Data Visualization for Storytelling and Discovery: Module 1 Introduction w/ Alberto Cairo

Welcome to week one in this week. I'm going to teach you the basics of the visualization. First of all, I'm going to Define what visualization is and what it is useful for then I will move on to the main principles of data visualization. I will also show you several examples that I believe can be inspiring of visualizations created by several media Publications that I admire. And then finally you're also going to have access to several a practical tutorials in which you are going to learn how to prepare your data before you visualize it.

Data Visualization for Storytelling and Discovery: Module 1 video, part 1: What Visualization Is

As long as you have signed up for this course, I guess that you are very familiar with the data visualizations that we see everyday in news media. There has been an explosion in the use of graphs, charts, and maps, and diagrams and infographics in media Publications. The reason for that is that these media publications have observed that readers really like information that is presented visually. At the same time, if you're familiar with these kinds of graphics, you may have always wondered well, how are these things done? Alright and one of the problems that I face in my own classes at the University of Miami. It's some of my students some of the students who come to my classes tend to believe that data visualization is sort of like magic. It's something hard to do. It requires the use of very complicated and complex software and that is only half true.

The purpose of this course is to teach you how to do data visualization and to prove that it is not magic. Before we get started though. I would like to show you several projects that I really like and I believe can be inspiring for your own work in the future. The first project is this data-driven, uh story done by Univision Noticias Online. Univision as as you know is a largest Spanish-speaking TV station in the United States, and this is a project that talks about how Hispanics voted in Florida during the 2016 presidential elections. As you notice it's a combination of tables a charts maps of Florida different kinds of graphical displays combined with text that put the data in context.

The second project that I would like to show you before we start with the actual content of the course is this very beautiful and very recent data driven story by The Washington Post titled America is more diverse than ever but still segregated. So again, this is another great great combination of beautiful visuals that actually look like um, a modern paintings or abstract paintings. Very very colorful and very very well-designed highlighting the fact as the title says that America is a very diverse country. The United States is a very diverse country, but at the same time in particularly in certain places is also a highly segregated country. So you have a country level map which you can view as a race map in which you can see each color represents one of the races or you can switch these map to a diversity map. And it will show you the level of diversity of each one of the counties in the United States.

Later on in this course, you will learn how to create these kinds of graphics. We can do this kind of map with a tool called Flourish, which I'm planning to teach you. Also you have another section in this same project that talks about the diversity in city-by-city so you can see both the less diverse and the most diverse cities, the less segregated and the most segregated cities. And then here at the end if you scroll down you have the interactive portion of the project in which you can you can input your own ZIP code and get an overview of how segregated or unsegregated your own area is. So I'm going to input my own ZIP code that will be 33186 in Florida in Miami. And this is how my district looks like.

Sometimes visualization can be used for artistic purposes or expressive purposes. It can be a little bit more innovative. Right? So I would like to talk a little bit about that side of the visualization as well. I am right now in involved in on an ongoing collaboration with the Google News Lab in which we are partnering up with designers from all over the world. Google gives those designers data and then Google asks those designers to come up with creative visualizations based on those data. I work sort of as the middleman the art director. Of these project. So the slide over here will show you the link where you can read more about this project and also the link that will take you to the portfolio of all the projects that we have created so far. There is one of them that I would like to show you though because it is a lot of fun and I also did it with in collaboration with my French Xaquín G.V., who used to be the graphics director at the Guardian. So Google asks, Xaquín if he was interested in doing collaboration with them. And they asked him well, what do you want to do your project about? What kind of data you're interesting in? Xaquín replied that he was interested in exploring what people search for in Google when they search for how to do something or how to fix something. So when we search for how to fix um what is that um? What comes next right.

And based on Xaquín created a very very interesting and compelling project titled how to fix a toilet because certain looking for the looking into the data that that Google. Xaquín discovered that one of the most common searches in Google is how to fix a toilet. Xaquín decided to uh create this project as a personal essay. So when you read it when you take a look at it, you will notice that Xaquín begins by confessing that he doesn't know how to fix a toilet. So he's one of those people who search for how to fix a toilet in Google. And you can scroll down and start seeing the project. This is a title how to fix a toilet. Another thing that we cannot do with search you have these very nice and cute animation over here. And if you keep scrolling down you will get to the actual data visualization, which is this one right? I'm going to explain what you're seeing over here.

Each one of these icons representing one thing in a common household represent the number of searches in a particular country for how to fix that particular object. So the bigger the object is the larger the number of people who search for how to fix that particular item. The pink line that you have surrounding each one of these objects represents the average worldwide interest for searching for that kind of that kind of object how to fix that kind of object. So if you see that the object these black object like an object is bigger than the pink line it means that in that particular country searches for how to fix that object were larger than the worldwide average. And if the

object is smaller than the pink outline, it means that searches in that particular country for how to fix that object where a smaller than worldwide than the worldwide interest.

Right now I have not put any country over here, but I'm going to do that. I'm going to move ahead. I'm originally from Spain. So I'm going to go over here. I'm going to type in Spain. In Spain we search for how to fix, um, let's say these things. And you notice that the sizes of all these objects have have changed very quickly, right?

So for example in Spain people searched a lot for how to fit a light bulb. This is not how to fix a light bulb by the way with I don't know how to do that. But how to fit a light bulb when the light bulb starts stops working. Um, I don't know why we search for these anyway, but it's still a lot of fun to discover. We don't search quite a lot for how to fix the refrigerator or how to fix a toilet. Right? It's how to fit a light bulb.

Anyway, let's search for other countries. There's a lot of fun. I'm going to go over here and instead of Spain. I'm going to search for for example, Russia. Let's see what people in Russia search for when they search for how to fix something and the patterns will be completely different because people in Russia for some reason when they Google how to fix something they think that they search for the most is washing machines. How to fix the washing machine? Washing machine as you see its enormous right now on that display and all the other objects are become much much smaller than the worldwide average.

Let's search for another country the United States. However, go over here and search for the United States you will see that. In their eyes stays like people in Spain they search for a lot for how to fit a light bulb when that light bulb stops working.

If you don't have any prior experience in data visualization whatsoever as I mentioned before all these projects may look to you like magic. You may be thinking. Well, I will never be able to create any chart or any map similar to these ones that you're showing to me right now, right? And as I said before, I don't think that that is true. I believe based on my experience as an instructor that anybody can learn good data visualization. If we learn two different things. First of all, we need to learn several principles and this is what this first week is about. It's about the main principles of data visualization design. And then we will need to pair these principles with the use of several software tools that you can also learn.

Data Visualization for Storytelling and Discovery: Module 1 video, part 2: Elements of Visualization

To understand how visualization works it is useful to think about the elements of data visualization. So I'm going to begin by defining what I mean when I talk about visualization. For me obviously is any sort of representation that is intended to let you see beyond what you can normally see. The joke that I usually tell my students is that a good visualization gives you superpowers in some sense. It gives you x ray vision. You may begin with uh with a large

amount of data. For example, if you show me those data, I cannot see anything in those data. But once you transform those data into some sort of graphical form suddenly my brain can start seeing the messages that are hid behind those data. So I usually say that visualization is that kind of representation that is intended to reveal insights from the data with the purposes of either exploration and discovery. Or the communication of the insights that you extract from the data.

Let's take a look at a practical example of data visualization. Let's imagine for instance that I show you a large data set of global temperatures from the year 1000 up to the year 2000. In this data set temperatures are measured in Celsius degrees and they are also measured in comparison to an average which is considered this zero point. Okay. So the baseline of the graphic that you're about to see is the average global temperature between the year 1961 and 1990 and that's considered this zero. That's the reason why in this spreadsheet that you have over here. You have positive temperatures and negative temperatures meaning above the average or below the average of that. Of those two years 1961 and 1990, right? So if I show you these if I show you these data set as it is right now the Excel spreadsheet, you will not be able to extract any meaningful from the from the data other than individual figures that's worth tables are for. So if the purpose of your graphic is to let people extract every single number then the table is fantastic for that. It works really well. I can just search for a particular year I can search for a particular area of the world for example, and then I can get the data point for that particular for that particular year.

But if the purpose if your purpose is to let you people see whether global temperatures have increased or decreased from the year 1000 up to the year 2000 then these data set are series right now in tabular form doesn't really work well. You need to represent it visually you can represent it for example through a time series line chart that show you the flow of global temperatures all over the world.

Actually the data set that I showed you before is the orient of these chart that you have over here, which I consider one of the most beautiful and one of the most relevant data visualizations ever created. It is commonly called the hockey stick chart because it has the shape of a hockey stick. The pattern that it reveals from the data that we had before is that from the year 1000 up to the year 1900 more or less global temperatures varied, but they varied within a particular range that was usually below the average from the 20th century, which is the black line that have there in the middle. But once we get to the 1900s to the beginning of the 20th century global temperatures started spiking up very very rapidly because of the effect of CO2 emissions and the increased use of fossil fuels and so on and so forth.

The chart contains an incredibly large amount of information. I'm going to explain to you how to read it. First of all, do you have the actual point estimates? So what scientists believe that were the temperature in that particular year? All over the world and that is the blue Line the blue Line represents the point estimates on the chart then in the middle of the blue line, you will see that there is a black line going across that blue line. That's that line is a trend line. It's so to speak the

average of global temperatures in all those years. So it's if you only take a look at that black line, you can get a good sense of how temperatures varied and then there is a red line also the difference between the blue line and the red line is that the red line represents actual records of temperatures measured with thermometers for example. But we didn't have thermometers in back in the year 1000 right or in the year 1200. Therefore scientists cannot rely on actual records of temperatures. What they do is to use proxy variables to estimate the temperature of the world at that particular point in time. They take a look for example at the growth patterns of tree rings. They take a look. Growth patterns of glaciers for example and based on those proxy variables. They can sort of estimate what the temperature was at that point in the past. Obviously all these estimates are quite uncertain and that's another another data point that is also made in the uh, in the chart the gray background that you see behind the lines. That's the level of uncertainty of the data. So scientists are basically telling you well, we are pretty sure that the global temperature in this particular year is here or was here was this particular temperature and as the blue line, but it could also be the temperatures were slightly higher or a slightly lower. That's the level of uncertainty represented by the gray line. Okay, the gray bar the gray area behind the blue line. As you may notice that gray area becomes narrower and narrower and narrower as soon as we start getting closer to the present just because these proxy variables are less than certain the closer we get to present times.

Anyway, based on these very simple chart. Let's think about what the elements of a visualization are. I usually say that a visualization has at least three different components. First of all, it has so to speak a container the scaffolding, the framework of the chart and that includes things such as for instance the scales that we use to measure the fangs the legends that we use to interpret the graphic correctly. Then we have the actual content of the chart. To talk about that content in data visualization. We usually speak about visual encoding. I will talk about encodings in just one minute and then as a third element in most graphics that we see commonly in the media or in scientific papers or in textbooks. We have annotations. We have the text that put somehow the data in context that could be the introduction to the graphic that could be the little call out boxes that we put on the chart to put the data in context in the middle or something like that, right? So the container, the content and then the annotations.

This is will be an example of what I'm talkin about. Right? So here you have two different charts. First of all data map and second of all traditional line chart. So first of all, I show you the entire graphic. Then I show you the framework or the container of the information that includes the scales and then I show you the content or what we usually call the encoding. Now, let's talk about the encoding because the encoding part is the most fundamental idea in data visualization. Just think about what data visualization is based on. Data visualization is based on the idea of visually representing numbers proportionally also. So what you do when you create a database is to get those numbers and then you map those numbers on properties of objects onto properties of objects. So for instance in a bar chart what you do. Is to vary the length or the height of different bars according to the numbers that you are representing. In that case we say that the method of encoding is length or height therefore the length or height of each one of the bars in your chart needs to be proportional to the numbers that you are representing. But in data

visualization, they are many other kinds of methods of encoding you can also have. Position right? The position of dots measured on common scales will be another way to represent data that will be another method of encoding. You can also use size of different objects. Areas of different objects such as bubbles or squares. You can use angle for example in pie charts and one of the methods of encoding in aperture is the angle of the segments that pie charts divided into. You can also use line weight, right? So for example, if you want to display how a total display branches out into its components, let's say for example that you do a graphic of exports from the United States to other countries, you may begin with a line of this thickness. And then you may subdivide that line into narrower lines right of different of different ways of different thicknesses representing the amount of exports going to different countries right. In that case your method of encoding would be line weight.

And then finally we can also use color right we can use color hue different colors to represent a categorical variable for example, or we could use color shade. The variation of the intensity of one color being proportional to the numbers that you are representing. Now the question that you may be asking yourself is how do I decide how do I know which one of these methods of encoding I should use in my charts and we will get to that very very soon.

Data Visualization for Storytelling and Discovery: Module 1 video, part 3: Identifying Encodings

To become a good data visualization designer is actually useful to learn how to identify different encodings. And also learning how to choose between them or among them. I usually tell my students that one of the best ways to choose ways of representing your data is always to think about the purpose of your visualization. For example, before I show you a graphic or a data set that was represented through a table and then later on through a line chart, right? Each one of those representations could be appropriate depending on the purpose that you have in mind. For instance. If you want your reader to be able to identify each specific figure on the data set you better design a table. But if the purpose of your graphic is to let people see the ebb and flow of a global temperatures. You better don't design a table. You better design some sort of chart. So purpose always keep purpose in mind.

Let me show you another example. Awhile ago I saw a chart designed by the European commission. I actually I did a little bit of work with them a little did a workshop with the European commission and they sent me examples of the graphics charts, maps, etc. They design every year to inform citizens of the European Union and they were absolutely wonderful. I really love the work but I always have a little bit of feedback about any chart that I see. One of the charts that they sent me was this one. This chart shows you where migrants arriving to Greece in 2016 came from. What percentage of migrants to Greece came from different countries. So for instance 47 percent came from Syria to Greece. 24 percent came from came from Afghanistan. 15% came from Iraq and so on and so forth. So it shows you among all the total amount of migrants arriving to Greece. The different portions represent the percentage coming from different countries.

And I was asked do you think that this is a good chart or about chart? And the answer that I always have to that question is that it all depends? It all depends what you want to show. It all depends on the purpose of the chart. If your purpose when designing this chart is to show that half of migrants coming to Greece came from Syria. 47% is close to 50%. Versus another half that came from other countries. Then the chart is completely appropriate. You show me half coming from Syria the other half coming from other countries. The only change that I would make in the charting in this case will be to color these segments differently. I would color Syria with one cube with one hue of color and then I would color all the other countries with the same hue. Just to emphasize that purpose: half versus half. So if the purpose is to show that half versus half. This is perfectly fine.

But this chart could have a completely different purpose, right? For example. It could be that your purpose with this chart would be to allow me to compare very accurately the percentage of migrants coming to Greece from different countries and to rank them. And in that case this chart is not that effective. If you don't read the figures the actual numbers on the chart. It's really difficult to estimate the size of the Syria portion versus the size of the Afghanistan portion. You basically need to use your fingers to estimate the size of each one of the segments right a pie chart or this variation of the pie chart, the donut chart is not great for accurate comparisons. Therefore you need to choose.

I propose to different redesigns two different makeovers for the chart. If the purpose is to show half versus half than you can just keep the pie chart as it is only you need to color these things differently. But if the purpose of your chart is to let me compare country by country one by one then you better design something more similar to a bar chart because a bar chart is great when you want to compare things with a lot of accuracy.

Okay, let's do a little bit of a game related to visual encodings. I'm going to show you again the uh, the slide that contains most of the visual encodings that we commonly used. We use data visualization size, length, area, line weight, angle, and so on and so forth. And I'm going to show you three different real data visualization and I'm going to challenge you to identify the methods of encoding that you see on each one of these charts ready. Here we go.

The first project that I would like to share with you comes from The Marshall Project, which is one of the media organizations that I would recommend that you follow closely because they produce very very good data-driven stories and graphics on a regular basis. A while ago back in 2016 the Marshall Project created a wonderful project titled crime in context. This is basically a description of what's going on with crime statistics in the United States. And as part of this story there is this wonderful interactive data visualization in which you can explore the rate of violent crime in your own city. So right now we are seeing the rate of violent crime was up 200 something percent in Milwaukee between 1975 and 2015.

Alright. So take a look at this chart. This is a line chart. Try to identify the methods of encoding. I'm good to give you just five seconds. You can stop the video here try to identify the methods of encoding and then you can resume playing the video.

The first method of encoding that you probably saw is color hue, right? We have red and gray that's identifying a categorical variable the city that I'm interested in Milwaukee. Although we can change that. I can go over here and change Milwaukee to Miami for example and see violent crime in Miami. So right now it's what matters to me, Miami. Versus what is secondary all the other cities in the data set so color hue is the first method of encoding over here.

Now, let's think about the methods of encoding that were used to represent the quantitative data the actual rate of violent crime, and this is a line chart. So you may be thinking well, perhaps it is a slope the slope of the line right the angle of the line. And that's certainly a method of encoding that gives you a clue as to how to read the chart right? So the steeper the slope the larger the increase or the decrease of the data, but there is a more a more important method of encoding at work over here, which is position over the two axes right? Just think about how how line charts are created a line chart is created by locating a point. First of all on the horizontal axis in time 1975-1976. So you position the dot according to the year. And then you vary the position on the y axis on the vertical axis of that point representing one year according to the metric that you're interested in. So the further you go the larger the rate of violent crime, so we are using position. The position of all those dots that will later connect with the lines representing the data. Let me make an aside over here because there's something else that are really like about these uh visualization by The Marshall Project, which is the way that they integrated the interface of the graphic with the annotation of the graphic. Remember that annotation was the third component of any visualization so they integrated both those things together because here the introduction to the graphic is the interface to the graphic. So here you can change for instance instead of viewing the rate of violent crime. Show me the total number of violent crimes. And when you change that the graphic changes. I can go back to rate then you can also search for a specific kinds of violent crimes. So if you're interested in learning not about violent crimes, but more specifically our homicides you can change these over here. And notice that every time that you change any of these items these percentage increase or decrease will also get updated. The next project that I would like to talk about is also a product of these ongoing collaboration with the Google News lab. In this case myself and the Google News Lab partner up with a visualization firm called polygraph to create a visualization to explain the level of gender balance or imbalance in US News rooms and also racial balance or imbalance in in these news organizations as well.

Anyway, so I'm going to show you the graphic. I'm going to just bring it up here on my browser. The title is how diverse our US newsrooms what is the percentage of women and the percentage of men in each one of these media organizations. And as I did before I'm going to challenge you to identify the methods of encoding. Please stop the video here try to identify them and then resume playing.

Alright, so we have plenty of methods of encoding at work over here. Right? So first of all, the first one that you have probably noticed is again color hue right blue and red. Blue representing men and red representing women, but there is another color method of encoding over here, which is color shade right. The bluer the color the large or the percentage of man and the redder the color the larger the percentage of women in each one of these media organizations. Another method of encoding that you probably saw is area right the bubble size. The size of each one of these bubbles represents the size of the newsroom's the amount of employees who work in The Newsroom. So each one of these media organizations, so you can immediately identify the larger ones right. The New York Times, The Washington Post, the Los Angeles Times, The Wall Street Journal, and so on and so forth. So the larger the bubble the larger the number of employees of the organization.

But there are more methods of encoding over here. There's actually a crucial one which is position again, the position on the X scale is also proportional to the either the percentage of men or the percentage of women in comparison to a 50/50 split. So the further to the uh to the left in my case, I'm seeing the graphic over here. One of those bubbles are the the larger the amount of men that you have on each one of these terms. And if the bubble disposition to the right, it means that there are more women than men in terms of percentage balance in each one of these organizations. So we have color hue, color shade, area, and also position. Most data visualizations that you will see out there will use much more many more than just one simple or one single method of encoding keep that in mind. It's really really important. And as an aside if you're interested in learning a little bit more about this project the data came from the American Society of News Editors. Those who this is organization who provided the data and the project contains many many other sections so you can take a look at this plate in leadership positions, for example, and you can also see the data in tabular form if you wish.

Okay third exercise in identifying methods of encoding. The next one comes from. NPR which is the public radio in the United States, which also has by the way an excellent data desk and graphics desk. So a while ago NPR was interested in learning whether the amount of people in the United States who don't have access to health insurance has increased or has decreased. And they decided to show that through a combination of graphs and charts and also maps. So I'm going to show you that project. The title of the project the title of the of the story is maps show a dramatic rise in health insurance coverage under the ACA. Otherwise known as Obamacare.

If you take a look at the past versus the present there is a large increase in the percentage of people county by county in the United States who do have access to health insurance, right? So you can see that the main method of representation is this beautiful map over here that you can navigate with a slider you can go back and forth in time to see how those percentages of a uninsured people, uh change. Now which methods of encoding do you see over here pause the video try to identify identify them and I will see you in just one minute.

Methods encoding here are a little bit easier to identify right. So in this kind of map this kind of map by the way is called a choropleth map. On a choropleth map the main method of encoding is color shade. In this case the darker the color, the larger the percentage of people who don't have health insurance or each one of these counties of the United States. But there are more methods of encoding at work over here. For instance, if you take a look at the bar chart here that you have on top that chart represents the amount of counties that have different ranges of percentages of uninsured people, right? So you see 0, 5, 15 10, 20 etc. and so on and so forth. And then the bar represents the amount of counties that are in that beam of uninsured percentage of uninsured people. This kind of graphic is called a histogram. Variation of the bar graph and the method of encoding over here is height. The height of each one of those bars is proportional to the amount of counties that have that particular percentage of uninsured people. And there is another method of encoding and this is a little bit trickier to identify. Think about this. Think about this way how our maps design not the data on the map. But the map itself. How is a map created? A map is created first of all by locating points on a geographic space, right? We use longitude and latitude to define the position for example of the boundaries of each one of these counties. Those longitude and latitude are also quantitative variables that we are using to draw the map itself. Therefore another method of encoding at work in any on any map is position. The position over an x scale horizontal scale and a wide-scale longitude and latitude also are used in these kind of data visualization.

Data Visualization for Storytelling and Discovery: Module 1 video, part 4: The Core Principles

You did a great job at identifying methods of encoding so I think that you are ready to learn the elementary and most fundamental principles of data visualization. I usually joke with students with my students at the University of Miami that I have written like more than 1,000 pages about data visualization. I have written three or four books. I'm in the process of writing a new one, etc. But I could condense everything that I have taught and I have written about data visualization throughout the years in just four main principles of data visualization design.

And these are the principles. First of all a good data visualization needs to be based on appropriate data. And this sounds like a no-brainer. Obviously, you don't want to lie, right so you need to use good data. But if you work with data on a regular basis if you are a statistician or a data scientist or a business analytics professional or a data journalist you probably know how hard it is to vet your data to verify your data and make sure that your data is actually measuring what you want to measure and present to the public. The second principle is that a good data visualization needs to be well designed. It needs to be visually attractive. The reason why I emphasize this point is that throughout my career, I have collaborated with both designers and journalists and also scientists and statisticians. And I have observed that scientists and statisticians tend to believe or data scientists tend to believe that good visual design is an afterthought. If I have time I will pay attention to typography, color. Those are things that just beautify the data. Well beautifying your project making your project look elegant and professional, it's extremely important. You need to make a thing approach your project look

good. If you want people to take a look at your visualization, that's a core principle. The third principle is the representation itself. So a good visualization needs to represent the data accurately. This is the whole principle behind the idea of visual encoding. And finally the fourth principle a good visualization needs to show an appropriate amount of data as sufficient amount of data. By sufficient amount of data I mean that visualization should never oversimplify things. And we journalists tend to oversimplify things. But it should also never overcomplicate things. Showing more data than it is needed to understand a particular story. We need to do we need to stay in between those two extremes: oversimplification and over complexity or over complication.

Let me show you a project that is say that I believe exemplifies what can go wrong when we don't pay attention at this principles. This is a map that hangs on the walls of the White House at the moment. So this is a photograph that was taken by a reporter visiting the White House. As you can see it's a map that represents the results of the 2016 presidential election at the county level. Red represents the areas that were won by President Trump and blue represents the areas that were won by candidate Clinton. President Trump is very fond of this map. He really likes this map. There's a reason why he told his aids to print it out and frame it and put it on the walls of the White House, right?

But he likes it so much that apparently according to several news stories. Every time that reporter comes visit him in the Oval Office at least in 2017. He handed out copies of these map. He had copies of this map on his desk and the reporter came in and president gave him a copy of this map. President Trump has also tweeted about this map. A while ago president Trump was trying to encourage, Texas' voters to go to the polls and vote in the primary for governor. And he tweeted something I'm not going reveal, but he says said something like I want to encourage all my Texas friends to vote in the primary to Governor Greg Abbott and so on and so forth.

And someone who opposes president Trump replied to him on Twitter, "you have no friends?"
And a supporter of President Trump's replied to this other person, "really do you think that we have no friends? Take a look at this beautiful map."
And actually president Trump retweeted this supporters saying, "such a beautiful map. Thank you."

Well, the map is certainly beautiful. I have nothing against that map. It's perfectly fine mathematically speaking, but is a map that is also highly misleading. It has also misled people who write books about the election. Year ago for example author Jack Posobiec published the book title Citizens for Trump in which he put that map on the front cover. Citizens for Trump and you have the map.

Well having worked in data visualization for more than 20 years and having taught how to do data visualization effectively for more than 10 and also, you know being a person who tries to be nice to other people. I like to offer my free advice to whoever wants to take it. And when I saw

Posobiec tweeting about this map, I replied to Posobiec saying, you know, I think that you either need to change the title of the book or change the map. Because the map is not showing what the title says. The title of your book is Citizens for Trump "citizens" for Trump, but the map is not representing citizens. So I am aware that changing a map can be a little bit complicated. Although if you take this course, you will learn how to design these kind of maps.

So perhaps instead of changing the map what you could do I told Posobiec will be to change the title of your book. Instead of titling your book citizens for Trump. Perhaps you should call it Counties for Trump. Because the map is not representing citizens, the amount of people who voted for President Trump and the amount of people who voted for President Clinton for candidate Clinton, it represents the amount of counties. The amount of territory that were won by each one of these candidates and that's not a good representation of the popular vote.
To understand why this map misleads so many people we need to go back to the principle that explained before the principle of visual encoding. So on one hand the title of this book is Citizens for Trump number of people who voted for either Trump or Clinton right number of people. But the thing that the map is representing the proportion that the map is representing is not a 50-50 split which is what actually happened in the 2016 election right. Around half of voters voted for President Trump and another half of voters voted for candidate Clinton, but what the map is representing is not a 50/50 split. It's an 80 percent versus 20% split because if you take a look at the sheer amount of red that you have of them on the map versus the amount of blue those amounts are 80% red and 20% blue. Therefore is not a good representation of the actual percentages of the popular vote.
So um actually said, well, you know, a better representation of the data could be something as simple as a bar chart. A bar in which we represent those numbers of votes, right. 40 I believe that the backdrop percentages where 46 percent voted for President Trump 48 percent vote for candidate Clinton and around six percent or something like that voted for other other candidates. Uh, but even even this chart doesn't capture the reality of the popular vote either. So if you want to title your book Citizens for Trump or Citizens for Clinton, it applies the same way. It doesn't really matter which candidate you voted for. This chart is not great either because it obscures a very important piece of information. Remember a visualization should show a sufficient amount of data, right? That's the fourth principle of data visualization.

Well, one thing that this chart obscures is that there is a high percentage of people who didn't even vote. When you take that into account citizens who citizens who could have voted but didn't vote. The chart changes completely because that percentage is nearly 40%. 40% of people didn't even vote. When you take that into account, you may discover that Citizens for Trump or citizens for Clinton are actually around 1/4 of the citizens who are voting age already. So the visualization changes completely.

Now let me tell you this country the United States the country were live in is becoming so ideologically divided the split is becoming so wide that is also influencing our preferences in terms of data visualization. So I follow both conservatives and liberals and progressives on social media and I have observed that President Trump supporters love the first map. Why?

Well because it speaks about what they want to believe right? They want to believe that President Trump won on a landslide. And he didn't win on a landslide. He won but he certainly didn't win on a landslide.

On the other hand when liberals and progressives who I following social media see that map tweeted by conservatives. They tend to reply with these other map. This bubble map using area as method of encoding. Now this map is showing the amount of votes won by the candidate who won on each one of these counties. So the bigger the bubble the bigger the vote for a candidate Clinton in Democratic countries, and in the case of red bubbles, the bigger the bubble the larger the amount of votes for president Trump in the on that particular county, right. So liberal tend to say, you know this map better represents the popular vote. So if you want to title your book Citizens for Clinton or Citizens for Trump, perhaps we should use these map instead.

Well, I would like to say to tell my liberal friends that they are also wrong. This map is also not good. It doesn't represent the data well. And it also doesn't show a sufficient amount of data. Why? Because if what you want to show is the popular vote you cannot obscure the fact that there are plenty of trump voters in Democratic areas and plenty of Democratic voters in Republican areas. Therefore you what you need to use or what you need to design is not just one bubble map. Or using the term that we usually use in data visualization proportional symbol map right. You need to use two maps one of them representing the amount of votes county by county received by President Trump and the amount of votes received by a candidate Clinton. Now the reason why by the way that the original map was so misleading and this is a fact that these two maps reveal. Is that many counties that were for that went for Trump during the 2016 presidential election are enormous in terms of amount of territory that they cover in terms of area. But they have very little populations very small populations. They look enormous when you represent them on the map, but in terms of population that are very small.
Democratic vote tends to concentrate on urban areas right in cities and this is what the second map reveal. Versus Republican vote tends to be more spread out all over the country and it tends to be bigger in rural areas. So these two maps are actually a little bit better if what you want to show is the popular vote.

But here comes the key question. If you want to talk about your victory, if you want to title your book Citizens for Clinton or Citizens for Trump, it doesn't really matter. Would you use any of these graphics? I wouldn't. Now let me tell you this. Let me make a tail over here. Imagine that. I run for president in the United States and that I win the presidential election. Now, this is an impossibility because I was born in Spain not in the United States so I cannot run for president. But just for the sake of argument imagine that run for president and that I win the presidency. I become president Alberto Cairo of the United States. If I win a presidential election, I would never print out any of these maps to pose to put on the walls of my white house or to print out a give to reporters. Because they don't capture the key metric in a presidential in the U.S. Presidential election.

To win a presidential election in the United States. Winning the popular vote is secondary that will not make you win the presidential election and certainly the amount of territory that you control in a presidential election is also beside the point. The metric that really matters in a presidential election, the U.S. is the number of electoral votes. Because the United States has an indirect voting system, right?

Whoever wins a plurality of the vote on each state that means more votes than any other candidate wins all the electoral votes of that particular state with a few exceptions such as Maine, but in general this rule applies. You in for example, 41% of the vote, but all the other candidates have fewer votes than you, you get all the electoral votes from that particular State and that's what makes you win the election.

So if I wanted to print something out and post on the walls of my white house after winning the presidency, I would actually print out something similar to these. First of all some sort of chart that shows the split in the popular vote, sorry in the in the electoral vote. How many electoral votes each one of the candidates got. In the case of the 2016 election, it was 300 and something for President Trump. 220 something for candidate Clinton. And then a couple of maps. One map that shows the results of the election at the national level at the state level. Who won where and then a second map that distorts the areas of each one of these states according to the number of electoral votes that they contribute to the election. When you pair when you combine all these graphics next to each other they provide a very good picture of what happened in the 2016 presidential election.

Continuing with the idea with over the four principles of data visualization right showing the right data. Making something that is attractive. Representing the data correctly and then showing sufficient amount of data. The original map is fine from the point of view of design. It is only that it was used for the wrong purpose, right. The purpose was to show the number of people. But I actually will represent the amount of territory. Therefore we are not using the right data in that particular case. The might is sort of attractive so it fulfills that purpose but then it doesn't represent the data proportionally, right? Because we are not representing people therefore instead of representing a 46, 48 percent of the vote. What we are actually showing is an 80% and 20% split. And then finally the most important point is that it doesn't show us efficient amount of data because it doesn't show the key metric that you need to win a presidential election, which is the number of electoral votes.

I could go a little bit beyond that. Let's imagine that for some reason I am not limited in terms of space in terms of wall space in the White House. Let's imagine that for example, I could use an entire wall of the White House to print out not only one map and chart but several ones. I will use all of them. I will print out all of them. Because when you combine them all. Alright, you get a better picture of one of the realities of the presidential election. I would begin obviously with the uh, electoral map. I will show the number of electoral votes. I will continue with the map and charts that show the split in the popular vote. And then I will continue with the county level vote or the map that President Trump loves so much because that map's is perfectly fine. It's only

that it needs to be used for the right purpose. Which is not to show the popular vote. It is to show geographic patterns in the data. Right? It's like Republican vote concentrates here. Democratic vote tends to concentrate here as well. And I will display them all together on the walls of my White House.

Data Visualization for Storytelling and Discovery: Module 1 video, part 5: What Comes Next

I would like to remind you of other things that you need to do this week besides watching these videos. First of all, remember that on every week of these scores that will be a series of readings that expand on the content of these video lectures. On this first week also remember that there is an exercise that will be described in the forum. And this is one of the requirements you will need to complete this exercise if you are planning to apply for a certificate at the end of these MOOC.

Also, there is a quiz and there's that's another thing that you need to complete if you are planning to apply for the certificate at the end. And finally there are also optional video tutorials about software tools. So those video tutorials would teach you for example, some elementary principles of spreadsheets to manipulate data and to clean up data. And there's another video tutorial about how to use a data tool that I use myself to transform data called Data Wrangler. This is uh, this is a free tool.

At the same time. I'm just to finish this video I would like to mention data sources because this is one of the most common questions that I usually get when I do a course like this. Where did you get this data? Well, I usually work with publicly available data, so I go to sources that are the usual suspects. So I get tons of data from the United Nations, for example, or from the World Bank.

I would like to mention just a couple of these sources because they may be helpful for you in the future. If you want to find data to play with. One of them will be Gapminder. Gapminder is the website of Hans Rosling who was a Professor of Global Health in Sweden who sadly passed a year ago. And Gapminder has a section that curates data sets. Data sets that come from places such as the the United Nations or the International Labor Organization or the International Monetary Fund. What the Gapminder Foundation does is to get all those publicly available data sets and they clean them up for you. So you can download them and start playing around with them in your data visualization tools. In some cases these datasets will need to be transformed before you can visualize them. So that's a little bit of a warning you will need to bring them up into Data Wrangler for example. The data tool that I show you before in order to shape the data correctly before you visualize it. That's something that you can learn in the optional video tutorials this week. And finally another source that I usually use our uh statistical agencies. The Census Bureau for example in the United States, uh the Statistical Office in Spain or Eurostat in

Europe also provides curated and very large data sets of European data that you can use in your own projects.

Thank you for finishing Week 1. I will see you in week two.