4x4 model for knowledge content

In this movie we're going to talk about something I call the four by four Model for Knowledge Content. It's really a basic philosophy for communications, and visualization is a key part of it. The big idea is it's about getting the right people to the right content at the right time. So. let's start at the beginning. If you go back to the beginning of human communications, humans have been telling each other stories forever. But, if you look forward to today, and you think about where we're telling each other these stories, where this content lives, it lives out in this Galilean Model of the Universe, as my friend Drew calls it, where all these conversations are happening out on the Web, on YouTube, on Facebook, on Twitter. In other words, your website is not the center of the universe where these conversations are happening, where this content is living. This is a key idea to keep in mind. The second key idea is the explosion of content. As Eric Schmidt said, we now create more content every two days than was created from the dawn of civilization through 2003. 2003, we're not even talking about 1990. This is after the dotcom boom. And by the way, this quote is from 2010, so it's probably every single day or every half day now. So that's the second key idea to keep in mind. We're in an age of data overload. But luckily there's a solution for this. There's a solution for this data overload problem which is that we can communicate for scanners, we can design for scanners. So for instance, if you've ever seen an eye tracking study where you have someone sit down in front of a computer and look at content, and there's a little camera tracking their eye movements. This is what it tends to look like. On the right hand side here, there's this very typical F pattern where people start on the left hand side and scan to the right, and drop and scan, and drop and scan, and drop and scan. Then, if you look at it like on Google search results, the scanning and the dropping just stops after the first few entries. But of course there is a design solution. We can put a face on a page to draw an eye to the face. We can put elements on the page to draw the eye. We can solve this problem through good design. Interestingly, there are extra tricks we can use. For instance, if you turn the face towards the content, not only will people look at the face, but they'll look where that face is looking. So we have techniques to solve this issue. And here's the third big idea. It's actually been shown by a study done at Northwestern University a few years ago that information overload is actually a myth. So, while we sort of feel overwhelmed at times, we actually are generally empowered by the choices we have available to us online. But, and this is important, we are generally questioning the accuracy of the information we are getting online. The four by four model was designed to solve and address these issues that we're facing right now. And it really comes down to this. You want to create content on four levels, and you want to bake into that content four key components as much as you possibly can. So let's walk through it really quickly. What you have are these four levels of content that I call the water cooler, the cafe, the research library, and the lab. So what is the water cooler? The water cooler moment in content online is, it's an image, it's a a tweet, it's like a 30-second video, it's the attention grabber. So like at a Monday morning you might be standing around the water cooler having conversation about that weekend's Boston Bruins hockey game, and maybe three of you are standing around, and two of you are big hockey fans. You're going to sort of chat and talk about all the great goal scorers, and penalties, and all the exciting stuff. And one person standing around the water cooler that morning really couldn't care less about hockey. And so, at that moment, that attention grabbing moment, two of you who have interest in the subject may say, hey let's go to the cafe, let's get a cup of coffee, and let's talk more about it. The third person's going to say, yeah thank you very much, I've had quite enough, and that person's going to leave. So, we've lost one person from that next step into the conversation, and that's okay. That person did not belong. They're not our audience for this content. So the two of you go to the cafe, and you have a longer conversation. So online that's the blog post, that's a short article, that's a three to five-minute video. And so, we're stepping into deeper content by enticing people, and they're going to self-select based on their interest in the conversation, in the topic. So once again, out of those two people at the cafe, one of them is really, really interested in the topic, and they're going to go to the research library and dig deep. They're going to read all kinds of articles. They're going to read long reports about NHL statistics, et cetera. And finally, again based on self-selecting interest in the topic, they might want to go to the lab where they can really dig in deep into the data, and play with data tools, and filter and sort. That's the data lab, the interaction data experience. So these are the four levels of content and, as I mentioned before, it's really all about getting the right people to the right content at the right time. You're not always going to create research library and lab-level content but when you do, you want to make sure you always include water cooler content, leading to cafe content, leading to those deeper experiences to get the right people to the right content at the right time. And as I mentioned before, the four key components, visualization, storytelling, interactivity and shareability, the more you can bake these components into your content the better off you'll be. Now of course here, today, we're mostly talking about visualization, but these are all aspects of it, and visualization's a key part of successful four by four communications.

Channel your audience

Any communications, whether it's a website or a video or even a circus, requires knowing your audience and adjusting the presentation to them. So, for instance, you're doing a circus, it's for kids, of course you're going to go full-on Ringling Brothers with lions and tigers and bears and clowns. But if you have a more sophisticated audience, you might have a little more nuance and artistry. A little more Cirque Du Soleil than Ringling Brothers. What do we need to know about our audience for visualization? It's really the same as for any communications challenge. We need to understand their culture, their level of expertise, and a few other things. I'm going to talk through them here one by one. First, culture. Your audience comes from a certain place in the world. It affects their language, it affects their perspective, it affects their context, so many different things. So, for example, think about, what does a wedding look like? Sometimes it looks like this or like this or like this. It affects the colors that are used. Now, color theory's its own big conversation that we're not going to cover in detail in this course, but culture's an important influence on color amongst other things. For instance, what does a wedding dress look like? Is it white? It is red? And remember that what looks odd in one culture might be completely normal in another one. Another thing culture influences is narrative context. Does your audience know the underlying story of what you're talking about? So, for instance, let's say you're creating a visualization of Wayne Gretzky statistics. Now, Wayne Gretzky was probably the greatest hockey player who ever lived. So, if you're doing this visualization for people in northern countries who play a lot of hockey, you probably don't need to set a lot of context. You don't need to explain who Wayne Gretzky is. But if it's a visualization for people in, maybe, Brazil or Madagascar, they might not have as much of that context. You might need to set up the story a little bit better for them. The next one is level of expertise. If you're creating a project and it's for people who are experts in an industry, you're going to approach it in one way. Less context, less hand-holding, maybe even use different language, different types of words, more lingo. If your audience isn't quite as informed on the subject, you're going to provide more background information, more hand-holding. It's going to be like, maybe, a shallower story with less detail. There's also what I call the consumption context or channel. Your work isn't just for fun or self-admiration, you want to publish something somewhere somehow. So, whether you're creating a visualization for the New York Times or the Daily Show or a Bill's Blog for data dorks, it's going to change your approach. One might be more serious. Maybe a higher standard of excellence, maybe more statistical integrity. The other might be more irreverent. You might have lower journalistic standards. You might round to fewer decimal places with your numbers. It's not about being less accurate, but you may be less detail-oriented in certain contexts. Another issue is accessibility. Now, data visualization is all about visualization, it's all about the eye. So, we're only dealing with sighted people, correct? Well, yes and no. You could say that you could create a, quote, unquote, data visualization but really it's just audio. Maybe there are other ways of thinking about this field. Data audiolization maybe is a word. But there are other accessibility issues also, even if you don't go quite that far, the big one being colorblindness. From one to 3% of people are colorblind, and there's an especially common form called red-green colorblindness where people have difficulty distinguishing between red and green. So, just think about stock market charts where red is always used to mean prices going down and green for prices going up. So, think about what that might mean for the one to 3% of people who have trouble distinguishing between those two colors. Luckily, in trying to channel your audience and understand what they're going through, there's a great website to help you with this. There's this website where I can actually take an image and upload it. So, I'm actually going to select an image from my hard drive here, and this is actually an image from Google Finance, so it is a stock chart image, and it's going to show me what the chart looks like to me. This is the original image over here. And then I can actually go in and click on these little radio buttons and see what it would look like for people with different forms of colorblindness. Really helpful tool, it really literally lets me channel my audience and see what they would see. It's a great way to take a look at your work and test it and make sure that you're being friendly to all of your audience. There are other things to think about when you're thinking about the visually impaired. A big one being contrast and font size, creating your charts with enough contrast to be easily seen, making sure the font size is large enough to be read. Forget about visual impairment, just think about your audience. If they tend to be over 40 years old, make sure your font sizes aren't 10-point. If your chart is for college kids, eh, don't worry about it so much. The next thing to keep in mind is whether or not your audience consists of true believers or skeptics, or more often than not, it's both, but you have to be aware of it. If you're visualizing data, let's say, about climate change, to climate scientists, it'll be received differently than if you're sharing the same data with a publication targeting the coal mining industry. This is a dangerous lesson in that you don't want to overthink how much convincing you need to do, otherwise it might bias what you share and how you frame it. On the other hand, being aware of whether you're sharing information to change minds or simply to provide, quote, the truth, is good to understand because it might affect how deeply you dive and how much interactivity you provide. And finally, there is the action that you want your audience to take. You're not sharing your content for your mother's approval or for the adulation of the public, maybe a little bit of both of those, but for the most part, you're looking for an action, a reaction from your audience. You need to know what you're looking for, what action you're looking for to design better outcomes. Do you want someone to call their congressman after looking at your visualization? Do you want them to answer a poll question? Do you want them to share it on social? Do you want to inspire them in some other way? If you constantly ask yourself if your design is leading toward this outcome, you're more likely to achieve it. So, these are all the things you have to understand about your audience when you're creating visualizations for them, but it's more than that. It's not just about understanding, it's about channeling them. It's about really getting inside their head. You don't want to ignore their culture and accessibility issues and all these other things, you want to actually feel their pain. You want to understand it as though you are them. You'll be less likely to be influenced by bias, for instance because you're going to feel their skepticism, their arguments against what you're showing them. It's going to help you articulate more clearly and factually if you can really imagine being them. Imagine that you are them and you'll communicate more meaningfully.

So what is data visualization?

So what is data visualization? It is data visualized. No fancy definitions here for me. But there's lots of things that go into good data visualizations but that's not the point of this movie. The point is to really just show some examples and try to set some boundaries for what we're talking about here. And you'll notice probably that the boundaries are pretty broad. I'm not going to constrain it too much. And you'll also note that this is not an exhaustive list. So let's start off with the poor lowly bar chart, right, one of the most basic and common forms. It's oft maligned but it's arguably the best form of data visualization. If you have one variable to communicate and the emphasis is on the magnitude of the values across different categories, and the goal is probably to allow comparison, you really can't do much better. So in this example, we're looking at a sales performance chart and you could easily see that 2006 was the best year, 2008 was the worst year, 2009 didn't do so hot either, and then the other years are sort of in between those extremes. It's easy to see the headline and the overall story here. In different chart forms, it might be hard to see the variations, for instance, between 2008 and 2009. The bar chart's great, it works fantastically. And what I would argue is that whenever you're doing visualization, you should almost always ask yourself, why should I not do a bar chart? That should be the first thing you think of. If you're looking at data over time, the line chart is a really great way to go as sort of by default. Or the timeline if you're doing an infographic, right, where it's more time-oriented content but not necessarily quote unquote data, right? If you're dealing with time, timeline, line chart, great defaults. By the way, one of the advantages of the bar chart, the line chart, and the timeline, and some of these others we're going to talk about is that your audience knows them so well, which means they can immediately recognize them, they know how to read them, they won't be confused by them, and that's a good thing. Another very common example of a chart is the scatter plot and the bubble chart, which is really just a scatter plot with a third variable added. Right now, the bubble size is sort of a third variable. It's telling us a little bit more than just the scatter plot, which covers two variables. There is the oft-reviled pie chart, radar charts and spider charts, tree maps which are an interesting way of showing hierarchical data, and we'll talk more about all of these in other movies. Stream graphs, matrices, these are all effective and interesting ways of showing and displaying data, and comparing things. Sometimes alternative forms really do help to tell a story. So this example is something created by Charles Minard in the 19th Century and what we're looking at here is the march of Napoleon's army onto Moscow and then the retreat from Moscow. So if we start on the left, the tan line is his army and the thickness of the line shows you how many people were in the army, so 422,000 or so people. And as they march eastward, so we can see geographic information on here, you can see how the army starts to shrink and shrink and shrink, and there're couple of offshoots where some people go off in different directions. The army shrink, shrink, shrink, shrink, shrink, shrink, shrink, shrink, shrinks until they get to Moscow and they're a shadow of their former selves. And then they retreat and they come back, they shrink, shrink, shrink, shrink, shrink, shrink, shrink, shrink, shrink, shrinks more. And notice when they're about two-thirds of the way back, during their retreat, and they hit the Berezina River, and look what happens there. Crossing that river, it was incredibly cold, which you can see down below where you look at the temperatures, and it was half yet again, half of their men. Again, back, back, back, back, back until you look at that thin line where at one point, right before the other offshooted people rejoined them, they were down to 4,000 people. Unbelievable losses, a great example of visualization telling an incredibly powerful story. Another example from the past, this is another 19th Century example, was this visualization created by Florence Nightingale. Now of course, she's best known for being a caregiver to soldiers, the lady with the lamp. But she was also a medical statistician and a data visualizer. She invented this form, which was called the coxcomb. This one shows monthly mortality rates from the Crimean War due to sanitation and other causes in wartime hospitals. Florence Nightingale often used visualizations when presenting to Parliament when she was trying to affect changes in these conditions in hospitals during wartime. Maps, of course, a very common form of data visualization. Showing geographic data, and also showing things like election maps and demographic trends and et cetera. A choropleth, this form that we're looking at here, is probably the most common one you see where regions are colored to represent data intensity in that particular region. So, the darker regions here mean that there's probably more of whatever it is being measured here, the lighter regions less. There's another category of visualizations that you might say are sort of useless, superfluous, what's the point, this is not serious data visualization. Yeah, okay, you could certainly say that. But that doesn't mean it's not data visualization and it doesn't mean that they're not useful in their own way. So here's an example of one, this is called a massive map of hip hop monikers. This is data visualization. So let me just click in here. And what we're looking at is a very large data set of hip hop monikers, right? The, sort-of the names that hip hop artists go by. And you see the connections between the names and sort of the categorization of them. So for instance, here we're looking at hip hop monikers that are mineral in nature. Goldie Loc, right, Goldielocks, Onyx, and you see this connection between that and the parent category, animal, vegetable, mineral. And if I go over here and I see this connection over here, animal, vegetable, mineral goes over to vegetable and I can see Blue Raspberry and Casey Veggies, et cetera. Interesting, not necessarily quote unquote useful. But very interesting and certainly data visualization. Another example of a visualization that again, you could sort of say what is this and why am I looking at it, what's the point? I actually personally love this one. So this is a visualization of pi. There is a computer that ran for something like a year and calculated pi to ten trillion decimal places. Now this visualization took the first four million of them and laid them out, so all these little dots, every single individual dot represents one of those digits of pi. And so of course I can come here and just scroll down this entire list to get all four million. And as you can see over here in the legend, each color represents what number is represented. I can roll over the list and see what numbers are beneath my mouse pointer, okay? So what is the point? It's kind of neat, it's kind of random. What am I looking at? Well, I can easily see based on this visualization that this looks like static. This looks completely random. And so while this is sort of a, huh, why is this, what is this for? It does reinforce that pi is random. It serves a purpose, and I think it's a really great and interesting way of doing it. So as I said, this has been a very limited definition of data visualization, mostly through example. Hopefully it's a good start for thinking about some of those standard forms and also some ways people have thought outside of the bar chart when approaching more complex problems.

ASK what makes a good data visualization

So in some ways, this might be the most important movie in the course. How do you make a good visualization? You might ask yourself that. And I actually use that as an acronym to help me remember. Really what it comes down to is accurate, story, knowledge. If you can create visualizations that are accurate, and tell a good story, even if it's not a linear story, and provide real knowledge to your audience, then that's a good target to shoot for. So let's talk about accuracy. This was published by a national news network and as you can see it's tracking the unemployment rate under President Obama. And, I'd like you to take a look at it for about five or 10 seconds, 30, however long you have. You can pause the video if you really want to think about it before I tell you the answer (laughs) And I'd like you to tell me what's wrong with it? So you might have noticed, that data point on the far right, in November, the most recent one is showing as 8.6% and yet look at where it is visually, it's on the same line as 9% and if you compare it to the 8.8% data point in March it's higher than that one. So, this is just wrong. This is what it should've looked like. Can you see any other issues with this chart that might qualify it as being inaccurate? Take a look at the scale. The highest data point here is 9.2% and the lowest is 8.6%. The scale should have been set from 8.6% to 9.2%, right? From the very highest to the very lowest. Or maybe 8.5% to 9.5% like these red dotted lines show. Or 8.7 to 9.3, you can play around with it a little bit, but showing all the way from 7.5 to 10. 5 is exaggerating to show a flatter line. It's accurate, technically, but it's misleading. This data comes from the Bureau of Labor Statistics, and this is how they provided it. Now if you look at one more issue here, take a look at the headline and the X axis label. This is supposed to reflect the unemployment rate under President Obama. That's what this is saying. Which sounds like this is for the entire length of his term. But of course, this isn't. This is just for the year of 2011. So the data isn't really providing the unemployment rate under President Obama, it's the unemployment rate for 2011. So that's accuracy. Let's talk a little bit about story. In the last movie, So What is Data Visualization, I pointed out Charles Menard's visualization of Napoleons march on Moscow. This is very frequently referred to by Edward Tufte and others as one of these great classic examples of visualization. And one of the reasons is it tells such a compelling story. Napoleon's army marched into Moscow and then retreated from Moscow, and lost something like 98% of it's army, 400,000 men. And while you can read that in a sentence and maybe you can see a bar chart that sort of reflects it, something about this really tells the story. You can see them moving to the East and the army just getting smaller, and smaller, and smaller. They hit the Berezina River on the way back and it shrinks by 50% yet again. And look at how thin that line is by the time it gets back. Meets up with a few more of its forces, but what a compelling story. And you can also track the temperature and how cold it was, which helps explain what was going on. Finally, knowledge. Another classic example of data visualization, this is a map created by John Snow, who was a doctor in the 19th century and he had a theory that cholera was a water-borne disease. At the time, people thought it was an air-borne disease. And when there was an epidemic of it, he mapped out where all the deaths were. You can see the little black lines next to the homes where people were dying from cholera. He was able to prove that they were all clustered around this one water pump here on Broad Street in London. He took this data and transformed the understanding of cholera for the entire world going forward. This imparted real knowledge. In fact, his work, including this map, was part of the origin of the field of epidemiology and public health. If you're careful and accurate, and you tell a good story and you impart real knowledge to your audience, you can transform them in ways that just informing them will never do.

Visual perception

The earliest forms of written language were visual. They were symbology, right? The image, the picture, represented the concepts being communicated. As communications evolved, our written language became more nuanced to communicate more complex ideas. But that hasn't helped us when communicating numbers. We can actually parse numbers much more quickly and easily in visual form, not text. Now, why is that? There are a bunch of reasons. One thing is that 30 to 50% of our brain is devoted to visual processing, and 70% of our sensory receptors are in our eyes. It's also been shown that it takes a tenth of a second to make sense of a visual scene, really just at the glance of an eye. It's also been shown that you have 323% better performance on tasks when you learn them accompanied by images versus just text. There's also something known as the picture superiority effect, that you'll remember 60% of something when you learn it with imagery versus 6% without, and this effect increases with age. There's also something called gestalt psychology, and these gestalt principles have a deep effect in how we perceive visual information. This helps immensely with data visualization to understand how it works. The first one is called figure/ground. It's most often used in logo design. So for instance, here you have the FedEx logo, one of the more famous examples of the use of figure/ground. You have the mark, FedEx, the letters in the foreground, and in the background, the negative space, you have that arrow, that white arrow subliminal effect, characterizing motion, which of course, is what FedEx is all about. In data visualization, figure/ground is the least actively used aspect of visual perception, but it is important to understand. What's interesting is that the bottom object, the bottom portion of whatever it is that you're looking at, so in this case we're looking at an area chart, the bottom is always perceived as being the figure, meaning in the foreground, and the top as the background, no matter what the color. So if someone were to look at the area chart on the left, they're going to perceive the white as being the data and the black as being the background. And the same thing on the right-hand side, the black would be the data and the white is the background. So you can't rely on color or contrast or anything other than bottom versus top. It's an important principle to understand. The next gestalt principle to talk about is proximity. So when you have items next to each other or near each other, so in this case we have three columns of dots on the left and two columns of dots on the right, they are perceived as being grouped, and therefore similar and together. So I immediately sense that all the dots on the left-hand side are together and with each other, yet separate from the two columns of dots to the right. Proximity is one reason why we recognize patterns in a chart like this scatter plot. The next principle is called a similarity, kind of an obvious one. Objects that are similar, the black dots, are different from the other objects, the white dots, straightforward. The similarity of things makes us categorize them, whether the similarity is in shape or size or color. The next principle is called parallelism. So the idea here is that when you have things in parallel, like the three lines in the middle, we assume that they're together and different from the other objects that aren't in parallel. So in data visualization, you might see patterns. So for instance, the overall trend lines for these different portions of this area chart, we see that parallelism and make judgments about these shapes because of that aspect of the shape. The next principle is called common fate, and I just love this title. I just like the name of this one. So here's how it works. Here you have a bunch of dots and there's no discernible pattern. But as soon as they start moving, I can see that these dots have a common fate, right? They're all moving in the same direction, so I can tell that they belong together. The ones that are moving are one group and the ones that aren't moving are in another group. That's common fate. Of course, you can see this in animated visualizations, or interactive visualizations where you have a rollover effect. You won't see this in static visualizations. Then there are the principles referred to as closure and continuity. These are interesting, although they're less relevant specifically for data visualization, but they play a role in really all design. So the first idea here is closure, the fact that we see things that aren't there. Our brain completes the picture. So for instance, in this image, in addition to seeing three Pac-Man, or three wheels of cheese, you probably also see the triangle in the middle. That's closure. Or here you have what's known as continuity, where you see the S. You don't just see two random arcs, you see a full S shape. Your brain just does that for you. Culture can play a role in continuity. So for instance, here you don't just see a random sine wave pattern, even though you do end up seeing that full pattern. But what you probably really see is the Loch Ness Monster. That's where culture can have an influence. So as I said, this isn't specific to visualization, but it's something to be aware of. As you visualize and as you're creating shapes, you don't want them to be misperceived or seen as something unintended by your audience. It could affect how your data is understood. The gestalt principles are key to how people perceive things visually. And there are lots of techniques to use to draw the eye, understanding how these principles work, and to help make your data stand out, whether it's tilting a line or making lines shorter or thicker, or dots fatter, using different shapes or hash marks, color and hue, et cetera. You don't need to understand this brain scan or every principle for how brains process information to do a visualization, of course. But isn't it nice to know a bit about it? I know it helps me think when I'm designing how to trigger the brain instantly to maximize effect.

Understanding your data

You don't have to be a mathematician or a statistician to do data visualization, but you do need to understand some basic concepts to reduce mistakes, increase your accuracy, and even to make more compelling visualizations. So I'm going to share three key concepts from math and statistics to help you get better at that. And here's the first one. And the picture may not immediately give it away, but you're going to remember this from your school days. There is mean, and then there's median. Of course, very different definitions of these words, but let's go to a better example using math. All right, so hockey player points scored. This is a list of individual points scored by individual players on a youth hockey team. And these are all fake. I made them all up except that last number on the right hand there. One kid on this actual youth hockey team, scored 517 points. That's actually Wayne Gretzky scored that number of points in one season as a kid. Crazy. Now, let's talk about the mean and the median here. The mean is the statistical average. That's when you add up all of those numbers and you divide by the total number of items in the list. In other words, here we have 15 kids who score different number of points, add them all up, divide by 15, that gives you the mean. The median is when you take the number in the middle of the list. So there are 15 kids here. So it's the number at position eight, smack dab in the middle of the list. So let's look what the numbers actually are here. If we added up all these numbers and divided by 15, cause there are 15 kids, the mean, the average is 57. Now look at that list. Did anybody score 57 goals or more? Yeah, one kid, Wayne Gretzky, nobody else got 57 goals. So does that seem like a representative of the sort of normal, the average, the middle-ish? No, it's a completely outrageous. Nobody got that number goals, therefore it's not a good way to look at the sort of typical. The median however, is the number right in the middle of the list. And if you take position eight in this list, so right smack dab in the middle of 15, you got 25, the median is 25. And by the way, if we took the mean, the average of this list after we removed Wayne Gretzky's crazy score, the mean without the outlier is about 24. So the median is a good representative of quote unquote "normal", all right? That's when we think about using mean versus median, when you have outliers that might skew your results in a weird way. Another thing to think about is when you should use the actual numbers that you're given in your dataset or a rank index or percentiles or some sort of ratio in order to represent your values. And they're not your only choices, but there are three pretty common ones to think about. And so here's an example of that. We have the GDPs of the United States and Singapore. So clearly the United States' number is way bigger, but what do these numbers actually mean? Well, if we look at them by rank, the United States is ranked number one, Singapore is ranked number 35. Okay, so US is much higher. Okay, now GDP per capita, let's take that GDP and divide by the population. And that tells us maybe a better indicator of how productive the economies are for these two countries. And look at this, the United States GDP per capita per person is $65,000 versus Singapore $102,000, rank 11 versus rank three. Very different point of view by simply changing the absolute value into a ratio in this case, a per capita number. Another thing to think about is when you might consider using the change in the values, as opposed to the actual numbers themselves. And so let's use GDP as another example, if we look at the countries of Cameroon and Tunisia, they have very similar GDP, is about $39 billion each, which countries should I invest in though? Where should I try to sell my goods and services, right? It's impossible to say at this point, but look at the rate of change in GDP, okay? Cameroon is growing at about 3.7% as opposed to Tunisia at about 1%. So maybe higher growth rate. Maybe also, if you look at how long it's had a good growth rate, a little bit more consistency over time, that's going to affect where you really want to put your money, even though the GDPs are pretty much identical. A couple other things to keep in mind are things like sample size and methodology, issues around quality and reliability. We're not going to talk about that here, go read more about them. But the basic idea is if you have a survey and you ask two people, their opinion about something, are you going to draw conclusions from a sample size of two? No. Or if the quality of that sample is not very representative of what you're trying to understand. No, it's not going to be very useful to you. And of course, there's the old correlation versus causation, correlation doesn't equal causation just because one moves and the other moves in tandem that's correlation. That's not necessarily proving that one thing causes another. These are the kinds of things you need to be aware of and think about when you're working with your data. And you need to know your data. You may or may not be the expert, right? You may be working with a colleague or some other sort of partner to help you really lean deeply into the data. And that's okay, but you need to know enough about your data to be able to work with it. One thing that I like to think about is at a minimum, you need to really understand the headline, right? The thesis of the findings in the data summed up into one or two key ideas. And in fact, I recommend that you figure that out first. It's also worth realizing that, of course, it's possible that there may be more than one headline, more than one way of presenting the data that you have. Now, let me share an example. This is a project that I created and I was looking at trends in jobs in the United States. The basic idea here is that I figured that back in the 1800s there were a lot of wheelwrights, people who fixed wagon wheels and probably fewer of them today, certainly as a percentage of workers. And so I wanted to look at, hey, how did jobs change over time? And I looked at the last 20 years of jobs statistics starting in 2000 to 2019. And the basic idea of this data story, and, you know, I had to think about it in terms of headlines was did jobs change? And I looked at it and sort of three different ways. I looked at the top 10 jobs, did those change? And then also I looked at beyond the top 10 jobs, did the employment rate, even beyond the top 10 change as a rate per 1,000? And then I also looked at it even as sort of more of a multiple rate change. And then finally I wanted to look at how about wages? Did wages go up or down, higher or lower for different categories of jobs and different individual jobs? So that's the gist of this particular project. So I thought about it in terms of headlines. Even before, by the way, I did my analysis, which is one of those things, that data analysts cringe a little bit when I say that, because you can introduce bias into your work if you have the headline figured out before you even done your analysis, but you know what? That's what we do in data analysis, anyways, we have a hypothesis. And so my hypothesis was jobs are pretty stable, they rarely change, but some jobs in this case, a blank. So I don't know the answer yet if I haven't done the analysis, gain or lose employees at large rates over time. So that was my headline. That was like the main thing I was wondering about. So that's the sort of gist of my story. But then yes, I also thought about how pay may or may not be keeping up with inflation in certain jobs. And therefore that was my secondary headline, right? It's going to be, you know, falling behind in some jobs and maybe doing better in other jobs. That was my headline. I knew that before I even did my analysis. And here's the thing, you need to know your data well enough to spot mistakes, okay? Well enough, I have to understand when you see things in the data that don't make sense, that you recognize that you spot them and you figure out what to do about it. As an example, when I was looking at those top 10 jobs, I noticed that registered nurses were a top 10 job, and then all of a sudden they'd just disappeared. And then there was another job that suddenly shot up to be in a very similar place that had a name, similar to registered nurses, but it was slightly different. What happened was the US Labor Department just changed the job code. So they looked like two separate jobs, but they were the same job. So there were a million things like that in the dataset that I had to tweak. And I had to know my data enough to spot those mistakes. And of course I have to do that without succumbing to bias, right? I can't say, oh, here are two jobs that sound alike, maybe, but they're really not the same, but I'm going to pretend they're the same anyways, because it fits my story. No, I'm going to be very careful about changes like that, make sure that I'm adhering to best practices and good principle. And as an example of that also , I would have guessed that management pay increase used more than most. And it did. Those are the blue rectangles at the top there, they were doing better than inflation over time. But I also would have guessed that computer and mathematical occupation would have done also pretty well, but the third row there that's that category. It actually performed very poorly compared to inflation really as a category among the worst. How could that possibly be? So this is an example where I know my data, I checked it carefully. I also knew my bias and my assumptions, really check those numbers very carefully. Cause how could that possibly be correct? But it was, so it's interesting, right? So by the way, this is another example, where I could have looked at the absolute values of this data. Did the job pay go up or down? And that would have been fine. And I could have even looked at that from a state of the standpoint of percentage change and that would have been something, but comparing it to inflation was what made this much more relevant. Because of course inflation goes up, therefore pay kind of always goes up, one would hope and think, but did it go up at the same amount as inflation, higher or lower? That's really what it's all about cost of living. So the basic idea here is you do need to know a few basic math and statistics concepts in order to do a really good job at this type of work, basic knowledge of your data, really work hard to keep bias from entering what you're doing. And if you know this stuff and just, you know, keep an eye on things like bias, then you're more likely to stay on track.

Explanatory

All data visualizations are what I call "expl_atory." Meaning, they're either explanatory or exploratory, and they can be both by the way. So when we think about explanatory visualizations, their entire purpose is to explain an idea to an audience. And whether it's to change hearts and minds, or simply to educate, it's just all about explaining something. So here's an example of a visualization that I created in the summer of 2020, looking at COVID-19 death rates. These were big numbers, and I think people had a hard time wrapping their heads around what those numbers actually meant. Is it a big deal that tens of thousands or hundreds of thousands of people are dying? And as hard as that is to say, some people maybe didn't quite understand why it was a big deal or what that really meant. And so I decided to explain it in the context of, how does that compare to other causes of death? Particularly, the top 10 causes of death. And so in April, we already surpassed the number of deaths caused by the flu in the prior year. We passed diabetes, then we passed Alzheimer's, and by the time I created this, COVID-19 had already become the number five killer in the United States, and numbers four and number three were quickly approaching. It was clear at that point that COVID would become the number three killer in the United States, a very big deal, particularly for an infectious disease, being really one of only two infectious diseases on the list. So when you think about an exploratory visualization, rather than just explaining an idea, you may provide context, you may also explain, as I said, but then you also let your audience dig in to find their own aha's in the data. So an example of this is this visualization looking at race results from a half-marathon from a few years back. And so what we have here is simply a swarm plot, a distribution diagram, which could be a nice beginning of an explanatory visualization. Maybe the point here is to say that there is a few people at the ends and most people sort of land in the middle, and maybe what the average time was, et cetera, et cetera. This could easily be just an explanatory visualization with some extra labeling. But this is also an exploratory visualization because I can roll over and investigate each individual competitor here one by one, and to learn a little bit more. I can find runners by names, so I'm going to type in the name Mike. This visualization was actually created by a guy named Mike Berry who actually came in second during this race, which is pretty impressive. Anyways, so I can search, I can explore each individual data point, and I can even group the data points by age and get more insights, more aha's. Aha, there were very few teenagers who ran the race, or at least 17 and younger. There were very few 60 plus year olds. The vast majority were in the 18 to 39 range. I can see the gender breakout, et cetera, et cetera. So I can really explore this dataset in addition to potentially with some context, getting that explanatory information from it. So as I said, a visualization must be "expl_tory." That is what it is, and so therefore it is either explanatory and/or exploratory. It has to be one, but it can be both. So when I'm doing visualizations, I always think about this. How far do I want to take it? Do I just want to explain an idea and that's it, and sort of provide my point of view to my audience? I do that sometimes. Or do I want allow them to explore? You know, to be honest, I tend to provide a combination of both to my audiences when I can. I like to explain things, data storytelling, and then allow exploration by creating interactive experiences where my audience can dig even deeper when they want to, sort of empowering them to find the treasure in the data that I'm sharing with them.

The six Ws

Information design and data visualization are really about focusing an audience on what's most important and only revealing more detail as you need to. So really it's all about information hierarchy. One of the great ways to help get to information hierarchy is using the six Ws. So if you remember from grade school, the six Ws are really the best way to organize and think about any story. And in data visualization in particular, you're not necessarily going to use all six, who, what, when, where, why, and how, but thinking about all six and figuring out which one or two or maybe three are the most important ones and eliminating the ones that you don't need is a great way to get at the hierarchy and the information that you're trying to show to your users. One of the more famous examples of visualization is John Snow's map of a cholera outbreak in London in the 19th century. And the idea here was that cholera at the time was thought to be an airborne disease, but John Snow's theory was that it was waterborne. And so he actually took a map and marked these little black marks to mark every death from this particular cholera outbreak, and you could very visually and quickly and easily see that everyone who was dying was dying around this one water pump on Broad Street in London. And so he was able to prove his point by focusing on the where, right. Where are people living who are dying? Showing that it was all around this one water pump and he was able to convince the city to pull the handle off the pump which led to the cholera outbreak diminishing. So focusing on the where in this case led him to a very logical conclusion that using a map was a good paradigm. So if you always think about the six Ws, narrowing down to the one or two that are important, it'll really help you in your visualization projects. So I was doing a visualization looking at hospital pricing data and I started off by thinking about what question am I trying to answer. The question is really simple. Where can I go for a specific treatment at a decent price and good quality? It's the logical question you might ask yourself if you needed to go and get your hip replaced let's say. So looking at the six Ws, I'm looking at a list and I say okay, I have a who, a what, and a where. It's pretty clear. So let's go through those. The where in this case is the answer that I'm looking for, right? It's built right into the question. Where can I go to get a certain procedure done at a good price and good quality? The what in this case is really the most important information. It's how I'm going to judge the places I'm looking at. Again, where can I find a hospital to get a procedure done at a good price and a good quality. And finally, the who is what I would call the granular answer, right? Where can I get a good price at a good quality? And in this case, it's specifically where, right? So I might narrow it down to a city, but then I really want to know which specific hospital so we'll call that the who in this case. That's the granular answer. In the end I don't really care about when, why, or how, right? I don't care why a hospital is lower priced. I don't care when I'm going to go. It has nothing to do with this data. It's really about the who, the what, and the where. The hierarchy of that then, if you think about it in sort of a structural standpoint, is I'm starting off with the what, right? I care a lot about the price and the quality. That's going to lead me to the where, right, maybe which city I can narrow it down to that has lower than average pricing and higher than average quality for the procedure I'm looking for. And finally that's going to lead me to that granular answer, the who. What specific hospital can I go to to get my hip replaced? So I did this visualization looking at a bunch of hospital pricing and quality data, and as I said before, the most important thing in this case was the what, right? I'm looking for low price and preferably high quality care. So the default view for this visualization is looking at procedure pricing. So it's sorting by pricing. So I can see that Los Angeles is the most expensive place to get a hip replaced, $223,000. And if I really just care about price, then I can go all the way to the bottom and find the least expensive place, Appleton, Wisconsin, and if I only cared about price, maybe that's where I should go to get my hip replaced. But, I do also care about quality. So I might then go in here and sort by quality and see that there are high quality hospitals in Miami, in Florence, South Carolina, and Portland, Maine, and the green bar indicates below average pricing in Portland so that's a good candidate. We find that promising. But in the end, I also might want to look at quantity of procedures performed. Who's done a lot of these which you would think probably correlates with quality. And I can see that in fact where I live, Boston, has done a lot of these, has below average pricing. So I'm going to click in there. I don't need to see it on a map, but it sort of led me to my home town which is kind of interesting. But if I click in there, I can immediately see the granular answer, right, the who. And each one of these dots represents a hospital. The large dot means they've done a lot. And green means above average quality and below average pricing. And I can see New England Baptist Hospital has a pretty good price, above average quality, above average quantity, looks like a good place to go to me. Hierarchy really is everything in information design and the six Ws are a fantastic way to get started thinking about the information you have, the questions you're trying to answer for your audience, and they'll help guide you to the right way to organize your data.

Three more Ws

So in the last movie we talked about the 6 Ws and how useful they can be to help you find the hierarchy in your information and figure out not only what story you're trying to tell but how to tell that story in your visualization projects. There are actually 3 more Ws that I think are equally as important. In this movie, we're going to talk about the other questions you should ask yourself when going through and designing your projects. The first one is what I call What's Wanting? What's missing from the data you have? In the end, you almost always are going to be missing data. Your client gives you data and theres maybe the granularity of the data is missing, or there's a column missing, or questions missing from a survey that you could use. And there are really three things you can do about that. The first one is keep calm and power on. The fact is sometimes that's all you can do. Sometimes you are not going to get the data that's missing. It's good to be aware of it and to know it maybe in a way to help you design around it so you know to not sort of, steer the visualization towards a question or answer you don't have. But the fact is, you're just going to have to kind of power on in the majority of cases. The second thing you can is you can go back to the source. You can ask your client for the missing data. If a field was left out of a data set, they may have that data. They may not have thought that you needed it. Ask them for it. Because sometimes they'll be able to give it to you. The final choice is really the least likely to happen. It almost never happens from my experience. Is you can generate more data. So if you're doing a visualization of survey data, let's say, and the question wasn't asked, you can always go back and ask that question, right? You know, redo the survey, or add a question to the survey. Like I said, it's not very common, but if you can do it, it's a great thing to do. The second one is what I call what in the world? And by that I mean, a lot of times you can find other data from other sources to bring into a project. A lot of times, the data that you have tells a very specific story, but it can benefit from context. And so if you can find data from other sources, maybe it's a census data, or World Bank data, it will help complete the picture. It will help complete the story you're trying to tell in a way that the primary data that you have just can't do. So in the example of the HOSPITAL PRICING visualization that I showed in the last movie, in this case, I was taking data from the Centers for Medicaid and Medicare services, which was really just hospital pricing data. But the question I was trying to answer was where could I go to get a hip replacement, or any other procedure done, for a good price and good quality. And quality was not provided in the data that I was given. So I went out and found also, hospital quality data, which happened to also come from Centers for Medicaid and Medicare services, but this data really provided the context, and in this case, really completed the picture of this. It's not just about getting a cheap surgery done, it's also about getting a good job done. So, quality was very important, and I could not have done it without getting that outside data to help. And so the third one is what I call what's wild? And by this I mean that sometimes you want to take a little bit of an out of the box approach, whether it's to capture a user's attention, or to really tell a new and interesting story. And it can really mean a couple of different things. One is it can mean that you're just going to take a very unique visual approach to your data. You're not going to show just a bar chart, you're going to try to do something very unique visually to show something and we'll talk more about that in other videos. Sometimes it's about bringing in some unexpected contextual data. So maybe instead of World Bank data, you want to bring in some data from another source that is sort of counterintuitive, or different, that might bring some really interesting insights to a project. Or, maybe you want to create a really out of the box interactive experience to help bring the data to life in a new and interesting way. A very simple example of maybe not what's wild, per se, but what's wild helped get me here, was I was doing a visualization, looking at some survey data. So in this case, you can look through and let's say, pick a question. So, in this case was, what percentage of consulting firms think that offline marketing is going to be more or less important over time. And it's about Leaders vs Laggards, alright? The people who lead the industry versus the people who are sort of behind the industry. What's the difference between the two? And as you can see, these are essentially bar charts that sort of slide in from the outside. And, the way I got to this was thinking about what's wild? What can I do that's a little bit different? I was actually drawing on a white board, and I was leaning sideways as I was trying to draw these bars. And it occurred to me that by looking at them sideways, it might bring an interesting perspective. Because in this case, I was really all about showing the gap between these two things. Showing it this way, how these things come to the center, and where the gap is between them, was just an interesting way of visualizing this data. It's not revolutionary, for sure, but its sort of an evolutionary visual display by thinking about what's wild. By sort of trying to get a little bit outside the box. Looking at, from another perspective, just the real percentages as opposed to emphasizing the differences, then more traditional bars that grow from the inside out. But again, are sort of flipped on the side, and sort of animating them to come to life a little bit. So the three Ws will really help you to push the boundaries a little bit, when you're brainstorming, to try to find the hierarchy that you're looking to tell in your projects. You know, what's missing? What in the world can I find to add context to it? And how can I push the boundaries a little bit? How can I bring a little bit of excitement and interest into a project? If you add these to the six Ws, and you will always be able to find your way into projects like this.

Explore your data: Visual exploration

Most of the projects I work on, my clients do the data analysis, but sometimes I have to explore the data and do analysis myself. And I'm not a data analyst. I'm not an expert in this field. But you know it's always a good idea to know how to explore your data to some degree and you really have to think of yourself as an explorer. You really are trying to discover things in the data that you don't know going into it and that really sort of requires an explorer's mindset. Your mind is open. You're wandering around just looking and trying to find new things. And there are all kinds of tools out in the universe to do data analysis and data exploration these days. Tableau is a huge tool that many many people are using across Enterprise and individuals. Tools like Plotly, sort of cloud-based data analysis tools. You know real statisticians and data scientists are using tools like Python or R. There are software tools like Gephi that anyone can download and run to do network analysis. And all kind of mapping tools like Carto which is an amazing open source online mapping tool. So the tools are out there. There's a million and one tools. You can find them in a whole bunch of different ways. I just strongly recommend you think at the beginning of your data exploration as being really about how to think about data and leveraging tools that maybe you already know. And of course the tool that most of us have used at least once if not for many many years is Excel. Or if you're on a Mac, maybe you've used the Apple equivalent which is called Numbers. And so we're going to explore this dataset in front of us and what this is is data from the Bureau of Labor and Statistics, and we're looking at minimum wage data from 1980 up until 2003. And I've also gone and found additional data to supplement what we're looking at to try to find an interesting story line. So in addition to this row which has minimum wage data, I also have data about the poverty line in each of those same years, the price of gas, the price of bread, the price of eggs, the price of electricity, and also the CPI which is the Consumer Price Index. It's sort of like an indexed average number of the cost of living, essentially the cost of sort of everything. And CPI minus which is sort of minus some key figures that they sometimes subtract from the CPI. But long story short, if I'm going to explore this dataset, I'll just usually jump in and start trying things. And so first thing I'll do is I'll select the minimum wage row, and I'll do command shift right arrow or control shift right arrow on a PC to select the entire row. And then I'm going to say go to insert in Excel, and I'm just going to say insert a line chart. Let's just see what a line chart looks like of this data. And so I'm just going to sort of drag this over to the left so it's all in the same area over here. And I can see that minimum wage has gone up and up and up, right? Essentially Congress will sort of pass a change in the minimum wage and then it stays the same for a long time, and then it jumps up and stays the same, jumps up, etc. So this line is interesting, but I don't know anything beyond what the shape is. I don't know what it means although I can start to make some guesses about it. So I'm going to do the same thing, but

I'm going to look at let's say the price of gas next. And so I'm actually going to start on the right hand side and command shift left arrow so just so I end up in the left side of the tab here so I can actually put the charts in the same spot. So I'm going to select that row once again, command shift left arrow since I start on the right. Go to Insert, drop in another line chart over here and boom I have a line chart. So I can see the price of gas has behaved very differently over time. I can see that it's actually stayed kind of the same for much of the time period, and then it sort of went up for you know about 10 or so years there, dropped down, went back up, and has sort of come back down again in 2013 is the dataset that I have here. So let's do the same thing. Let's look at the price of let's say eggs. And I say insert line chart. And what do I see here? I see again sort of you know it's own set of information here. The price stayed pretty stable, went up, came back down a little bit. Let's do the last one. Let's look at the price of electricity out of curiosity. Actually let me do the CPI. So if I go to the CPI and I select that row and I say Insert chart. This one's going to look very different because if you think about it, what I'm looking at now is the overall cost of living when I take the price of literally everything in the economy pretty much. And so yeah of course, this says what inflation is. I can see the price of everything just sort of rising steadily. So this is interesting. I can see stuff here, but I have no idea what it means other than the basic conclusions that I've just sort of explained. So another way to explore this is to look at things in a little bit more of a sort of a structured, maybe strategic way. And one way to do that, especially when thinking about something like minimum wage, is to think of it about the minimum wage not as a dollar value, but as a purchasing power value, right. So in other words, how much gas can I buy with minimum wage? And so the way I do that is I just created a ratio. I just literally did a formula. So if I double click on it, you can see what I did is I took B2 which is a minimum wage and divided that in this case by B4 which is the price of gas. And so then I did that for every single column. So for 1980, I could buy 2.79 gallons of gas for minimum wage. For 1981, it was 2.63, etc., etc. And I did the same thing for the price of bread, eggs, electricity, and all of my items here. And I just charted each one of them. So I can see that minimum wage could buy me more and more and more gas until eventually it was less and less and less gas with some spikes along the way. Bread has also sort of pretty much steadily gone down with a couple of minor spikes. Eggs interestingly has sort of gone up and down and up and down, went way up at one point, went way down. And unlike bread and gas, I can actually buy a tiny bit more eggs, at least in 2013 dollars, than I could in 1980 dollars. So charting these things tells me something interesting. I'm beginning to get to something pretty interesting. But I get into trouble when I try to do a line chart of all of the data points all at the same time. And so this is the line chart that I generated to do that and as you can see, the reason it's problematic is that I have essentially all of my elements down here. CPI, price of gas, eggs, etc. But then electricity is way up here just because the scale is completely different. So if I scroll back up to my data here, the price of electricity, you know minimum wage to electricity, I could buy 58 units, whatever is being measured here on minimum wage versus two units of gas, six units of bread, etc. So the numbers are just so different that this chart of all of them together is just really hard to read. So instead what I do is I generated spark lines and spark lines are kind of a cool feature in Excel, and I'll just show you how to do that real quick even though I'm not really here to teach you Excel. If I select the entire dataset, and I can't use the labels, I have to just use the numbers here. And I command shift right arrow and then command shift down arrow, I get the entire range of data, the whole sort of table of data. And then I can go up to Insert and I can choose over here Spark lines and I'm going to do a Line spark lines. And so now it knows what data I want to select. Then I just have to select six rows to place it into because it always has to be the same number of rows, and once I say okay, I get these little spark lines. Now if I scroll over and zoom way in on these, spark lines take the whole scale problem out of it because essentially it's like they're all on the same scale. So I can see in this case that the numbers sort of went up and then came back down. In this case they just sort of went steadily down. And down here I have two lines that look pretty much the same, and as you can imagine, these are my CPI scores, my CPI values. So what does this tell me? It tells me that on minimum wage compared to the entire consumer price index, my ability to purchase against all things in the economy, has sort of steadily gone down, went up a little bit here, and then back down. So I haven't really done deep analysis here, but sometimes just creating visualizations, looking at data using different chart forms, in this case a bunch of line charts for the most part, has really started to reveal some trends in the data. And so next time, I'm going to explore other ways of looking at data in Excel.

Explore your data: Indexes and ratios

So I've talked about how sometimes exploring your data is best done just by starting to generate charts, really just starting to look at the data, and sometimes a visual representation of the numbers is the best way, often times, the best way for you to see what's interesting to see. For instance, so here we have our minimum wage data. There's no way I could have scanned across this row of numbers and seen what I see here, that it goes up and stays and goes up and stays and goes up and stays. Maybe I could have seen that because these numbers are literally the same, but I certainly couldn't have told you what the pattern is in gas, that is sort of goes up and down and up and down and up and down and then shoots up and shoots down, et cetera. So visualization is a very powerful tool to explore your data, which is why I start generating charts. So first I generated charts of just the actual numbers themselves, then I looked at the ratio, how much gas could minimum wage buy me over each year, and the other things, charted those. But as I mentioned previously, when I try to chart them all at once I ran into a problem. And that's because the price of bread and eggs and gas are all sort of similar numbers, two, four, six, down here, but the unit of electricity is a much bigger number, so looking at them on a chart and making comparisons between them gets very difficult. Which is where indexing can sort of come in and help save the day. So when I explored this data I decided to do an index because an index essentially will turn all of my values into a value between zero and one. And I'm not going to explain exactly how to do indexing, I'm going to do that in a convert your data movie, but the basic idea in indexing, if I double click you can see the formula, I take my number, how much gas I can buy on minimum wage in 1980, and then I divide that number by the maximum value for how much gas I could buy in the entire data set. And that turns it into, as I said, a number between zero and one, so here, it turned it into .556, or sorry, .56 essentially, which tells me it's sort of in the middle of the pack, and the maximum value in this entire set is wherever I see a one, because I'm dividing that number by the maximum number and so it's the same thing, divided by itself, therefore it's a one. So you can always see where the maximum number is when you see a one when you do an index. Long story short, if I chart these indices, now I can see how much minimum wage could buy me in terms of gas on an index from zero to one, and the chart in this case looks the same, but it's just, it's the exact same values but turned into ones instead of turned into the actual dollar values. Why did I do this? Wny does it matter? It's because when I want to chart them all at once, now they're all in the same scale. Now they're all at one for a maximum value, and zero for a theoretical minimum value. So I can easily see the red line here, which is my price of bread indexed, has just sort of gone steadily down, versus my purple line, which is the electricity index, came down, went up, came down, went up, and then down a little bit. So it's easier to see them, even though this chart is hard to read, it's sort of easier to see them all on the same scale. Indexing is actually very similar conceptually to the spark lines that I showed you earlier. Because the spark lines essentially also turn all of these values into the same scale, in fact what it really does is removes scale. They're all in a scale only relative to themselves; whereas, now I can see them compared to each other. It's sort of different scales for each one, although they're all on a shared scale; whereas, a spark line literally sort of removes all scaling and I can just see the trend in the line itself. So as I'm exploring the data I do it a bunch of different ways, I visualize different charts, I try different things like ratios and indices, and then the last thing I will always do is I will try different chart types. So I generated a radial chart of all of my values, which is a little bit harder to read. I generated a scatter plot of all of my different values where they're all really clustered together. If I were to sort of shrink the scale on this I might be able to see patterns in here, maybe not. Or this crazy donut chart, which really tells me just about nothing, but you know, you try it, you try the different charts you have available to you in your tools and see what you can see in your data. I always recommend you don't just try the charts available to you in your tools. Once you sort of think you have an inkling and a hypothesis about your data set, go out, find different tools, try different charts, get inspiration about different charts that will help you see the things in your data to actually explore them further. And of course you might need to use more sophisticated data analysis tools than what I'm showing here. So, this is not a class in data analysis, this is data visualization class, and so I'm using a visualization as my primary mechanism for exploring my data set.

Convert your data: Indexes and ratios

I think we can all agree that life doesn't always work out perfectly, right? Your prom may have been a disappointment, your boss may not praise you as efusively and frequently as you would like, and maybe every now and then, you get a flat tire, and when you do, it's probably raining out, right? But, the thing is, I believe, and I think evidence has shown me personally, that life usually does generally kind of work out okay and for the best. However, despite this rosy unicorn-and-rainbow-filled outlook, unfortunately, in data visualization, data never comes in the form that you need it in. You will always, almost always, have to transform your data, convert your data, in order to make it usable for you to do the work that you need to do. So, there are a bunch of different things that you tend to have to do, and essentially, I find these five to be the most frequent, and I'm going to go through them one-by-one, starting off with indexes and ratios. So, what we're looking at here is data from the Bureau of Labor and Statistics, and what we have is, essentially, at the top, we have minimum wage data. So, this is literally the minimum wage for every year from 1980 until 2013. We also have the poverty line. We have the price of gas and bread and eggs and electricity, as well as the consumer price index, which is just sort of an aggregated value of the cost of living, essentially the price of all kinds of things all mixed together, and then, an alternative to CPI, which essentially is just like the CPI, but minus some key numbers. Long story short, this is all about how much stuff can I buy with minimum wage, you know, what is minimum wage, that was sort of my idea behind collecting this type of data together. And so, I looked at this data initially, and I said, "Hey, let me just sort of explore it a little bit," and I'm now going to just sort of jump to the right, Cmd right arrow, and then, Cmd-Shift-left arrow in order to select that entire row, and then, I can quickly generate a chart. This is what I did in Explore your Data. And I can see, essentially, what minimum wage was in 1980, and how it went up and sort of stayed the same, and went up and sort of stayed the same, all the way up until 2013. And, I can do the same thing with the price of gas. Go to the right, Cmd-Shift-left arrow to select that row, insert chart, and I'll do the same thing with eggs, why not? Take the whole row, insert a chart, and I get some stuff. I can see these things, and I can see that minimum wage went up. I can see gas sort of generally stay the same, and then, sort of spiked up and spiked down and spiked back up, and eggs have also sort of generally gone up with some ups and downs along the way. So, this tells me something about the prices of these things, but I don't know what it means in the context of, "Well, is this good or bad? "Is it easy to live on minimum wage," etcetera. So, this wasn't enough to look at the data, so the next thing I did, if you remember in the data exploration video, is, I converted these numbers into ratios, right? I said, "How much gas can I buy on minimum wage?" And I did that by dividing the two numbers, right? That's what a ratio is. A ratio of minimum wage to gas. And so, by dividing minimum wage by the price of gas for every single year, I got these numbers. So, I could buy 2.79 gallons of gas for minimum wage, you know, one hour of minimum wage work, in 1980. And when I chart that, it's a very different line, right? I could buy more and more and more gas, until I could I buy less and less and less gas, except for a spike, and then, less again, versus eggs, which sort of went up and down, but actually ended up slightly higher in 2013 compared to 1980. So, I did that for all of my different numbers, so these ratios were very helpful, to look at these numbers, until I ran with this issue of scale, and then, I said, "Okay, how am I "going to solve the scale issue?" And I tried spark lines, is one way. And then, I also looked at creating an index, and indexes are a great way at looking at values like this, and I'm actually going to sort of delete the index that I precreated here and recreate it for you, and just show you how to do this, because, you know, you may find yourself having to create indexes fairly frequently when you look at data like this, and so, it's a good skill to know how to do. So, to create an index, and I explained this briefly in the other video, but I'll explain it in more detail now, you select the value you want to index, and then, you divide by the maximum value of all the values that you're comparing it against. So, in this case, we can actually use an Excel formula called MAX. And so, if I say, "Okay, I want to select the maximum value "of the same exact row." I can just click on that row again, after typing in MAX, parenthesis, and then, right-click, sorry, Shift-right click, to select the entire range, and you can see up here, it said MAX B14 to AI14. Finish the parenthesis, hit Enter, and now, I have that index value over here. Now, watch what happens if I try to click and drag this. It might do what I want it to do, but it might not, and the reason is, if I double click this guy, you'll see that now, it's taking the maximum value of all these, the same range, but it's moved over by one, because I haven't locked the cells, and this is one of those tricks to Excel. If you're not a heavy Excel user, you may or may not be aware of it. But when you create a formula like this, if you're going to copy-paste or click or drag, you have to realize that, I want, when I click and drag, I want this number to change, the first number, because I always want it be, you know, in 1980, I want to divide this number by whatever, but in 1981, I want to divide this number. So, it's okay if my B14 changes to, in this case, C14. However, I don't want the maximum range to change, because I always want to divide by the maximum range from B14 all the way to AI14, and the way I can do that is simply by changing this, so I add a little dollar sign. It's always Column B, and it's always column AI. Now, I didn't add dollar signs in front of the numbers, and you'll see why in a second. So now, if I click and drag, now, you'll see, if I double-click on it, I'm dividing this number, the blue, by the same denominator all along the way. So, I'm just going to click and drag this guy all the way to the right-hand edge of the screen here, and now, I'm going to go back to the left, I'm going to select this entire row, and I'm going to click and drag all the way down, six rows, and this is where those dollar signs that I did not add come into play, because essentially, by not putting dollar signs in front of the B19, it allowed that value to also shift down, just like I was letting it shift to the right, and by not putting dollar signs in front of the 14s up here, it allowed, when I clicked and dragged down, for it to change row by row by row. It's still using the same column values, because the dollar signs in front of those guys, but by not putting a dollar sign in front of the 19, or in this case, the 14, when it was up here, it allowed it to sort of pull down. Definitely, this is not an Excel class, so if you don't understand what I'm talking about, go take an Excel class, and you can understand what cell locking is all about, but it's a very important part of the work you will do in Excel, so you don't make weird mistakes, cause it is easy to make mistakes in Excel. So, ratios allow us to see things about the data, and then, indexes allow us to compare the numbers within a much tighter scale, because now, all of these numbers have been converted to a number between zero and one. So now, my charts all look exactly the same, the individual charts, as they did before, although the numbers on the axes are different. But now, when I look at a chart with them all together, they're all on the same scale. They're all between zero and one, so I can see the difference between the price of bread in red, versus electricity, all in the same scale, and so, it's much easier to make comparisons between them. So, you know, spark lines can do something similar, but you may find yourself using ratios and indexes quite frequently, so I definitely recommend that you take a look at, you know, really get familiar with how to do this type of work in Excel, or whatever tool you're using.

Convert your data: Percentiles

Another thing that I do very frequently to convert my data, is I turn numbers into percentiles, because often times when you have a list of things in, let's say, rank order, the rank is helpful. Like in this case, I have GDP, right, the gross domestic product, essentially the size of the economies of 181 countries, and so I can see that I'm sorted by the GDP in this case. So I can see that the United States is number one, with a 16 point something trillion dollar economy. China is number two with an eight point something trillion dollar economy. Now right there I can see why rank order might fail in certain types of analysis, because China is number two, so it's almost as good as the United States, great. But you know what, it's literally about half the size. These numbers are from 2012 or 2013 by the way. So it's literally half the size, so being two is not close to the same thing as being number one in terms of the overall size of economy. Another example of that is if I take just Germany in this case, which has a three point something trillion dollar economy, and Sudan, which is a 58 billion dollar size economy. Germany is number four, so they're right up near the top even thought it's literally like just barely over 20% of the size of the United States. Actually maybe a little less than 20%. Sudan is actually number 68 on the list at 58 point something billion dollars. So, even though I can look up those rank orders, as I said, the rank order may not be that informative, and also rank order out of 181 is hard for me to exactly figure out what that really means. Now, 181 isn't a crazy number, but what if the number was 4,273, if that was my list count? Now if I had a value that was 1,267 that out of 4,000 whatever, you know it just gets harder to sort of understand what that really means. And so percentiles can help solve that problem. And so I'm just going to show you how I would calculate a percentile in this example. First of all, to do a formula I start with the equal sign, and percentiles are always one minus something. And that's because what we're trying to do is to figure out essentially a decimal out of the maximum value, the count, and then we subtract it from one because the highest value is essentially the closest to 100%, and the lowest value is the closest to zero percent. So essentially what I'm doing is I'm going to say one minus, and then I'm going to say the row number of this thing that I'm looking at, divided by the count. And so if I hit enter on that, I see that the Unites States is in the 99th percentile, meaning that it is higher than 99% of other countries. Now, if I change the formatting, I'm going to hit command one, or I can also go to format cells in the menu, and if I changed this instead of being as a percentage, if I just say show me as a number, you will see that the number is actually, it is a decimal. So one minus the row number, which in this case is number one, so it's one minus one, divided by 181. That's sort of what the formula is doing. So if I do command one, or control one on a pc, and turn it back into a percentage, and I'm going to get rid of the decimal places, it's 99%. Now, I've talked about cell locking before, and we're going to see a symptom of the problem here. If I just click and drag this down, I've got problems here because the first row, instead of dividing the row number by this number, it's dividing the row number by this number, which is not what I want to do. I always want to divide by my count, the count of things in the list to figure out what percentile this thing is in that list. So I have to do the row locking, in this case, of the column and the row, just by putting those little dollar signs, I don't even really need to do the column, I could just do the row, because I'm not moving left or right. But now if I click and drag this all the way down, I can see essentially where everything belongs. So if I click and drag all the way to the bottom here, I can see, as I should, that Tuvalu, being the last on the list, is in the zeroth percentile, zero percent of countries are below it in the list. Okay so this is all it's doing, it's telling me where something lives on the list. So Germany, at number four, is in the 98th percentile, Sudan at number 68 is in the 62nd percentile, meaning 38% of countries are lower than Sudan on this list. So percentiles are a great way to sort of get an understanding, in addition to rank order, where something lives on the list relative to its peers in a sort of a number scale, zero to 100 that we're used to thinking of things like this.

Convert your data: Aggregating

One of the most frequent tasks that you have to deal with in converting data is taking transactional data, where each row of data represents like one data point essentially, and aggregating it, so you can actually perform calculations and visualize the data in summary form. So a great example of this is I had this data in front of you, which is Congressional voting data, and so this was many thousands of rows of data. I can't remember how many, but essentially, each row of data represents a bill that was being voted on. That's what this number is here, this bill_id. The vote, Yay or Nay, yes or no. For each individual person, this person here, what party they belong to and what state they're from. So literally, this is one person's vote on one bill during this session of Congress. Like I said, thousands of and thousands of rows. And I wanted to do a visualization looking at partisanship. So what I needed to do is perform a calculation to figure out, okay, so for each bill, did each person vote the same as his or her party, and do that for each and every bill that they voted on. And so I literally wrote a script that just went through, looked at every single row of data, figured out for each person which way do they vote, yes or no, and was it the same as the majority of their peers, essentially a sum of votes for all of the Rs or Ds and do they match it or not match it. And I used a script to essentially convert that into one row per person. So I can see that this person voted with their party 504 times, voted against their party 29 times, so therefore, I can tell sort of their percentage of partisanship, as compared to some other people in the database. Just literally by creating a script that did a sum, like literally how many times did they vote the same as their party, row by row by row. So I turned that, essentially, many thousands of rows of data into essentially five hundred and something rows of data, one for every Congress person, just by sort of aggregating all those numbers all together. This isn't the only way you can do this type of activity. I was working in SQL, in this case MySQL, so I was writing SQL script, but in the next example I'm going to show you, there's a way to do this within Excel, using something called a pivot table.

Convert your data: Grouping

A lot of times you get data that's really sort of transactional and it looks like this. I have a row of data for every sort of data point along the way. I'm back to my minimum wage data. I have minimum wage at the category. I have the amount, so three dollars and ten cents and the year, 1980. This was the minimum wage in this year. Here is the minimum wage in 1981, et cetera, et cetera, et cetera. Now, I have the same data for the price of bread and each one in a row and the price of gasoline, et cetera. This data is the same data that we've been looking at in some of the other videos in this course, but long story short each row is a transaction it just makes it harder to do certain things. It's not impossible, but it makes it a little bit harder for instance, for me to generate line charts for every single on of these elements because unlike before where I could use keyboard shortcuts and jump around and generate charts. Now if I want it to do a line chart I have to sort of click and drag and make sure I stop at the right place and manually do it. It's just harder to do and there's certain things that are almost impossible to do. It's also really hard for me to get a sense of the average values or the sum of values with things like this, but by thinking in terms of aggregation and also in sort of regrouping the data, it makes it much easier to perform tasks like this. One of the great things built into Excel is something called a pivot table. I can go into Excel and if I wanted to create a pivot table I can literally just click anywhere in this range of values. I don't even have to select the table. I can go up to insert and say pivot table. It now knows what data I want to use and I can say put it in the same worksheet or in a new worksheet. I'm just going to put it in a new worksheet for now. I get this thing. This is where my pivot table is going to live and I get these controls over here which essentially are what I want to do with my data. This is not a course in pivot tables. There are courses on pivot tables in Excel all throughout the LinkedIn Learning Library so check those out. At a minimum, the thing I always remember which I think I learned from one of the LinkedIn learning courses was that you should always think of your value first. Essentially, what I'm going to be doing is I'm going to be dragging fields into these sort of buckets in order to generate my pivot table automatically. The first thing I think about is my value meaning the number I'm thinking about and I'm worried about. In all of the cases, I'm thinking about the amount. How much is minimum wage? How much does it cost to buy a gallon of gas, et cetera? If I think of my value first and I drag that into my values field, what happens is my pivot table actually starts to create. I'm just going to zoom in so you can see it. By default, it thinks I want a sum of all of those values. Well, that's kind of weird. That's not really what I want. What if I wanted a row for every year? A single row for every single year of my data set. If I now click and drag and put year down in the rows, now I'm saying give me the year in the rows, a year for every row. That's exactly what I see. For 1980, the sum of the amount for that entire year for whatever it is I'm looking at which is still not a very useful number, but I can see the sums for every year. If I wanted to then say by category, give me columns per category, now what I can see is that for bread in 1980, the sum of the values is .501. For gas it's this, for minimum wage it's that. So long story short, it makes it a very quick and simple way to essentially take my transactional data and turn it into an organized, aggregated view of my number. By the way, you don't have to only use sums. You can do things in here and say show me the average values, show me the max values, show me the data as percentages or difference from a row total. There's all kinds of ways you can perform calculations while you're reorganizing and aggregating your data.

Convert your data: Data formats

One of the other most frequent tasks you're going to have to do is, you're going to have to convert data from one format to another. And so for example, what you can see here is JSON, Java Script Object Notation, and this example looks pretty crazy, this is actually the GDP data that we've been looking at, and Java Script Object Notation, JSON essentially has these curly braces on either end of a data value. Sometimes is has straight braces, straight brackets also, which I'm not going to get into why that is. It has field names, and then a colon, and then that fields' value, and each individual value is separated by a comma or each individual field rather, so I have an ID, SL, and I have a value Sierra Leone in this case. And so, long story short, it's just organized this way, it's a just a text file. There are other types of text-based data storage formats, CSV files where every value in every field is separated by a comma, TSVs every value is separated by tabs, et cetera, et cetera. You have XML, so many different data formats, and often times you get data in a format that isn't the format you need to work in. Now, luckily there are all kinds of tools to make changes that you need to make in order to look at your data. So for instance, if I were to Google CSV to JSON, 'cause maybe I got a CSV file, but I need to work in JSON 'cause JavaScript likes JSON format, I can actually copy and paste in CSV data, like this example data here, comma delimited fields and values, and if I hit convert, it's going to convert it into a JSON format. Languages, most modern languages also have built into them, the ability to work in different formats. So like PHP as an example, doesn't read JSON natively, but using json_decode, I can take a JSON file and work with it in PHP. It'll turn into an array, and then I can use json_encode to save out a JSON file back to a web browser for usage with JavaScript, or some other code. Long story short, this is just part of what you have to do very very frequently as part of your data transformation process, and so you can expect to run into this quite a bit. So that was essentially five different ways you may have to convert and transform your data. The five most frequent that I find myself running into, long story short, learn to embrace it, learn to accept it, because it is nearly inevitable that you'll have to make transformations of your data before you can work with it. So, learn the tools and techniques, and more than I've shown you here really will be helpful for you, but this is a good start.

Sketches and wireframes

I'm a computer user. I spend 99% of my day at my computer. I am digital all the time, okay? I am no Luddite. However, when doing data visualization, I always, always start with pen and paper or on a whiteboard. It is absolutely critical to work this way in an analog way. And I want to talk about why? There are a few key reasons, okay? The first one is speed, okay? I'm going to talk about all four of these in detail. Next is flexibility. There's also scale and a body mind connection bit which is a little bit harder to articulate, but we'll talk about it. So let's talk about speed, okay? You can whip up a sketch super fast when working in analog, okay? You can just really just draw a rectangle, put some lines on it, draw a little line chart, et cetera, ridiculously fast. You don't have to fight with a mouse or with software features, bizarre curves in an illustration tool, et cetera, et cetera. Now, the quality isn't always good as you can see here, but we're not going for quality. This is purely ideation. This is brainstorming. Just trying to get to an idea that might work. You can iterate on multiple ideas and multiple ways very, very quickly when working with analog tools. So, then you can find maybe an idea that after a very quick sketch seems like it might be worth pursuing further then you can go to software to start to think about how you might refine it. The second argument for using analog tools is flexibility. You can push the envelope without knowing exactly how you're going to pull it off technically. And that's okay, you can figure that out later. So for example, you may use complex shapes or weird gradients of color. You can whip up these concepts without worrying about Photoshop, Adobe illustrator, tutorials, animation software, whatever it is you're using or if you're using code especially, how do I do a curve in this code again? How do I do a gradient of color? It doesn't matter. You'll figure that out later. First, let's see if it's even going to work conceptually, okay? You don't have to think about how you're going to code it. The idea is to get ideas down on paper quickly and ignore feasibility questions. You don't need to create layers in Photoshop and label things, fight with text tools versus drawing tools, et cetera. You don't have to worry about alignment, getting things perfect, right? You'll get to a perfect version later on. You get to revel in the temporary nature of this medium. The idea that you're just testing, okay? Analog lets me mock up opacity with just hash marks or dots or just scribbles much faster than selecting a color, going to options, changing a Pasadena, et cetera, et cetera, new layers. I don't have to fight my software to try stuff out. The third reason is scale. I like to draw large, which is another advantage when you're working on a whiteboard. I have a full size painted whiteboard on my wall. I can make drawings that are several feet long and wide. And this one's hard to articulate. I don't know why this helps, but it does. Being able to stand up to the board and sort of step back and look at it, big picture. Look at several visualizations at a time, getting closer, work on the details. I can sort of see the trees and the forest, like I said, hard to articulate, but I do feel that this adds value when I'm ideating on something. And then the last one is the body mind connection. I don't want to go all Zen yoga on you or maybe I do, but there's something magical about the physicality of drawing with a pen and paper, whiteboard, et cetera. Taking abstract data, making sensory experiences, visual experiences and essentially making those ideas become sort of physical, right? My theory is that I find it much more satisfying when I brainstorm physically in this way. And it helps me actually understand and really, I don't have to imagine the idea I get to see it and there's something to that. I would also say that I think I create better work because maybe my whole body is involved not just my brain. But again, that's just sort of my theory. So, great, maybe you believe all these arguments that I'm making and that's wonderful, but maybe you say, oh, but I can't draw, okay? Well listen, you've been watching what I've just been showing you, these are all actual sketches from me. They're terrible, right? I can't draw a straight line. I can't draw my way out of a paper bag. So this is not about drawing. This is about brainstorming visually capturing concepts. And even though I can't draw either, right? I have to sort of say the obligatory, you know what? Actually, everybody can't draw. It really is true. We have the capacity to communicate visually. Yeah, sure, some can do a more refined, detailed photorealistic work others of us, maybe can't quite pull that off fine. Doesn't matter. Some of us can just approximate an idea that we can then perfect in other ways, right? But we're just getting the idea down. We don't need to be able to draw well to do that. And second, by the way, here, it really doesn't have to be analog, right? Your old tools aren't necessarily truly better. I'm a digital guy. I occasionally use an iPad or now I have what's called a remarkable tablet. The point is to use analog methods, right? To use sketching software, not design the software. As many barriers as you can remove, to iterating quickly to testing ideas quickly, move quickly, move flexibly, bring your body into it, generate as many ideas as you can and then refine from there. Do what's comfortable and you can't go wrong.

Defining your narrative

So, visualizations are more than just data or charts or maps, at their best they're a story. Humans really have been telling each other stories as their primary means of communications and education for, really, tens of thousands of years. Whether around a campfire or using modern media, the fact of the matter is that there's a lot of power to the phrase once upon a time. We're wired for stories. And visualization takes work to create stories out of it, and some of that work really has nothing to with the visualization. It really is about the storytelling aspect to it. Just like great design is really about strategy and thinking and planning before the first stroke of the paintbrush, it's no less so in visualization. Visualizations, especially interactive ones, aren't necessarily consumed in a linear way, like a book. You can't necessarily control how your user processes the information you give them, but you can structure a story in a narrative way with a narrative process. You could encourage users to walk through the information in a linear, progressive way even though you can't force them to do it. And so, let's look at stories. The fact is that stories have a very simple structure. Every one has a beginning, a middle, and an end. A more nuanced view includes a few more elements. The beginning of the story and then the protagonist faces a serious challenge, and then there's the middle of the story and then there's some sort of climax. Then there's the denouement, that's sort of like, after the climax, things start to wrap up all the way down into the grand finale where the book closes. If you think about your visualization in these terms as much as you can, you really can't go wrong. And so, we're going to design a visualization together using this structure as our guiding principle. So, the first thing you do is you define a headline, and there may be more than one headline. In fact, I recommend that you go through the process of thinking of more than one. I like to think of it as coming up with a New York Times headline, which is a very serious and journalistic and straightforward headline. And then there's the New York Post version, which is a little bit more salacious and crazy and scandal-ridden. Sort of the fun one versus the serious one. And sometimes for your visualizations, you may lean one way or the other, but play around with headlines. The headline is the most important top line summation of the story you're trying to tell. You can also think about the headline in terms of, what will people tweet when they share your visualization? And you also want to, when you're coming up with your headline, actually use blank spaces when you're thinking about it for the data parts. So, in other words, you want to minimize bias. So, if I'm coming up with a headline about the best place to get a hip replacement is blank, 'cause my visualization, I'm going to be finding the answer to that question, but the headline, the main point of it is the best place to find a hip replacement is. That portion you want to lay down pretty hard and fast on your infographic. So, then your next point in the hierarchy of your data visualization is your introduction. So, in an infographic like this, it's, of course, just a paragraph. It's establishing the premise and the context for the visualization before diving into the data, and it belongs below the headline at the top of the page to draw the eye there first. So, again, this is conceptual and structural, using the story as our guiding principle. This is, of course, not meant to be the design, quote, unquote. If you think of it in the story standpoint, that's the beginning of our story, is the headline and the introduction. Now let's move on to the challenge. So, there are a lot of ways to add the, quote, unquote, challenge to a visualization. One is simply to add a callout, big, bolded text like this that calls out the main idea of the story you're trying to tell or the problem that's being addressed or the question that your visualization is meant to answer. So, in this wire frame, this bold type will really draw the eye. In a real graphic, of course, it could take different visual forms. And so, we move on to the meat and potatoes of the story. So, we've set up the premise, we've gotten the user into it, now we're going to load them up with data, the meat and potatoes, and here's where the chart with all the detail and the interactivity and imagery is going to go, in this gray area in the middle. This is where the user is going to spend most of their time really exploring and playing around and looking at the data. So, how do we introduce the climax, the sort of pivotal moment in the experience? And we're going to look at two different scenarios, one is a static infographic and one is an interactive chart. So, in a static infographic, you have information here, in this example, this wire frame, that's really meant to be consumed linearly. Of course people are going to start on the left if they're left-to-right language readers. They're going to follow the arrows, and more likely than not, end up on the right-hand side. But if I had a climax, a big story, a special thing that I wanted to draw attention to, I'm going to highlight it, maybe in a visually interesting way. So, I'm going to create a focal point that's going to contain the climax. It could actually contain the conclusion or the challenge, too, this type of a visual focal point treatment. This might draw the eye there first, of course, and now I've killed that linear progression experience 'cause someone's going to read this first. But again, the idea isn't necessarily to enforce linear consumption of the content, it's just to use that structure when you're thinking about it. Now, if I was trying to do this in an interactive graphic, the approach might be a little bit different. You can enforce a linear consumption of the content by having a Next button and making your users click Next and Next to reveal one piece of the story after another exactly how you want to tell it, and then you can do the big reveal at the end of that process. So, now we have the exact same visual structure as the static infographic, but we're forcing the user to get to it piece by piece, bit by bit. Or another way to do it is to have sorting and filtering buttons. So, you might put your filters in a certain order, and again, assuming people are going to go left to right, which is a relatively safe assumption, but not always. You can encourage a user to use certain filters first, clicking on the left, and then when they click on that filter, you might be able to highlight certain things. When the data reveals, one bar is taller than the others and you can use a callout to highlight something interesting that happened there. So, again, the user is progressing through the content, and there isn't necessarily a denouement. Sometimes you have that shiny object syndrome where they'll come into interactive, filter, filter, find something interesting, and then they just leave. You don't get the opportunity to conclude things for them. It's hard, in that situation, to offer conclusions and to deconstruct things for them sometimes, but it's a really good exercise to try to think about how you would do that. So, if we go back to our static infographic, it's really easy to offer a conclusion on the page. So, we're used to seeing it in the lower right-hand corner. You have a box with the small print, the footer. It's very easy to see structurally. I would also strongly recommend that you always include your sources. Where did this data come from? Why should I believe you? Why is it valid? That's sort of part of the conclusion in a story structure. In an interactive infographic, you can also offer more conclusion-like things. So, after your users go through this deep exploration process, they're trying to find deeper information and their own a-has, they can deconstruct the info themselves quite a bit, but then maybe at the end, there's a Next button and they click through and you can sort of offer them this summed up view of the data. Maybe you have a spokesmodel celebrity who's going to come on camera and speak to the users about the most important conclusion from the data. So, there are a lot of ways to introduce conclusions and denouements into interactivity even though you're never guaranteed that people will experience them. I think it's really helpful to think of every project, even your non-linear interactive ones, within that linear storytelling framework. It helps you work through functionality logic as well as the hierarchy of information, and it also helps you measure your success. Did I present the conclusion, the finale to my users in a way that they're most likely to consume it? Even though you're not always telling a linear story, it's a really helpful way to think about things.

Making everything relatable

Imagine you were dropped into a village in the Amazon, and you had to describe to the people there who you were and where you were from, would you start with a joke about the most recent celebrity scandal? Maybe segway to your resume and maybe your academic career. Never mind the contents, just speaking English, of course, would be a problem, right? Like in all communications, in Data Visualization you have to connect with your audience, you have to speak their language, literally and figuratively. For DataVis, there's a very subtle analog to that, in that data is inherently confusing. It can be overwhelming and unrelatable, even if you share the same culture and language, you really have to take extra care to make the abstract data that you're talking about relatable to people. In fact, it's really your primary job to make that data relatable and understandable, which are very strongly correlated. So I was doing a visualization on forest sustainability and it had few different stats in it. This is one of them. So, the stat is, that the demand for wood worldwide is going to triple by 2050 to more than 10 billion cubic meters. So, numbers like this are very recognizable to people who are in the industry, but this infographic was for public consumption. And so, how do we take this number, 10 billion cubic meters, and turn it into something that people relate to? They can sense that it's a large number, but they don't really know what it means. So maybe 10 billion cubic meters is half the forests on Earth, or maybe it means all the wood burned in every fireplace in the United States for the past twenty years. I'm making these up. Maybe it's all the wood you would need to replace the carpets on every floor of every house in Cleveland. These are very tangible, metaphorical ways to relate that number to something. So, my client gave me a different metaphor. And their metaphor was that it's the equivalent of 10, 000 Empire State Buildings full of wood. Now, this is a pretty good start. I can actually visualize that to some degree. I know that the Empire State Building is a very large thing, I've stood next to it. And I can sort of imagine, or maybe it's the fact that I can't imagine it, but I can imagine how overwhelming 10, 000 of them would be. So, it's a good start, right? What does it mean, 10, 000 Empire State Buildings full of wood. I can definitely visualize it, which is half the battle, but I would argue that it's not the best metaphor in this case, because it's not really on topic. Saying that there are 10, 000 Empire State Buildings worth of wood, doesn't help me think of it in terms of sustainability. So let's move on to the next example from this project. The next stat was, that global forest carbon stocks are estimated to be 861 billion tons. So, what this actually translates to, is that all of the forests on Earth store the equivalent amount of carbon as 861 billion tons. Okay, so that's what this means. So again, thinking about it, the client gave me this metaphor, this comparison. They said that that number is the equivalent of 27 years worth of fossil fuel consumption. So in other words, 27 years of people driving their cars and heating their homes etc. So now I get it, I now know that if I took every forest on Earth, and cut it down and burned it, it would be like immediately releasing 27 years' worth of driving and heating your homes etc. So, coming back to what does that mean, that number? 27 times the world's carbon emissions, it's on topic, this is about sustainability so I've got the other half of the battle won on this one. But I would argue again, it's not really visualizable. I can relate to it, but I can't quite see what 27 years means. But it's almost there. So, the third step that they had was that the world currently has two billion hectares of land that is degraded or deforested. And this point for this stat was that degraded and deforested land is therefore available to be reforested, right. So this is something that I can work with in terms of sustainability. So it's really an important idea. But what I can't remember, is what is a hectare? Right, a hectare is something to do with a certain number of acres, I certainly can't visualize what two billion hectares means, I just don't have a concept of what that really means in real terms. But my client, again, gave me the example. And so, what they said, is that's the equivalent of the land mass of The United States, plus China. So that feels like a lot of space. I know what the Earth looks like, I know now what sort of, roughly, percentage of the Earth's land mass that means. So the fact that there's that much land that is ready to be reforested and made into sustainable forestry, is a really big deal. And so, in this case, it's both relatable and visual, as well as on topic. So I think that this is one of the better examples of how to make abstract numbers and abstract data relatable to people. This project actually had a lot of examples of data that were sort of hard to understand and make relatable to your average person but, if you think about your audience, you think about what they know, and more importantly think about what they don't know, assuming they know less than you, and provide extra context, provide extra information to make sure it's clear. Remember to speak their language and provide references that they can relate to. So just remember, there are no Empire State Buildings in the Amazon. So, if you were dropped into the Amazon and had to explain to the natives, where you came from, talking about the Empire State Building might not do the trick.

Illustration and iconography

In this video, we're going to talk about the use of illustration and iconography in data visualization. I can really sum it up in one sentence. Visual elements, illustration and iconography, are really essential for visualization in many cases. They help make your content relatable to your audience, but you do have to be careful. You have to resist the temptation to overdo it and understand the time required to do it right. We communicate visually for many reasons, and in visualization, I think it comes down to three things primarily. One is tangibility. We're trying to make the intangible tangible. We're trying to make data, numbers, into things that people can relate to and understand, and visual elements really help do that. Whether it's a chart where you can easily visually compare datasets or illustration to help bring themes to life, it's really an important part of what we do to make the intangible tangible. Second is making the complex simple. Illustration and iconography can really help you reduce text. You don't have to explain something in long paragraphs of text when an image will help bring it home. They can help convey meaning quickly and easily. And finally, it's about context setting. Visual elements can add a lot of value and they can reduce distraction. They can actually help grab attention, and of course, they can really help establish and emphasize themes in your infographics and visualizations. So, let's say we're going to create some healthcare infographic, and of course we want it to be shared widely on Twitter and Facebook and Pinterest, so we want our imagery to really jump out and grab attention. So, our first temptation, of course, is to use a very large illustration or a photo of a theme-oriented image. So, in this case, a healthcare image. We want it to be immediately clear, as soon as someone sees this graphic, this is about healthcare. If you see it on Facebook and you're a doctor or you're in healthcare or you care about healthcare, as soon as you see this stethoscope, you're going to say, oh, this is all about healthcare, this is about me, I want to look at it. So, as you start adding content on your infographic, and let's say we have four pieces of content we want to share, of course you're going to start to want to add imagery to go with the elements of content that, again, reinforce your visual theme. And probably, those little images that go with each little content bite should be relevant to the content that it's related to. Both uses of imagery, the large, thematic image as well as the smaller more focused images, of course, are very valid uses of illustration and iconography. They really will help capture attention and help your users understand the meaning in the graphic. They'll draw the eye and they can help reinforce your themes. But you have to be really careful in what you do and how you do it. So, for instance, in the lower left-hand corner here, we've added a bar chart made out of hypodermic needles. And while I can certainly see that the third syringe is the tallest and the fourth syringe is the shortest and I can read the data to some degree, it actually has been proven that it's harder to read the real relative values when you're using irregular shapes like this as opposed to just the trusty old rectangular bar. So, in this case, we're actually detracting from understanding rather than adding to it, even though I can sort of read general trends in the data. Now, if you have to use and you want to use theme-based imagery, which I do recommend, just be careful about how you do it. So, here we have pills, which, of course, are related to our theme of healthcare, but they're much more uniform. So, the shapes of these bars now are much more uniform and clear, I can really more easily tell the relative values of the bars as opposed to just sort of generally the values of the bars. If you also note now how the main image can really draw the eye from the top left to the bottom right. So, our image is really reinforcing our linear storytelling structure. We're drawing the eye through a linear progression. One thing to keep in mind is that illustration and iconography can be difficult. It can be hard to create even very simply concepts in icons, and even sourcing icons from third parties like from stock photo and stock illustration sites, it can be hard. So, that's not to discourage you from doing it, I actually strongly recommend it, but you do want to be careful about it. You want to make sure you factor it into your schedules and your budgets on the projects that you do.

Typography

Typography is at the foundation of great design and it's at the foundation of really good data visualization as well. All the standard design typography principles apply but it's extra important here because every emphasis that you make in data visualization is so important. It can really change your audience's perception and understanding of the data your communicating if your typography is done weirdly or if it's not done thoughtfully and strategically. Accuracy is the key, so you have to be extra thoughtful about your typography to call attention to things to emphasize the things that you want to emphasize. This video won't cover the fundamentals of typography but there are other great courses like this one from Ina Saltz here on Linkedin. So you should definitely check that out. But I will point out some specific ways typography applies to infographics and data vis specifically. So, there are basic types of type in charts and graphs. They may seem obvious how to approach but there are some nuance between them. And so you want to think about these categories and use your type design skills to help users understand what type of information they're looking at, right? Legends and axes labels versus data labels versus callouts, et cetera. So in other words, you want to sort of keep legends and axes in one style for instance. Labels in another style, callouts in another, et cetera. Makes it easier for your audience to quickly understand the visual language of the visualization that you're creating for them. Axes and legends should always be labeled, right? Here we have Snoods and Whatchamacallits and if there were no labels on the y and x-axis how would I know what I was looking at? Right? The norm is to use small text, you don't want to draw attention to it. And I always recommend that you use gray if you're on a white background, right? You want your legends and your labels to be faded back. They have to be legible but they're not what should be drawing the eyes, so they should not be high contrast. You very typically see labels like this where we have a lot of stuff going on so people will turn the labels on their side just to squeeze them in. While not strictly forbidden, it's not really a good idea. Not to mention in this example, you don't really need to label every single bar. Once I know that the black bar is Whatchamacallits and the lightest bar is Gizmos, I don't have to repeat that labeling. Might as well reduce visual distraction and not label it all the way across the chart. But a better way to do it is to not force your audience to tilt their head to read your labels. So, you know, not that this is the only way to do it but make 'em, you know, horizontal so they're legible and readable. Don't really make your audience tilt their heads. Now, of course, usually, almost always in a chart, you need numbers on it, right? I need my y-axis to tell you that it goes from zero to 1,000 in this case. No numbers doesn't make sense. Sometimes you need to add more numbers, right? There may be an argument for including 250, 500, 750, et cetera in this chart. You know, it's not always the right choice, I always say at a minimum you have to have the bottom number and the top number. But if there's compelling argument for including numbers in the middle, include 'em. And maybe 750 is a really important number. So in this case, add 750, make sure it's really bold and red and maybe even draw a dotted line. If it's a really important to see that third black column to the right that it's almost at 750 but not quite, then that's a compelling argument for making sure 750 stands out. Maybe 750's a benchmark, a target, another argument for making sure that it's really there and bold and bright and visible. As with axes labels, there's a constant tension with data labels, right? There's a balance between being accurate and also remaining legible and readable and beautiful, right? Because, you know, a designer would say let's label nothing, it's much more pretty and aesthetically pleasing. And a data person says, label everything 'cause every single data point is important. And you have to find that balance for yourself. I would certainly tend towards labeling fewer things rather than everything because when you label everything you're saying, everything is important. Really what you're saying is that nothing is important. So, I can label everything or I can label nothing or I can just label maybe the few things that are actually important to this audience. The things that I'm really focused on. Maybe, sometimes, it's really just one thing that needs a label. So I can just label that one thing and make it really big and bold. Or, use design, even better, to really make it pop and really make it stand out with background color, with contrast, with design, et cetera. Good design will always find that balance between accuracy, readability, storytelling, the granularity of the data you need to communicate, as well as aesthetics. You can make the case that this really isn't even a label anymore. This is more of a callout, right? This is really drawing attention to something. You can use different font face, weight, color, backgrounds, images, color, et cetera as I have here, to really make a big impression. To really make a callout draw the eye to something. With, you know, not really crazy adjustments to your typography, it's sort of like a label, it's sort of like a callout. And, you know, in this case, it's really making it clear that I want you to look at that one data point. Here's an example from the real world. A project that my company did for a client. And just about every example of typography is at play here. And there's a lot of variety in the typography without it feeling like it's 800 fonts and all kinds of things going on. You know, we have a title which is the very large, bold type in the upper left hand corner. We have a callout quote which is clearly different typography than above and draws the eye in a very different way. Obviously, the most important data point that we want you to see is that 5.69 in the center of the image there. That's the number that is the most important number here so I'm really making it clear to look there. But then using a very similar graphical style, with a little callout box, for the other similar types of data points but lighter typography to make it clear that they're sort of less important, in addition to the size of the callout box. You'll also notice that we have two different types of labels on our axiss. So we're showing both the percentiles, the 25th, 50th, et cetera percentile of the scores. As well as the actual values of the scores the 5.5, 5.7, in the lighter gray. And you'll notice in this case, you know, you can make the case, hey why is the value, the 5.7 a lighter number when the actual number down below, the 5.69 in the big callout, is the number we wanted you to see. And that was just a client decision. They wanted to sort of emphasize in the labels, the percentiles but emphasize the value in the callout. So that was a conversation that we have, it was a very strategic decision. Even though the decision seems like it's sort of at odds, you know, internally within the same graphic. And again, typography, just from that standpoint, on the far left side in the bottom we have our sort of legend explaining the relative value of the scores, one versus seven. And then finally, the very small type, you know, very deemphasized typography where we sort of have our footnote. So a lot of different examples of typography all in the service of a data visualization using a lot of different sort of basic principles of typography but nothin' fancy goin' on here. So I started this with a list of four things but really only talked about the first three. We talked about axes and legends, data labels, callouts. You know, infographics are sort of their own thing, right? They're like any design project. All of your type design skills and experience will go into creating a full infographic. Great infographics always have great type design. So, I recommend looking at the typography courses here on LinkedIn for a deeper dive into the basic principles. It's a really fascinating and important topic for designers. It's a very powerful tool to make your work more beautiful and to help improve understanding and impact for data visualization specifically. Follow the rules and best practices for typography in your data visualization and you'll be doin' really well. And then think about the nuances to type that are specific to visualization, such as, making your axes labels gray and desaturated, your footnotes really small of course, and really calling attention to the data points that are the ones you really want to call attention to.

Position, size, color, contrast, and shape

There's a limited number of ways to show differentiation between objects on a page or a screen in order to reveal trends, patterns, and outliers, which of course is the primary goal of data visualization. There may be an infinite number of ways to execute on these things, but it's really a pretty simple list of ways to do this. And it's limited to position size, color, contrast, and shape. It's maybe surprising that there are so few options. This is what data visualization is really all about. Your goal to share knowledge based on data usually what's interesting is where it varies. And so there's a valid story in this chart, but if all charts look like this, the field wouldn't be very interesting. Usually we're looking at charts that look more like this, right? Or maybe this or this. What's interesting are the outliers, the trends, the patterns in these different views. So there are five primary ways to show that differentiation to allow your audience, to see these things. And some of these are more powerful than others, by the way, And let's start off by talking about position, okay. So position is interesting, position in many visualizations is actually driven by the data, right? You're not going to change the position of these dots in the scatterplot. The X and the Y coordinates put the dot exactly where it belongs based on the data, right? You can't change them manually. By the way, position is a very strong pre-attentive trigger. We talked about visual perception, the idea that you react subconsciously very quickly, pre-attentively to certain types of triggers. Position is very, very powerful. So human beings are very good at seeing when the dots are very close to each other, when they're far apart, when they track and certain sort of patterns. Now in an infographic, you can use position to provide emphasis, right? If you put things at the top of the page, you're telling your audience, look here, these are more important than the stuff at the bottom of the page, or you can draw the eye with the position towards the center of the page or towards the upper left-hand corner which is a good spot for left to write language cultures like English, but maybe less so in other cultures, right. In infographics position is an arbitrary decision to draw attention, to tell your audience something is more or less important. Of course, in data visualization. It may or may not be possible to do that. Size can also be data-driven right by the way, size is also a very good, strong pre-attentive trigger. We immediately pre-attentively see large dots and small dots. Now in this case, a bubble chart, the large dots are driven by some variable, right? And so again, the data decides the size. You can't manually change it. Now you could on a chart, theoretically, have a scatterplot, just X and Y coordinates a bunch of dots and manually change the size of one in order to draw attention to it and maybe label it. But boy, you should do this very cautiously because most people might assume that this is a third variable, right? It's not really the same thing as an infographic where yes, you can make certain things really big to create a focal point to draw the eye and to again, communicate emphasis or importance to your audience. Color is another method. It's very, very challenging, but it can be pretty powerful. Now it's not as good a pre-attentive trigger as some of these others that we're talking about, but it also depends on what color you're using. And other attributes of the color, such as the intensity of the color, the luminance, which we'll talk about in a moment. But like size and position, it can be used to draw the eye manually for emphasis on say an infographic, or it might be used to represent a variable, for instance categories, a fourth variable in this bubble chart, right? The green categories, one type of thing. And the gray is another kind of thing. Now you have to be cautious with color for a few reasons. One, like I said, it's not a very strong pre-attentive trigger. Now, here we have green versus dark gray, light green, dark gray. This is easy to see pre-attentively, but if you had a bunch of light blue dots and a bunch of light green dots, especially if they're smaller, it gets harder to pick up on those differences pre-attentively. So color is an interesting one. You have to be careful about. You also have to worry with color about accessibility. There is the most common form of colorblindness, which occurs in up to 8% of men, by the way, is called red, green colorblindness. For those people, they will have difficulty distinguishing between certain shades of colors, particularly reds and greens. And so here we have a rainbow run through three different colorblindness simulators, and you can see the orange Stripe and the green Stripe in the lower right-hand example. They're barely detectable that they are different. So even if you have a legend, this can be a problem. In addition to colorblindness, there is also the issue of contrast sensibility with accessibility. So if you use a very light shade of gray against a white background, people with certain visual difficulties may have trouble picking up and certainly reading that text. This affects by the way, a much larger number of people, the contrast issue up to 30% of people, right? So I can tell you from experience as I get older, it's harder for me to see smaller type and or less contrasty type. It just happens. So something to be very aware of when you're creating visualizations. Now, luckily for us coming back to colorblindness, there are websites like this one colorbrewer, which allow us to identify colorblind friendly pallets. And so I can find collections of colors that will not cause problems for people with different forms of colorblindness. And there are different options, different decisions I can make along here for addressing different types of color palettes, such as sequential, pallets versus diverging pallets and other things. Not going to get into these here, but long story short, there are tools that can help us avoid some of the challenges with color. And by the way, coming back to this image here, you know, this visual on the left was run through simulators. These simulators are available on various websites and or as extensions now in the browser. So you can test images, you create and see what they would look like, simulating different types of colorblindness and even other vision deficiencies. So we can solve for these problems with tools like color brewer and some of these filtering tools that help us sort of see what our work will look like. So we can change them, of course. Another really good way to emphasize things is with contrast. And so contrast unlike color, although clearly they sort of go together to some degree. It's about the intensity of the color is a very strong pre-attentive trigger. So light blue, light green, not so good, dark blue, light green, much more effective. And that's more about the contrast than the color differentiation. And of course we use color in data visualization in a bunch of different ways, sometimes like with size or position, it can be simply data-driven. For instance, in this choropleth map, we have geographies and the color intensity within the geographic shape tells us what the measure is. So the darker color means a higher number, the lighter color means the lower number, et cetera. And of course we can use color manually too. We might fade back all, but one data point or all, but one geography and have that one pop out in color just to bring attention to it. So maybe it's data-driven, maybe it's not. So here's an example of a project that I did that uses contrast. And so what we're looking at here at the moment, we're just sort of passing time. You can see the keyboard typing as you have an opportunity to read on the left and what's happening is we're looking at one letter being typed at a time from the book, Anna Karenina, and then eventually it just speeds the way up. And what we're being shown here is actually the concentration of usage of different letters on a keyboard from the book, Anna Karenina. So if you took the entire text of Anna Karenina, there are a lot more, A's very few q's, okay. And no surprise there, but it's sort of the, the, the project, the goal here was to look at whether certain keyboard layouts maybe are more efficient than others. So this is the Cordy layout. You also get to see the pattern usage of different fingers, but the transparency, the contrast tells us which letters are used more often. And we can see that in different patterns, different keyboard layouts, some are much more efficient than others. Obviously you want your fingers to be all sort of stuck in one area, the more your fingers jump around, the less efficient it is. And so that's the purpose of using contrast for this particular visualization. So it's, data-driven use of contrast in this case, but the other use of contrast is simply to allow you to see volume. And so here's an example from the same project here, we're looking at those keyboard jumps. So how often do you jump from the top row to the bottom or the bottom to the top, if you were typing out all of

Anna Karenina in this case. And so it's a bunch of lines all across, you know, sort of all on top of each other, if these were all opaque, it would be harder to see the patterns just by making them translucent. We can see where all the lines criss-cross, which ones are thicker, which ones are thinner. It's just easier to just sort of pick up the actual pattern through the spaghetti, through the noise. And another example where we use contrast to simply draw the eye to things here, we have a visualization, which is all about job trends in the United States. And I'm already part way down the visualization at this point, where I talk about how wages have increased or decreased relative to inflation for different categories of jobs. And this is sort of the overview look at all of the data. I don't even need to explain what we're looking at here necessarily, but as you scroll through this story and we highlight a portion of the story, for instance, this category of jobs management occupations, well, we can see that category lights up and then all the other categories fade back. So just using contrast to draw attention to things in that case, an entire job category in this case to three specific jobs whose pay went up the most in this case to another category in this case to other specific jobs, et cetera, et cetera. So contrast outlining can draw the eye manually to things, to help tell a story. And last but not least there is shape. And actually, maybe it is the least, it's my least favorite way to distinguish between different types of data. Because couple reasons, one is it can be visually confusing, especially if it's not used sparingly. For instance, if you had a scatterplot with five categories of things and use five different shapes, that's not going to work well. And the reason is the second reason I don't like it is because it's a very weak pre-attentive trigger. So here's an example, pre-attentively you don't pick up on where all the squares are versus where all the circles are. Is there a pattern here? Is there a difference between the two? I don't know. I have to really pay attention and look around and investigate. Whereas if I add in this case color and or contrast, now I can actually make a pre-attentive judgment. And by the way, circles and squares are very different shapes. One has pointy corners, the other one doesn't. So even with very well-differentiated shapes, they're not well differentiated pre-attentively. Now imagine if these were squares and diamonds, right, a square and a diamond are the exact same shape. Just one is rotated 45 degrees, right? Or at least close enough, long story short, very weak pre-attentive trigger. Therefore not a powerful tool to use in the tool belt. If you're going to use shape though, make sure you dual encode it double up on the pre-attentive triggers by using the shape as well as we're doing here. Something like color and or contrast. So yes, we have five primary techniques to show differences for our audience. Experiment, figure out what works for your audience. Think about how many different ways, how many different types of things you need to reveal to them, experiment with it, really test it, mix and match, find rules that you can use over and over, but really see what works and make sure you're, you're a thoughtful and careful about your audience's pre-attentive response to your visuals.

The importance of scale

Visualizations consist of two components. There is the space where the data is displayed, and then the objects that represent the data. So a lot on this course is going to be about the data itself, right? How do think about it, how to display it, how to make it compelling, etc. In this video, we're going to concentrate on the space where the data is displayed. We're mostly going to be looking at 2D planes here, although there is such a thing as 3D visualization also, but that's outside of the scope of this conversation. The point is that the size of that plane is very important. The scale drastically affects the visualization and the story that it tells. So here we have a bar chart with some interesting and easy-to-see patterns, right? What's going on? A is doing well, B is doing pretty well, and the rest are doing less well. Here we have another chart. No one's doing that well in this one, right? A is sort of slightly better than the rest, but they're all kind of doing pretty poorly. It's also hard to see the variation between the groups, D, E, and F look identical here. In this chart, all of them but A and B are doing really pretty poorly. Now maybe you noticed that these charts are actually all the same. If you look at the Y axis, you'll see that the data values are all consistent, it's just the scale that's changing. So in the left-hand chart, the scale goes from zero to 250, in the right-hand chart, it's from 100 to 240, and the middle is from zero to 2,000. And that dramatically affects your perception of the data that you're looking at. And this is really true with all charts. There's no way to display data visually and divorce it from the scale in which it's being displayed. But you might ask, how you can know how to set the scale for a chart. Which one is right when you look at these three charts? Which is the correct way to do it? It sort of depends, but there are definitely some good rules to follow, and in this particular example, the one on the left is quote unquote, right, and by the end of this video, you'll probably agree. So, the first thing to consider is, am I using a chart type that requires a certain approach to scaling? And for the most part, the answer is no, but there are some chart types that do have specific rules to follow. So, as an example, bar charts really should always start at zero. And here's why, the height of the bars in a bar chart actually means something. So if you look at the right-hand example, if you look at F, for instance, that data point looks like it has practically no data in it. But the fact is, we're already above 100 at this point. The left-hand example is an accurate representation of it. I can see that F has a bunch of data, it's above 100, it has less than A maybe, but that missing data in the right-hand example just isn't fair, it just doesn't accurately represent how much data is in that F category. There are plenty of other chart types that don't have to start at zero, but bar charts really always should. The second question to consider is, am I comparing things in a self-contained context, within the context of the chart? So, in this example, let's say you have a chart that has no external reference, there's no need to put these numbers in the context of other numbers outside of the data that I'm actually displaying? So here, we're looking at widget sales for a company. And I actually generated this chart in Excel, and it's automatically exported it with a scale that started at zero, and went up to $12,000. But as you can see, the numbers are only from 5500 to $10,000 or so. So since I don't need to compare this to any other numbers, and I'm in a bar chart, I don't need to show the numbers below 5500, I don't need to show the entire scale. I'm just showing a change over time. So I can actually set the scale here to 5,000 to 10,000 if I want to, right? I have a nice round top and bottom number on my graph, and I'm telling a complete story. Or, I could actually add the exact same value above and below the minimum and maximum values so that there's the exact same buffer above and below at the bottom and the top values in this chart. So this is actually the most balanced visually, right? There's exactly the same number of pixels above and below the line. But it's kind of weirder numbers on the axis, right? 4472 to 10,472. I'm not a big fan of this approach. More importantly, research has shown that people are much better at remembering round numbers, so using random-seeming non-round numbers like these won't help your audience understand or remember your data. I could also set essentially an arbitrary scale. So let's say that I'm a sales manager at this company, and I have a sales target for my people of selling $20,000-worth of product, and I want to show these numbers in the context of that reference. I'm manipulating the scale for a valid purpose. I'm not intending to influence the perception of the data for evil purposes, but within the valid context of this sales target, I'm just showing people where they stand. It's okay to change scales for good reasons, but you really do have to have a solid reason and you have to be consistent with that. Another influence on scale is whether or not you have an external reference, right? You have some arbitrary number outside of the context of the data that I'm referring to. So here we have the same chart, and this is very similar to the last example, where we were showing it in the context of an internal sales target. Now sometimes you might want to set the scale based on an external target, right? This is not about showing my salespeople how they're doing compared to a target that I set, but it's more about an external number, right? This is the total widget sales and entire marketplace, so I want to show them within that context. It's a very similar motivation, it's a very similar thing, I'm just sort of setting context and comparing it to a number, but if it's external to you and external to your data, you have to do it that much more carefully and thoughtfully so that you avoid the perception of bias on your part. And speaking of bias, that's always a great question to ask yourself. Am I being fair and unbiased, especially when thinking about scale, you have to think about this very, very carefully. You don't want to be that guy, right? You don't want to put on a suit and look all pretty and legitimate and then play games behind people's backs. No one likes a cheater. So it's a really good exercise when you're creating visualizations, to look at your chart with different scales. I would recommend that you experiment with scales, change it up, see how it looks, think about whether or not you're being accurate, think about whether there's a reason you're setting a particular scale if you are, and above all, look for bias, and eliminate bias whenever you can. And really when in doubt, channel your audience. Think about a few things. One is think about people who don't know the data and don't know the story you're trying to tell, and make sure that your scale is going to help them understand your data and help them get the story that you're trying to communicate. But especially, think about your audience in terms of two categories. Think about your skeptics and your believers. When you set a scale, ask yourself are the skeptics going to believe this? Are they going to buy it, are they going to think that this is an honest representation of the data? And the same thing on the believer's side. Make sure that when you present data in a certain scale, you're not just reinforcing the believers, you're not just giving them the data story that they want, make sure that it's a really valid, accurate representation of the data that you're sharing. Use your powers of scaling for good, not evil.

Legends and sources

In this video I'm going to talk about the necessity for clear legends, so that users understand what they're looking at in your visualizations. And also the importance of including your sources. These details can make or break an information design or data visualization project. Legends, sometimes called keys, are those explanatory queues that are often found in the bottom right-hand corner of a chart, and so even when you create a chart in Excel it sort of puts them over there on the right-hand side, it helps the viewer understand what they're looking at. In the most basic form of a chart, like this one here, where everything is labeled, there's really only one thing, I know how many Snoods each one of my Whatchamacallits has, and Whosawhatsits, et cetera, I don't really actually need a legend in this case, it's one of those rare times when I don't, but for the most part, you're going to need legends, especially if you start adding weird shapes or colors, or, more data, more layers of data, in this case I need to know what the Tally Hos are, versus the Hither Tos and the Be Bops. If we don't include a legend, if I can't understand what these different colors mean, then I'm creating art, it's not pretty art, but it is just a picture, there's no knowledge, there's no value here. But my job is to inform people, not just to create pictures. Better than a legend or a key that's off to the side, by the way, is in-line labeling. If you can include your labels identifying color values, shapes, and the like in this way, you're helping your audience keep their eye on the data, which will always be better than forcing the tennis match watching, back and forth, swivel head behavior a legend requires. Zen Buddhism has a concept called beginner's mind. And the idea is that you should come to everything with an open mind, without any preconceptions. This can be hard to do when you're doing data visualizations, when you're working on something for a long time, it's very detail-oriented, you know everything about it by the time you're done, and trying to think of it as someone who knows nothing about it can be difficult. But if you can channel a child's mind, or a novice's mind in the topic area that you're discussing, if you can get to a point where you have great user empathy for people who know nothing about what you're showing them, then that will help you understand what's hard for them to understand, which will help you figure out what you should include in the legend. Sometimes you need more than just a legend, especially in interactive graphics, so this hospital pricing visualization that I shared earlier, there's a lot going on here, there's first of all a ton of data, there's different colors, there are these different-sized bars, there are two different types of bars, if I click into this thing, I get a lot more detail, what do these dots mean, what are these axes, what are these other bars, there's a lot here. And so, no little legend or some labeling on this screen is going to do me justice to understand it. Even if I was fairly knowledgeable about this topic area. So what I always do, especially for interactive graphics, is I actually create an entire how-to, I'll take a screenshot of the interface, and I'll look at it, again with that beginner's mind, what might I not understand, and I will draw little lines, and I will label everything, I won't leave it up to anyone's imagination, what's going on here, make it very, very clear of what all the details and shapes and colors represent. The other thing that I always include in screens like this, usually at the bottom, are the notes, sources, sometimes it's just about the data, just where the data comes from, oftentimes I'll also include notes on the technology. One of the most important reasons to provide sources is for credibility, so if you're creating a visualization on a topic, especially if it's a controversial one like politics or climate change, then when you provide your sources, it'll allow your users, whether they're believers or skeptics, to look at the data themselves. And this has two advantages, one, credibility, as I mentioned, in that they can disprove your thesis or confirm your thesis, or at least get the sense that they have the opportunity to judge you based on the merits of the data, and not just have to take you at your word for it. And the other reason is that a lot of your users, if they're interested in the topic area, might want to dig deeper into the data themselves, so by providing a resource, you're actually being a good citizen and giving them access to the information that they can go play with on their own. If your mission is to inform your audience, which I would argue it should be, if you're in data visualization, then that's great. If not, then you might want to try another calling. Data visualization can be a long and complicated process, and when you're finished, the last thing you want to do is all the busywork, going through the details, like the legends and the sources and the how-tos, but it really is as important as everything else that you're doing, don't rush it, don't neglect it, give it the time and effort that it deserves.

The right paradigm: Basic charts

One of the most difficult things to do when you're starting in data visualization is figuring out which charts to use in which situation. Now eventually you're going to want to push the envelope and try different forms, different chart types, really out there alternative approaches to visualizations. But before that, I'd recommend having a really good handle on the basic charts, and when you might use them. You know the saying, before you can walk you need to learn how to crawl. This video is a high level overview, but hopefully a really good introduction to the topic of when to use which basic chart forms in which situation. These charts, by the way for the most part, allow you to easily display one to three variables. So these are the most basic chart types that I'm going to be talking about in this movie. You have bar charts, line charts, area charts, timelines, scatter plots, bubble charts and pie charts. These seven forms are very straight forward and basic and for the most part are forms that people naturally understand. So lets start talking about the bar chart or you could call it the boring old bar chart. And the reason it's the boring old bar chart is because it is used so widely, because it is so effective. The fact is that humans have a capacity, a built in capacity to easily parse the differences between these rectangular shapes. We're wired to see this type of chart. So while it is boring and old, the fact of the matter is, it's extremely effective and what I would argue is that anytime you're doing a visualization you should start off by thinking of it as a bar chart and ask yourself, is there a reason that this should not just be a bar chart? It's effective, it's easily understood, you won't be confusing your users, etcetera. Lack of confusion is a good thing. Now, a bar chart is really great at showing those just one or two variables, you can add more data to it. So here we have what's called a grouped bard chart, so we have two different data points, the gray and the black and it's very easy to understand still. When you start adding more data points, it can start to get a little bit harder to read. In grouped bar charts, this certainly gets into the category of maybe I want to try a different form. And when you have a whole bunch of them, even if they're separated into different groups in this way, with a lot more spacing etcetera, it can get overwhelming quickly although it is still decipherable. If you're trying to make an emphasis on a comparison within groups, if you want to think of the elements of data as part of a whole, like a category, then a stacked bar chart might be a better way to go. Here you can see that the whole bar represents the total value for each group and each segment represents the category value, sort of, the proportion within the group of the data. If you want to emphasize not the total value for each but the relative value, so how much each category influences the total value, then what's called a stacked percentage bar might be a good choice. So here each bar represents 100% of each data point, so each segment within, each color, represents the relative value within the whole as a percentage. It's often easier to see relationships like this in a stacked percentage chart. If I wanted to show the relative strength of a category within the whole. A bar chart can't convey all types of data. So if you look at these two charts, both are showing changes in values over time. The problem is that the bar chart really only shows each value at a single point. So for instance, maybe this is telling me the change in value of a stock price every January 1st over X number of years. So I get that sort of snapshot of a moment in time. Whereas the line is really great at showing me the trends over continuous time. Line charts are a great idea for a default choice of chart type, when you're showing things over time. Information designers telling content driven stories with time based elements, will often use what I call timelines as a good default paradigm, right? Each one of these dots is on a timeline, at a certain point in time, and might have more information inside of them. An area chart is like a filled in line chart. Line charts are often better than filled line charts like this because where the lines cross it can be hard to see, with the filled in area charts, where the dark gray's covering up the lighter gray and the medium gray behind it. It's hard to tell where the relative values are. I don't know where the bottoms of those troughs are, in the light gray. But this is an interesting way at looking at data. There's also what's called a stacked area chart. Here you have the filled in line charts and they're treated sort of like the stack bars where they're on top of each other. Again, this is good at showing categories of data over time and how they relate to each other. And finally, like the stacked percentage bar chart, we have the stacked percentage area chart, which again, is a really effective way of showing the relative strength of categories across time, or whatever the X access represents, but as a portion of a whole. So again, the top here is 100%. Another great chart type for showing two variables is called a scatter plot. Scatter plots are great at showing correlation. So here you can see that as X increases, as things get further to the right, Y also increases. They also tend to go further up. You could have shown this data in a bar chart but you wouldn't as easily see the correlations. This one shows what's called a positive correlation, where as things go up on one axis, they go up in the other axis. This one has a negative correlation. As one increases, the other decreases. Here is a scatter plot with no discernible correlation but there are some interesting patterns and certain types of patterns will show up much better in a scatter plot than in a bar chart for instance. Bubble charts are great at showing three variables. It's really just a scatter plot, but now we have a third variable. The size of the dot is representing that third variable. And so we can add more interesting layers to the data by looking at it this way. And once again, in this example, we have correlation. As X goes up, so does the size of the dot, generally. But again it's easy to see the outliers. So we have that one giant dot over on the left hand side, that's something worth investigating, it sort of, it bucks the trend and it's very immediately visible. Lastly we're going to look at the pie chart and you can spend a lot of time reading about the intense debate about how worthless the pie chart is. There are plenty of detractors of this form and a few defenders, but really, to my eye, the pie chart has two major problems. One is that, it's really hard to parse when there are more than a couple of data points. So here we have two, four, six different pieces of data and it's a little bit hard to parse. I mean I can certainly see the smallest slice and the biggest slice, but three of the medium slices, I can't tell really anything about them. And that gets to the second point, is that the pie chart is really bad at showing slight variance between data points. So those two top wedges look almost identical in size. Again, human eyes, human brains, have a hard time parsing circular shapes and arcs, whereas if these were bar charts, I could probably immediately see the difference between those two data points. But with all that said, the pie chart actually is pretty effective at comparing the difference between two data points and especially if it's just one variable. So if the point is really just to show that this is a lot more than that or that this is pretty much the same as that, then I would say that the pie chart works pretty well. So these are the most basic chart forms. I'm sure you are already familiar with all of them or certainly most of them. Hopefully this video has helped you understand specifically when to use each form.

The right paradigm: Alternative charts

Basic charts and graphs are familiar to everyone and they're great at showing certain types of data. One limitation though, is the number of variables. What if you have more variables or more complex information than you can do in the most basic charts? This video goes through some of those alternatives and when you might use them. One of the alternatives is a chart type called a box plot. And, you know, it's interesting, you actually won't see these in the real world that often, although if you're in the financial world you see them very frequently in use in showing stock market data. They're actually really good at showing multiple variables. So, let's say that this is stock market data. These box plots are showing me the open and close price. Let's say that this is the open price and this is the close price of this stock. As well as the average stock price throughout the day. These things are called whiskers, these little dotted lines, the lines above and below. These might be showing me the top price for the day and then the bottom price for the day. You can actually change what the whiskers represent. I could say that this is showing me the top 90 percent, 90 percentile for the day and this the bottom 10 percent price for the day. You can decide what you want it to represent. You can even include dots, so sometimes you'll see these little dots show up. To show outlier data even beyond. So let's saying if this was showing me the bottom ten percent. These dots will represent two things that are sort of outside of that scale. Box plots are really great for showing a lot of data in a very simple form. And even if you're audience isn't necessarily familiar with them. A good legend can make it easy for users to parse the information quickly. I also really like heat maps for showing a lot of data and a lot of variables at once. It's a very effective form for seeing trends in the data. It's easy to see the correlations and the overlaps between different items. So here we're looking at the time of day and the day of the week for web visits on a particular website. So it's easy to see that in the middle of the day, during week days, this particular website gets a lot of traffic. Early in the morning, no matter what day of the week, not a ton of traffic. Heat maps are a great way at looking at really a limitless amount of data, across a number of variables and to see how one compares to the other or how groups compare to other groups. Radar or spider charts fall into a category that I would label, use with caution. They're good at showing multiple variables at a time and showing their relative strengths between items. Especially if they're transitioning like this where I can see how things are changing over time. They're kind of like pie charts, it's actually hard for the human eye and brain to parse the absolute values of these things as they're changing. So for instance, right here in this data point, it's below one but how much below one and how much above or below these other ones? It's a little hard to tell. So the way this charts works is each one of these spines coming out of the center represents a single variable. So what we're looking at here is data from the Eurozone crisis. We're looking at, this pinkish line represents Greece and the dark blue line represents the Eurozone on average and the entire size of the blue band represents all of the data points in the Eurozone. So for instance if I look at 2012 I can see that Greece by far had the worse debt ratio as a percentage of GDP, the highest interest rate, et cetera the lowest growth rate, so it's very clear from this chart why Greece was struggling through the crisis. But again I have a hard time parsing the absolute values of these numbers. But as I turn on all of the countries, this shows one of the great flaws of radar or spider charts, sort of like pie charts, when there's too much data it's extremely overwhelming and not very helpful. Now if it's animating when all of these numbers on here it's interesting and if I track one line carefully I might see it. But not the most useful form for something like this. So if you're goal is to compare whether A is better than B and you have a bunch of variables this form can work. But again, be cautious if you have a lot of variables and, or, if you're trying to show absolute values, if knowing the absolute values of the numbers is important. Another chart form that I really like for looking at multiple variables is something called parallel coordinates. And so here we're looking at four different variables. We're looking at sepal length and petal length and sepal width and petal width, which are botanical terms, these are referring to the lengths and widths of parts of flowers. And so what's interesting here is that each one of these lines that goes across, so I'm tracking this blue line, shows me the data point. So where it crosses the vertical line that's the actual value. So this is an eight here and this is a six point four-ish here, et cetera. And while, when there's a lot of data, over all the experience can be somewhat overwhelming, it is easy to see the trends, it's easy to see for instance that there is decent amount of variety of sepal length amongst this red category. But then there's much less variety for the red category's petal length. It's also really good at showing relationships between variables. So for instance I can see that there is some sort of correlation relationship between the second variable and the fourth variable, petal length and width, amongst this red category. But less correlation when looking at sepal length versus sepal width for that category, there's a lot more spread here than here. One draw back of this particular data form is that it can be overwhelming when there's a lot of data to look at. But when using an interactive version, like this one I'm going to show you, there's actually functionality sometimes available called scrubbing. So in this case I can actually click and drag and select just a portion of the data to get a much more limited view and focus in on what I really want to look at. So say I wanted to look at just the red lines or just some of these green and blue lines, I can really narrow it down and it's a little bit easier to see. If I want to narrow it down further, I want to say, gee, three point two over here that's really interesting. I can actually click and drag over here and scrub just those two, so all of these and over these, where is the cross over? What are the data points that are within here and within here? And now it's really easy to see the patterns that I'm looking for. So this example is sort of cheating. Really I'm just throwing a bunch of standard charts, a bunch of scatter plots onto a page. This is called a scatter plot matrix. So I'm able to look at a bunch of variables and bunch of different views of those variables at the same time. This is actually the same data as was in the previous example, the parallel coordinates example. But what you're looking at here is a scatter plot of each of these four related just to each other. So for instance, in this corner, I have sepal length and sepal length, I have a row of sepal length and a column of sepal length. So you'll see it's a perfect correlation here that as one goes up the other goes up, 'cause it's the same. But over here what I'm looking at is sepal length versus sepal width. Sepal length versus sepal length. And so just by looking at it in this way I can see the different correlations and patterns and how different they look. So for instance if I look at petal length versus petal width there's a really interesting correlation pattern here that I might not see if looking at this data in a different way. There really are many ways, you can even say infinite ways, to look at data. So between the basic charts and these alternative forms you can cover a lot of ground discovering patterns and trends and making comparisons in your data even when you have what might seem like an overwhelming amount of data.

The right paradigm: Hierarchical data

Data visualization is often about focusing on showing the connections between and the hierarchy of objects. This unique category requires a different way of thinking about and displaying this in useful understandable and meaningful ways. This video's going to look at a few types of visual approaches to this type of data. So I'm going to be using a bunch of examples from the D3 website. If you're in data visualization, you may already be aware of D3, and if not, you very soon probably will be. D3 is Data-Driven Documents. It's a JavaScript library for doing data visualization. And it's most often used for interactive graphics, but you can just create any visualization out of data using this incredibly easy-to-use, open source library. There are a lot of great examples on this website. And in fact, I use it very frequently not just to find specific examples and code examples to work with in D3, but for inspiration, as you can see there's just tons of things to look at on this website, all kinds of forms. The first one we're going to look at here is called a tree diagram. So it's sort of the default form for doing hierarchical data. So whether you're thinking about an organization chart or anything where you have a parent object which has children and maybe those children have more children, et cetera, it's just sort of a very easy understandable, and simple way to think about displaying this hierarchy of data. It's sort of like the bar chart for hierarchical data. So when you have data and you want to show the connections between the different objects and maybe the hierarchy of the objects, the default is going to be a tree diagram. It's sort of where you can start thinking about it at least and then you can experiment from there. You can also use interactivity to bring it to life, so for instance, if I click on these nodes, I can collapse it all. I can get the entire thing to collapse into a much smaller form and then some of these are clickable to open up more details. Another form to show more complex data where there's a lot of connections in between objects including connections from one object across nodes to other objects is called a node-link diagram. Like a tree, these have objects, the nodes, these little circles, and then links which are the lines connecting the nodes. Like a tree diagram, it's very easy and simple to understand. But like other hierarchical data forms, you can get into trouble quickly. The default in these types of forms is to show everything. I'm trying to show you data about all of this stuff. So by default, I want you to see all the connections, but when there's a lot of data like this, it can be difficult to parse, difficult to understand, difficult to see how this little dot over here relates to this dot over here unless I very carefully follow the chain. So once again, interactivity can help solve those problems. So if in this particular example, I can click on these dots and collapse them. And if I do that enough times, I'm going to get way less detail to look at. And again, if I want to open up more detail, of course, I can do so. As you can see, this diagram, sort of floats around and naturally finds itself in space to help reduce the number of lines that cross each other which is one of the other problems with this form is that as lines cross and more detail gets obscured, it gets even harder to read what's going on. This is called a force-directed layout when the graphic sort of settles in a position to best show what all the data is. Interactivity can also help in other ways when you have node-link diagrams where there's a lot of information, and it's a little hard to see. So in this example, it's using what's called fisheye distortion where as I roll the mouse over the data it sort of brings the focus, it sort of zooms in a little bit on it, again, making it a little bit easier to see some of the details in between the data points. Another visualization method that I like for showing both the forest and the trees in terms of, how to look at the data, both details as well as big picture view, is something called an adjacency matrix. So in this case, this is a 2D display of many, many data points. So each column is one variable, and each row is the same variable. So sort of like in the alternative charts movie before this one, we showed the plot matrix, the scatterplot matrix. Here you have variables compared against each other where, of course, you will always see a correlation. And then this variable, Child2, is then shown compared to the other variables. In this case, we're looking at cast members of Les Miserables and when they're on stage at the same time. So Child2 was only on stage with two other characters. Other characters in the play, of course, were on stage a lot more. This character, of course, is on stage with a lot more characters. In this particular data form, one of the most important things about it is the sorting. Sorting really plays a big role in how you see data using this form. So for instance, right now, we're sorting by cluster, the order of list is by cluster. And that's sort of a way of grouping these characters. But if I group them by frequency, meaning the more frequent ones are at the top, the ones who appear most frequently, and the least frequent characters appear on the bottom, then I get a different view of the data. I'll see different patterns in the data. Or if I view just by name alphabetically, I'll see different patterns yet again. So this is a very interesting way of revealing patterns in data that you might not see using other data forms. One of my favorite visualization methods for hierarchical data is called the tree map. This isn't so much about the connections between items as it is about the hierarchy only of the items. So this form's also very conducive to interactivity. So over here in this orange category, I can see a bunch of items that live within this category and how much relative space they take up in the entire dataset. And then I can click into this to zoom in and see more details. Sometimes you can click and click and click and go deeper and deeper into this type of visual display. Another form that you see more and more of is called the core diagram. So you have a circular display along the outside of the display is each variable. What we're looking at here is Uber rides by neighborhood around San Francisco. So rides to and from the financial district using Uber cars. One of the things about this form that makes it popular these days is that nature loves a curved line, right? There are no straight lines in nature. And so, we're attracted to this type of display naturally, but again, sort of like in pie charts, circular lines can be harder to parse. Now, I can easily see, for instance, that in the Financial District because of the thickness of the lines there are a lot of rides between the Financial District and South of Market. But it is a little hard to parse some of the other things about this data because of the curving nature of the line. So you have to use this form with caution. Also with this type of display, you can have an issue with an overwhelming amount of data, and that's where interactivity, once again, can really help. Because if I roll over each data point, I can just filter out those lines and much more easily see the details that I want to see. Finally, another old standard is the Venn diagram. And so this is hierarchical data, right? It's about how many things fall into this big red circle? How many things fall into this blue circle or the green circle? And where they overlap, of course, those are sort of, you could think of as child objects of those combined datasets. This is a three set diagram, right? I have three datasets, and I see where they overlap. I recently saw a seven-set Venn diagram which was really interesting to look at but incredibly hard to understand what I was looking at. But it was very fascinating. The Venn diagram is sort of like a bar chart. It's kind of a norm. Everyone is familiar with it. People understand what it means when they look at it. It's so common nowadays, of course, that it's even used as a form of comedy. (laughs) You can do funny little Venn diagrams. You probably see these on Facebook all the time. Like a pie chart, this form is really good at showing the big picture, where things overlap, but not necessarily the specific data, how much the overlap is. People spend entire careers thinking about how to visualize hierarchical and relational data. I hope this was a good introduction to some of the basic forms. As always, I recommend that you start simple. Use interactivity to make the overwhelming less overwhelming and experiment with different forms. Play with your data, try these different ways of looking at it in hierarchical forms and just find the one that works best for you.

The right paradigm: Maps

In this movie, we're going to talk about Maps. You can't really talk about map-based data visualization without talking about Google Maps. They pretty much single handedly brought maps back to the masses. Well, I guess, I mean that's not entirely fair 'cause they were preceded by some others in the field, right? But Google made Maps great and ubiquitous in a way that really no one else has. And you can really visualize so many different things. So here I am searching for hockey rinks in Boston, and as you can see, what Google Maps shows you is so many different things, right, there's so much data in this view. You know I see the roads. If I zoom in, I can click on things and see markers. I can see the thing that I've searched for. In fact Google Maps is so great that the more I zoom and the more I zoom, I get more and more detail. Different colorings to represent different types of things like this is a university campus. And if I zoom in far enough, I'll even see the shapes of the buildings, right. So I can recognize this u-shaped building as opposed to these other buildings I might be walking by. It's an incredibly rich visualization of unbelievable amount of data. So as a developer, you can do a lot of things with Google Maps. And you know you have the base layers, the base maps that you see as a regular user using Google Maps. And of course you have Satellite view and Street view. You have access to all of those different views of the maps, and you also have the ability to do other things with your data. So there's the ability to show places and do custom markers on those different places. You can use the routing information. How do I get from point A to point B? Google Maps will figure that out and how to show it. There are also some other really interesting data visualization tool sets available to you via the JavaScript API. So this example we can see this is tracking flights to and from London, and so you have the ability to put objects on Google Maps in layers and show them interacting with real data in real time. Or in this example where we can see earthquakes. And so the data is represented as dots on the screen and the size and the colors all represent something about the earthquakes. This example is a heat map showing population density around the world. And finally, you can do other interesting things with Google Maps such as this which is actually a puzzle where you have different shaped objects on the map and you have to click and drag them and put them into place. And if you get it right, it snaps into place and changes color. There's so much you can do with Google Maps API. I encourage you to play around with it and experiment and try different things. When it comes down to it, there are really five of what I would call standard ways to show data on maps. And we're going to walk through them one by one and show some examples. So the first one is Markers. Markers are for pointing out points of interest on a map, located at a very specific latitude and longitude on the map. So this is the example I showed earlier where I searched on Google Maps for hockey rinks in the Boston area. And you can see the markers are dropped in place at a specific location to show every search results. These little red dots with the white squares represent each hockey rink that showed up. Platforms like Google Maps have builtin functionality like these little roll-over callout windows. And you can actually custom design callouts in Google Maps and lay them on top of it as well. Sometimes you want to put a marker in a specific spot such as this marker that is for the TD Garden where the Boston Bruins play. It's in a very specific spot. It's at a very specific latitude and longitude, exactly where that building is in Boston. But sometimes you want a marker just to sort of generally represent like a region like Boston. And if we were to do that, it would drop a marker right here in the heart of Boston which isn't meant to really be at a specific latitude and longitude. It's more to show a general area and labeling. If you don't know what the exact latitude and longitude is for a location, whether a specific address or a general place like Boston, Massachusetts, you can actually just Google it now. Just search for Boston, lat, long, and Google will just give it right to you in an easily digestible format. Okay, so right there in the upper left hand corner, we have latitude and longitude. You can search for a specific address or a general region, even something as broad as Massachusetts, and it'll give you the latitude and longitude for the geographic center of that location. Another basic approach to showing data is layers on top of a map to indicate the data associated with the region. It's great for showing the data itself, not just the content that's associated with the region like markers do. So markers show you there's a hockey rink in this place and here's information about it, maybe in a callout. In the case that I'm describing using layers, you actually have the data built into the map itself. One way to show data on top of the map is in these layers that are in this case dots, point clusters. So the circle size indicates the amount of data. In this case there's also a label showing me the number. Interactivity can do a lot to add functionality. So for instance as I zoom in on this map and I zoom in on this dot here, it'll actually break up. So that 210 which was aggregating a bunch of data points breaks out. And I can see as I'm zooming in closer to see more detail, 40 of them are over here, 31 over here, 10 over here, etc. And once again, the more I zoom in, the more granular I get until I get to the point where I can see just where all the individual dots are that make up those aggregated numbers that I was looking at before. Another very common approach in mapping data is called a Chloropleth. So you see these all the time during elections, right? So you see a map with red and blue states and that tells you who voted which way in that state. A lot of times they're binary, on and off, or categories like red versus blue. But other times they're used to represent variations in data values. So for instance in this map, we're looking at country scores in terms of their resilience to various risks that they face. This was created for the FM Global Resilience Index, and we only have four shades of blue. So the most resilient countries are in the darkest shade of blue and the least resilient are in the lightest shade of blue. The shading makes it relatively easy to see the variance between the values. I really like chloropleth maps because you know the country itself is the color. It does show you the value that you're looking at. But of course there are risks with chloropleths. For instance, you know, looking at very small countries, so for instance some of these islands over here in Indonesia, now in this case they all belong to one country so maybe you can sort of tell what shade it is. But sometimes you have countries that are just a single island like some of these Caribbean countries. If they're the lightest blue, I can't easily tell if it's blue or gray. So you know the risk with chloropleth is you're relying on the size of the country to show things and if the country is small, it can be difficult to see. So use it with caution. Sometimes a cartogram is another option. But in general, chloropleths are great. Another way to show data on a map is called a Heat Map. And so this is similar to a chloropleth in that you're showing data associated with regions, but this is really great to show the concentration of data on a fairly micro level even from a macro view. So I'll show an example to explain what I mean by that. So what we're looking at here is the use of the hashtag MTVhottest when tweets are being sent out around the world. So the size of the dot represents the volume of tweets and you can see the clock running down here so you can see the timing of these tweets happening. And so this dynamic animated view, I'm seeing them happen not exactly in real time but I can see them happening over time. If I switch over to the static view, what I see is an overall heat map. So I can see that overall of the entire time period being tracked how many tweets happened in one region compared to another region. Once again, if I zoom in on the map, that aggregated view of the heat map starts breaking into its component parts, and I can zoom in and I can continue to zoom in and zoom in and zoom in and get more and more detail. And so I can see the relative strength of certain regions, the relative number of times the tweets were shared, and see it in sort of an overall intensity view rather than a specific geographic share of the view like I would in a chloropleth. Finally maps are very often used to show Flows between regions. And the way this is often done is simply by putting arrows on things, right. So here I'm looking at the exports or imports of some goods between various countries. So in this case, the thickness of the line indicates the volume of the flow. And in this particular example, China is highlighted because that's the focus of this particular data display. If the object is to show movement of something between regions, then flows like this are a basic default to use. Maps are by definition data visualization. They're showing you location data in a visual form. So laying more stuff on top of a map is a very effective way to share information that's geographic in nature. These five forms that we've gone through are a great start. And like with a bar chart, I challenge you to always think about when it might not be a good idea to use them, when you might want to try something else. And like going beyond the bar chart, you need a compelling reason to do so.

The right paradigm: Creativity and innovation

Once you get past basic charts and graphs and maps, hopefully, you'll get to the point where you're trying what I'll call lazily, just creative and innovative visualization forms. Now, innovation can be a force for good, certainly, but sometimes, people are trying new things just for innovation's sake, and of course, I'd recommend that you concentrate on the former. Really innovate when you need to to try to find better ways to communicate the data that you have. In this video, I'm just going to share a few examples and partly, it's because you have to master the basics before you start breaking the rules. But also, because I don't want to overwhelm you with too many things. This is mostly for inspiration purposes. When you are trying to get creative and innovative, of course, you want to be thoughtful about how and when you push those boundaries of the well-established norms. You don't want to reduce the clarity of the data you're explaining. You don't want to confuse people. So, the first example in this category I want to show you is a static infographic. And, in fact, when you look at this one, you'll see it's really actually not innovative, per se. This is a standard chart form. This is really just a bubble chart, right? So, I have an x-axis, I have a y-axis, I have the size of the bubbles, which, of course, in this case, are flowers to indicate the data. But what I like about this chart and the reason I put it in this particular movie is just that it's very creatively and interestingly done and there are a few things about it that I like. So, first of all, what are we looking at? We're looking at the number of deaths from wars during the 20th century. Really from 1900 or just before 1900 up until 2010. And what you can see is at the bottom of the flower, the stem is the start year of the war and the top of the flower, the bloom, where it ends up to the right, is the year that the war ended. And so, it has a really evocative, interesting look and feel to show me how long a war lasted. And so, for instance, the Israel versus Palestine conflict has that very long, windswept look, 'cause it started so long ago and of course, is still going on today. There are four variables that we can look at here, the start year, the end year, the number of deaths, and also, the color to indicate the region where the wars occurred. And just the creative use of poppies as the visual metaphor is very evocative and on-topic for this particular data. The one interesting thing about this data, on the y-axis, you can see that that represents the length of the conflict, which, of course, is also visible in the x because of how far the poppy flows, that windswept effect that I mentioned. So, I would loved to have seen this graphic use, actually, different variable for that axis. But other than that, I really, really like this graphic. So, this is another static infographic that I'm putting into this category of creative and innovative for a few reasons. What this is showing is Nobel Prize winners from 1901 to 2012 and you can see across the left-hand edge, these are the categories of the Nobel Prizes. Chemistry, economics, physics, et cetera. And what you'll notice is that there's essentially a line chart for each category with dots indicating each winner. And so, each dot is placed on a y-axis above or below the dotted line to indicate whether they're older or younger than the average age winner for that category. If the dot has a little circle around it, that means it's a woman, the other ones are men. And really, though, in the end, even though this is a very innovative and beautiful and very interesting and deeply explorable infographic of quite a bit of data, really, what we're looking at is very standard forms. Once again, just like the last example. So, we have a line chart across an x and a y-axis. It's also, you could look at it as a time line. We also, in the end over here, where we're looking at degrees, this is showing you what degrees various winners have earned, and we have, essentially, bar charts. Right here, it's easiest to see here in the literature section where I have a bar chart. It's flipped on its side, but that's what it is. Or over on this side, where it shows us which top university winners went to in the various categories. That's essentially a Sankey diagram, these flow lines indicating how many people from each category went to each of these top universities. It's a really beautiful and interesting and innovative and creative example using, once again, very standard forms. This is another one of my favorite examples. This is a visualization of drone strikes. And as you can see, it starts off with really telling a story. So, if you go back to our defining a narrative movie, it's a great example of how you can set up a very linear, standard storytelling structure. Then, in this case, it leads right into this first animated and then interactive graphic showing drone strikes. So, once it gets past the initial introduction, it starts showing us one after the other individual drone strikes over time. And the use of this arched line to indicate the strike, very simple idea but it's really evocative and very on-point. You also notice these moments where it pauses and text comes up and tells you a little bit about that moment in time. Some significant event that occurred. And then, of course, the data that follows that are put in context. So, it's a great way of telling a story, showing you a lot of data, a lot of details, and yet really providing all the context and insights that you might want from this type of information. Once the animation is done, it's now a fully interactive experience that I can engage with. So, I can roll over each one of these and get more information about the details behind each one of these drone strikes. You know, the number of dead, who, what type of person that they were, et cetera. I can also flip the entire experience on its side and just sort of look at it from a different direction. Get sort of a different view of the numbers. I still have access to the same information in the end, but it's sort of a slightly different perspective on the same idea. This next example is also really interesting. If you just listen for a second. (calming guitar music) What you'll notice is that you have sound in the background and what we're looking at here is live, real time edits to Wikipedia. And so, every time someone at this exact moment is editing an entry on Wikipedia, this application is throwing a dot on the screen. The dot is related to the size of the edit and the tone that you hear is directly related to the exact same thing. So, in other words, if it's a bigger edit or a smaller edit, then the tone goes up and down accordingly. So, this is a really interesting example of, I guess you could say, data auralization or audiolization, depending on how you want to phrase it. I don't even know if there's such a word. So, the audio doesn't add new data, new information to this experience, but it adds a level of depth and interest and context that I find really helps the experience overall. This is another example of the use of audio in a data visualization. (cheery tones) As you can hear, as the stock price goes up and down, the notes go up and down. The sound doesn't necessarily add anything new to the information, yet again, but it really is essentially creating a song out of the performance of the stock market. It's another rich, interesting and innovative experience. So, in this final example, what we're looking at is something called the racial dot map. And what it is is a single dot placed on a map of the United States for every single person in the country as of 2010. This is from the 2010 census. So, when you're zoomed out like this, of course, what you can really see is population density, right? I can very easily see Chicago and Detroit, New York City. You know, the east coast is extremely densely populated. The mountain west and the desert west are pretty much empty, right? So, it's really fascinating and interesting to look at just at the highest level. But let's say I start zooming in and what happens is these aggregated views of the dots start to break up a little bit. And eventually, if I keep zooming in, I'm going into Boston here, eventually I'm going to end up at a point where I can see every single dot, again, broken down so that I can get to the point where I can see individual dots, like over here, just west of Boston. And once again, every single dot represents a single person. You'll notice, also, the different colors. Each dot represents a race. So, it's a really interesting look at racial segregation also. It's very easy to see which races live in different neighborhoods in different cities across the country. What's really most innovative about this, though, is the technical achievement. Being able to place 300 million dots onto a map that actually performs in a very reactive, quick way in a browser on the web is quite an achievement. So, this example belongs in this category, for me, not only because it's really interesting and creative but also quite an innovation in technical achievement. This movie is as much about inspiration as anything else. My advice for when you're thinking about getting innovative and creative is to just think about, really, everything in this course. Figure out the basics first. Think about layering things on top of your basics, like illustrations and metaphors, like the poppies and the arched lines for the drones. Try to start thinking about innovative forms on top of your more basic forms. Make sure that the creativity and innovation add to the experience and they add to the understanding of your audience. Don't be afraid to try new things. Just make sure that it really helps.

Challenge: Improving on "The perfect report"

So, we're going to take literally everything we've covered in this course and put it to use. Now, obviously, that's no small task but don't be overwhelmed. Think of this as just a quick thinking exercise really. First, you're going to have to understand your data. That's a very important component to all of this, of course. And you may also need to reorganize your data, so keep that in mind. Second, I want you to think about a visual approach. Of course, that's the goal for this entire course, right? Third, once you've done that initial thinking and planning, I'd like you to take that visual approach that's in your mind's eye, I want you to sketch it out, and to sketch it in a form that you could present to a client. It doesn't have to be perfect, it doesn't have to be beautiful, but it has to be cohesive and finished enough that you could stand in front of a client and explain it. Which gets us to the fourth objective, which is that all of your ideas should be defensible. You should have solid reasons for every choice that you make during this challenge. So, here it is. This is a spreadsheet that I was given by a client of mine and he referred to this as the perfect report. He loved this thing. So, what is it? This is a sales performance ranking report for Acme Widgets company for the month of May 2014. We have five salespeople in this company. Each one gets his own column and then there's a totals column. And the report is broken into four categories. First of all is unit count, and so we have a row representing total units. So, person one sold 60 widgets in May. Person two sold 74, et cetera. And then, you'll see below that these different categories. And you'll notice that these numbers don't all add up to the total, right? So, what this is is think about it like add-ons or accessories for cars. So, out of 60 cars that were sold, 54 of them had a whatsit and 28 of them had a doohickey. The second category is penetration. What that represents is the percentage of total units that were sold with that particular add-on. So, 90% of the units that were sold had whatsits, 47% had doohickeys for this person number one during this month. You'll notice at the bottom of this section, there are these two rows, product index and combined index. All that these are is the addition of these values. So, the combined index is just the sum of .9 plus .47 plus .38, et cetera all the way down the line. The product index is the exact same thing but without the whatsits category. If you click in, you'll actually see the formulas in Excel and you'll see how the math was done. The third section here is income. Pretty straightforward. We sold $23,000 worth of whatsits this month for person number one. $26,000 of doohickeys. You'll also know, at the bottom of this one, in addition to a total row, another extra little bonus row called PRU. And what this represents is the entire sum, so $63,000, divided by the total units. Per retail unit, they sold $1050. Finally, we have the rankings section. Here, you have person by person rankings across the different categories. So, in terms of sales of gadgets, person number three was the best salesperson. In terms of babooms, it was person number two. And person number one in this category was the worst. We also have a total score and then the total rank overall. And once again, person number three is doing a really good job, person number one probably has some work to do. So, that's it. That is the data that I was given. Again, my client told me this was the perfect report and so, your challenge is to take this perfect report and make it more perfect. Take 20 minutes. You don't have to do a perfect representation of this data in 20 minutes, no one could do that. But I want you to do your best job. As I mentioned, keep the objectives in mind. You want to understand your data. You might need to reorganize it. Think about a visual approach. Sketch it up in a way that your client will understand what you're trying to present and make sure you can defend your ideas. Have fun and come back and you'll see how I solved this challenge.

Solution: Improving on "The perfect report"

So I hope you enjoyed the exercise of going through the challenge. As you remember the goal was to take every single thing that we've talked about in this course and apply it all, you know, no big deal. But no, mostly this was supposed to be just a quick thinking exercise. And we of course had four primary objectives. Understand, maybe you need to reorganize the data, think about a visual approach, sketch it in a form to present to a client, and make sure its defensible. So if you remember, this was the report that we were given. This was the quote unquote perfect report. Sales performance ranking report, I can see people, I can see different categories, how much different people sold, how well they did compared to each other, but it's really just a big giant pile of numbers. So, in my mind, this is no perfect. Of course, my goal is to make it more perfect and here's how I did it. One thing I like to mention is that, I can give you 20 minutes to do a quick sketch and as you can see here this is far more than the quick sketch, this is an intermediate version of the sketch, done in illustrator in order to show my clients a fully color and a little bit more polished version. So the first thing you'll notice about what I've done here is as I looted to in the objectives, I thought a lot about reorganizing my data. One of the big changes I made was I took the four things that my client told me were the most important aspects of this data, and I moved them up to the top. So, product index, combined index, PRU, and overall rank. Rather than burying those at the bottom of the report, or in some cases, right in the middle of the report, they're right up there on top. And I also added ranking to the indexes as well as the PRU. So right out the gecko, you can see how these varies sales people are comparing to each other across these important categories. One other thing you'll notice is that for the PRU and the indexes, really the most important thing is who's doing the best and who's doing the worst. And that is carried through most of this report. There's red and then there's green. But you also see that for the overall rank, my client did want to see how everyone's performing not just the best and the worst. So that the only place in this entire report where you'll see the different variations of color. The very best are number two, three, four, and five. The rest of this report is showing the different product categories, so once again, rather than have the report broken into categories like, total units, penetration rate and then income, rather than that it's more about show me what's-its at once, show me do-hickies at once, et cetera. Now, before you get too confused by this, remember that this is a sketch, and you'll notice that the data values are the same for what's-its as well as do-hickies, that's not reality, it was just my way of quickly sketching something together in illustrator without having to create a whole bunch or bars and size them and use real data. So focusing in on what's important here, in the what's-its category, you'll notice that I've flipped it on it's head a little bit, while I'm looking at one person at a time, as I mentioned before, I am not showing you units and then penetration rate, et cetera, it's all in one. And so here I have the total units sold were 60, out of those 47 of them were what's-its and here I have my penetration rate, which of course matches that math, and the total income value to the right of that. You'll also note that I am using the good ol' fashion, boring ol' wonderful bar chart because, in this case, I thought it was a great way to show the data, very understandable, as I've been arguing throughout this course. And as I mentioned earlier, throughout this report you can see the best, you can see the worst, both in total units, as well as in the particular category, units sold, as well as in percentage, penetration rate, and income rate. So all that data in a very dense way, but easily understood by the audience. I really hope you enjoyed this challenge, I know I had a lot of fun doing it, and I hope also that what you came up with in the end, was an improvement over what you might have come up with before you took this course.

Introducing motion

So, this video we're going to talk about movement. This isn't necessarily better than this, but sometimes movement is a good thing. So let's say you're taking this guy for a walk in the woods, and along comes this guy. Then, in this case, of course, movement might be a very good idea. But we're talking about data visualization. Static infographics can be really compelling and interesting and informative, there's no doubt about it, but sometimes animation can really bring data to life. So whether you're talking about video, or an interactive experience, motion can add a lot of visual interest of course, at a minimum, and it can also add real understanding. It can really help you see the data better and understand more. Motion should always be part of your thinking in doing data and visualization, unless you're working in print. I would say there are four key arguments for using motion in data-vis. The first one may seem less important quote unquote, but it's definitely not unimportant. So, the first argument is really not all about eye candy, but that's sort of a part of it. I would argue that eye candy actually is important, that it creates a better user experience for people when something is beautiful, it increases engagement. When people are happy playing with something, they're more likely to share it, it'll have a bigger impact overall. So, it gets sort of downplayed and poo-pooed but I think eye candy is important. The first reason to use motion is for transitions. So I showed this example earlier in the course. What we're looking at here is race results from a half marathon, and when I click on this button to sort these overall results into age groups, watch what happens. They transition, right? The dots float up and down and out and about into the different categories. So, yes, this is eye candy. It's sort of fun and sexy to look at, which, as I said before, is important. I think it's a very useful thing. But in this case there's more to it. That transition is actually helping me track individual dots. So, for instance, this guy over here. He's the fastest guy on the course. Watch what happens when I click, and keep your eye on this dot. I can watch that dot as it moves up and down. Now, for the fastest guy on the course, of course, it's easy, right, I can easily find him again but let's say I wanted to see these people. Sort of, yeah, there's a big bulge of people in the middle and then it sort of narrows down and then a little bit of a bulge. Interesting, let's see where this person ends up. So I'm going to keep my eye on that dot as I click into the age group results. And I was very easily able to watch her land right here, so this person is 18 to 39. So that transition being animated really helped me track what was going on. It goes beyond just being eye candy. The next reason to use motion is for interaction feedback and interactive experiences. So this is another example that I have shared before in other videos in this course. This is a visualization of hospital pricing data. And in this example, the interaction feedback that I'm talking about is, as I roll over these bars of different regions and their pricing data, I see that the bar fades back a little bit, right, the color changes. So I immediately sense that there's something more to this. It's just a little bit of feedback to say this isn't static, there's something behind this. And of course, yes, if I click into it the thing opens up and I get more data. So the third reason to use motion is to draw the eye, to draw attention. So we've talked about, you know, eye candy, we've talked about transitions, we've talked about interaction feedback, this is just about grabbing the eye and drawing to where you want it to be drawn. So if we go back to this example, you know, I'm looking at hospital pricing data, I get the interaction feedback, and then, as I mentioned before, I'm just going to scroll up so you can see it when it happens. When I click into Boston, watch what happens. Did you see these dots down here that sort of bounced into place? Now, again, you could say that's eye candy, you could say that it's sort of sexy, it's sort of engaging, but most importantly, it drew my eye down to that place on the page. That movement caught my attention. And so again, it's sort of telling me there's something going on here. These dots aren't just drawn on the page, they're not static, they're interactable. So it's sort of helping me understand that if I roll over these, something might happen. And finally, one of the big reasons to use motion is to show change over time. If you've ever seen Gapminder, this is from Hans Rosling, a professor, and this is a really fantastic example of use of motion. What this is is looking at data on income and life expectancy over the years from 1800 until the modern day. Each one of these bubbles represents a country. And so you can see they all start in the lower left-hand corner and then, as time goes by, the countries in orange, which are the western democracies, start to float up and to the right, so they're getting richer and they're getting healthier, they're living longer. And you notice that the countries in red, Asian countries, are sort of lingering behind, lingering behind, but then you see China, the giant bubble of China sort of starts shooting up and to the right, up and to the right. And a lot of the African countries are still lingering far to the left and further down than the rest of the world. Really fascinating and very effective use of motion in, what is in the end, a very standard form, right, just a simple bubble chart. But it's such a compelling story and so well told. And could not have been told as well without that movement. So, as I said at the beginning, there is nothing wrong with being still, but motion is a very powerful tool in your arsenal. It adds to the user experience, it can add to the understanding. If your medium allows it, if you're not working in print, then use it. Use it in all the ways that I've described, if appropriate, given the data that you're working with.

When to go interactive

So I want to discuss when you might think about ticking things and going interactive. And listen, the spoiler alert on this one is, my answer is kind of, always, sort of. If you can, if you have the time and the budget to make interactive things, you should. So that's the short answer, but here's the longer answer. I like to start off this way. How did you learn to count? When I was a little kid, I learned using these, they're called Montessori counting beads. And so when I was three years old, someone handed me three of those little beads in the upper-left-hand corner and said, that's three. And that's not that weird. Everybody learns to count that way probably with blocks or something like that. But then they handed me this single-wired together object in the far-upper-right and said, that's 10. Then they handed me the flat object in front of those in the lower-right-hand corner, which has 100 beads all wired together into a single object. So when I was three, I think I understood that 10 times 10 was 100. Okay? So that's very tangible. This physical experience of learning numbers, made numbers very tangible to me, and interactivity makes your data tangible to your audience. They're actively participating, they're touching and affecting and filtering and sorting the data. So they have the experience really more deeply than if they're just passively watching. So that's the first argument. The second argument is this. Everyone's a narcissist. Okay? And I mean that in the nicest way possible. When we're consuming content, if it has nothing to do with us, then we can't relate. Okay? So we want to make our content relatable, make it as much about our audience as we possibly can. So the example I like to give is this. If you are from deep in the Amazon, you might not enjoy watching a hockey game on TV. Maybe at first, it's interesting, right? Like, hey, why is that stuff so slippery? And what's that black disc everybody's chasing around? But after a little while, it might be boring because you don't understand the rules. Everything is so foreign, there's not enough about me to keep my interest. So as an example of a project that I created where the interactivity really helps bring this data to life, it's something created for the Boston Public School system for their 10-Year Master Planning. And so we have all the schools in the Boston Public School System, and you can go in, you can see a map of the schools and where they exist in the city. And each school has an individual profile. So I can find a particular school by name and click into it and see its data and have measures on different scores, et cetera, et cetera. And this is all well and good. It's good data. But this is essentially sort of replicating what was in the printed report, which was a fine document that had a lot of good stuff in it. But in addition, we have the ability to go in and analyze the data in more depth and look at it from a different direction. So I get these essential distributions to look at all the different data points as far as accessibility goes, where how many schools fall into different categories on accessibility or the quality of their heating distribution systems, et cetera. And I can highlight a specific school, make it about me. Where does Bates School live? Aha. I see which category they're in all along the way. And I can of course filter by different things as well. The point is, as a parent, as an administrator, seeing the overall data is nice, hearing what you're going to do with this data as far as the city is concerned, is helpful and interesting. But I want to see where my school is. I want to see how this is going to affect my family and my kids. And so if I can make it about me, the way I like to think about it is it allows you to turn your data story into your audiences data story, which is very powerful. And a related issue by the way, is credibility. So in this particular example, parents and administrators might question the logic. If say the outcome was, the recommendation was to close a school. Why? Well, I can go through and dig through the data and see how my school compares to all the other schools in context. And another example in terms of credibility is transparency. And so one of the features built into this tool, is that the map, as I zoom in, more information is revealed. Specifically the plan. Okay? And I see not only the layout of each individual school building, but also the layout of the land that the city owns around the building. Owns or, and or leases, I think. And so, a long story short, if the argument was to say, close this school that's an orange here, or that, you know, hey, we can't expand it, so we have to do something somewhere else. So I can see that the city doesn't own enough land around that building to do much else versus this red building where there's a ton of land, plenty of room to expand, et cetera, et cetera. Now that transparency might enable parents to make an argument such as, okay, well, why don't we take that orange school and build a version of it on that land that the red school occupies? Hey, that's the thing with transparency, right? You give your audience the ability to make arguments with your data, which is not a bad thing overall. So interactivity really is about letting your users explore data, which enables things like making it tangible, making your data story into my data story, and increasing transparency and credibility. Simply put, you should go interactive when you can. Allow your audience to find their own Aha moments in your data, expand the reach and the depth of your content. And by the way, get past the editorial constraints of a non-interactive thing. Even a deep, deep report has stuff left out of it, has sorting and filtering decisions predetermined. Allow your audience to get past that and find their own pathways through your data.

How to think interactively

So, you've decided to create an interactive experience, which is good, as I said, I think you should pretty much do it if you can. But if you're new to interactive, it can seem like a new world. You had to think about being interactive versus creating static experiences if that's what you're used to doing. But, you probably use interactive technology all day, every day, right? Your phone, you're on the web all the time, you're on Facebook, et cetera. So the leap really isn't that huge if you think about it as a user. In this movie, we're going to cover a few ways to take the information you have and sort of flip it on its head a little bit, think interactively, and create productive experiences for your audiences. The first rule in thinking interactively is don't think interactively. So it sounds a little counterintuitive, but I recommend that you don't think about features and buttons and actions at first. The first thing you want to do is look at your data, think about the story that you want to tell, think about your audience, all the things we talked about at the beginning of this course. So thinking interactively really is sort of like a microcosm of this entire course. It's a microcosm of the whole design experience. We're just sort of taking it to another level. Creating ways to expose more knowledge for your users through touching and clicking and rolling over. Really want to go back to the basics in thinking about your design in general, before you think about features. Good place to start is to think about, when I look at the data that I have, what's missing? What's not there that I wish I could see? And I don't mean what's literally missing, but what's not available to me? What's not easy to parse? This is where I usually start to think about features to add to an interactive experience to expose what's missing. In addition to what's missing, what can't I see? It may be there, but it's sort of like a forest versus the trees issue, can I bring more to the surface and show both the forest view and the trees view for my users? Interactivity is great at shedding light on what's hidden or occluded in your data. So here's an example of interactive experience that I created, and I'm going to start off by admitting to you that I did this in exactly the wrong way, I did it completely backwards. I followed none of the advice of the rest of this course. I didn't start thinking it through, I didn't plan it, I didn't think about my audience, I didn't design it, I didn't go through the whole process. I literally started coding first. Now, that's the wrong way to do this, except that it's a great way to explain how to think interactively, 'cause with a very pure experience of looking at the data I had, thinking of features and how to release the story from that perspective. So, I don't recommend this, but I'm going to talk about that way to help you think about interactivity. So what we're looking at is one dot for every single team in the NHL, the National Hockey League, and every year of play. So as you can see, a couple labels here, the 1929, 1930 Bruins are here. We're looking over here at the 1995, '96 Detroit Red Wings. So you have about 1,300 data points in this one bubble chart. Now when I first created this, I'm going to scroll down a little bit so you can see, there were no filters at the top, and if you can imagine, there were no labels and no colors and no legend, and nothing, I just had an axis and a bunch of dots on the screen. This is the data that I had available to me. So, as I mentioned, there are 1,300 or so rows of data, one for every season for every team. And back up to the top here you can see I had a code for the team, season, how many games they played, wins, losses, ties, points that they got for the season. And just for point of reference, you get two points for a win, one point for an overtime win or tie. And points percentage is how I'm measuring quote on quote, the best team in hockey, and that is the total points of team one divided by the total that they could have won if they'd won every single game. And we also have a couple more data points here, one of which is strength of schedule, which is, were they playing a lot of hard teams that year, or a lot of easy teams that year? And then some other data that I'm mostly not using. But there's a lot of data in here to look at. So, again, back to the scatter plot. As I mentioned, I had a lot of data to work with. I had a simple scatter plot that had a lot of dots on it, but my first question was, what's missing? What's hidden? What might I want to know if I'm a user? And so, first thing first, what do the dots mean as a user, if I'm just using a bunch of gray dots? So the first thing I added were these labels, and I decided of course not to label every single dot, 'cause that would have been insanity. I just labeled the best and the worst, right? Red and green, best, worst labels. By the way, side note, I double-checked and I used a color blind friendly red and green, these will both work well for all different types of color blindness. So, now that I have the dots labeled and I know that the best team ever was the 1929, 1930 Bruins, which of course made me happy, being from Boston. And the worst team ever, 1974, '75 Washington Capitals. The next thing I might ask myself is, okay, I know the best and the worst, but what about all the rest of the teams? Okay, so thinking interactively, of course it was logical to then add rollovers. So, as I interact with a dot, I can now roll over any dot and find out the details of these teams. And so I go down here and I see these two dots that were just above the '74,'75 Capitals, and they're the 1992, '93 Ottawa Senators, and the same year, San Jose Sharks. What happened that year? Two teams that really struggled, interesting. So, back to what's missing, what might I want to know as a user, again, driving me towards functionality. I'm looking at the Bruins up there and I'm happy to see that, but I know that my friends in Montreal are going to say to me, Bill, come on, the Bruins are not the best team ever, it's the Montreal Canadiens. They've won 26 Stanley Cups over the years, the Bruins have only won six. And so I said, alright, let's give tools to allow people to filter the data to find what they might find useful. So that's when I started adding these filters you see at the top here. So I can go in here and filter by teams. Alright, my guy from Montreal, let's let him sort and find just the Montreal Canadiens and look at them. And yes, see the dotted outline means Stanley Cup winners. And sure enough, there's a lot of Stanley Cup winners in there, he's right. I also might want to say, okay, let's just look at Stanley Cup winners. What does that look like? Okay, by that measure, the '43, '44 Montreal Canadiens were the best ever. Look at this, '37, '38 Chicago Blackhawks. They only won 14 games, they really struggled that year as a Stanley Cup winner, that's a very interesting find. I also then said, okay, what about seasons? I might be interested to see different years. So I can go in here, and let's say look at that year where we had those two teams that really struggled. 1992, '93, look at how big that spread is, that's interesting. Or I might want to look at 1996, '97, let's see what we see in there. Okay, a much more tight spread of dots, Boston Bruins at the bottom that year. Well, let's look at 2010, 2011, another year the Bruins won the Stanley Cup, all very interesting stuff. Maybe I want to look at decades, maybe I want to look at the 1920s, interesting, a really big spread of numbers, versus the 2000s. Much tighter spread, seems like a more competitive era. It's an interesting way of looking at the data. Finally, and this is where we get to that credibility question, I added a bunch of these, what I called superlatives. So, let me look at who won the most games in green, versus who won the least games in red. And this is just sort of the top 50 and bottom 50. And sure enough, the top 50 had very high points percentages, and the bottom 50, for the most part, didn't do so well, with some exceptions, of course because there was some short seasons over the years. But then I said, alright, across these different superlatives, there was one down here called Strength of Schedule. As I mentioned before, that means teams that had either a really easy schedule, that played a bunch of bad teams, or a really hard schedule, they played a bunch of really good teams. And so interestingly, the Washington Capitals, remember the worst team ever by points percentage? Happen to have one of the hardest 50 schedules of the entire 1,300 dataset. And by the same token, those Boston Bruins who I was so happy were quote on quote, the best team ever, happen to have one of the easiest 50 schedules in the entire NHL record. So, from a credibility standpoint, this interactivity has exposed something that I don't necessarily want to share with you. But by doing so, you're probably more likely to buy the rest of what I'm trying to show you with this data. So, thinking interactively, again, what else might I add to this? What other features might I add? Well, let me start by saying this, in interactive data visualization, the one thing I don't think you should do is try to recreate Excel. You don't have to add a filter and a button and a feature to expose every single, itty bitty, little bit of data. You're supposed to be, remember, thinking about your data, your story, and your audience. You want to think about the insights you're trying to bring forth and only add features and buttons that expose exactly that information. So what do I want to see as a user of this? I might want to be able to compare two teams. Show me just the Bruins and the Canadiens. So I haven't added that feature, but I might do that at some point. Or I might want to look at cross seasons, I want to look specifically at 1926 and 1942. Or I might want to say, show me teams with franchise players, I would love to see how the Bobby Orr era of Bruins would compare against the Wayne Gretzky era, Edmonton Oilers. So those are things that I would want to get out of this data as a hockey fan. So those are features that I might want to add to this interactive experience. Interactivity let's your user dig deeper and find their own stories in ways that you might not have thought of, and/or that your editorial process makes impossible to include. So channel your audience, know your data, think about what's missing, and then make functionality to give that to them. Think of yourself like an outfitter for an adventure company, you're providing gear. You're letting your users go knowledge spelunking on their own, you're just providing the tools. But don't imagine those tools until you've envisioned the answers that you're trying to help them to find.

Finding the right technology

If you do decide to go interactive, you have to think about what technology you're going to use. And this is a very largest subject, a lot of choices that are going to bring forward a bunch of different features and benefits and really this could almost be its own course, but I'm going to try to just outline some of the big issues and point out a few technologies within the current library available. So, this, in the first rule when thinking about technology, is don't really think about technology. Okay? So, essentially you want your goals and your audience to drive all of your decisions, including your technology decisions. Think about your audience, okay? And yes, of course, you have to think about your technical capabilities. If you do that, the options will kind of reveal themselves, right? You'll be reduced to a manageable list to choose from. So, things to focusing on, one is the vision for the platform. Are you building a platform that needs to evolve and change over time? Or like a one-off, that's just going to be thrown out? If it's a one-off, maybe you find the simplest technology, you ignore re-usability and scalability. Otherwise, you might pick a more robust, modular, reusable, scalable technology. You also think about your audience. What's their demographic, right? Are they technophobic and under gadgetized or techno driven and gadget laden? Are they somehow, I don't know how this could possibly be, but still stuck in a world of IE6 and blackberries, or are they more current in their technologies? Right? Many choices revolve around audience device compatibility. Another thing to think about, is should you be using straightforward charts and graphs or maybe pushing the envelope both visually and in terms of interactivity and complexity of the visual experience, et cetera. So there are a whole bunch of tools out there, tools like Tableau, which is a a platform for doing data analytics and visualization and creating dashboards, Power BI, which falls into a similar category. You have web-based tools like Datawrapper, where you can upload your dataset and create interactive visualizations that can be embedded into websites or exported as images and put into graphics. Tools like Flourish also, which allow you to create fairly customized animated visualizations with no technical experience within a limited set of templates. You also have tools like Carto, which are cloud-based mapping tools, which allow you to create very sophisticated maps with again, very little technical expertise although you can add some technology to it and take it even further. And you have tools like D3, which is a JavaScript library for creating all kinds of interactive animated visualizations, it's programming. So you can really create anything you can imagine within a set of established simple charts and graphs, sure, but because it's code, you can do literally anything you want. And speaking of which, there's a platform called Observable. It's a cloud-based tool, a web-based interface for programming in all kinds of different libraries, including D3. And what's neat about this, is that the code, which by the way you can find other people's code and then play around with it, allows you to just make a change in the web browser like I'm going to change the duration and the delay for this visualization. And now I can just make it happen and do exactly what I asked it to do. So you can experiment with your visualizations live in browser without setting up a server, without really even knowing too much about the code you're playing around with, because maybe it's something that somebody else created. And then you can download that version of the code and expand upon it on your own. So tools like Observable and a lot of these other software packages are really amazing. Some of these are fairly new and bring a lot of life to the data visualization community and for your options when picking tools. Other things to think about is, do you have the technical ability to pull off your vision? Right? If you're going to work in D3, Observable, et cetera, are you a programmer? If you got it, fantastic. Or if you need help, great. Do you have access to people to help you? And you also think about, who's going to be maintaining this long-term? If it's no one but you, you can do whatever you want of course, but if you're handing it off to another team, it's of course conscientious to think about who might be picking it up and maintaining it down the road. If you know who it is, you know their capability is great. If you don't know who it might be, of course, you might want to lean a little bit closer to standards, technologies that are more abundant to be essentially more likely to be able to find somebody down the road who can maintain it. Another thing to think about, are the existing technical requirements. Does your client or your company have the technology in place that might be driving the decisions you need to make? Right? Maybe they are an all Microsoft team. Therefore, Microsoft technologies might be preferable or maybe they're all open source, et cetera, et cetera. Yes, it's about device limitations as well. So, by the way, if they have a tech team that's devoted to a particular type of technology, odds are, it's going to make sense to adhere to that. Otherwise, maybe you have more freedom once again, to do whatever you want, or maybe that team wants to transition to something new and you can help them essentially accomplish that through the project you're working on. In the end, finding the technology to use can be a very nuanced process. And by the way, it's only getting more nuanced. There are more tools, new things coming out all the time. So just remain client-focused, thoughtful, pick what works for the audience, the people managing the technology and collaborate with those folks as much as you possibly can.