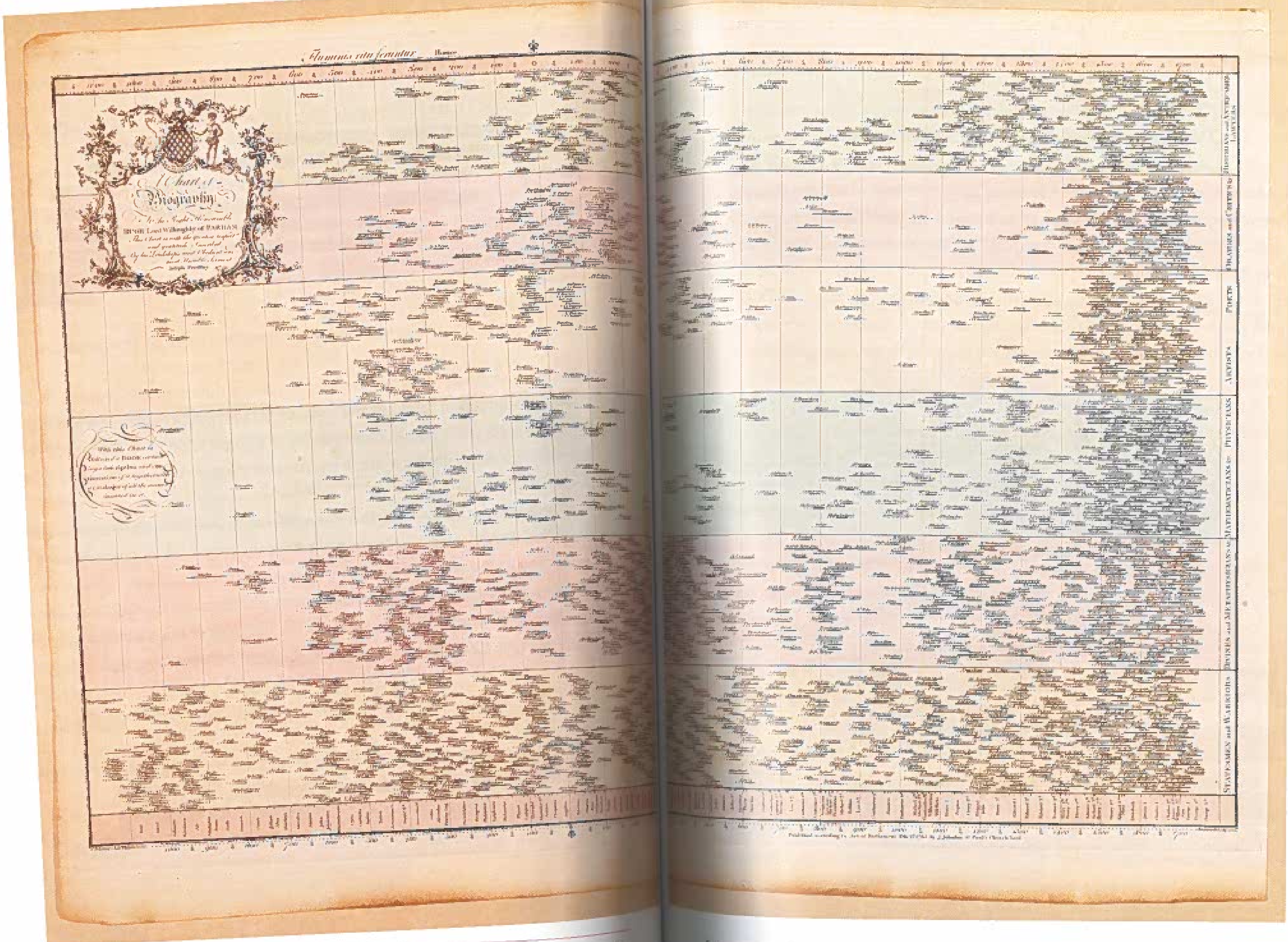
This Gantt chart shows staffing shifts at a coffee shop over the course of a day, denoting breaks with the white gaps, lunch breaks with the gray breaks, and other time away from the store with stripes.

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.

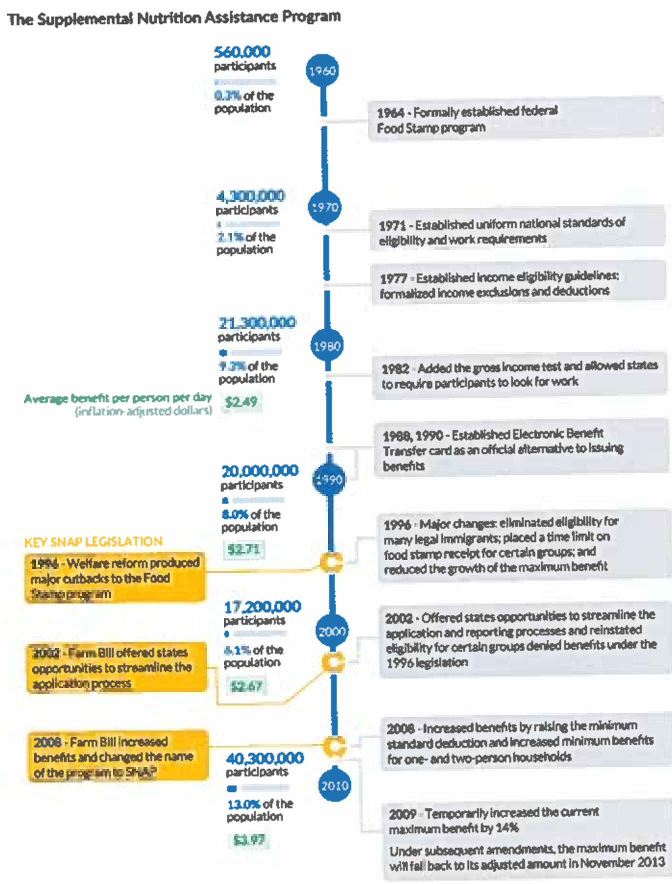


Joseph Priestley's *A Chart of Biography* (1765) shows the lifespans of about two thousand statesmen, poets, artists, and others. It covers an enormous time period from 1200 BC to 1800 AD. A horizontal line for each person shows his or her lifespan. Dots indicate wherever such there is uncertainty around 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



Source: Based on Congressional Budget Office (2012)

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.

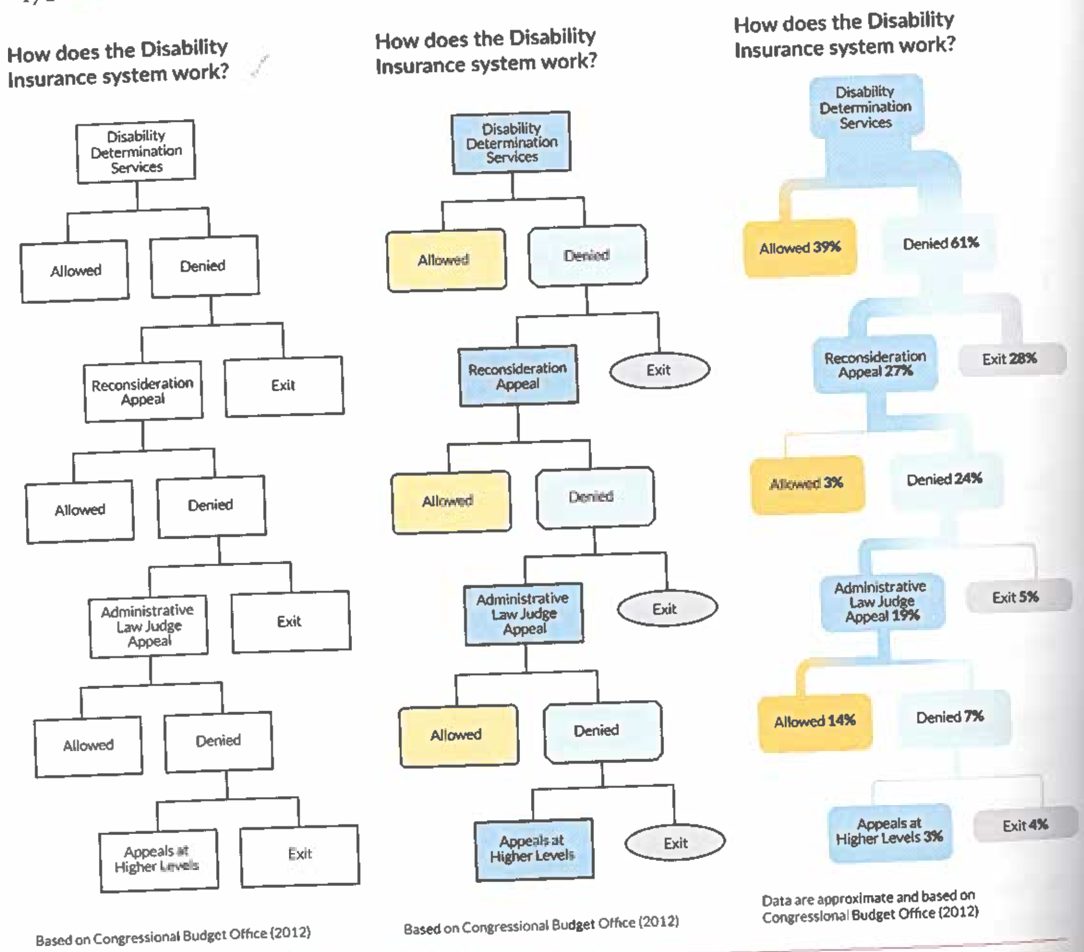
A timeline shows when certain events take place. It can be basic and flag events with a line, icon, or marker, or it can be more involved and include annotation, images, or even graphs. Though horizontal timelines are common, timelines can also be vertical or even a variety of different shapes. This 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.

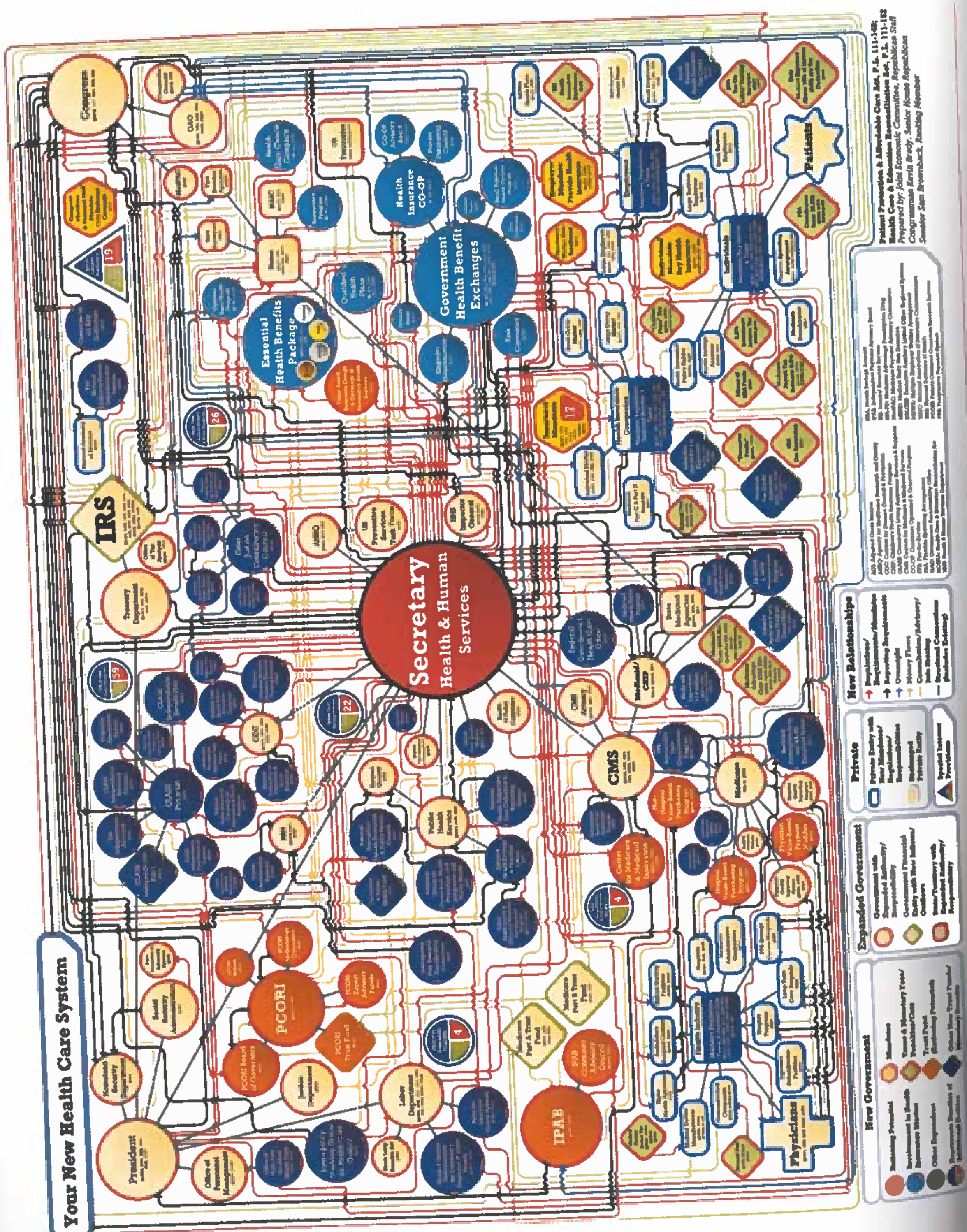
TOTALS VS. PER CAPITA

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.

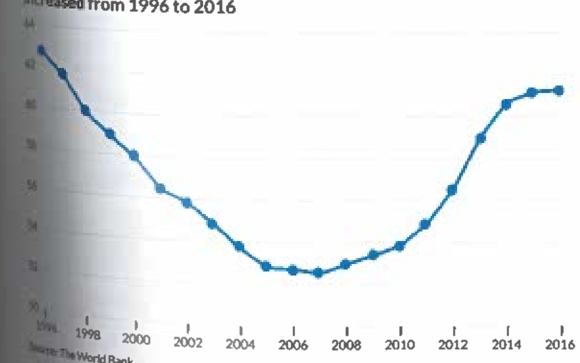
CONNECTED SCATTERPLOT

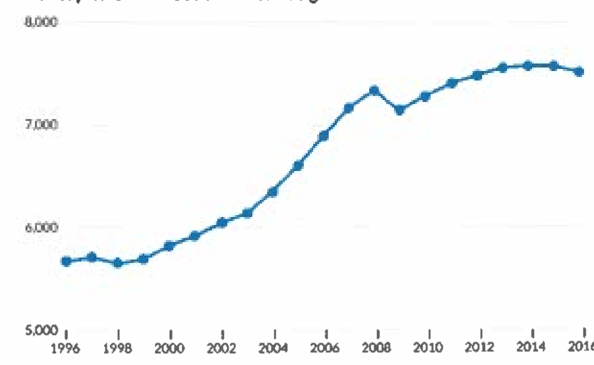
Imagine showing two line charts side-by-side. You may be asking your reader to examine the relationship between the two. Do they move together? Do they diverge or converge? How are they related?

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.



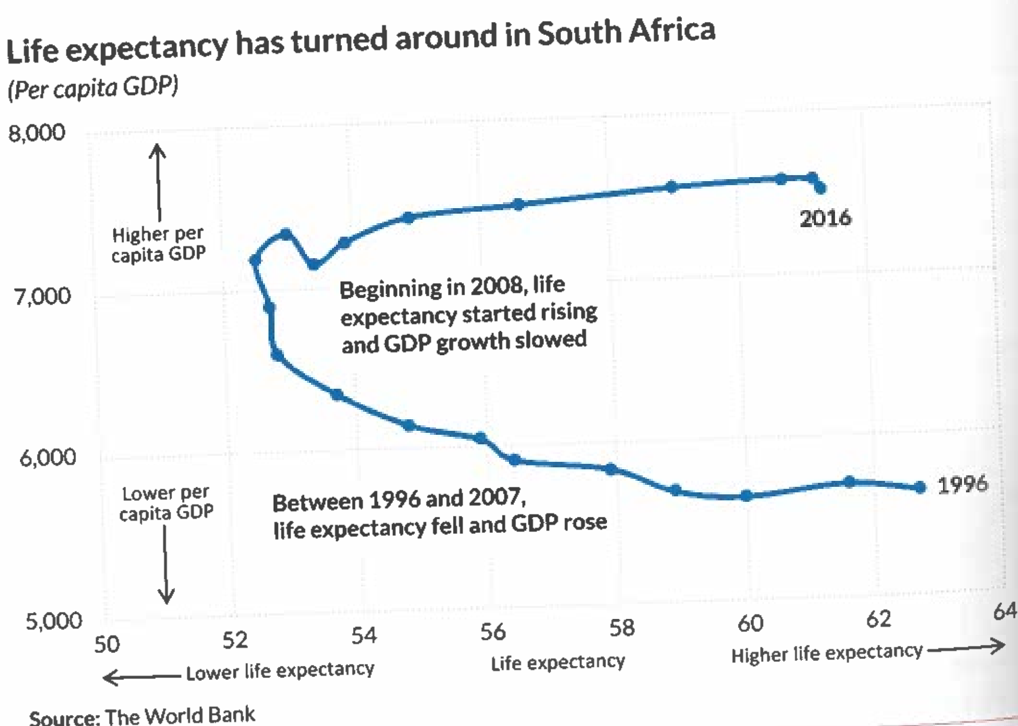


A common challenge is how to clearly show the association between two time series.

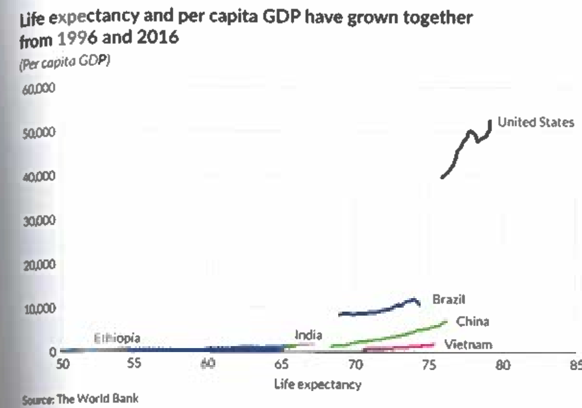
Now, notice what happens when the two lines are combined and plotted in a single graph. Here, life expectancy is shown along the horizontal axis and economic growth is plotted along the vertical axis.

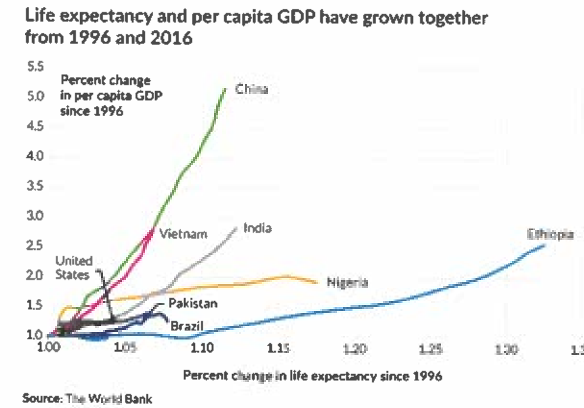
Instead of jumping back and forth between the two charts, we can see that during the first half of the period life expectancy fell (moved to the left along the horizontal axis) and the economy grew (moved up along the vertical axis). Starting around 2006, life expectancy started to increase again (now 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.

6

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