Data Visualization for Storytelling and Discovery: Module 3 video, part 1 - Choosing Encoding

On week 2 we use visualization for exploration and discovery. And now we move on and we start learning how to use graph charts and maps to communicate insights to people.

Let's go back to the idea of encoding. You remember the core idea behind data visualization is data visualization consists of visually representing numbers through the variation of certain features of objects such as the length of an object, the height of an object, the the position of that object, size ,a line weight, area and so on and so forth now. I'm going to give you a few. I'm going to show you a few techniques that can help you decide when you have a data set and you don't know which in which way you should represent the data which method of encoding may be more appropriate. What are the techniques that we could use to decide on that.

As a reminder of what I said and week number one the key word is purpose. What it is that you want people to see. What is the graphic is for. Do you want people to compare things? Do you want people to see change over time? Do you want people to see geographic patterns in the data? That will be possible purposes of any visualization. And according to those purposes you will use one kind of chart or a completely different kind of chart. For example, if you want people to compare things perhaps the bar chart will be the best way to do that not in every case but in most cases. If you want to show geographic patterns of distribution of a variable over a geographic area, right perhaps a map is better than a bar chart, right? So you need to think about what the purpose of the visualization is.

As I said choosing encoding and organize your data in a way that enable enables you to complete certain tasks. Those tasks would be compared see geography panels and so on and so forth. There are several resources that I usually recommend people to consult when they are going to visualize data. Resources that can help you decide which graphic form better suits your data and better suits your purpose. One of them is called the data visualization catalog datavizcatalogue.com. The dataviz catalog is basically a website that contains most graphic forms that have have been invented so far in data visualization. Each one of the icons of the little blue circles. Are you have over there is a button and you can click on them every time that you click on any of those buttons. You will get a pop-up window that will contain a little article that explains when and how to use a bar chart when and how to use a scatter plot when and how to use a data map and so on and so forth.

And then we have annkemery's website Ann is also a researcher a practitioner in data visualization who has a wonderful a web blog that I recommend that you visit. One of the sections in his website is called essentials. And the essential section of her website is very similar to the data visualization catalog. It's another data visualization catalog. It works the same way. You click on any of those buttons and you will get a little article in which Ann explains when and how to use a scatter plot a bar chart, etc. Etc.

Another resource is the visual vocabulary in which you can find online. The visual vocabulary is a huge poster in PDF format designed by the graphics department at the financial times. The financial times is

another organization that produces fantastic data visualization on a regular basis. So I would recommend that you follow them closely because they're charged are amazing. Awhile ago they decided to educate the entire newsroom into how data visualization works. And instead instead of just telling other people what data visualization is for. They decided to design this poster in which you have several columns, as you can see. Each one of these columns corresponds to a particular purpose. What it is that you want to show. Do you want to compare things? Do you want to see correlation between different kinds of things? Do you want to display the distribution of the data? Do you want to short change over time? Do you want to show spatial data? What it is that you want to show? And according to what you want to show you have several options. Each one of the uh, a graphic forms underneath below each one of the columns contains an icon representing the graphic form itself and then a little bit of a little bit of text that explains how and when to use each one of these graphics.

Data Visualization for Storytelling and Discovery: Module 3 video, part 2 - The Encoding Hierarchy

There is another resource that I use myself to choose graphic forms to encode data that I described in my book The Truthful Art. This is a hierarchy of methods of encoding it's a hierarchy that was devised originally by several statisticians. William Cleveland,. Robert McGill is a hierarchy that shows you how graphic forms are organized according to accuracy. I'm explain what what I mean. By explaining a little bit of the background of where this hierarchy come from comes from.

First of all, I'm going to show you the hierarchy so you can see it. All right. So this is what Cleveland McGill did. Back in the 80s Cleveland McGill did several experiments in which they encoded this the exact same data set in multiple ways. Imagine that we have numbers such as 20 40 60 and so on and so forth. They encoded those numbers as a bar graph. They included those numbers as a bubble chart. They encoded those numbers as rectangles of different colors. I'm making up the details but this is the gist of the story.

And they showed each one of these representations of the same data to different groups of people. To ask them for example, if this bar is representing 40, how much is this other bar representing right? You can put question mark in there. They asked people to make an estimate of the size of this bar in comparison to this bar. And I was like people were very good at that if you see this bar and you know that represents 40, you can probably estimate that these other bars close to 80, but then they did the same thing with bubbles. To another group of people they said, you know, this bubble is representing 40. How much is this other bubble representing? And here the estimate is a little bit harder. It is harder to see that this bubble over here is actually double the area of this other bubble over here. You can insert two bubbles of this area inside the area of the second circle. So people were less accurate in their estimates right. And then they represented the data also with color shade, right. If this color that is dark that is lighter represents 40. How much is this darker color representing, right and they discovered through this process and this experiments that it is possible to create a hierarchy of methods of encoding according to accuracy. The further up you go on this scale the more accurate the judgments that you can make based on this graph on these graphics. And the further down you go the less accurate the judgments that you can make.

We can extract several techniques to choose graphic forms from this hierarchy. The lesson that extract is that for example, whenever I want to represent my data very accurately and whenever I want people to see things very accurately to compare things very accurately or to see change over time very accurately and so on and so forth. I go up in this scale I will go to graphic forms that are built on to using methods of encodings such as length height position measure on common cartesian scales right such as for example the line chart, the bar chart, the scatter plot, the slope chart, and so on and so forth. All these graphics have in common that they are built on common scales and at the same time that they use methods of encoding such as position, length and height. These are the methods of encoding and the techniques of design that enable more accuracy. So you want you want to see very accurately it's better to go up in the scale. But that doesn't mean that the methods of encoding that are on the bottom half of this scale are useless right there. They can be really really useful in certain circumstances. Right? For example when you want to give just an overview of the data. When you don't care that much about the whether people can compare numbers very accurately. When we want to do just provide a big picture of the data. Methods of encoding such as color hue, color shade, area and so on and so forth may come in handy. I already show you one example of the use of these methods of encoding .I'm going to show you just another one.

This is a graphic that was published by the New York Times after the 2008 election in the United States. As you can see the data is represented through several maps some of them using bubbles sizes bubble areas. Some other maps using color hues and color shades right. Now why do they use these methods of encoding in this particular case? Methods of encoding that enable a lot of accuracy that belong to the bottom half of William Cleveland and Robert McGill's scale. Because the purpose of this visualization is not to enable you to compare and rank Miami versus Austin Texas vs. Los Angeles versus San Francisco versus New York. The purpose of this graphics is not accuracy. It is to provide a big picture of the data. To provide to give you the overview of the main patterns in the data. Democratic vote tends to concentrate on the coast. It tends to concentrate also in urban areas. Republican vote tends to be a little bit more dispersed over more rural areas and the interior of the country. That is the overview of the data.

So going back to the idea of purpose. If the purpose of your visualization is to show at a very very very broad level. The bird's eye view of the data if that is a purpose of your visualization and to show the overall patterns in the data these kinds of methods of encoding work really well. On the other hand if the purpose of your resolution were to lead people compare and rank county-by-county according to a Republican vote or two Democratic vote. The map will not be the right solution. You will need to use other kind of graphic forms such a table pair with a bar graph or something like that representing the percent of Republican vote or the percent of Democratic vote.

Sometimes particularly when you create interactive graphics. You can let the reader choose how to visualize the data. Show me the data on a map. Show me the data on a bar chart the map will show me the geographic patterns. The bar graph will let me compare the numbers more accurately. When you start using for example Flourish, which is the software tool that you are going to learn this week through the Practical videos, you will notice that a Flourish will let you visualize the same data several times and

in multiple ways. And you can combine those charts on those maps into stories, which you can navigate through buttons. You can click on next and see the data in another form and then click on next and see the data represented in a different way. So choice is another very important principle in data visualization when you have that luxury. In particularly when you're doing an interactive graphics.

Data Visualization for Storytelling and Discovery: Module 3 video, part 3 - Design for Understanding

When it comes to designing a good visualization for communication, there are more things that we need to consider other than the encoding of the data. Remember the elements of any visualization that explain in week number one, right? We have the scaffolding we have the frame. We also have the content the encoding. But then also we have the design itself of the visualization and also the annotations and we will get to those in just one minute.

For now. I would like to talk about how important it is to pay attention to design elements in your visualization. How important and relevant it is to think about how your data is organized for understanding. And how relevant it is also to think about how to declutter your graphic. How to remove features in your display that don't have any sort of purpose. Let me show you a couple of examples of what I mean just to illustrate this principle. This chart that you have over here coming from the European Commission shows a particular variable. It doesn't really matter what it is. Is the revenue to GDP ratio blah blah country-by-country. Each one of these lines represents one country in Africa, right? So we have the variation of this variable on each country in Africa. Year by year. From the year 2007 to the year 2014. And then also we have the average of all these countries which is this dark pink line or purple line down here that says sub-Saharan Africa that is the average value of all these countries.

Now there's nothing wrong in this graphic in terms of choosing encoding. This is a writing coding for these data, right? We are we want to show change over time. Usually a line chart a time series line chart is the best way to do that. The challenge over here for this chart I believe is not the encoding itself. It Is the way that things are arranged. The fact that all the lines are cramped into the same chart makes the chart hard to read because the lines overlap and obscure each other, right?

So when this happens we need to think about an alternative strategy. Usually what I do by the way in my own graphics is that first of all, I try to show everything on the same chart. If that is possible, that's the right way to go. But in many cases you have so much data that you try to cram it all into the single into a single chart. The chart will be hard to read. In those cases, it may be better to separate. Organize data in data in a different way. So an alternative design that I provided for this chart follows that principle. This chart that you have over here encodes exactly the same data that we had before. Each one of these panels that you have on the screen corresponds to one country. Each one of the panel's reproduces all the lines that we had before, right? So they repeat the same lines over and over again, but here we use color to emphasize some countries in certain cases. For example, the first panel is a panel for Mauritania. We have all the countries in the background greyed out. Then we have Mauritania a highlighted in green the line from highlighted in green. And then also we have a black line and that's the average for all the countries in the mix. And then we repeat the same pattern over and over again

and each one of the panels will highlight a different country Mozambique, Cape Verde, Senegal, Mali and so on and so forth. Each one of the panel shows you the data for one particular country. But we still show all the countries on every one of those panels it's only that we grade them out a little bit. So they stay in the background. This is a graphic that displays exactly the same data as the one before but it enables better understanding because it separates that a large amount of data and it chunks the data. Into smaller portions that are a little bit more digestible. It lets you focus your attention country by country and then see the patterns and trends on each one of these countries in comparison to the average line, which is the same on each one of these patterns.

Another principle or rule of thumb in data visualization is to try to show things as directly as possible. Let me try to explain what I mean. For that I'm going to show you another graphic design by the statistics that I mentioned previously William Cleveland, Robert McGill from another paper that they wrote together. So they were making these point. Let's imagine that you are comparing two different variables. The change of two different variables over time for example. Remember imagine that these lines are not labeled curve one on curve two imagine that these are for example, the unemployment rate in one country versus another country. Country one and Country two rate. So unemployment as you can see goes up and then it goes down later on right. So the point that Cleveland made over here is that if you care about showing the unemployment on each one of these countries separately then this is fine. You show one line for one country another line for another country. But what about if the purpose of your visualization is not to show each one of these each one of these countries separately. What about if the purposes that we sort of system is to show the difference between unemployment rate in one country versus another country. Then this graphic is not that effective because it forces you to squint a little bit. And use your fingers to estimate the distance between each one of these lines on each one point in time. This is a great example to illustrate this principle of show things as directly as possible according to the purpose of the graphic. If the purpose of the visualization is to show the difference between the two countries or the two curves. Plot the difference don't plot each one of the curves separately right. As you can see the difference between those two curves or those two countries in my made-up is example goes down or the different strengths a little bit between one country and another country.

Another principle of design in visualization is that we should always strive to declutter our graphics a little bit. Right? So the way I do this is that I think whenever I design a chart in a software tool such as Flourish the ones that you are going to learn this week or iNZight or many other tools that I use every single day. Is to take a look at the chart that comes up from the comes out from the software and said okay.

So let's take a look at all these features. Does this feature have any purpose? Does this particular feature of object in the chart a helps me understand the data better or makes the graphic beautiful more beautiful because in that case I will keep that element or that feature in the chart. But if a feature of element element in the chart has no purpose whatsoever. It doesn't make your graphic more beautiful. It doesn't make your graphic more understandable. Then that particular element in your chart has no purpose and you can safely remove it because it has no purpose whatsoever.

Let me show you an example. So these graphic over here shows you some sort of distribution of data, right? You have several bars that add up to 100% right? And then you have the components of that one hundred percent, right? So those little subdivision show you how that 100% divided up into several into several different portions, right? This is called by the way. I stacked bar graph. It shows how everything adds up to a total. Now, there's nothing wrong with the method of encoding over here. Nothing wrong with the way that the chart has been created itself. But there are things that we could improve design wise and in terms of decluttering the graphic. Just think about the many things here that we could safely remove without compromising the graphic at all. For example, the third dimension. There is not a reason why adding a third dimension of a graphic to a child like this if that third dimension is not encoding anything. It makes the graphic actually harder to understand. Because in order to estimate where each one of these bars sits in the background you need to mentally project the boundary of each segment on to the background in order to calculate where that segment is stands. So that makes the third dimension makes your heart your life harder. It makes the chart harder to decode and harder to read. Therefore in my opinion we can get rid of the third dimension in the chart. In general. I would say 99% of the time third dimensions 3D FXs In charts have no purpose whatsoever. And you can safely remove them. They just clutter chart unnecessarily.

Now another thing that we could remove or at least downplay a little bit or make make a little bit more a little bit less visible. Is grid lines. Grid lines in many software tools such as Excel or even Flourish or iNZight. Sometimes they are over emphasized. They have the same visual weight as the actual content of the data and that should not happen. I am a fan of grid lines. I like to include include gridlines in my graphics, but I try to use first of all as few grid lines as possible, right. And second of all I try to deemphasize them so they stay a little bit in the background. So the actual data pops out and the grid lines is saying the little bit of the background, visually speaking. So if you take a look at this chart for instance, you can compare it to the quick makeover that I made based on it in which I proceeded this way. First of all, I remove the third dimension to make all the bars flat. And that makes things easier to understand. I created a better visual hierarchy on the on the legend. So I used two different font styles right. One for the title and another one for the for the copy itself. And then I demphasize the gridlines by making them much thinner much lighter and also this continues these are dotted lines. They are not continuous lines that demphasizes the grid lines. We could also send the grid lines to the background, but I wanted to make the point that sometimes you want to be bring the grid lines to the foreground if that really matters to you. There are certain circumstances in which this may be appropriate. In this case it doesn't really matter if we could have sent the grid lines to the background as well and it will be equally fine. But at least they are demphasized. They. don't matter that much in comparison to the actual content of the graphic which is the subdivisions of the graphic.

Seeing these kind of before-and-after redesigns or makeovers can be highly educational. I have I have read a lot of them. I have learned a lot from other people's makeovers. If you want to take a look at more examples of these kinds of this kind of process. I would recommend that you visit our website call Makeover Monday. Makeover Monday is a website that is run by several people who use a software tool called Tableau. And also people who work for Tableau software itself .

And what they do every week is suppose a challenge to the community of people who follow this website. So they say well here's a data set right and try to come up with a good way of representing this data set and then they post the results of the of that challenge and that can be highly educational to see what solutions other people found to those challenges.

But in other cases they begin with an actual chart they pick up a chart. That for some reason they believe could use I could use a little bit of improvement, uh some changes and they made those changes. They make makeovers, right? That's a whole reason why this website is called Makeover Monday, right? So and they show you the before and the after here's how the chart looked like before . Here's how it looks now. And they explain the reasons for those for those changes. So this could be very very educational.

Data Visualization for Storytelling and Discovery: Module 3 video, part 4 - Words

When you begin practicing with Flourish, there is something that I would like you to pay attention to which is that visualisations are not just made of visuals: graphs, maps, charts, etc. Words matter! There's a reason why I said at the beginning of the course that I see at least three different components in any visualization. The framework the content and encodings, but then the annotations that we put in the graphic itself and this includes headlines, introductions, little pointers inside the charts or colored boxes and things like that. Those elements are really relevant in any data visualization that is intended to be shown or to educate the general public.
So please pay attention to those things. Don't just for example when you design your project, in Flourish or any other software tool don't just put random title on the chart. Put a title in there that really tells you what it is that you are about to see. So pay a lot of attention to the writing.
To give you some inspiration. I would like to show you several projects that I believe do a great job at combining words and visuals and making those words and visuals play together well. Interplay a quite well. So in all these projects words and visuals are perfectly integrated. They feed into each other. They strengthen each other.

The first project that I would like to show you comes from the Tampa Bay Times. The Tampa Bay Times is a is a newspaper here in Florida in the Tampa area. A while ago the Tampa Bay Times won a Pulitzer Prize. Thanks to a project called failure factories. And what I'm showing you right now is the introduction to that project that won the Pulitzer Prize in 2015. The title says why Pinellas County is the worst place in Florida to be black and go to public school? Well, the project is basically some data exploration that shows that Pinellas County in the past few years has become the schools in this County have become increasingly segregated. More and more racially segregated in the past few years. And in parallel to that process is certainly schools in this in this country have also become worse. Students have started performing worse and worse and worse. And those two lines go in parallel.

Now all the data for this project could be shown at once. I could show you everything at once in there with no annotations whatsoever. But what the designers of this project decided to do was to chunk the information and help you navigate the information little by little. So they don't show you everything at once. They show you a little bit of data on each step of the project and then they pair the data that is

being shown visually through the charts and maps with text that puts the data in context and points out what it is that you need to care about from that data. Hey, pay attention at this feature. Hey, take a look at this. Hey, this is what really matters? That's the role of text in our visualization.

I'm going to just show you the first step of this project. So why Pinellas County's the worst place to infrared to be black and go to public school and you have the byline there click or press continue. So the project Begins by saying 84 percent of Black Elementary School students in Pinellas are failing State exams, and they highlight Pinellas County on the map. Then you continue and give a little bit more of information. Another nugget of information.
Almost every other County does better. So now they compare Pinellas County to all other counties in Florida, right? Then you can continue. Says only 7 of Florida's 67 counties do worse. All our poor on rural places and they highlight on the map all those seven places. This kind of chunky this kind of chunking of information this kind of narrative structure can really help people I believe understand complex data sets a little bit better. And as you can see copy text plays a key role in data visualization for communication.

Another project that I would like to show you to illustrate the importance of pairing text and visuals comes from the New York Times. The opinion pages of the New York Times. Believe It or Not perhaps you're not aware of these by the new but the New York Times has one of the best graphics departments in the world. That create charts and maps for the newspaper itself. But recently they also hired one or two designers I believe to create charts and maps for the opinion pages. So right now some of the columnists who write for the opinion pages in The New York Times can illustrate their columns with charts and graphs designed by these people.
One of my favorite ones in the past few years is this opinion page title Our Broken Economy in One Simple Chart, which is a basically a column that explains the growing inequality in the United States in terms of income and in terms of wealth. And the chart that begins this Begins the the column basically tells you the entire the entire story. The x-axis is the income percentile people in different income percentiles in the population, very poor people. Versus very rich people over here. And then the method of encoding over here is position. Each one of these dots represent the income growth of people on each one of these percentiles of the population over the previous 34 years. So the further down a dot is the smaller the gains that these people in this percentile made in the past 34 years and the higher up a dot is on the chart represents a larger amount of income increasing the past 34 years. So basically the people who have gained the most in terms of income in 2014. Are the people who are high already high up on the on the income ladder? They pattern was completely different in the 80s in the 80s people who gained the most in terms of wealth and income per year where the people who were down on the income ladder and very rich people also made gains their income also their wealth. I don't remember this is wealth or income this actually income growth not wealth. So the people who observed the the lower growth in common term in terms of percent change are the rich people. So the patterns the lines basically cross each other. And then the story continues so we have the column itself has describes the pattern of the data and then you can see that if the designers inserted other charts that provide more detail. And there are related to the content of the column itself. So you when you read it, you will notice that the paragraphs of text and the charts basically speak to each other quite well, they are perfectly integrated seamlessly integrated. I believe that this is what you should try to accomplish to think not

only about the visuals, but also think about the text that you can write about the visuals that you're presenting to people.

One final example to talk about the importance of text is this data-driven project by The Washington Post, which also has an excellent Graphics department. So the stories titled Virginia's uneven recovery mirrors is growing political divide. They're basically telling you that. The more Virginia's economy grow more unequal. Right? Also, the political divide has increased and they proved that point they corroborate that point through different charts and different Maps. I'm not going to go into detail of this project. I would just encourage you to navigate it because again, I think that is a great example of how text and visuals can speak to each other and can interact with each other in a wonderful wonderful data project like this .This data map for example is something that you will learn how to do this week thanks to thanks to Flourish.
I stopped on this map in particular because I this week the Practical videos deal with how to use Flourish for data visualization and Flourish will let you design both these kind of data map and also different kinds of graphs and charts Scatter Plots line chart bar graphs Etc. So I would encourage you to take a look at the Practical videos this week. Don't forget to take a look at the other readings try to participate in the discussion forums complete the exercise this week and then post it in the forums for other people to comment on and also try to comment on other people's exercises providing constructive feedback. Congratulations on completing week 3, we are ready to move to week four, which is I believe the most exciting week of all because is the opportunity to create your own personal project.