Defining visualizations

Here's how I'm going to organize this class. First of all, I am going to define what I mean by visualization. Then, we will explore the different components of any data visualization. In our third step, we are going to talk a little bit about how visualization may lie. And this is connected to my newest book, "How Charts Lie," or better said how we sometimes lie to ourselves with the charts that we see every day. And then finally, we are going to learn certain, very important principles of design that apply to visualization when the purpose of that visualization is to communicate ideas.

So let's begin by defining what we mean by data visualization. Data visualization, as I said before, basically consists of mapping data onto objects. And when those objects, when the data is mapped onto those objects, certain properties of those objects will vary according to the data that you're trying to represent. We will get to that in just one minute.

But the purpose, the reason why we do that, the reason why we visualize data is that we want people to be able to extract meaning from those data. A data visualization is, above all, a tool that we can use to extract insights from the numbers that we see every day, or that we see in a spreadsheet or whatever. Right? The way that I usually explain this is that a visualization is like a pair of glasses. So if I take my glasses off, I am not able to see things very clearly. But if I put my glasses on, this tool that I put in front of my eyes allows me to see better.

A data visualization can be conceptualized in a very similar way. It is a tool that enables us to see beyond the complexity of the data. And what are the purposes of data visualization? Well data visualization can be used to explore data, and you have already learned a little bit of this in previous modules of this course. But it can also be used to communicate ideas based on those data, which is what we are focusing on in the current module.

So visualization can be used for exploration and discovery, or it can be used for communication. Although in both of those cases, the idea is basically the same. The idea is that we human beings have a very hard time extracting patterns and trends from large amounts of data. For example, the data that you have on the screen right now is a very large data set of global temperatures from the year 1000 up to the year 2000, measured in Celsius degrees in comparison to an average. The average of the 20th century between 1961 and 1980. if I am not wrong. Right? So that's the reason why you will see negative value and positive values in this data set.

If I showed you this data said, if I give you, for example, the Excel spreadsheet, the data is put into, and I ask you questions based on this data, it will be very hard for you to answer those questions. For example, let's suppose that I show you this data, and I ask you, "Was the average global temperature in 2000, in the year 2000, higher or lower than it was in the year 1000?".

Well, if you want to answer that question based on the dataset, you will need to open the Excel spreadsheet. Go to the top of the spreadsheet. Take a look at the temperature of the year 1000. Then scroll down, all the way down, in the spreadsheet. Take a look at the last number on the data set, and then compare mentally – in your brain – one number to the other number and see which one of them is higher than the other.

Spoiler alert: The average temperature in 2000 was higher than the average temperature in 1000. Right? But the question is hard to answer. I'm forcing you to open the data set and then scroll up and down. Right? It's hard.

Now let me pose to you an even harder question. Let's suppose again that I show you this data set, and I ask, "Has at any point in the past 1000 years the average temperature been higher than it was in the year 2000, which is the latest year in this particular data set?" That is a much harder question to answer. I am forcing you to again open up the data set, scroll up and down, read every single number, and see if whether any of them is higher than the latest number of the data set. That's very hard to do.

But what about if instead of showing you the data, I transform, I map these data into certain objects. Right? And I transform those objects in proportion to the data that I have on my data set. In other words, if I transform these data into a line chart. Right? The data that I was showing you is the data behind one of the most famous data visualizations designed during the 20th century. It's commonly called the hockey stick chart.

And the story that the hockey stick chart tells us is that in the past, between the year 1000 up to the year 1900 more or less, global temperatures varied. That would be the black line that you see in the middle. Global temperatures varied, but they varied within a certain range. And it's only after the year 1900 that global temperatures started spiking up very, very rapidly. Right? That's the story that the chart conveys. It's the pattern that we didn't see before when we were looking at the numbers. But that pattern becomes evident, becomes visible, becomes unavoidable once we map those numbers into objects, and we transform all these numbers into a particular chart.

This chart, by the way, contains tons of layers of information. First of all, you will see that there is a blue line. The blue line are our estimates of temperatures, right, for every year in the past. And I say estimates because obviously we don't have records of temperatures from the year 1000 or from the year 1500. What climate scientists do to estimate what the temperature was in the past is to take a look at proxy variables.

Variables that we can explore and that can tell us something about the variable that we want to learn about. Right? If we don't have direct access to the temperature of a particular year, what we could do, for instance, is to take a look at growth patterns of tree rings. If you cut a tree and you take a look at the rings inside, based on the width and the distances between those rings, you can sort of estimate what temperature that particular tree was experiencing at each point in time. So if you have a very old tree, you can sort of have a map of global temperatures of the area where that tree grew. Or what the temperature was in that particular place.

In any case, blue line is the estimates based on these proxy variables. And then the red line. The red line is actual records of temperatures, beginning I believe in the 18th century. In the middle of the 18th century, 19th century, we have actual records of temperatures, and that's the red line.

The black line that you see in the middle, that's called the smooth. Right? It's sort of the average direction of the data. It's a line that was added to make it more evident what the directionality of the data is, whether it goes up or it goes down. Because as you can see, both the blue lines and the red lines, they're both very squiggly. Right? They go up and down like that. It seems almost random.

And then the gray area behind the lines, that's the level of uncertainty of the data. Obviously, if you're using proxy variables to estimate something that you want to analyze. If you use a proxy variable, there will be a lot of uncertainty around that. But once you have, that's the reason why the gray area is so wide, so large in the past. But that gray area, that level of uncertainty, becomes narrower and narrower and narrower the closer we get to the present because once you have actual records of temperatures, that level of uncertainty becomes much, much smaller.

This is the power of data visualization. The power of letting you see patterns, trends, stories that may go unnoticed if you only take a look at the row numbers or the actual figures on your data set. Once you transform those into some sort of chart, you will be able to see those stories immediately.