Identifying encodings

Let's take a look at a couple of data visualizations and see whether we can identify the encodings that were used in those graphics. The graphic you have on screen right now, it was designed by The Marshall Project. The Marshall Project is a news organization, a nonprofit news organization, and a while ago, they did a story titled Crime in Context. And within that story, you have a line chart, a line chart that shows right now the rate of violent crime in different cities between 1975 and 2015. And then you have the variation of that rate over the years.

Before we get to the encoding, by the way, I would like to point out a feature that I really like about this project, which is the menu. All right. The navigation menu on top. You may notice that the menu has been integrated, it has been blended with the introduction to the graphic itself. You can vary the different elements in the introduction to the graphic, and the graphic will change. So if you want to see, for example, instead of rates, you want to see the total counts, the total number of violent crimes, you can select that. And then the intro to the graphic will change, and the chart will vary accordingly. And then if you want to see specific kinds of crimes, for example, if you want to see specifically homicides, you can select that from the top menu. And then all the elements in the introduction text will change accordingly. You can also change the city, so if you want to see those numbers not from Milwaukee, but for example, Miami, where I live, or Miami-Dade County, so I'm going to select Miami-Dade County. I can see that now, right now, that the rate of homicide in Miami-Dade County was down 53 percent between 1975 and 2015. And the time range can also change. Anyway, this is a very interesting way of presenting the data.

But in any case, let's focus on the encoding, so let's focus on the chart itself. Which encodings, among the ones that I showed you before, can you identify here? I'm going to just give you a give you a clue, or give you an idea. First of all, focus on the line itself, right? The line that represents the change, or in this case, homicides in Miami-Dade from 1975 and 2015. The encoding there is position, why position? Well, think about how a line chart is created. A line chart, the scaffolding layer of a line chart is usually a vertical axis, which measures what it is, what you want to measure, more homicides, fewer homicides, and then the horizontal axis is in this case, it is time, right? Is year by year, right? What you do in a line chart usually, conceptually is first of all, place tons of little dots, each one of them corresponding to a year, over the horizontal axis. One dot from 1975, 1976, 1977, and so on and so forth. You put those dots on the horizontal axis, you change the position of those dots according to the years. And then what you do is to change the Y position, the vertical position over those dots according to the metric that you want to measure in this case, homicide rates. The higher the position of a dot is, the higher the homicide rate in that city is in that particular year.

But then what we do is to connect those dots with lines, but the encoding is still the position of the dots that mark the years, right? As a byproduct, though, as a byproduct of connecting those dots, we could also say that a secondary encoding in a chart like this is the slope, the slope of the line, the angle of those lines is also sort of a clue as to how to, you know, interpret or to see the variation of homicide rates year by year. But the primary encoding is still position, right? The position of those dots identifying the homicide rate for each specific year. We could also say, by the way, in this chart, that color hue is used as an encoding because color hue, in this case red and gray, color hue is used to identify the city that you are highlighting, in this case Miami-Dade and the other cities. All the other

gray lines that you're having, though, in the background, all those lines correspond to all the other cities in the data set. They are still visible because the designer of these charts wanted to put the line that you selected in contexts. Allowing you to compare the line that you're selecting with all the other lines in the data set. But the color hue was used to identify, again, the line that you're interested in and all the other lines that should stay in the background.

Another chart or another data visualization that we could use to test our ability of identifying, to identify encodings is this produced by NPR, National Public Radio. So this is a chart that shows a map that shows, as the title says, a dramatic rise in health insurance coverage under the ACA. The ACA is the Affordable Care Act, also known as Obamacare. Take a look at the map that you have in there. That kind of map, because you are interested in terms, the terms that we use in data visualization, is called a choropleth, choropleth map. Choropleth maps are maps that use well, sort of spoiler alert, the encoding of this map is shade of color, right? The darker the color, the higher the number that that color is representing. So in the map that you have on the screen right now, the main encoding of that map is color shade, right? As you can see, the counties on the map, the ones that have darkest colors are the counties that have highest rates of lack of health insurance. The darker, darker the color, the more people you have in that particular county who didn't have health insurance in 2015. And the lighter the color, the fewer the number of people who didn't have health insurance in 2015. By the way, this is an interactive graphic. So you can you know, you can hover over and you can also use the time slider to go back in time and compare the current colors, the current shades of colors to how the map looked like in 2010. And you will see that more counties have much higher rates of uninsured people, right? Before the Affordable Care Act was approved and was passed. All right. So it's shade of color, right? Color shade is the main encoding here.

But there is another one because there is another graphic on the screen right now. Take a look at this sort of bar graph that you have right underneath the color, the color legend. This is not really a bar graph, although it uses bars to represent the data, this is called a histogram. A histogram is a kind of bar graph, a variation of the bar graph that is used usually to represent the distribution of a variable. The number of counts, that in this case, the number of counties that you have within each specific range of an insurance rates, right? So basically what this chart is telling you, if you pay attention to it and considering that the height method of encoding here is height, the height of these bars is proportional to the counts to the number of counties that you have within these sort of beans of data. This is basically telling you that in 2015 there were 110 counties that had a rate of lack of health insurance between 0 percent and 5 percent, more than 1,000 counties that have a rate between 5 percent and 10 percent, around 1,000 counties that have 10 percent to 15 percent, 500 counties or something like that between 15 and 20, 182 counties between 20 percent and 25. And then 48 counties that have a rate of more than 25 percent of people lacking health insurance. Take a look at how this histogram looks like. If we go back in time to 2010, you will notice that the bars on the right end of the spectrum become much higher, becomes much taller, showing you that there are more counties with higher rates of lack of health insurance. Again, the method of encoding in this particular chart, this histogram, is height, the height of the bars is proportional to the counts.