Visualizing with Clarity

During the exploration phase, you get to look at your data from a variety of angles and browse various facets, without having to dwell on charting standards and clarity. You understand a chart better because you know more about the data after you examine lots of other quickly generated charts. However, when you use graphics to present results to other people, you must make your graphics readable to those who don't know your data as well as you do.

A common mistake is that all visualization must be *simple*, but this skips a step. You should actually design graphics that lend clarity, and that clarity can make a chart "simple" to read. However, sometimes a dataset is complex, so the visualization must be complex. The visualization might still work if it provides useful insights that you wouldn't get from a spreadsheet.

As an effort toward clarity, people often preach removing all elements of a graphic that don't help you interpret the data. When you "let the data speak," you have done your job. This is fine, but it assumes the only goal of visualization is quick analytical insight, which is a small subset of what you can get out of data. It's okay to ponder and reflect, and elements that are not helpful in one situation might be helpful in another.

That said, whether it's a custom analysis tool or data art, make graphics to help others understand the data that you've abstracted, and try your best not to confuse your audience. How do you do this? Learn how we see data, and use that to your advantage.

VISUAL HIERARCHY

When you look at visualization for the first time, your eyes dart around trying to find a point of interest. Actually, when you look at anything, you tend to spot things that stand out, such as bright colors, shapes that are bigger than the rest, or people who are on the long tail of the height curve. Orange cones and yellow signs are used to alert you on the highway of an accident or construction because they stand out from the monotony of the black pavement. In contrast, Waldo is hard to find right away because he doesn't stand out enough to stick out in a sea of people.

You can use this to your advantage as you visualize data. Highlight data with bolder colors than the other visual elements, and lighten or soften other elements so they sit in the background. Use arrows and lines to direct eyes to the

point of interest. This creates a visual hierarchy that helps readers immediately focus on the vital parts of a data graphic and use the surroundings as context, as opposed to a flat graphic that a reader must visually rummage through.

For example, Figure 5-1 is the scatterplot from the previous chapter that shows NBA players' usage percentage versus points per game. The dots, fitted line, grid, border, and labels are of the same color and thickness, so there is no clear visual focus. It's a flat image, where all the elements are on the same level.

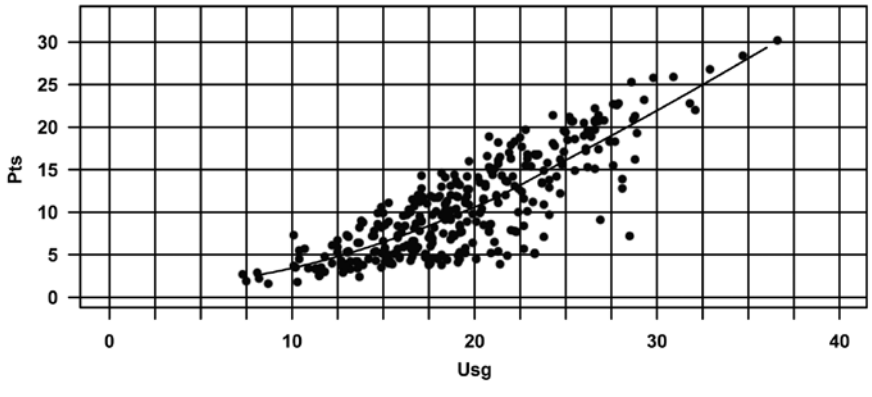


FIGURE 5-1 *All visual elements on the same level*

This is easily remedied with a few small changes. In Figure 5-2, the line width of the grid lines is reduced so that they are no longer as thick as the fitted line. In this example, you want the data to stand out. The grid lines also alternate in width so that it is easier to see where each data point lies in the coordinate system, and there's no imaginary blur that you get in the original chart.

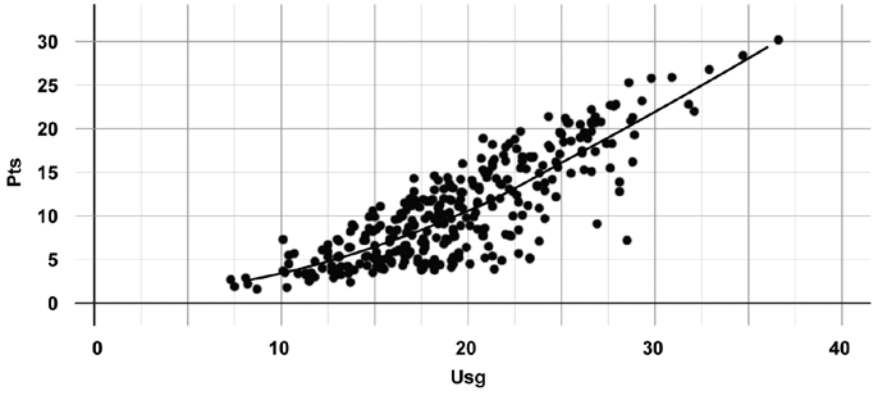


FIGURE 5-2 *Width of grid lines reduced to fit in background*

Still though, the fitted line is obscured by all the dots, because (1) it's thin compared to the radius of each dot and (2) it still blends in with the grid behind it. Figure 5-3 changes the color to blue to make the data stand out more, and the width of the fitted line is increased so that it clearly rests on top of the dots.

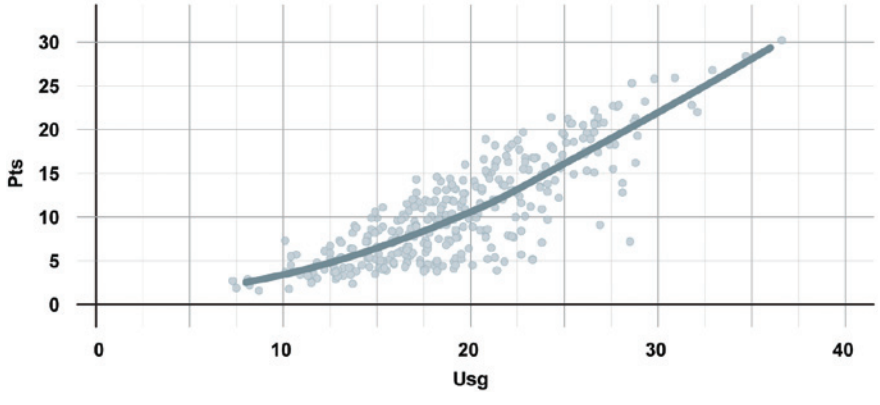


FIGURE 5-3 *Focus of chart shifted to fitted line with color and width*

The chart is a lot more readable now, but if you imagine people viewing the graphic like they would a body of text—from top to bottom and left to right—more descriptive axis labels and less prominent value labels can help, as shown in Figure 5-4. The text within the chart works similar to how it does in an essay or a book. Headers are often printed bigger and in a bold font to provide both structure and a sense of flow. In this case, the bolder labels provide immediate context for what the chart is about. Also, notice fewer and less prominent gridlines, which directs focus further to the upward trend.

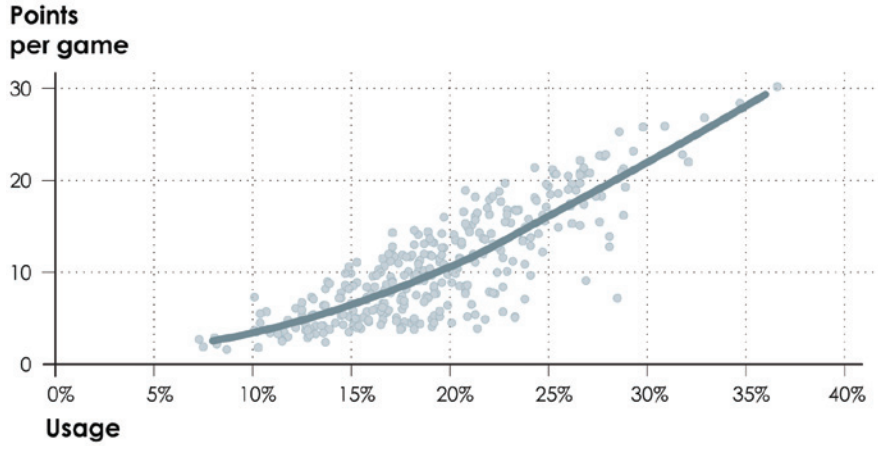


FIGURE 5-4 *Grid and value labels adjusted and fewer, less prominent gridlines*

Even if a graphic is exploratory or made to show an overview rather than a specific point or story in the data (such as a trend line), you can still use a visual hierarchy to provide structure. A presentation of a lot of data at the same time can be visually intimidating, but a breakdown by category helps readers browse visually. For example, Figure 5-5 shows 2,000 films across 20 genres over 100 years.

FIGURE 5-5 *(following page)* The History of Film *(2012) by Larry Gormley, http://historyshots.com*

Each layer alternates in color to separate genres, which makes the chart easier to read left to right, even if the name of the genre is not within view. Font and color separate genres (medium and red) and film titles (small and black), and the timeline on the bottom shows a division of film eras with tick marks. Had the same colors and fonts been used throughout, as in the scatterplot in Figure 5-1, it'd be a headache to browse this poster.

Sometimes visual hierarchy is used to show process or reflect how you might explore a dataset. Imagine you generate a lot of charts during the data exploration phase. You make a few graphs to see an overall picture, and in that summary, you note specifics and then make charts that focus on those. You can design your graphics to follow this same logic, basically taking readers on a tour of your analysis.

The bottom line: Graphics that follow a visual hierarchy are easier to read and can be used to guide readers toward points of interest. In contrast, flat graphics that lack flow make it harder for readers to interpret results and discourages closer looks. You don't want that.

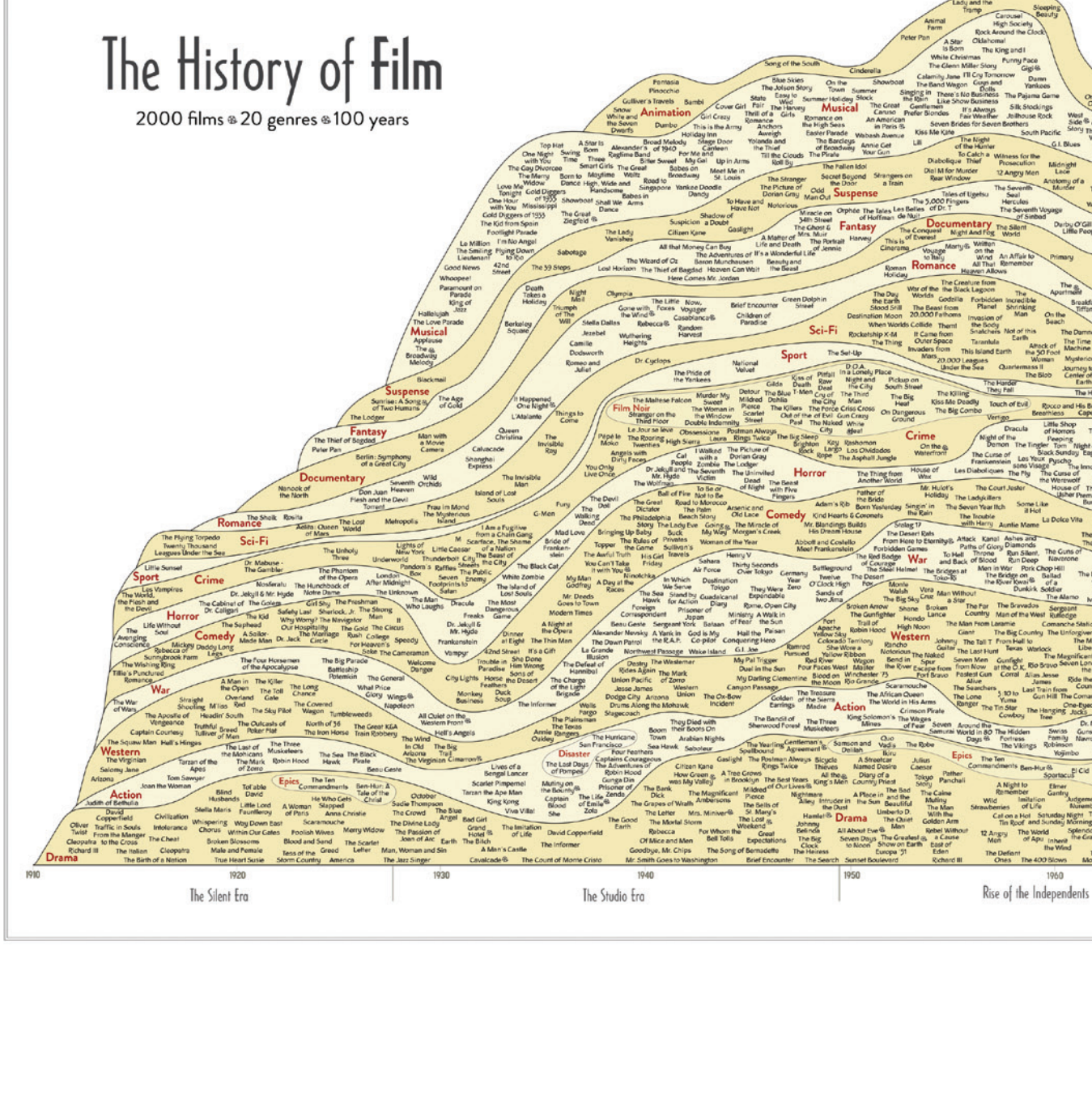
READABILITY

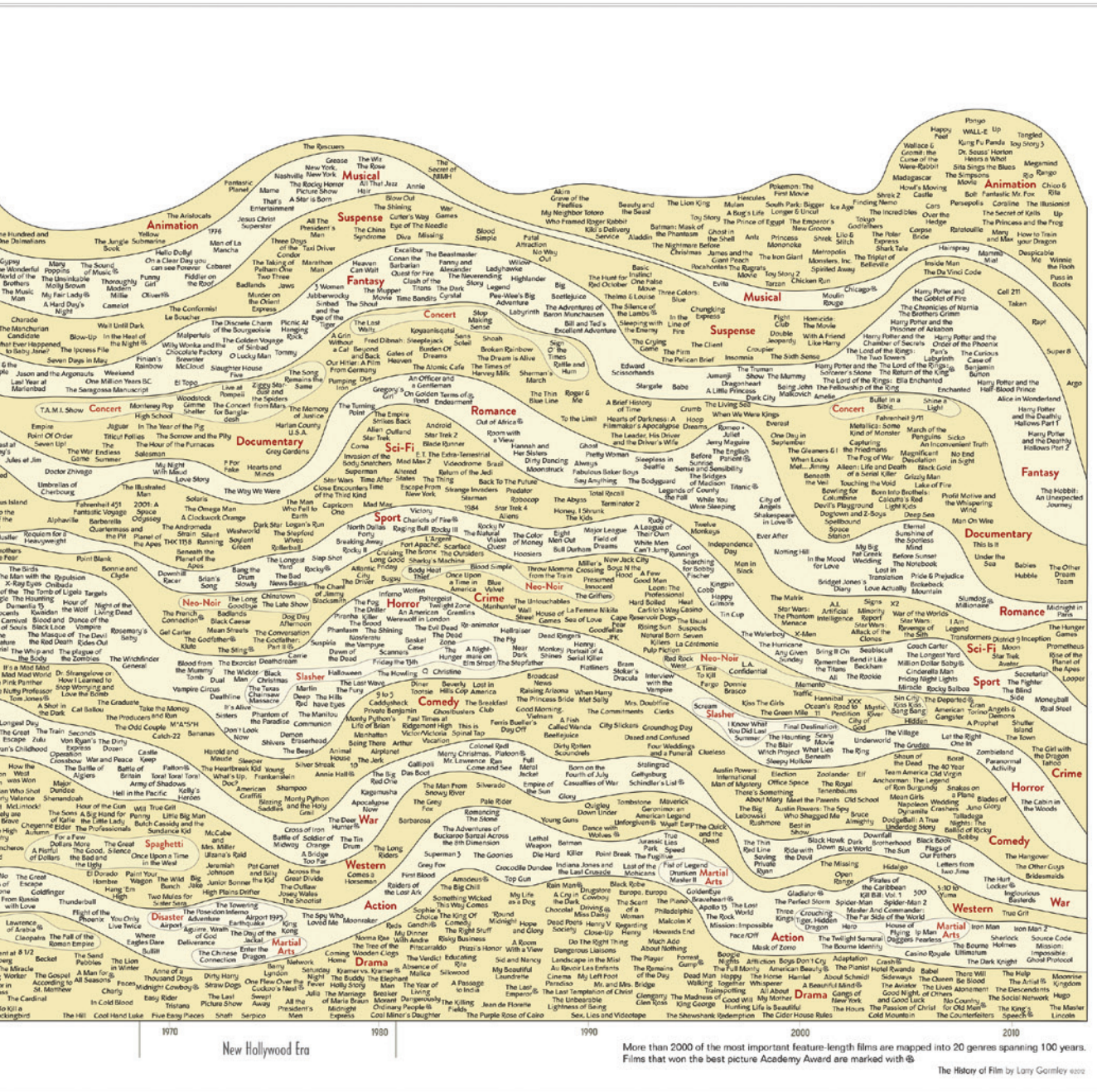
An author who uses words to describe a world or character interactions makes abstractions so that a reader can picture what's going on. Poor descriptions and little character development challenge readers to make sense of what seem like obscure clues. If readers can't connect the dots and understand what the author tries to describe, the words lose their value.

Similarly, you encode data with visual cues when you visualize it, and then you or others have to decode the shapes and colors to draw insights or to understand what a visualization represents, as shown in Figure 5-6. If you don't describe the data clearly, which makes a data graphic readable, then the shapes and colors lose their value. The connection between the visual and the underlying data is broken, and you end up with a geometry lesson.

The History of Film

2000 films ❀ 20 genres ❀ 100 years





More than 2000 of the most important feature-length films are mapped into 20 genres spanning 100 years. Films that won the best picture Academy Award are marked with ⊕

The History of Film by Larry Gormley ezra

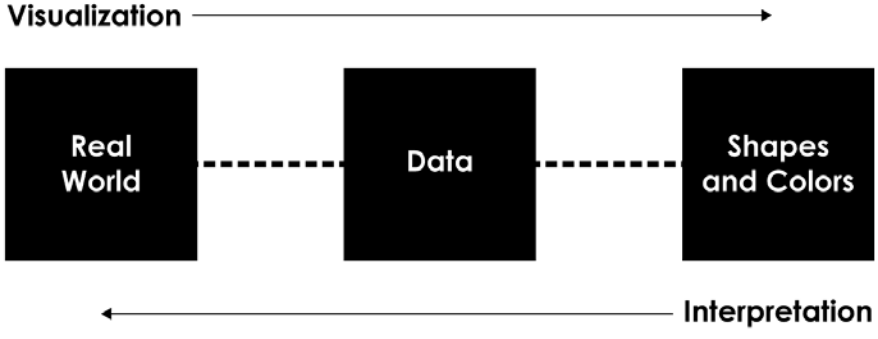


FIGURE 5-6 *Connecting visual cues to what data represents*

You must instead maintain the connection between visual cues and data because the data is what connects a graphic to the real world. So readability is key. Allow comparisons, consider the context of your data and what it represents, and structure shapes and colors (and the space around them) for clarity.

ALLOW COMPARISONS

Allowing comparisons across points is the main purpose of visualizing data. In table form, you can compare only point by point, so you place data in a visual context to see how big one value is relative to the rest and how all the individual data points relate to each other. As a way to better understand data, your visualization isn't useful if it doesn't fill this basic requirement. Even if you just want to show that values are equal across the board, the key is still to allow that comparison and conclusion to be made.

Traditional graphs, such as the bar, line, and dot plots that you've seen throughout this book were designed to make comparisons as straightforward and obvious as possible. They abstract the data into basic geometric shapes so that you can compare length, direction, or position. However, as shown in Figure 5-7, you can apply small variations to these charts that can make a graphic more or less challenging to read.

You saw how area should be used as a visual cue in Chapter 3, "Representing Data." When area is used to indicate values, determine the size of shapes such as bubbles and squares by their total area rather than the length of radius or side length. Essentially, the size of shapes are based on how people interpret them visually.

However, also keep in mind that it can be harder to see small changes between two-dimensional shapes than it is to see differences between position or length. This is not to say to avoid area as a visual cue. Instead, area is more useful when there are exponential differences between values. When small differences are important, look to a different visual cue, such as position or length.

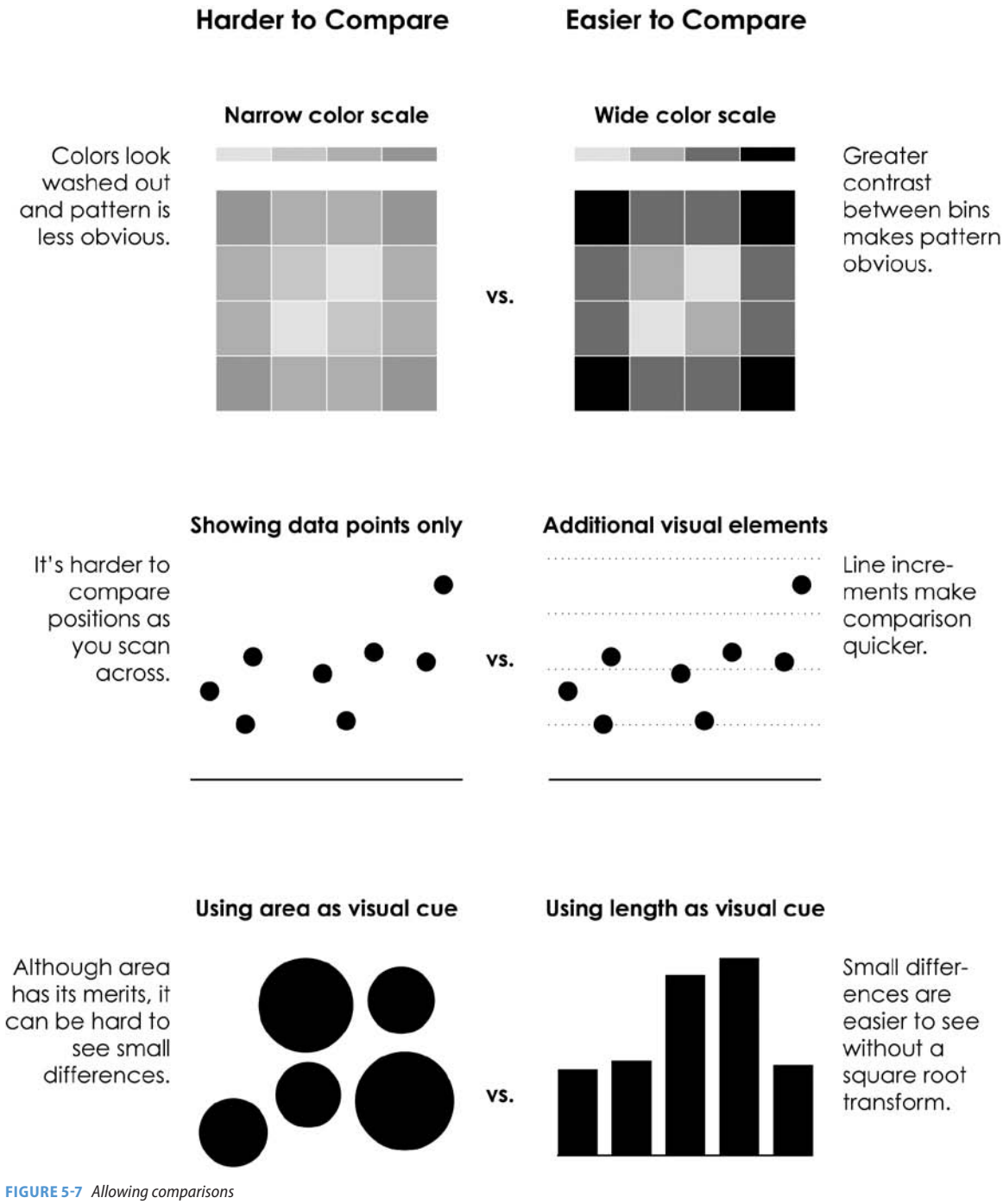


FIGURE 5-7 *Allowing comparisons*

For example, Figure 5-8 shows a number of identified species of invertebrates and vertebrates. The bar chart on the left and the bubble chart on the right show the same data, but because there so many more identified species of insects than vertebrates, the bars for the latter are dwarfed. They are barely visible, and the bar for corals is also just a sliver.

On the other hand, the bubbles let you put large and small counts in the same space. The downside is that you can't visually compare values as accurately as a bar chart, but in this case, the bar chart doesn't even give you a chance to compare the values. So there's a trade-off.

Note: Area can also make data seem more tangible or relatable, because physical objects take up space. A circle or a square uses more space than a dot on a screen or paper. There's less abstraction between visual cue and real world.

The graphic in Figure 5-9, made in 1912 when the Titanic crashed into an iceberg in the Atlantic Ocean, also places information within a familiar geographic context.

Each layer from top to bottom represents the time it would take to travel across the ocean via a 17th century ship (40 days), the Titanic (4 days), and by a not yet realized airplane (1 day). If only you could fly the Atlantic! Grid lines separate the modes of transportation as well as provide estimated travel times. Grids can also improve readability in more traditional charts because they dictate spacing and reflect scale.

Introduce color as a visual cue, and there are additional considerations. For example, you saw how those who are color-deficient see shades of red and green. If you use red and green hues with the same saturation, the colors look the same to those who are color-deficient. Color options also change based on what scale you use for a chart or what you want to show. As shown in Figure 5-10, there are three main categories of color scales, with variation within each.

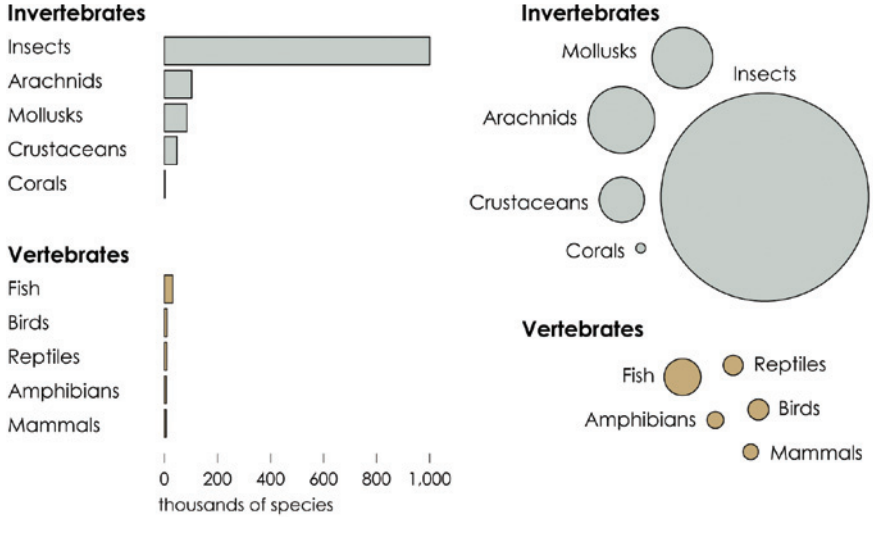


FIGURE 5-8 *Bar chart versus bubble chart to show large counts*

IF ONLY WE COULD FLY THE ATLANTIC!

ONE WAY BY WHICH THE ICEBERG DANGER WOULD BE AVOIDED

This diagram tells its own story of how we have conquered time and space. The problem of flying to America is now well within the bounds of possibility. Mr. James V. Martin, in an aeroplane fitted with floats, proposes to attempt a flight across the Atlantic, from Newfoundland to Ireland, next August, and, if successful, believes that he can cover the two thousand miles in forty hours. Once accomplished, it will not be long before the journey is made in a single day.

DRAWN BY SID TREEBY

FIGURE 5-9 If only we could fly the Atlantic! *(1912) by Sid Treeby*

FIGURE 5-10 *Color scale options*

A sequential color scale is used to represent a single variable without a separation requirement (positive versus negative, for example). Darker shades typically represent higher values and lighter shades represent lower values. You essentially choose a saturated hue and then decrease the saturation in increments to create a scale. With the sequential scales in Figure 5-10, the saturated hues are on the right and saturation is decreased as you shift left.

When you do have a natural or defined split in the data, such as increases and decreases or political leanings toward two parties, you can use a diverging color scale. It's like a combination of two (or more) sequential color scales with a separator in between to indicate a neutral value, such as a change of zero or a balance of political favor.

Qualitative color scales are useful when your data is categorical or non-numeric. Each color might represent a category, and the varying shades should provide visual separation.

Regardless of the type of color scale you use, there should be enough variation between your choice of hues and saturation so that you can see differences. Choose shades that are too similar and it's a challenge to make comparisons. A narrow color span restricts the amount of difference between shades, as shown on the left of Figure 5-11, whereas a wider color span on the right makes it easier to see differences. This works in the opposite direction, too. A color span that's too wide can exaggerate differences, and if

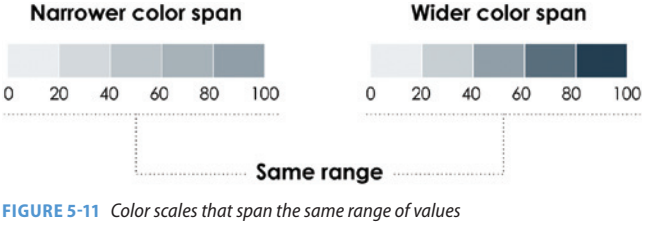


FIGURE 5-11 *Color scales that span the same range of values*

you don't pay attention to the context of the data, you might show patterns that look obvious but are not significant.

Figure 5-12 is the Cartesian equivalent. The space between each tick mark on the vertical scale is tiny, but because the span of the values is also small, the change in the line's position looks big.

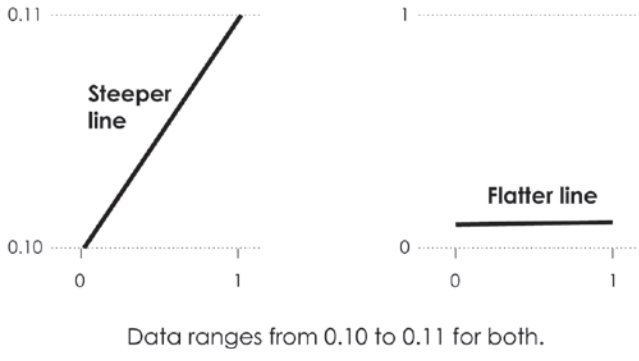


FIGURE 5-12 *Zoomed in on a Cartesian coordinate system*

Sometimes it makes sense to do this, and other times the zoom exaggerates what's actually there. A rule of thumb: Match the amount of visual change to the significance of the change in real life, and as always, represent the data fairly so that others can make fair comparisons.

REPRESENT CONTEXT

Context helps readers relate to and understand the data in a visualization better. It provides a sense of scale and strengthens the connection between abstract geometry and colors to the real world. You can introduce context through words that surround a chart, such as in a report or story, but you can also incorporate context into the visualizations through your choice of visual cue and design elements.

As shown in Figure 5-13, Stephen Von Worley showed the increased variety of colors in the Crayola crayon spectrum. In 1903, on the release of the first wax crayons under the brand name Crayola, there were just 8 colors. Over the years, Crayola inherited and created other colors in between the existing hues, and by 2010, there were 120 shades offered. In addition to red, there is now also bittersweet, brick red, mahogany, maroon, orange red, red orange, violet red, wild watermelon, radical red, razzmatazz, fuzzy wuzzy, and scarlet.

It makes sense to use the actual colors to represent the shades each year, to show the increase in diversity. A grayscale version would require a label for each shade and would quickly clutter by 1949.

Often your choice of visual cues changes based on the expectations of those you make a graphic for. A graphic that does not fulfill expectations can confuse readers. (I of course, mean this from a design perspective rather than a data one. Unexpected trends, patterns, and outliers are always welcome.)

Note: Choose geometry and color based on the context of your data. Software defaults are rarely, if not never, the best option.

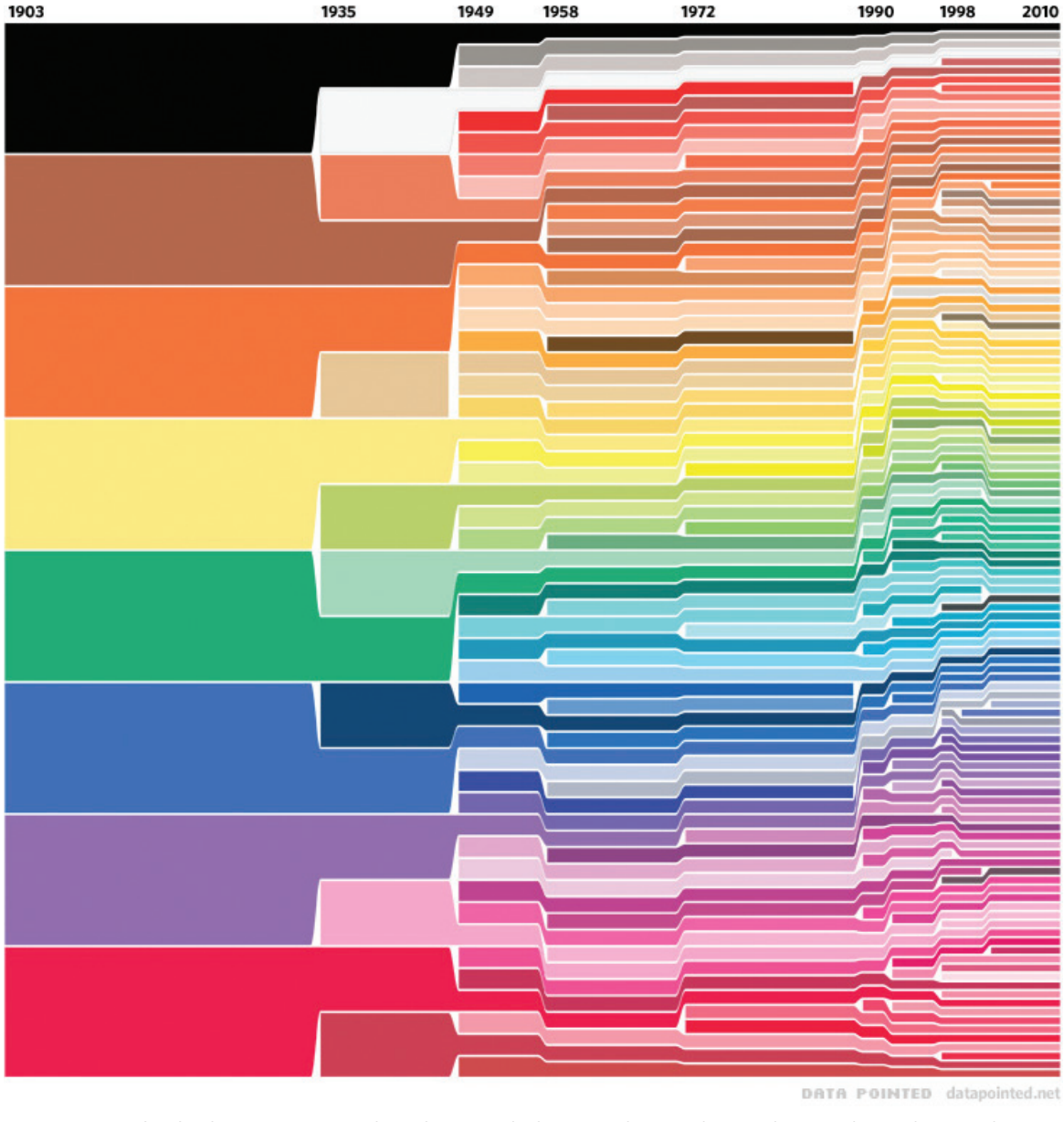


FIGURE 5-13 Crayola Color Chart 1903–2010 *(2010) by Stephen Von Worley, http://www.datapointed.net/visualizations/color/crayola-crayon-chart/*

For example, the United States has a two-party system, with Democrats and Republicans. Blue is the color of the Democratic party and red is the color of the Republican party. Therefore, a map, as shown in Figure 5-14, should reflect the party colors. Flip the colors, and the proportions between two groups would be the same, but because the party colors are so commonplace, it's probable that readers would misinterpret the results as a Barack Obama win in the Midwest and Southeast and a Mitt Romney in the West and Northeast.

FIGURE 5-14 *Color by expectation*

The charts in Figure 5-15 show movie trilogy ratings from the review aggregation site Rotten Tomatoes. On the site, a ripe red tomato is used for movies that earn at least 60 percent positive reviews (fresh), whereas a splattered green tomato is used for movies below the 60 percent threshold (rotten). The graphs match the site's color scheme so that you can easily see which movies were fresh and which were rotten. The length of each bar provides a more exact value.

Context can also affect your choice of geometry. For example, the Bureau of Labor statistics releases monthly estimates for number of jobs lost and gained. Figure 5-16 represents jobs lost between February 2008 and February 2010. More jobs were lost than gained every month during this period. The taller the bar is, the more jobs that were lost on the corresponding month.

The chart with values in the positive makes sense, but consider the context the chart is usually presented in. People expect to see bars in the positive for jobs gained and in the negative for jobs lost. However, the coordinate system in Figure 5-17 would put jobs gained in the negative. Negative jobs lost means new jobs.

So instead, it's more intuitive to frame jobs lost as negative values, as shown in Figure 5-17. It makes more sense to show something lost moving downward, when that something is looked at negatively. On the other hand, decreased weight, when the goal is actually to lose weight, might work better on the positive side of the axis.

FIGURE 5-15 *Color based on where the data comes from*

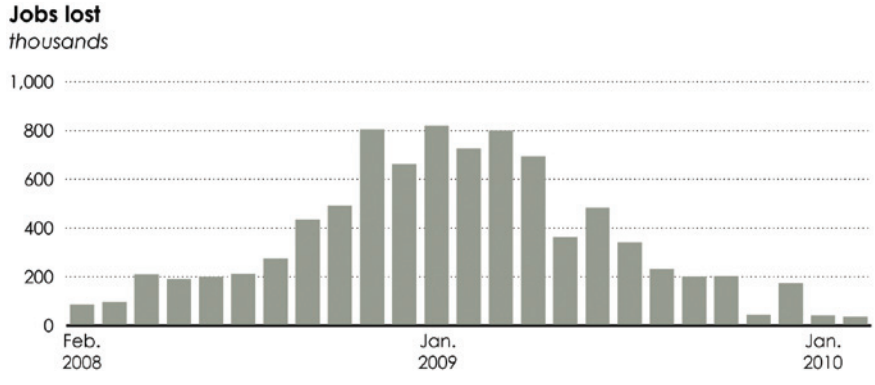


FIGURE 5-16 *Visualizing data generically*

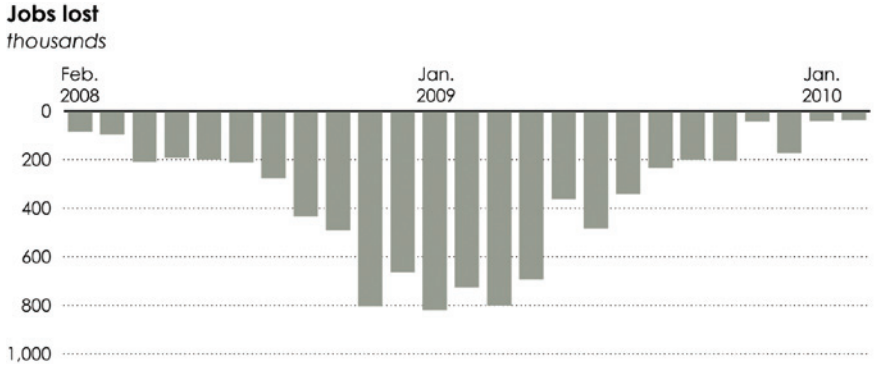


FIGURE 5-17 *Visualizing data in context*

NEGATIVE SPACE

Clutter is the enemy of readability. A lot of objects and words packed into a small area can make a visualization confusing and unclear, but put some space in between and it's often a lot easier to read. You can use space to separate clusters within a single visualization, or you can use space to divide multiple charts, so that they are modular and don't all run together. This makes a visualization easier to scan and mentally process piece-wise.

Figure 5-18 shows equally spaced rectangles, which appear to be in the same cluster, followed by ways to separate them with space and other elements, such as lines and contrasting colors. The space implies division (which you should keep in mind when you don't want to separate visual elements).

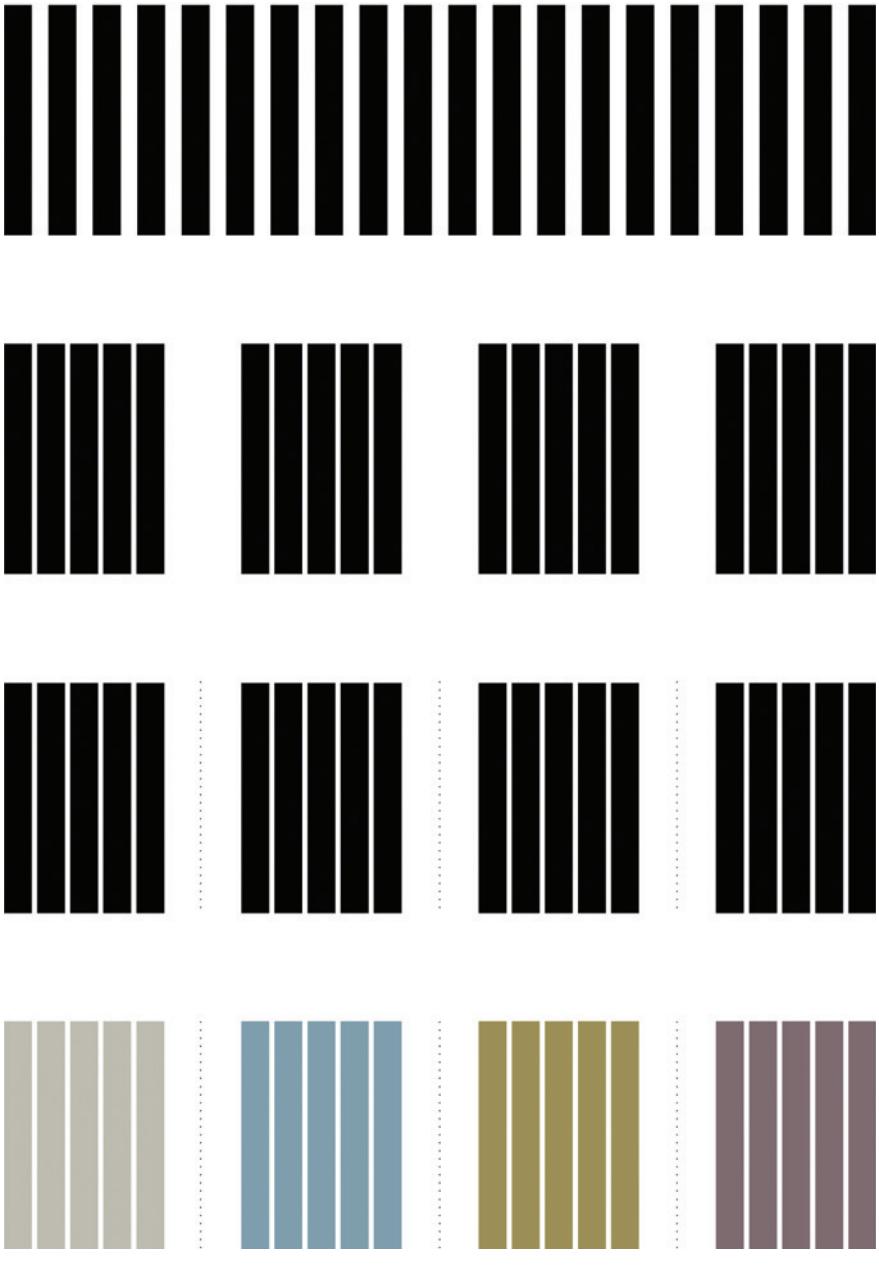


FIGURE 5-18 *Grouping shapes with space and other elements*

It's easy to see how this works in practice. Decrease the space between the labels and small charts (refer to Figure 5-15), and you get Figure 5-19. Although you can figure out which bars correspond to which labels by their positions, it is not immediately clear.

The same applies even if you don't want to show specific groups. Figure 5-20 is the map from Chapter 1, "Understanding Data," which shows fatal crashes in the United States. The top version uses small dots to show each accident, and the bottom version uses larger circles.

Because small dots are used in the top version, it's easier to see the pattern of roads and city centers. The negative space in between points help show where there are no roads or where fewer people drive cars. Places where there is no data is just as important as the places where there is data. On the other hand, the bottom version uses large circles that are relatively large compared to the size of the United States and the total number of crashes during the selected time period. There is practically no negative space, so roads and city centers are hidden by the data, and you only see country boundaries. Without a sufficient amount of negative space, the visualization is useless.

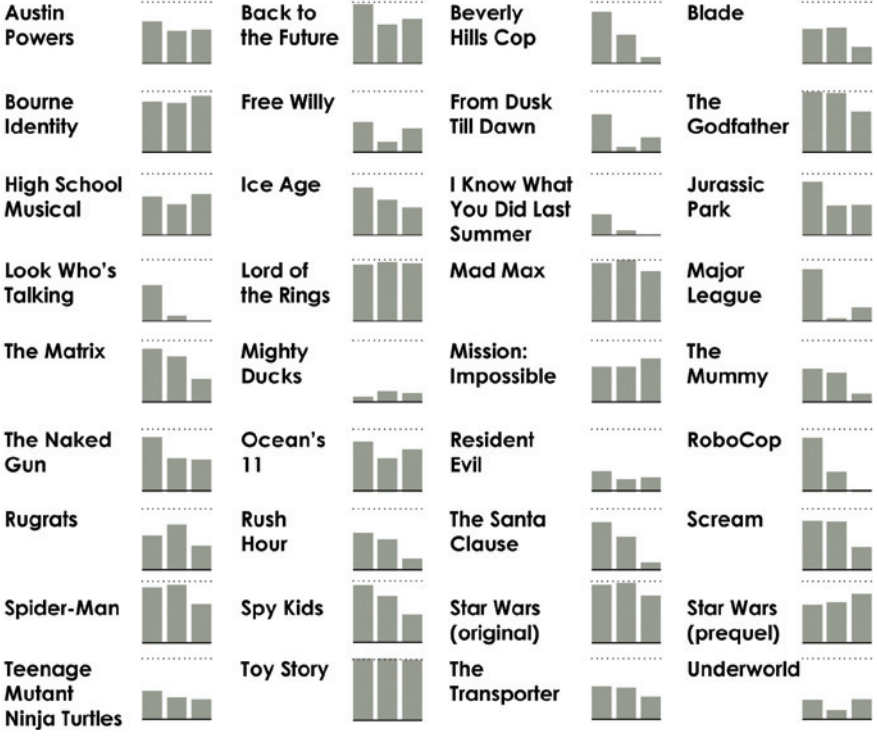


FIGURE 5-19 *Decreased negative space, decreased readability*

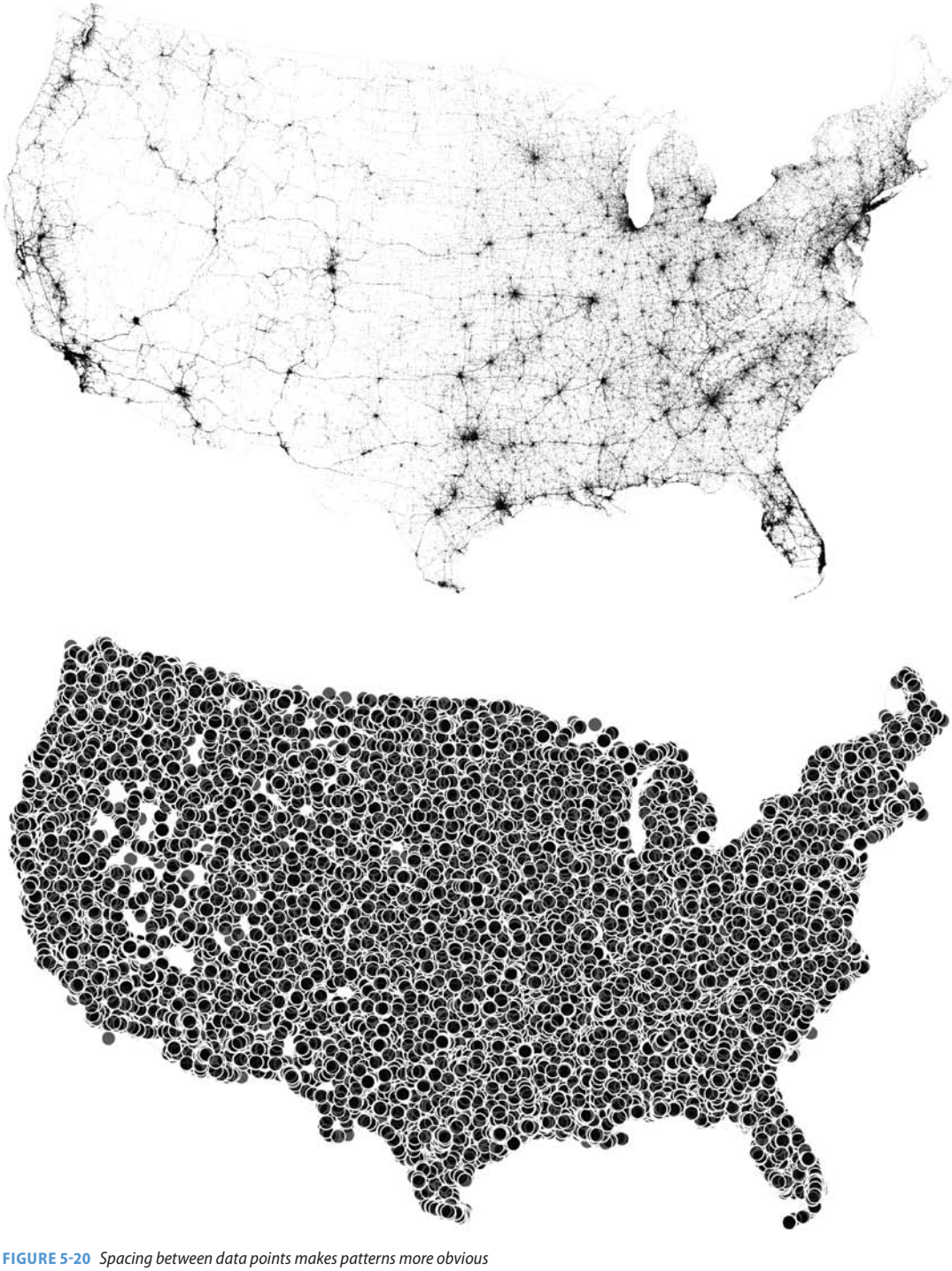


FIGURE 5-20 *Spacing between data points makes patterns more obvious*

You can also have too much negative space that interferes with the primary elements. Figure 5-21 shows a series of bars where the space in between each one is about the same as the width of a single bar. Look at the bars individually, and it's easy to see the separation between each gray bar, but look at the group as a whole, and you perceive a visual vibration between gray and white that almost makes the figure appear blurry. The equal negative space confuses your brain about which part to focus on.

FIGURE 5-21 *Visual vibrations from equal negative space*

On the other hand, reduce the negative space and increase bar width, and the image appears less blurry, as shown in Figure 5-22. The bars clearly dominate focus, and the negative space serves as thin separators in between.

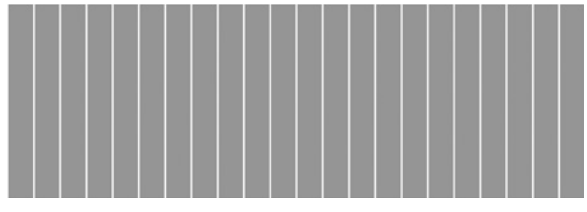


FIGURE 5-22 *Bars visually dominate with little negative space.*

As shown in Figure 5-23, the opposite direction works, too. The bars are thin strips and negative space is relatively large. Like in the earlier section on allowing comparisons between shapes and colors, contrast is the key.

With little differentiation between negative space and the elements of interest, a visualization is less clear, so experiment to find the right balance.

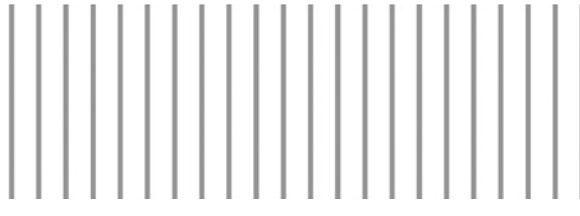


FIGURE 5-23 *Less vibration with a contrast between bars and negative space*

HIGHLIGHTING

Readability in visualization helps people interpret data and make conclusions about what the data has to say. Embed charts in reports or surround them with text, and you can explain results in detail. However, take a visualization out of a report or disconnect it from text that provides context (as is common when people share graphics online), and the data might lose its meaning; or worse, others might misinterpret what you tried to show.

Highlighting can guide readers through the data and direct eyeballs to the most important parts in a graphic. It reinforces what people might already see or draw attention to areas or data points that people should see.

To draw visual attention to a data point, you simply do what you would in real life. You make it stand out. Speak a little louder. Make it a little brighter. Edit an area or point in a visualization—while keeping the data, its visual cues, and readability in mind—to differentiate it from the rest. Use a brighter or bolder color, draw a border, thicken a line, or introduce elements that make the point of interest look different.

For example, Figure 5-24 shows how to use color to highlight a specific point. Most of the shapes are a neutral color, and the point of interest is purple, so attention immediately focuses on the parts that stand out.

Visualize time series data, and you might focus on specific years, such as in Figure 5-25. As you know, America loves their competitive eating, and no contest is more important than the annual hot dog eating contest on Coney Island. The top bar chart shows the number of hot dogs and buns that winners ate each year, but you can highlight bars to shift focus to years when someone broke a world record or when a certain person won.

On to more important matters: Figure 5-26 shows the world life expectancy chart from Chapter 2, "Visualization: The Medium," categorized by geographic regions. Each line represents a country's time series. The graphic shows all the countries that data was available for but shifts focus for each region. So the current point of interest is highlighted and brought to the front, and the rest are moved to the back and made a light gray, which remain for a sense of scale and context.

Again, to highlight elements, you make points of interest more visually prominent than the rest of a graphic. You place it higher in the visual hierarchy, so you either move the point of interest up or move everything else down. Elements on the same level get the same attention.

For example, Figure 5-27 shows the availability of movies that led at the box office, via streaming rental on iTunes, Amazon.com, and Vudu or the subscription-only Netflix. DVD availability is provided for reference. Availability is the point of interest, so a brighter color moves it up in the hierarchy, whereas neutral colors move other areas down. More specifically, rectangles highlighted yellow indicate that a movie was available via a service; an empty rectangle means not available; and a gray rectangle means the movie was only available for purchase.

FIGURE 5-24 *Examples of highlighting with color*

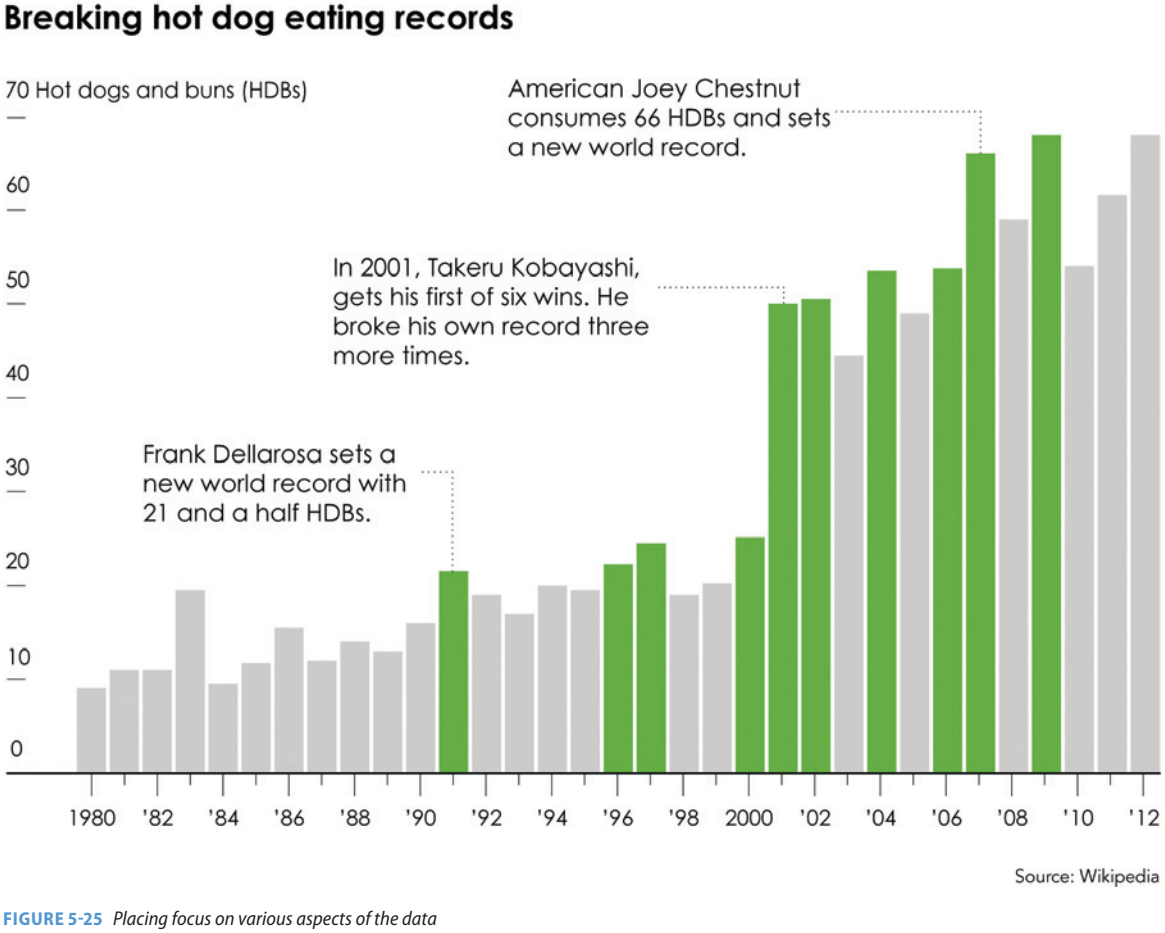


FIGURE 5-25 *Placing focus on various aspects of the data*

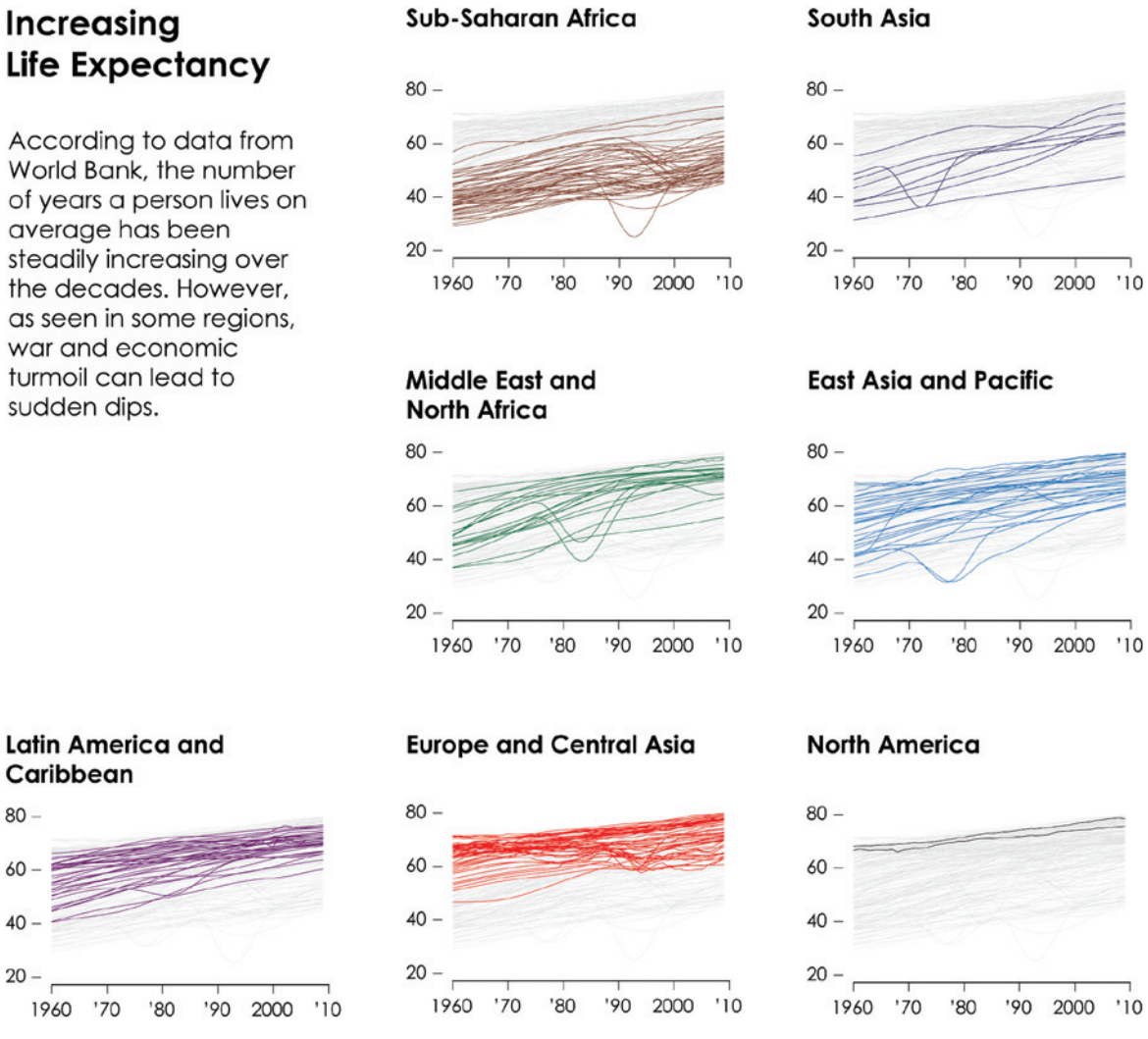


FIGURE 5-26 *Showing all data, but shifting focus with highlighting*

As shown in Figure 5-28, the focus easily shifts. A brighter color for nonavailable or shades of equal brightness take away focus from the main point of the graphic. A poor choice of color can also lead to misinterpretation, where it looks like Netflix has the most available streaming movies, even if the legend indicates otherwise.

Highlighting doesn't always have to be front and center. You can put it in the background, as shown in Figure 5-29. The unemployment time series data still keeps focus, but gray bars highlight periods of recession and provide information outside the primary dataset.

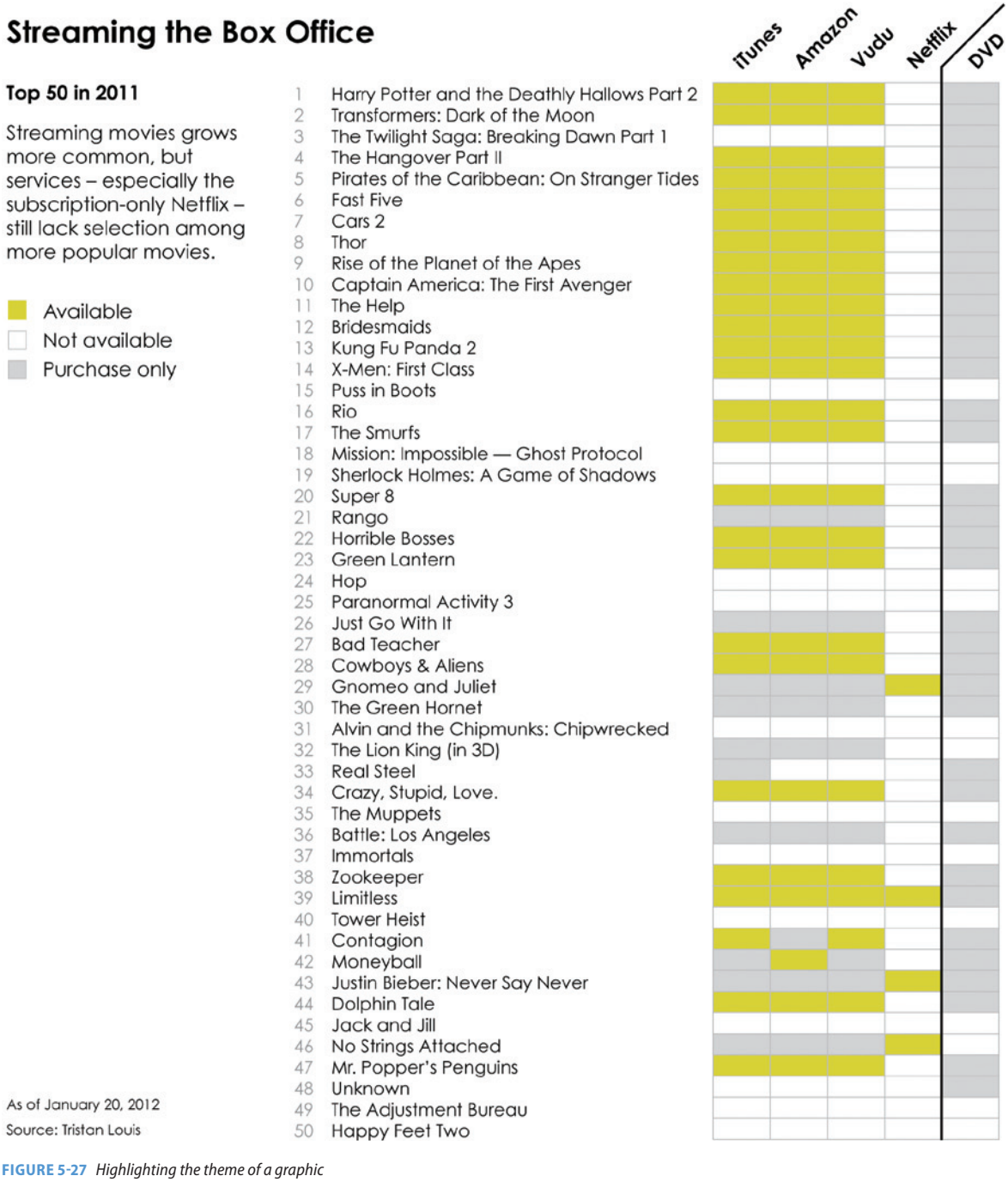


FIGURE 5-27 *Highlighting the theme of a graphic*

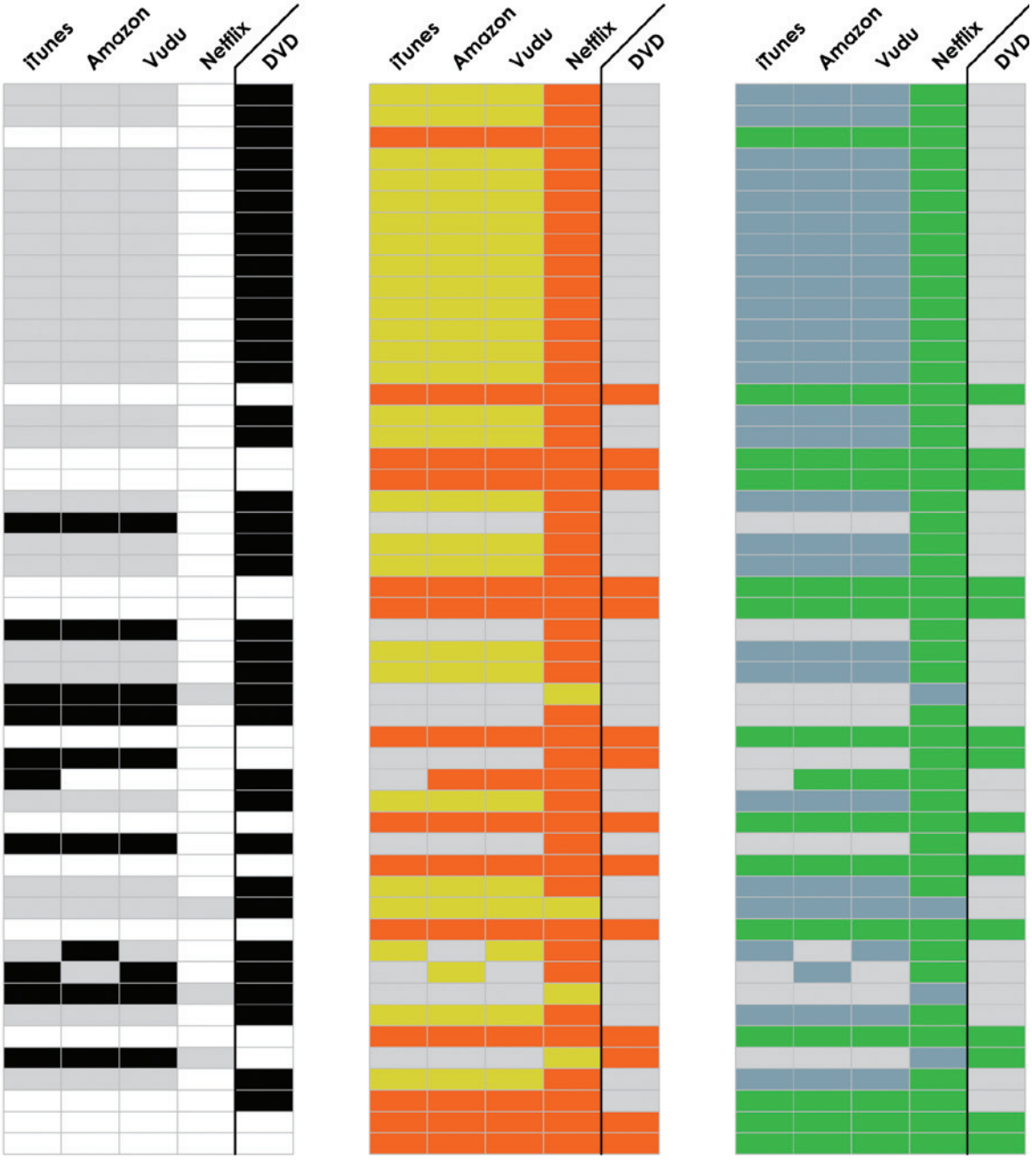


FIGURE 5-28 *Various color choices change the visual hierarchy.*

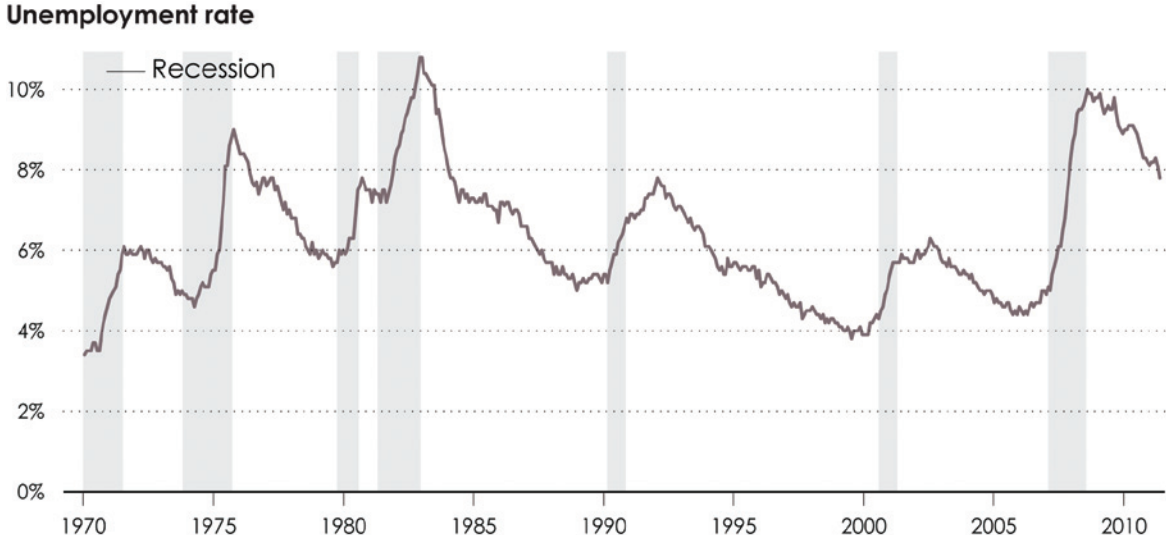


FIGURE 5-29 *Highlighting in the background*

However, wherever your highlighting fits in the hierarchy, be sure the new visual cues don't conflict with existing ones. If you have a bar chart that uses length as a visual cue, you obviously do not highlight with length, too. Have a scatter plot? Don't highlight with position. Heat map? Highlight with the color palette rather than introduce hues that change visual patterns.

Ask yourself how people decode information via shapes and colors in a visualization, and then don't get in the way. For example, in Figure 5-30 on the left is a bar graph with no highlighting, and the charts on the right show unsuccessful attempts to highlight. Why don't they work? A bar chart uses length as its visual cue, so when you extend a bar, the new length changes the value. Change width, and the new bar fills more area. (Bar charts actually use area to encode data, but because width stays constant, you can decode values via bar height.) Then on the far right: A shift up doesn't exactly change the value, but it makes the chart less readable.

Note: These conflicting visual cues are in the context of how you use bar charts, but you of course must consider conflicts within the context of your own visualization and how you encode your data.

In contrast, Figure 5-31 shows highlighting with visual cues not used by the bar graph. The color, border, and pointer send focus to the bar of interest but don't change the overall visual pattern.

Note: Highlight with unused visual cues. Otherwise, you change perceived patterns and make it more difficult to interpret the visualization.

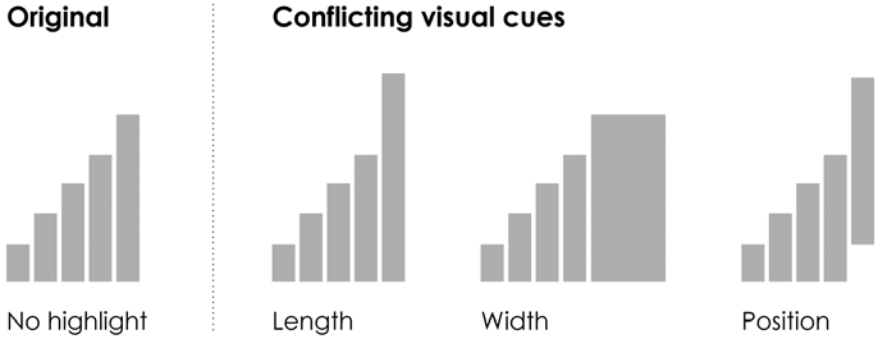


FIGURE 5-30 *Using conflicting visual cues to highlight*

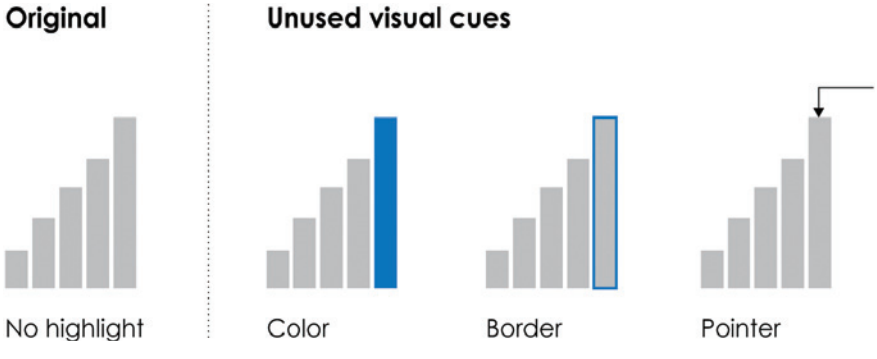


FIGURE 5-31 *Using unused visual cues to highlight*

ANNOTATION

When you highlight elements, it is not always obvious why, especially when readers aren't familiar with the data. (And they aren't most of the time.) Annotation within a visualization can help clearly explain what a visualization shows. What is that outlier? What does that trend mean? This might be left to text outside of a visualization, but when you put explanations within a graphic—as an additional layer of information—the visualization is self-encapsulated so that it's useful on its own.

EXPLAIN THE DATA

Like everything discussed so far, annotation follows a visual hierarchy. You have headers, subheaders, subsubheaders, and explanatory text. As shown

in Figure 5-32, size, color, and placement dictate how much attention annotations receive.

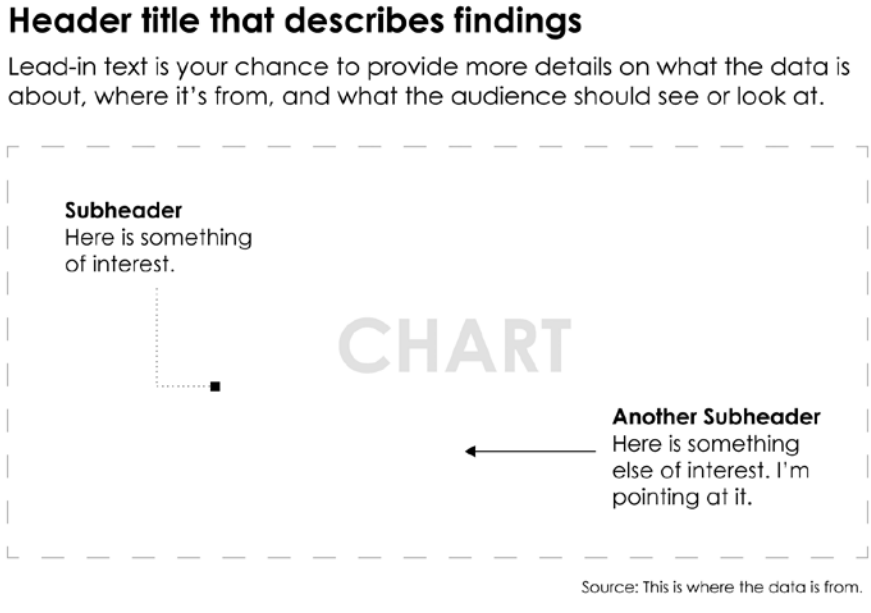


FIGURE 5-32 *Annotation for a chart*

The header is typically printed with larger and bolder fonts to set the stage or to describe what people should see or look for in the data. If the header is small and blends with everything else, people might skip it and look straight to the more visual elements. A descriptive title also helps. For example, "Rising Gas Prices" says more about a chart than just "Gas Prices." The former presents a conclusion immediately, and readers will look to the chart to verify and see details. The latter leaves data interpretation to readers and places them in the exploration phase. Then again, this might be your goal, so describe accordingly.

Lead-in text, like the header, is used to prepare readers for what a chart shows, but in further detail. The text is typically smaller than the header and expands on what the header declares, where the data is from, how it was derived, or what it means. Basically, it's information that might help others understand the data better but often doesn't directly point to specific elements.

To explain specific points or areas, you can use lines and arrows and use annotation as a layer on top of a chart. This places descriptions directly in the context of the data so that a reader doesn't have to look outside a graph for additional information to fully understand what you show.

For example, returning to the scatter plot in Figure 5-4, a layer of annotation is added, in addition to highlighting of specific points, as shown in Figure 5-33. Dark circles and pointers highlight specific players, and lines connect annotation to dots for the lowest scoring player with the lowest usage percentage, DeSagana Diop, and the highest scoring player with the highest usage percentage, Dwyane Wade. The point for Will Bynum, who somewhat strays from the trend, is also highlighted and annotated. There is also a pointer for the trend line and an explanation of usage percentage, which isn't common knowledge for most.

The key to useful annotation is to explain or highlight a chart as it relates to the data (and your audience). For example, the explanation for the trend line could be, "There is a positive correlation between points per game and usage percentage." This is true, but the generic statistical description doesn't relate to the context of the data. Similarly, you could describe Dwyane Wade as the player with the highest usage percentage and points per game, but

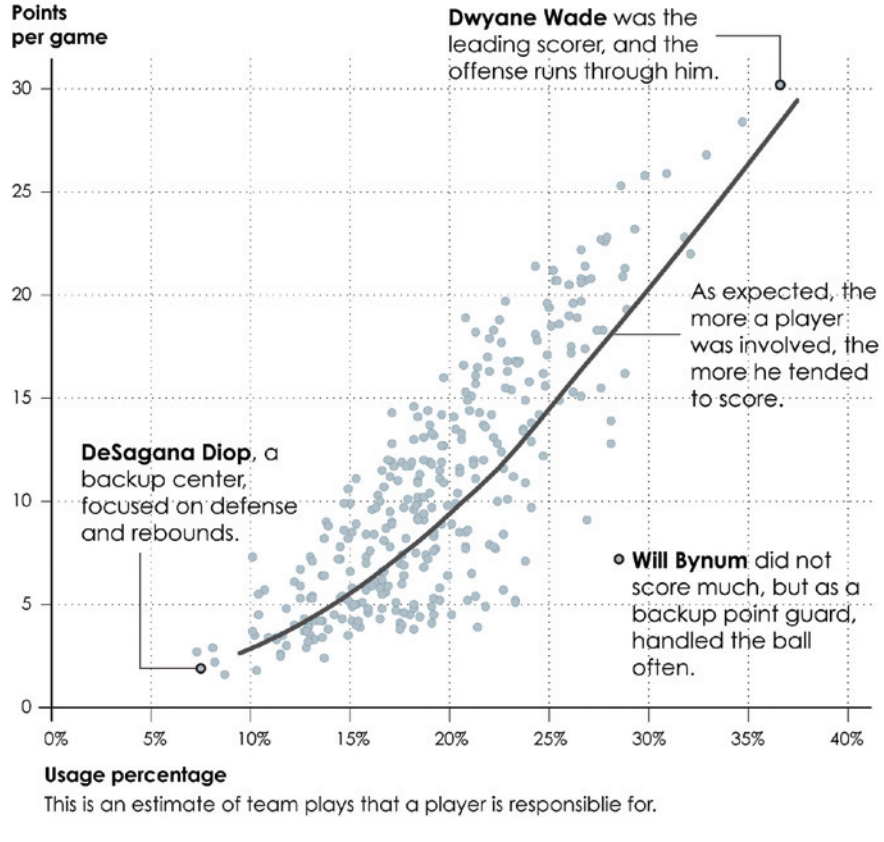


FIGURE 5-33 *Annotation added to scatter plot*

what does that say about him as a player? These are subtle changes that can greatly decrease or improve readability.

EXPLAIN STATISTICAL CONCEPTS

If a large proportion of your audience is unfamiliar with statistical concepts, you can annotate to explain or help them relate. The descriptions in the previous scatter plot of basketball players are an example. They don't just point out Dwyane Wade, DeSagana Diop, and Will Bynum. They also help explain what the corner positions, as well as a partial outlier, on an x-y plot mean so that readers can infer what positions in the middle represent. The pointer for the trend line is a description of correlation.

Figure 5-34 is another scatter plot, but it focuses on the gender pay gap in the United States, based on median salaries, according to the Bureau of Labor

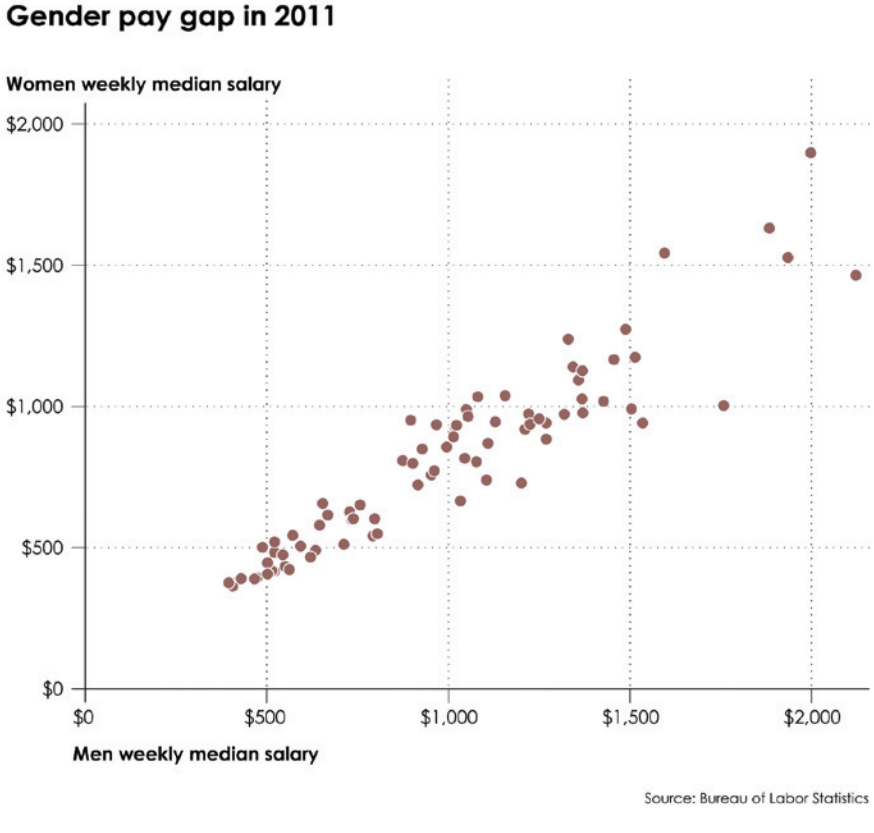


FIGURE 5-34 *Unannotated scatter plot*

Statistics. Each dot represents a profession, and men's median salary is plotted on the horizontal axis versus women's median salary on the vertical.

Without annotation, it's clear there is an expected upward trend between the two. With professions where men tend to make more, women tend to make more, too. If you look closely, you can also see that the dots tend toward the horizontal axis, which means men tend to make more with the same occupation.

The annotated chart in Figure 5-35 makes the pay difference clearer. A diagonal line through the middle represents equal pay, which is marked as such. Dots below the line are jobs where men make more than women, and dots above the line are where women tend to make more. These areas are also labeled.

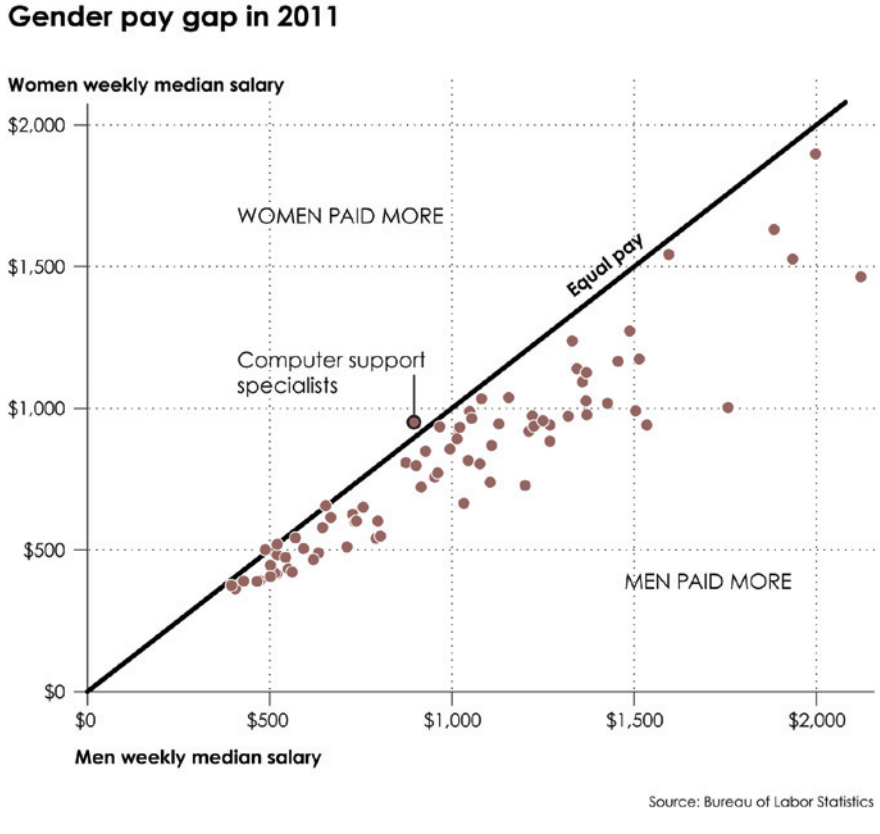


FIGURE 5-35 *Annotated scatter plot*

Computer support specialist is the only profession in this dataset where women tended to make more than men.

The annotation explains how to read the scatter plot and what the data means. Sure, many people know how to read a scatter plot and interpret relationships between two variables, but many don't, and it doesn't hurt to clarify.

Distributions are another challenging concept. People have to understand skew, mean, median, and variation, and that observations are aggregated across a continuous value scale when visualized.

For example, it is common for people to interpret the value axis of a histogram as time and the count or density on the vertical axis as a metric of interest. This leads to confusion, so it is useful to explain the various facets of a distribution.

In Chapter 4, "Exploring Data Visually," you saw distributions for flight arrival delays. Figure 5-36 shows the distribution of delays for Southwest Airlines. A negative delay means an early arrival, and a positive one means the plane arrived late to the destination airport. A delay of zero means an on-time arrival.

To clarify, simply add those descriptions as annotation on the histogram, as shown in Figure 5-37. Avoid jargon and explain in the context of the data.

In the end, you must consider what your audience will or might not understand graphically and statistically, and annotate based on that. Single variables, time series, and spatial data are easier to understand visually because they tend to be more intuitive than multiple variables or more complex relationships.

Note: Show people your visualizations to see how they interpret results. If they're confused, explain the data clearly.

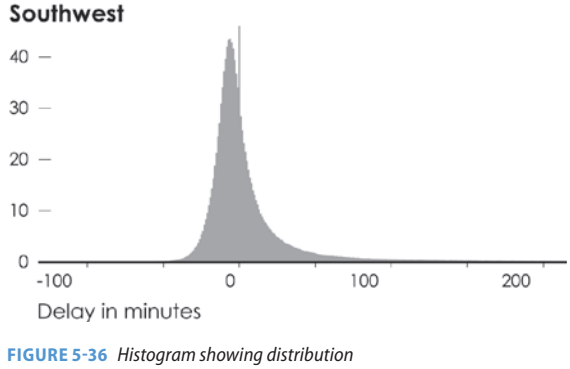


FIGURE 5-36 *Histogram showing distribution*

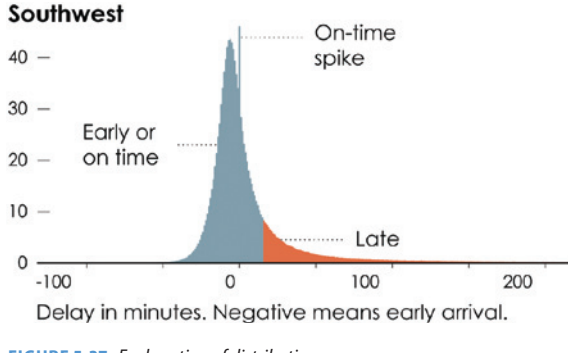


FIGURE 5-37 *Explanation of distribution*

EXPERIMENT WITH TYPOGRAPHY

There have been plenty of polls and questionnaires that ask what the best typeface is for visualization, but there's always a ton of variability, and there's never any consensus. This might be because taste in typefaces has a lot to do with personal preference. Nevertheless, it's worth exploring various fonts for labels and annotation, outside of software defaults, which are generic and less refined.

Note: A typeface is a design for text, such as Helvetica or Baskerville, and it pertains to the appearance of the characters. A font is a specification for a typeface, such as 10-point bold Baskerville.

The effect of typeface choice is most obvious at the extremes. Figure 5-38 shows the same graph with various typefaces, and you can see how readability and feel changes for each. For example, Helvetica, a sans-serif

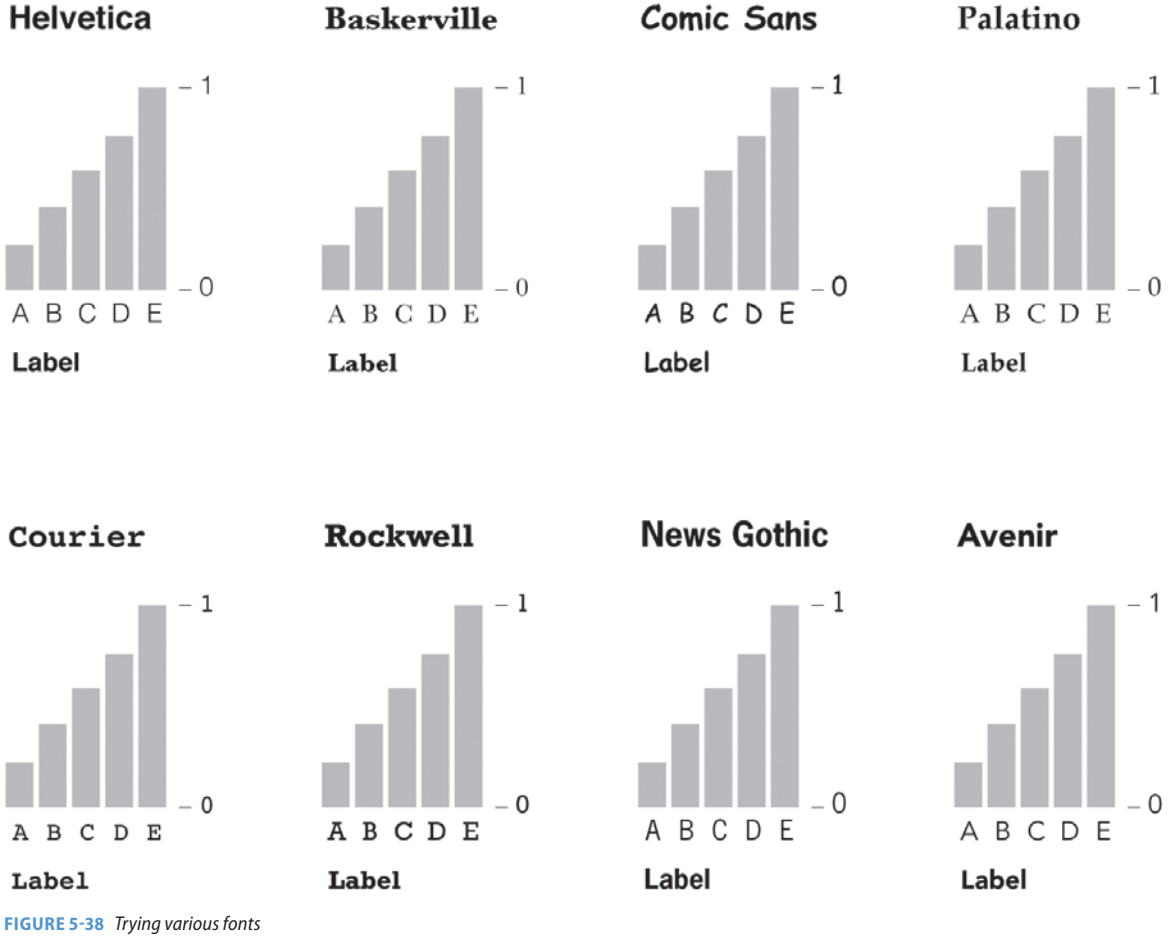


FIGURE 5-38 *Trying various fonts*

typeface, is known for its neutral look that lends to fact presentation, whereas Comic Sans has developed into a meme and a way to avoid being taken seriously. Serif typefaces such as Baskerville and Palatino are reminiscent of vintage graphics. You should probably avoid Wingdings for most practical uses.

Again though, a lot has to do with personal preference, so experiment and see what you like—especially because it's so easy to do with modern software. Remember the visual hierarchy, though. Headers typically stand out visually, so a larger, bold font often works, whereas tick labels are usually smaller and demand less attention, so the typeface should still be readable at a relatively small size. Sans-serif fonts often work well for the latter because serif fonts with a lot of flourish can be harder to read in confined spaces. Although, this is nowhere near a rule.

DO THE MATH

After you get data, the natural first step is to visualize it directly, but after that, it might be useful to do some math for a different point of view. This can shift focus toward something more interesting in the data and in some cases, avoid guesswork as readers try to interpret your graphics.

For example, summary statistics, such as mean or median, can serve as a quick point of reference or to provide a sense of scale, as shown in Figure 5-39. Violent crime rates for each state are shown, and bars are colored based on whether they are above the national average. The distributions of rates isn't especially complex in this example, but it helps you get a sense of where each state lies relative to the national average.

As an additional step, you can transform the data based on a reference point, rather than just show it in the context of the raw data. Figure 5-40 shows global gas prices, which you saw in the previous chapter, relative to average gas price in the United States. Purple indicates higher gas prices, and green indicates countries where gas prices were lower. The two maps show the same data but tell different stories via subtraction and division. The first map focused on worldwide comparisons, whereas this map provides a simple connection between the data and U.S. readers.

What about Figure 5-29 that shows the unemployment rate over time? Maybe you're more interested in annual changes than you are monthly unemployment

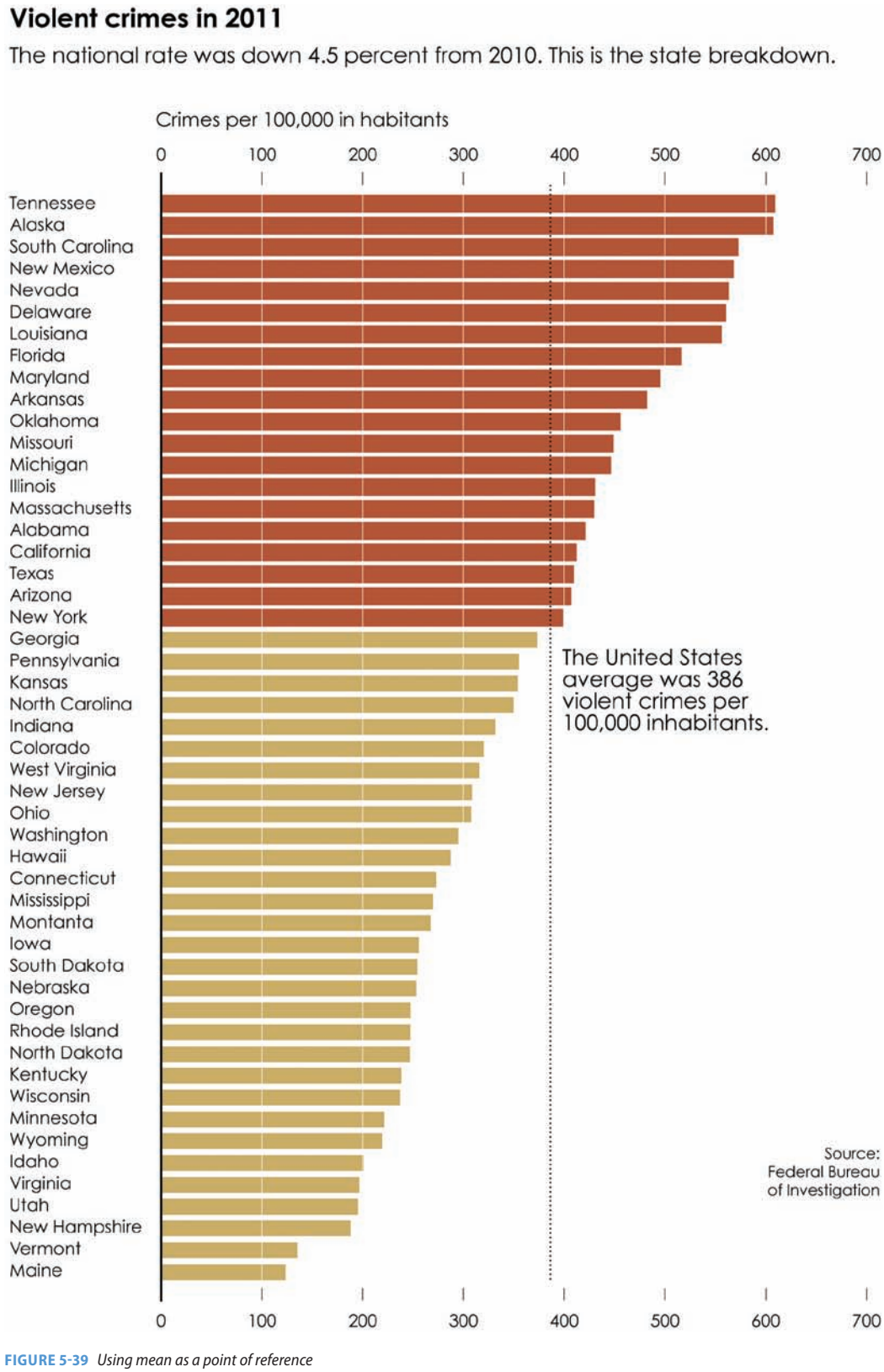


FIGURE 5-39 *Using mean as a point of reference*

rates. From each rate, subtract the rate that came the year before, as shown in Figure 5-41.

You can take it the other direction, and add values, as shown in Figure 5-42. A step chart shows the monthly cumulative cost of cable over a year versus the modest cost of Hulu Plus and Netflix. An aggregate at the end shows total annual savings if you were to switch to the latter.

Straightforward math operations can help you see your data from a different angle or bring focus to a graphic. Of course, the more statistics you know, the better you can process and analyze your data, which in turn can lead to more informative graphics. Account for how people might interpret a graphic, and if they have to do math in their head to make inferences, it might be worth the effort to do the math for them and translate the results visually.

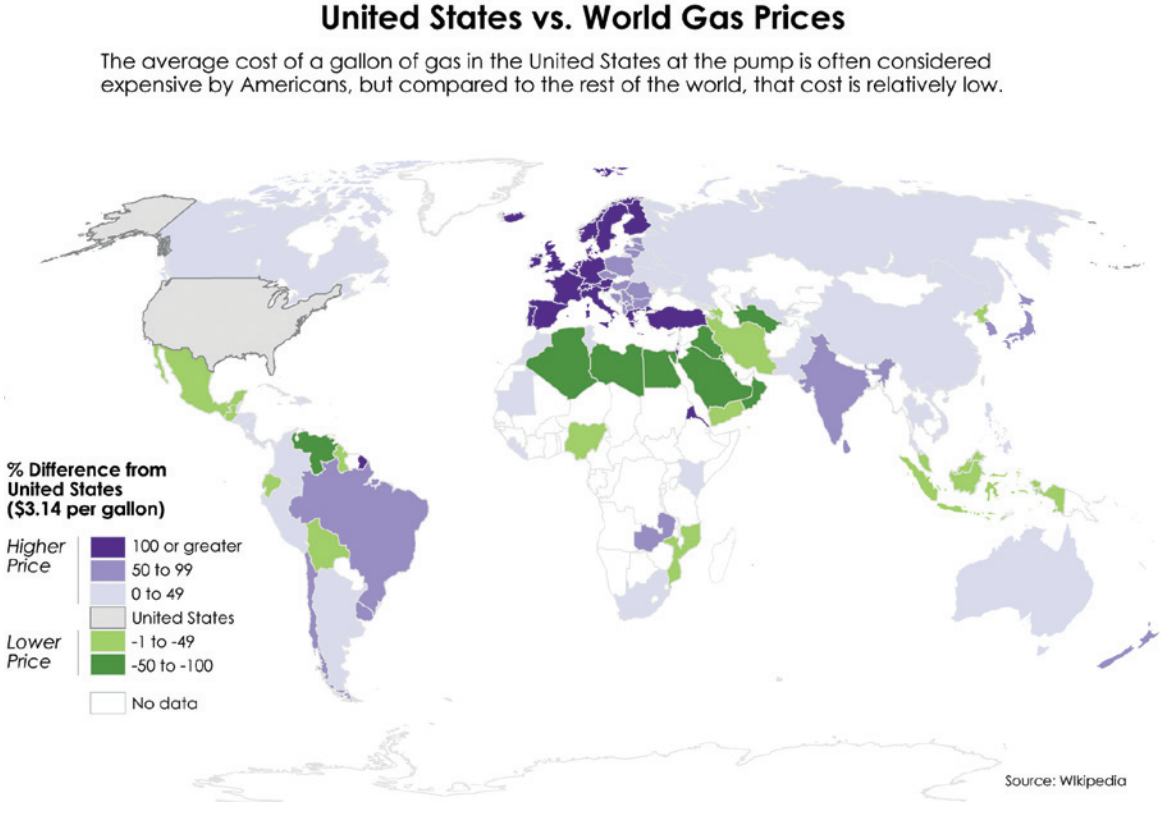


FIGURE 5-40 *Transforming data based on point of reference*

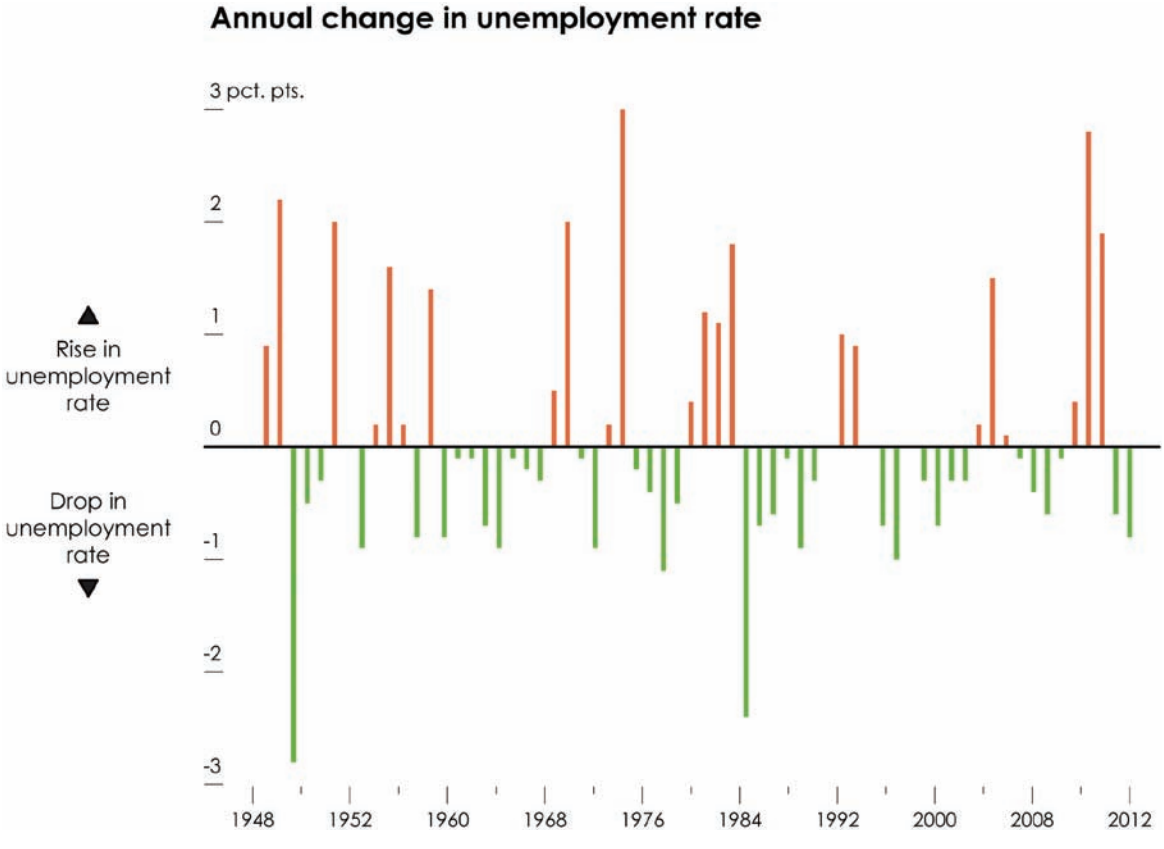


FIGURE 5-41 *Showing changes instead of raw values*

WRAPPING UP

You get leeway as you explore data on your own. However, if you want to make actionable insights based purely on your graphics, you must make sure you're seeing the right thing. You wouldn't want to make a poor decision because you visualized your data inaccurately.

Similarly, as you present your graphics to others or pass them onto the rest of the world, where others might make decisions based on what you show, it's your responsibility to display the data accurately. Differentiate elements, highlight important bits, and annotate to explain and describe the data.

Of course, as you saw in Chapter 2, visualization as a medium creates a wide spectrum of applications. How much you highlight, how much you explain to viewers and readers, and what you display depends on what you want to show and who you present to.

Warning: Make sure your data is comparable when you transform multiple datasets. Are they from the same source? Is the methodology the same? What level of uncertainty is attached to the estimates? If you're not sure, you should find out because incorrect math can lead to incorrect conclusions.

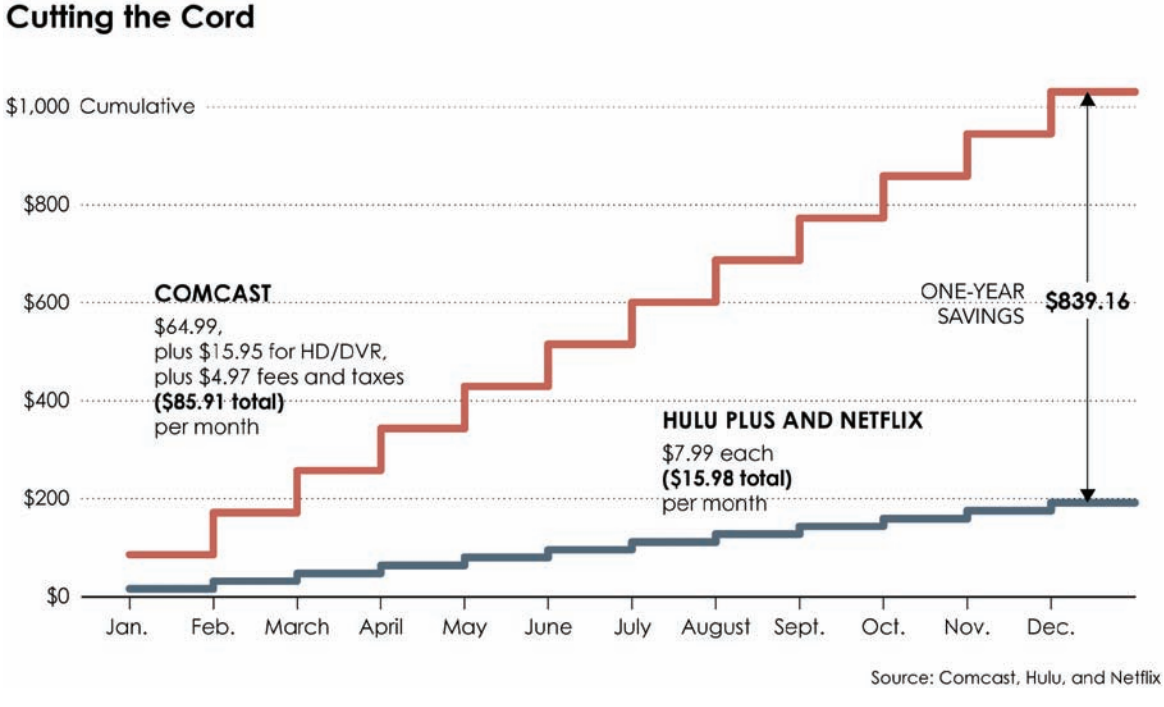


FIGURE 5-42 *Cumulative values and totals*