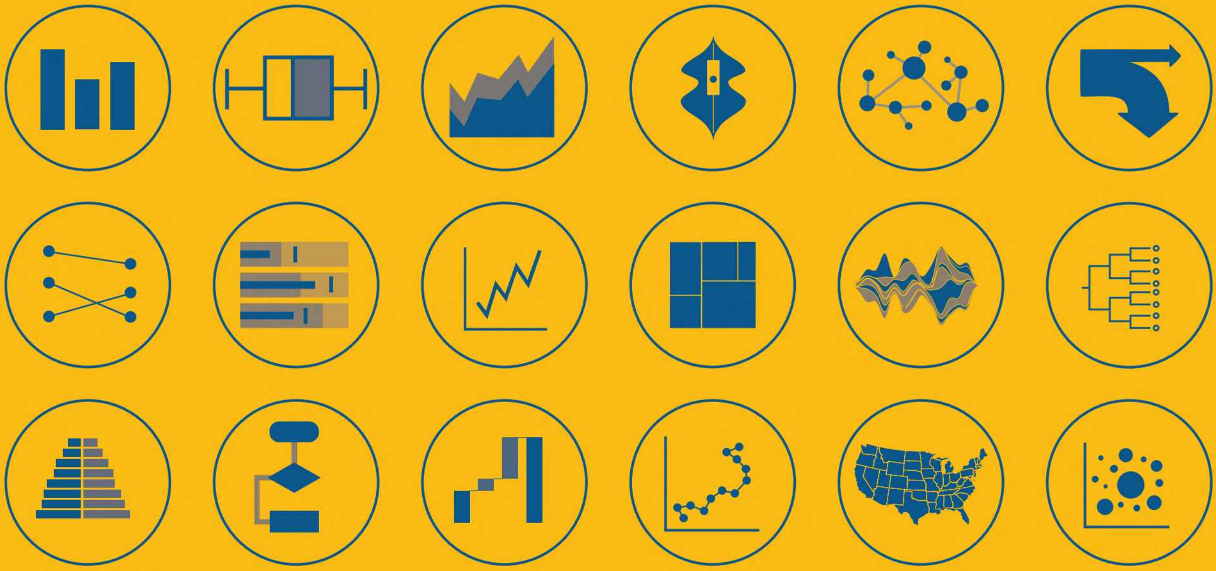
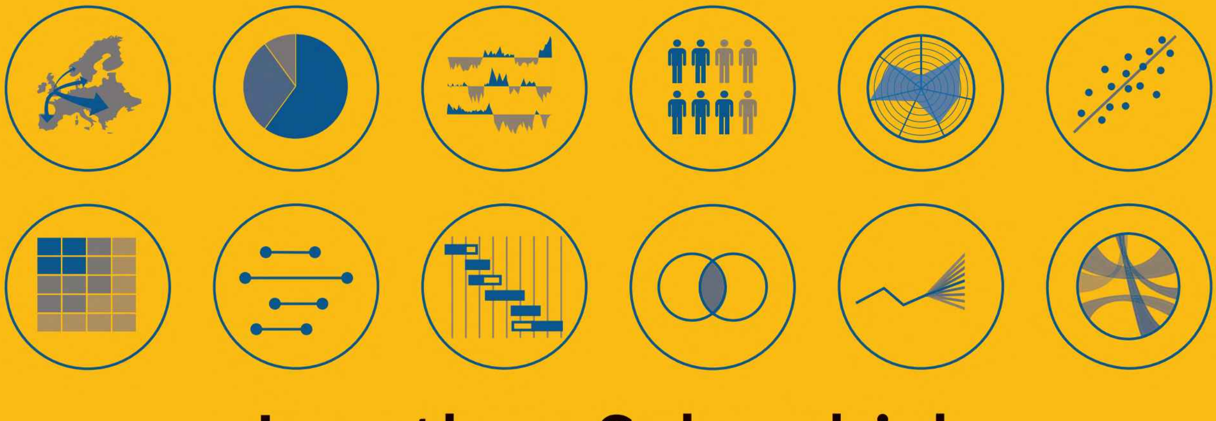


BETTER DATA VISUALIZATIONS

A Guide for Scholars, Researchers, and Wonks



Jonathan Schwabish

BETTER DATA VISUALIZATIONS

1

VISUAL PROCESSING AND PERCEPTUAL RANKINGS

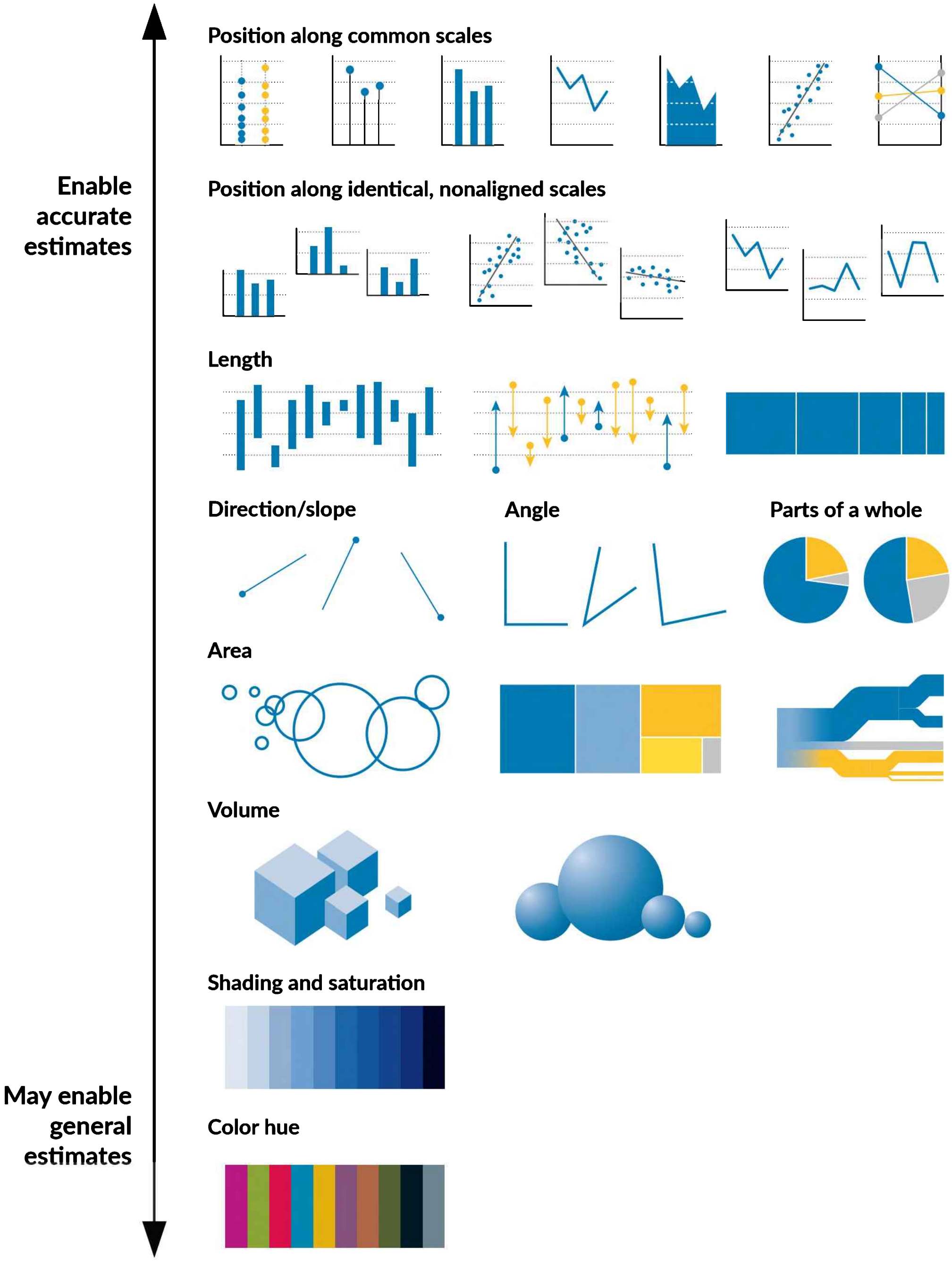
Before we start creating our charts and graphs, we need to cover some basic theory of how the brain perceives visual stimuli. This will guide you as you decide what chart type is most appropriate to visualize your data.

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

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

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

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


Enable accurate estimates
May enable general estimates
Position along common scales
Position along identical, nonaligned scales
Length
Direction/slope
Angle
Parts of a whole
Area
Volume
Shading and saturation
Color hue

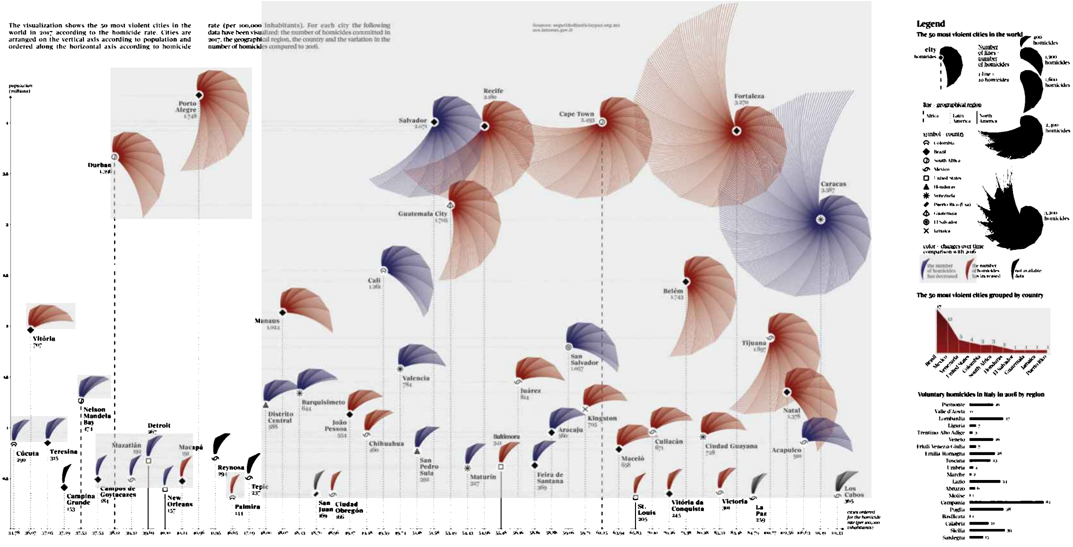
Perceptual ranking diagram. What kind of data visualization you choose to create will depend on your goals and your audience's needs, experiences, and expertise. This image is based on Alberto Cairo (2016) from research by Cleveland and McGill (1984), Heer, Bostock, and Ogievetsky (2010), and others.

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

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

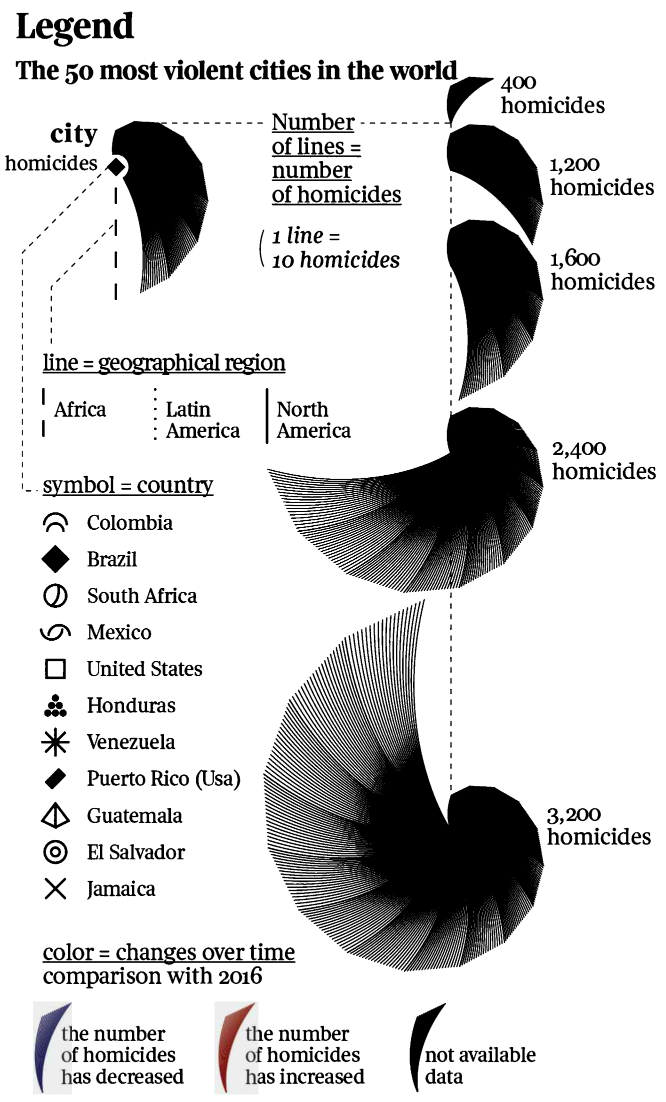
Spurring readers to engage with a graph is sometimes just as important. Sometimes, it's more important. And nonstandard chart types may do just that. In some cases, nonstandard graphs may help show underlying patterns and trends in better ways that standard graphs. In other cases, the fact that these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

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



The visualization shows the 50 most violent cities in the world in 2017 according to the homicide rate. Cities are arranged on the vertical axis according to population and ordered along the horizontal axis according to homicide rate (per 100,000 inhabitants). For each city the following data have been visualized: the number of homicides committed in 2017, the geographical region, the country and the variation in the number of homicides compared to 2016.

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



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

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

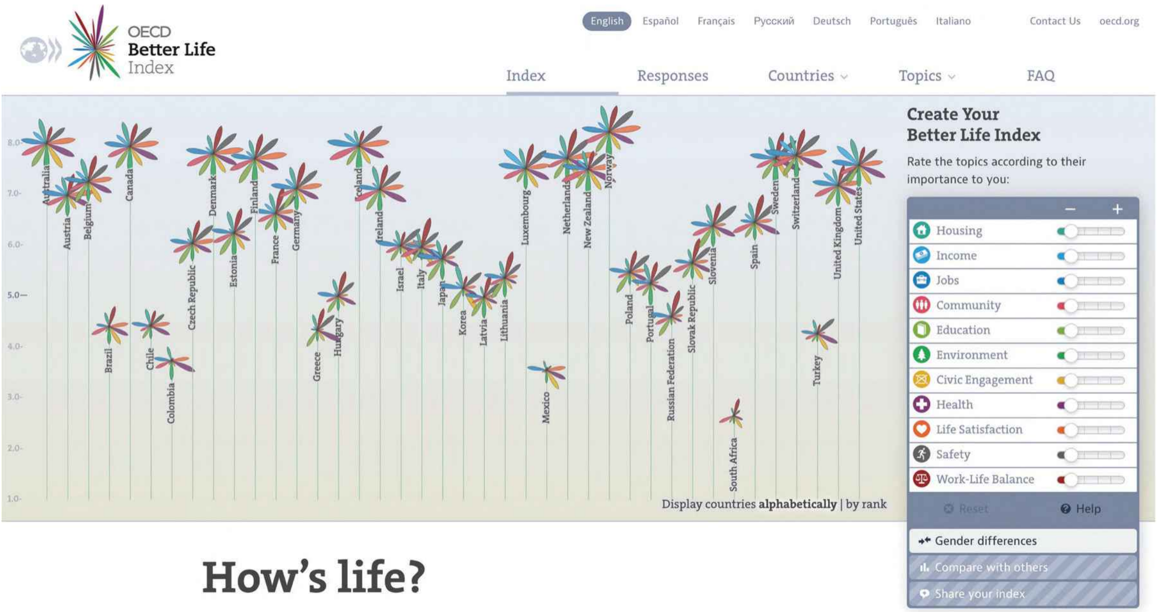
Data visualization is a mix of science and art. Sometimes we want to be closer to the science side of the spectrum—in other words, use visualizations that allow readers to more accurately perceive the absolute values of data and make comparisons. Other times we may want to be closer to the art side of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

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

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

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

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



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

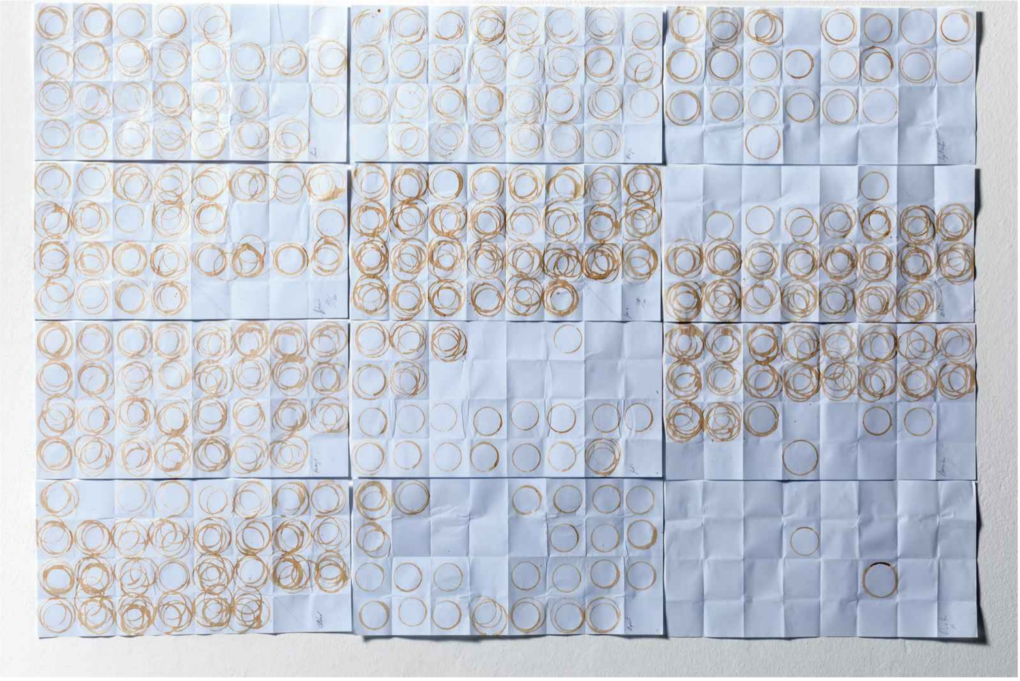
Source: Organisation for Economic Co-Operation and Development

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

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

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

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



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

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

ANSCOMBE'S QUARTET

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

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

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

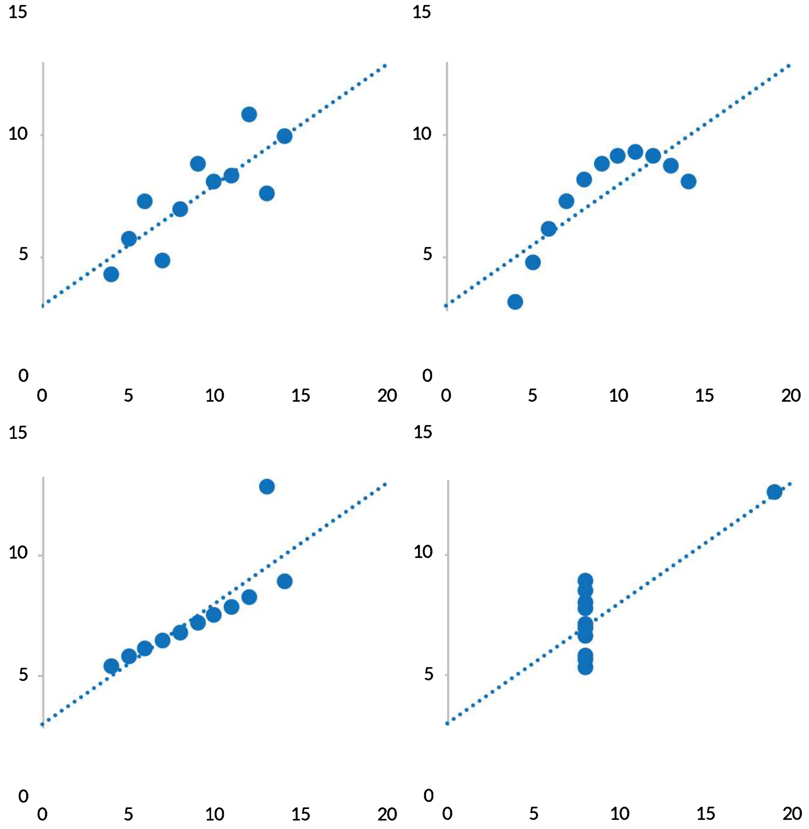
Data set		1	1	2	2	3	3	4	4
Variable		x	y	x	y	x	y	x	y
Obs. No.	1 :	10	8.0	10	9.1	10	7.5	8	6.6
	2 :	8	7.0	8	8.1	8	6.8	8	5.8
	3 :	13	7.6	13	8.7	13	12.7	8	7.7
	4 :	9	8.8	9	8.8	9	7.1	8	8.8
	5 :	11	8.3	11	9.3	11	7.8	8	8.5
	6 :	14	10.0	14	8.1	14	8.8	8	7.0
	7 :	6	7.2	6	6.1	6	6.1	8	5.3
	8 :	4	4.3	4	3.1	4	5.4	19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

Source: Francis Anscombe

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

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

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



The data visualization representation of Anscombe's Quartet. Notice how much easier it is to see the positive relationship between the two variables, the curvature in the pattern in the top-right graph, and the outliers in the bottom two graphs.
Source: Francis Anscombe (1973).

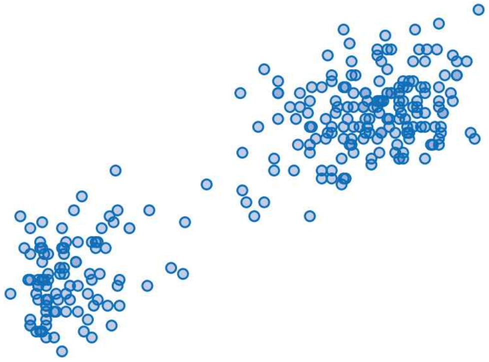
GESTALT PRINCIPLES OF VISUAL PERCEPTION

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

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

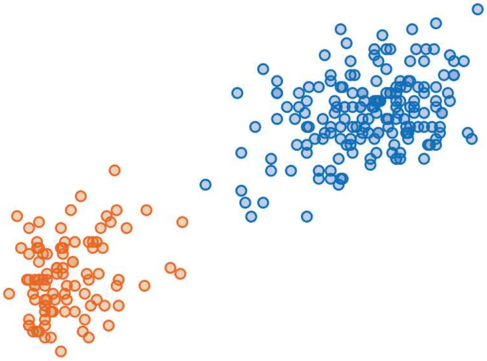
PROXIMITY

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



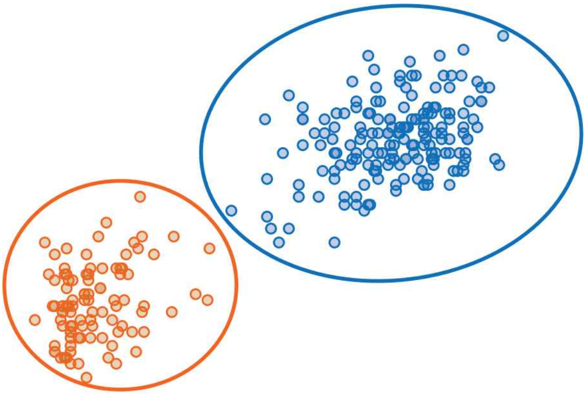
SIMILARITY

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



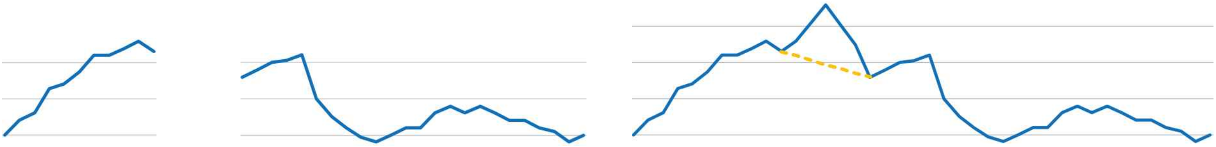
ENCLOSURE

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



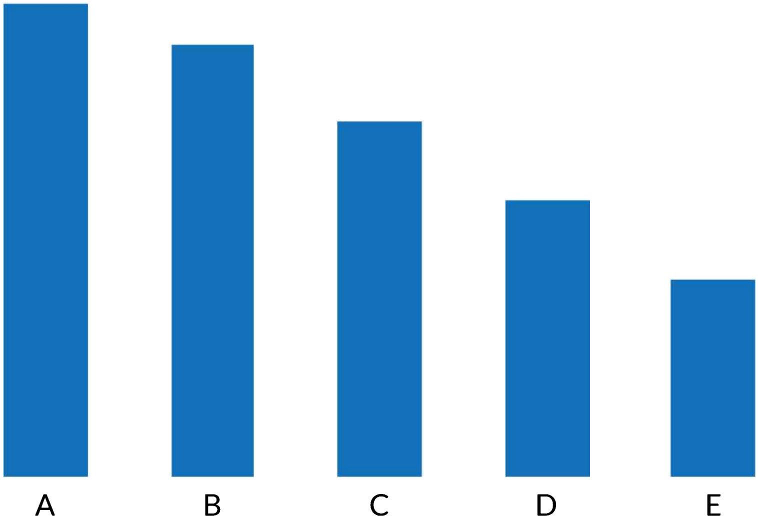
CLOSURE

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



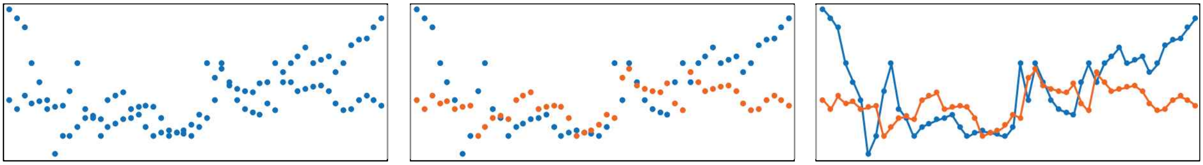
CONTINUITY

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



CONNECTION

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



PREATTENTIVE PROCESSING

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

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

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

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

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

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

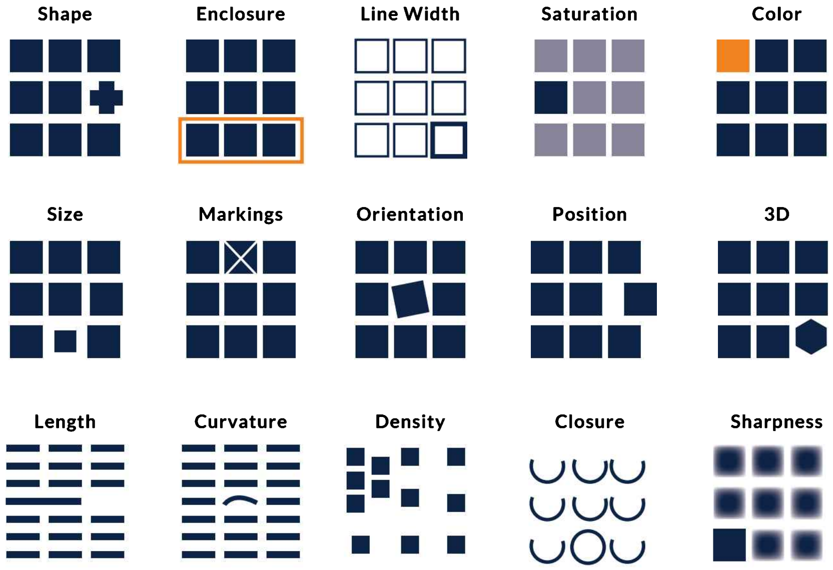
	Q1	Q2	Q3	Q4
Bob	26	35	72	84
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	95	13	64
Jon	83	62	46	48
Karen	30	65	98	82
Ken	38	28	45	71
Lauren	98	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

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

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

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

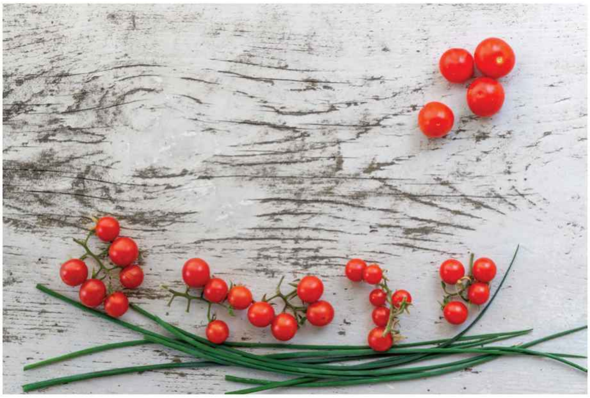
It's easier to find the numbers in these two tables than the first because the numbers are encoded using *preattentive attributes*: color and weight. Each distinction helps us effortlessly identify the key number.



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

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

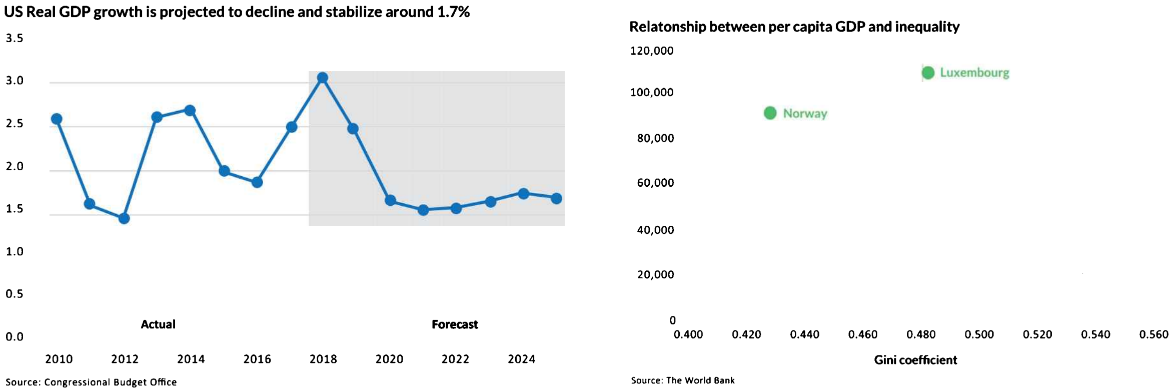
Preattentive processing works in photographs too. Consider these images of fruits and vegetables. In the photo on the left, the eye is drawn to the upper-right corner. The group of tomatoes is slightly larger than the rest and positioned away from the group. In the photograph on the right, however, the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



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

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

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



US Real GDP growth is projected to decline and stabilize around 1.7%

Source: Congressional Budget Office

Relatonship between per capita GDP and inequality

Source: The World Bank

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

WRAPPING UP

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