[SQUEAKING] [RUSTLING] [CLICKING]

ESTHER DUFLO: So our last lecture, we decided to do a sort of a very short introduction to visualizing data because it's very important. And in fact, I don't think it's usually taught. We don't teach it formally to graduate students. We don't teach it formally to ourselves. In fact, my first caveat is I'm not really a data visualizer. I mean, not any more than the fact that Mr. Jordan doing pose because he's talking, right? We all do it as part of the scientific communication exercise, but I'm not trained as one.

And one thing that I realized in the process of preparing for this lecture is I surely don't think hard enough about graphical representation of my data. And yet, a lot of people do think about these issues right now. There are a lot of people who are thinking about data visualizations. There are a lot of resources on the internet. In fact, there are entire courses on data visualization. As I was preparing for an hour and a half, I couldn't figure out what they could be teaching for a whole semester, but that's a different story.

And there are many, many resources that are available on the internet, and I'll point you towards the end of this lecture-- during the lecture and towards the end of the lecture, I'll point you to the key things. So I'll try to point you some useful resources. There are two, three blogs that are really good. And in particular, there is a very good blog-- it's more than a blog. It's like a website where there is a lot of tutorials and examples of how to do it in R.

So the good news is that this is, I think it's a fairly recent area of excitement, how to visualize data. And therefore, a lot of the work is recent, and therefore a lot of the work gets incorporated into R. And so it's constantly getting better. So having learned R in this class, having learned how to open it and do program with it, you're in the right place because I think this is where most of the development will get done. And I'll give you some examples that are produced in R to prove my point.

So there are really two different goals of data visualization. And then within the second goal there are subgoals. One is for yourself. So maybe what you started doing in your project once you get a data set is plotting some things to see how the data might look like. So you would do a histogram, you'd do some non-parametric regression, you'd do some density estimation, and that's going to guide you for future analysis. And for that you might produce any number of graphs and look at it yourself.

And then there is another goal that's telling a story about the data and your results to communicate your results to an audience. That's what I want to talk about today. What I could see on data visualization course-- actually, I took one on Coursera, and then I realized this was all about the former. After I was well into it, I realized this was all about how to plot histograms and this kind of stuff. So there is sometimes a little bit of confusion on the two things.

But what I do want to talk about today is how we use data to tell a story to communicate your results and why it's important. So as I said, it's not necessarily emphasized in academic settings. You don't necessarily spend a lot of time on how to represent the data. In fact, we're spending just one lecture having spent a whole semester on how to produce the data, and yet, it's essential.

As Angela and Ludo can tell you, the communication of scientific results is critical to success in academic settings and non-academic settings. So of course, in non-academic settings I think it's plainly obvious that a lot of the work we do in meetings and consultants prepare presentation and how good your deck of presentation is essential because often that's the only output that really you're going to have with your clients.

But what people don't realize, or at least maybe I didn't realize till reasonably recently, partly because you just learn it as you go along, is that in academic settings it's essential as well, because the ability to present your result quickly and to present your result effectively orally and in written form is essential. The graphical representation of results is a key piece of that. And what is particularly useful is that tables usually need some work to go from what you're writing in the paper to what is going to be in the presentation.

You're going to have to reduce the size of your table because people cannot see all of the numbers at the same time. The text, of course, is for the written communication. The oral presentation, of course, is for the oral presentation. So are the slides. But the graphs, actually, usually good graphs will be the same or pretty much the same in the written form of the paper and in the presentation.

So your ability to produce this good graph is going to be central. So for example, in *Science* today, *Science* used to only publish very, very short articles. But the way they do it now is that they will allow you actually to submit longer articles, which will be online. But then what is going to be published in the print version is one page. And out of the one page, about one third of it is going to be taken by one graph. This is to emphasize how important that one graph will be.

So that's why it's important to spend some time on this and to try to-- what I try to do is boost from reading what people say about visualization and introspecting on what I would put in a graph myself. I try to give you a sense of some mistakes to avoid. So there is actually an emerging science of visualization with some cities and everything. But in fact, many of the principles we hear about are more like common sense than based on any kind of science.

Some of them are-- it's more pretend science than science. When you read this blog, they are pretty clear that sometimes we just don't really know. So what I'll talk about today and also within communication for other, there is another subdivision, which is are you writing a graph for the New York Times or are you writing a graph for scientific colleagues? And those are also different because the audience is slightly different and the level of sophistication of the audience is slightly different, what they want from the graph is slightly different.

I'll show you actually one graph, what happened to the graph when it moved from the scientific publication to The New York Times, and we can discuss on whether The New York Times version is superior or not. But for the most part, I'm going to talk today about publishing your results towards presenting in seminars and stuff like that. And in fact, it's not that different, but it's a little bit different maybe because aesthetic choices might be somewhat different. Yeah?

AUDIENCE: I also wanted to add that usually you also have discussions about generalizations within the scientific journals, whether we're accomplishing [INAUDIBLE]

--even in that one visualization. I've seen some journals that are a little bit more technical others, so even within--

ESTHER DUFLO: Yeah, even within there might also be a difference. What I would say, though, is that probably the objective of a good visualization should be that it should be at least acceptable to the range of scientific journals, because you should have-- there's no point to go-- it's not art, and at the same time it has to be readable and aesthetically pleasing, et cetera. Then, of course, there will be some differences from place to place.

Different journals have different codes, et cetera, that they can help you with anyway. But when you go to, say, present an academic seminar for the first time, no style figure specialist is going to help you to do your presentation, and there I'm going to try-- we're going to try and discuss what needs to go into these graphs and what needs not to go into this graph.

I think one of the-- there are really two guys who seem to be very good in this field. I mean, I'm sure there are very many, but there are two that I was very interested in reading their stuff. One is Robert Kosara. He has a blog called eagereyes.org. And one is Few. His name is Few. And his website is called Flowingdata. And I'll show you some evidence of both.

So Kosara has a lot of more very no nonsense discussion of what all of these things means and Few has a lot of tutorials on R and a lot of examples on R. Also both of them point to some examples, and of course, they have a great deal of examples of visualizations. So here's the definition of what visualization. It seems a bit obvious a priori, but it's good remembering what it is. So first of all, it's based on data.

So the objective of visualization communication of data, so it's transformed the invisible to the visible. A photo is not a visualization even if it's an informative photo. Then it should produce an image. That sounds obvious, but it's not necessarily the case. So a table is not visualization. It should be that the image is the prime-- when we're talking about visualization, it's like the image is the prime means of communication.

So there is some text that will go with it, obviously, but the text is supporting of the image. The result must be readable and recognizable, so we need to provide a clear way to learn something about data from the graph. That means the data, of course, will be less rich than what you have in the original data set, but some aspect of the data should be left. The data should be readable. The reader must be able to understand what's in the data without going back to the text.

So what comes out of this is that a graph needs to be self explanatory. That is something that people tend to forget very quickly. The graph doesn't come on its own. It's part of a paper or it's part of an oral presentation, so it's going to be discussed. But nonetheless, it should be that-- and that is true, by the way, for tables as well. That in scientific communication, someone should be able to take your tables and your graph and read them in isolation.

And the reason why it has to be the case is that someone will take your tables and graph and read them in isolation. So if they cannot read them, they are going to be irritated and I'm going to stop reading your paper. Likewise, when you present, some people, present company excepted, might actually lose focus for some time to what is being said, but the image stays with them. So the image has to be readable on its own. Of course, there will be additional material coming from the text, from the commentary on the image.

But the commentary on the image must not be essential to the understanding of the image and the image must also not lie on its own. So I'll give you some examples of data can lie. So we are trying to show the data. We're trying to show the actual data, not lie about it inadvertently or advertently. There is many, many examples that I'm not going to show you about how people lie with data, in particular on television and stuff like that. That's not what we are interested in. That's not the most important.

When we are producing-- so another thing is that in a paper, you're not going to have 150 graphs. At least you shouldn't. I'm an editor of a journal. I send back papers. When I send back promising papers to authors to review, I always say I don't want to see more than eight tables-- and this is economics, where papers much longer. I don't want to see more than eight tables, I don't want to see more than five figures. So in scientific journal, of course, it's going to reduce to two figures and one table. Yeah?

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Yes. I mean, people are more or less-- it's not a usually written in a style guide of the journal and editors are more or less stringent, but that's roughly what you should think about.

AUDIENCE: I was asking because I remember in a lot of electrical engineering journals, it's about six figures usually and then anything more than that.

ESTHER DUFLO: Yes. So this is same thing for economics. It's, like, I think five figures, 10 tables is what you should aim for. That means that you have to make a lot of choices about what the figures might be. So what are the figures doing for you? And in tables you can put a lot of numbers, but in figures, well, you can too, but generally then they will need to be disciplined, in some sense, for the data to be readable. So you have to be strategic about what the figure is going to do for you.

And in particular, it shouldn't replicate the tables. So it's not just like another way to see the data. In that case, if the table really replicates the figure, the tables can go in an appendix. If the figure really replicate a table that you want in the main text, then presumably this table is not needed. In particular, you don't need to put in papers figures that are just going to present. You have a table of means, say, in your comparison of means and your table.

You don't need another with three outcomes and treatment and control. You don't need another table that shows outcome, treatment, and control little bar charts. People are not idiots. They can read the table. So that's actually a very good-- this guy that I'll talk about, Peter Tufte's-- Edward Tufte, sorry. One of his big advice is your audience is intelligent. So generally keep that in mind. In whenever you're doing, your audience is intelligent.

That means that the graphs are not there to dumb down the tables to make it attractive or accessible to a dumb audience. Your graphs are there, any charts that you're putting out there to add meaning that is really not easily conveyable to an intelligent audience by tables, by way of tables. So what do we do with the graph? How do we choose the graphs?

We choose them so that they illustrate the story that we want to tell in the graph. So there is some amount of marketing in producing a graph. In tables as well and in text as well, but in graphs even more because the graphs are attractive and that's what people are going to go to first and that's what they might keep in mind. So there is a moment of-- an element of marketing, which has to be, of course, true to the science that's behind.

But what are you trying to tell with that graph? Are you trying to say that-- are you trying to show some basic pattern of the data? For example, some distributions that would be more difficult to show otherwise? Are you trying to imply that there is a causality between two variables? Are you trying to show interesting breakpoints? Are you trying to show that some regions are different than others, et cetera?

So you have to make your decisions about what is my graph trying to show before you start constructing the graph. So you both need to convince the reader and to convey information. That's why it's very different to these type of graphs which are meant to illustrate your stories are very different from the type of graphs you're going to draw for yourself. So a bad graph is one where you are drawing from yourself and you're saying, well, I'm just going to put it.

Typically that's not the graph you're wanting because writing a paper is not telling what happened to you in your summer vacation and how you progressively arrived at what the final form is. Writing a paper is the most efficient, concise way to reveal what you found at the end of the day. So what I want to do first is to go over-- there is an article by a guy called Schwabisch in the *Journal of Economic Perspective* a few years ago.

The reference is at the end of the lecture. It's called an Economist Guide to Visualizing Data. He's actually an economist in government, so it is important for him to-- he keeps visualizing data not only for academic audiences but also to show people, convince policymakers, et cetera. So he gives a number of examples of not so good graphs that he then changes. But before I go to him, here is one graph that Sarah sent me this morning as an example of a bad graph.

And it's pretty remarkable. It's a bit hard to see maybe for you, so it's about banned books. And it's pretty remarkable because it shows that-- so this is the example of the act superseding the substance because it really means nothing because you have the books that are-- in columns you have reason why, then title and writer, where banned and when banned, and for whatever reason these things are connected. And they flow as if they had any relationship with each other.

So it's drawn as in some kind of chronology, maybe. But in fact, these are different things. So it just really makes no sorts of sense whatsoever. Then the line goes all over the place. So this is an example. I think this was-- if I understand proper, this was an example of how infographic is great. So this is an example of how you can really go too far when privileging form over function. You can study it in detail. There is actually some interesting fact in that graph. It's just that you can--

AUDIENCE: It's hard to find.

ESTHER DUFLO: You cannot detect them very easily. They are very efficiently disguised in there. So here are much better graphs but still problematic ones. I love that Schwabisch gives us as examples and that we can maybe discuss a little bit, see what seems right or wrong about them before me giving you any of the general principles, and then we'll go back on how we fix them with the general principle.

So it's an interesting article because it gives you a-- I mean, many people do that in the different sides of giving you a bad graph and fixing it, but this is the type of graphs that you will actually see very frequently in our profession and that you might like to fix. So this is paper about-- and this is all from published paper. This is how the graph got published, so this is not outlandish. So this is a paper about a disability. What is wrong about it? Yeah?

AUDIENCE: For a lot of the graphs there's just a lot of white space at zero. It constrains your vision of the actual information. Just took out the [INAUDIBLE]

ESTHER DUFLO: Yeah, so on the one hand, we have that. On the other hand, sometimes the data is leaving the graph. So we don't actually show actually all of our data. So that's maybe one problem. Yeah?

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Yes. So this is not self-explanatory. You have no idea what this is about. And in particular, good luck to you. What's NC? North Carolina, maybe. But then A is like Alaska and Oregon, or? Probably not. So we could try and guess, but we have no idea what these things are. The labeling is poor.

AUDIENCE: The dot with two hyphens is a little strange at the very beginning. I don't know if these are extrapolations. You see the dot with the two hyphens?

ESTHER DUFLO: The dot?

AUDIENCE: The dot with the two minus signs at the beginning.

ESTHER DUFLO: Yeah, the tick marks in front of the-- the tick marks in front of the things are-- yeah. What other things that seem-- yeah?

AUDIENCE: The title place [INAUDIBLE]

ESTHER DUFLO: Yes. So the graphs kind of is-- there is that. And then also caseload is repeated four times. But sometimes we don't know what AO is, but caseload we really have a clear sense of what it is because we hear it many times. What else sounds a little bit not great? Yeah?

AUDIENCE: There are no tick marks on the x-axis but then there are those other dots that go along the zero axis. I don't know if they're marking half, one half year--

ESTHER DUFLO: Yeah. So the tick marks are-- there are these weird tick marks in the middle but no tick marks in the x-axis.

AUDIENCE: Is it plus 1.5, 2.5?

ESTHER DUFLO: Maybe that's what it's supposed to be. Yep?

AUDIENCE: Can I ask for your opinion for the lines going this way. I've had professors say they're really distracting and you should get rid of them, but I actually kind of like them. They kind of help me.

ESTHER DUFLO: OK. Well, those are called grid lines. Those are called grid lines. So it depends. So in this case, do you think these things are they doing something for you?

AUDIENCE: In this case not really, but I feel like sometimes they're helpful. I've had like at least one professor who was like, never put grid lines. They're distracting. So I feel like it's an opinion thing. I don't know. Unless you say now that it's not.

ESTHER DUFLO: I don't think there is a rule that you should never put grid lines or you should always put grid lines. I think the rule that maybe the principle that maybe we should get to is-- the question we should ask is are they helpful or are they distracting. Are they enhancing our understanding of the graphs or not? And so in this case, in this case, my instinctive reaction is that there really are very many grid lines. Maybe a bit too many.

And another problem is that, for whatever reason, the x-axis is the thicker line. The thickest line in this entire graph is the x-axis, which is not data. I mean, it's useful to know where the x-axis is, but why does it need to be the thickest thing that there is no information in there? I would like the information to be-- I mean, there is another x-axis but this is a zero line. So he might be interested in people crossing the zero line, but why is it the thicker one? It's not very clear.

AUDIENCE: One part that isn't entirely self-explanatory is years since what?

ESTHER DUFLO: Yes. So actually there is a title. I was a little bit unfair because there is a title to this graph which is somewhere which says something like impulse response. So it's basically after the shock, what happens to the famous caseload. So you would know that this is zero year after the moment of the shock and then what happens to the caseload several years after that.

AUDIENCE: Would it be OK if, in the context of the paper, we actually reference [INAUDIBLE]?

ESTHER DUFLO: They certainly do. So in that sense, if you read the paper and you read that, you will see what this AO is. But it violates what we just said at the beginning, that the graph has to be self-explanatory. So it would be OK-- I mean, nobody is going to go to jail for that graph. Remember, it was published. There is something that is good about it. One thing that is good about it is that it's four panels instead of trying to put everyone in the same graph, which would be very, very busy.

So it's a nice thing that it's four panels so we can actually see the difference between those things, whatever they are. Surely they are referring to these things so we know what AO and NC and W is in the text, but you cannot go back and refer to this. You cannot just open the paper, start by the end-- if it's a working paper the tables and graphs are usually at the end-- and try to understand what the paper is about from the graph. So it is not self-explanatory.

Generally I find it quite busy. There are lots of numbers that are always the same, so maybe we should find a way to not repeat the percentage sign. Do we really need to know every time that it's percentage? So here is another one. That's actually a very popular-- it used to be very popular when gross regression were very popular, although that's not one. Education and export of office machines. So what's the problem with this one?

AUDIENCE: Text is squished up.

ESTHER DUFLO: Yes. So you don't show the data because there is so much data that it's all on top of each other so you can't really see the point. Yeah?

AUDIENCE: It's funny that they have to add a label to point to China.

ESTHER DUFLO: Yes. Yes. You have to label China if they want to label China. Why do they do that, you think? Why do they point China in China?

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Yeah, so because in the text they are particularly interested in China. So one of the things they want to point out is that China is above that regression line. They tend to sell a lot of office machine compared to their years of schooling. Yeah.

AUDIENCE: Are colors taboo?

ESTHER DUFLO: So the colors is-- no, not at all. So colors are fine. You have to be careful because with colors for the following reasons. A lot of journals today have both colors and black and white in the sense that the electronic version and working paper is, of course, the same. The electronic version will be in color and the paper will be printed, but then when you print out it's going to be printed on black and white. So when you think about your graphs, colors are fine.

Absolutely, in fact, in all presentations you will see people using color effectively. But you have to think of the black and white version as well and thinking about how it's going to look in black and white. Sometimes you need-- you need the color to look good in black and white. So the color is not an issue, per se.

Yeah. So here you have to add a label of China because China-- because the people want to show that China is above the regression line. Well, if that's the point they were trying to make, it's not clear that we need to have every single-- do you know-- Unless you know all of your country code, it's not clear that's going to help you. OK, this one I'm guessing is Kazakhstan. Ukraine. But some of them I just don't know. What do I not know. M and G, I don't know what M and G is.

AUDIENCE: Mongolia?

ESTHER DUFLO: Say?

AUDIENCE: Mongolia?

ESTHER DUFLO: Mongolia, maybe. Yes.

AUDIENCE: So in this case, it's clear that adding all labels is very counterproductive. What could be another way of presenting this information where we don't add all labels but we still communicate what we're trying to?

ESTHER DUFLO: Yes, so I'll give you how-- I'll show you how he fixed it. Maybe I should have done it this way of first showing the problems. I want to go over the problems then give you a few principles and then see how he applies the principle to this. But clearly there's no point giving people all these codes that most of the people don't know the codes anyway, so they are not able to do anything about it.

By the way, most cross-country growth regression used to have all these codes, perhaps, because if you think that your audience is other people who run cross-country growth regression, then they know the codes. Maybe that's fine. But for the most of the people, knowing the codes are not very helpful. Then you could say, well, someone might be interested in looking over all this data, but then maybe there is another way to present that for someone who is interested in looking at all the data.

And in addition here, it's too many because even if you knew all the code by heart, there is this big cloud over there to the top where clearly even if you knew the code, there's nothing you can do with that. So it's just not at all helpful. On the other hand, you want to insist that there may be some of the data or the names there that you are interested in, that you want to make sure to-- this one we call the clutter graph.

Here is another one that has some issues. So this is about discounted expected lifetime earnings by different type of education. So what are the things that seems a little bit not great in that graph? You like it? I think it's a cool graph. Yeah?

AUDIENCE: There's two different metrics but they are plotted on the same graph, and somehow comparing GPA versus number of schools which will be [INAUDIBLE]

ESTHER DUFLO: So I think it's fine, actually. At least that I don't have a problem with that because it's saying this is the expected lifetime earnings for someone with no schools, with one year of school, with three years of schools. This is college, like finish-- no college, one year of college, two years of college, and then graduation but with different. I don't know if it's college or high school. Doesn't really matter, but it's different type of graduation with different GPAs. So maybe it's OK to put them.

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: No, I think you finished with four years. If you finish, you have done four years. You've done none, you've dropped out after one year, after three years, you've graduated with a 2, 3, or 3.75 GPA. So I think the actual data that he presents, there is no issue with that. Yeah?

AUDIENCE: I take issue with the colors inside the bar graphs. I think they're a little too pretty. I think maybe I understand they wanted to do a legend maybe to smush the lines together so that you could compare them more easily, but if it were me, I'd prefer breaking them out with the labels underneath.

AUDIENCE: [INAUDIBLE] the whole black and white and color thing, [INAUDIBLE] black and white, the color is really-- it's a pattern.

ESTHER DUFLO: I agree. I agree with you. This is dreadful. People tend to agree with that. So moiré, little dashes, et cetera inside bar charts are completely-- let's say recommended against. You can just not do-- don't do bar charts. Don't do patterns. It will look good. Bar charts can look perfectly well in black and white of different colors.

It's not clear to me that you really need the different colors as long as you-- you could do everything in one color as long as you separate them a little bit and you put-- because now it's a bit difficult because you have to go to the-- fortunately they are ranked, but otherwise you have to go from a legend to the bar.

AUDIENCE: The legend title, put it under--

ESTHER DUFLO: Take the legend and put it under the bar, and then people will immediately see what they need to know. Now they need to run from the legend to the coding and try to figure out, and then plus you have to deal with these kind of shades and little waves and stuff like that that are just hard to see. That's the first point.

This is completely gratuitous-- to have the legend on this particular graph to have the legend on the side rather than to have them under and to have all the bar smushed together and to have this kind of patterns not a good idea. These are all, by the way, default in Excel, so they are easy to find. So that's the first problem. What's another issue in this graph? It's actually something in this graph and like the other one that we're just maybe not aesthetically pleasing that is profoundly misleading. Yeah?

AUDIENCE: It doesn't start at zero. So you normally add a squiggly thing

ESTHER DUFLO: Yes. This one doesn't start at zero, and honestly, doing a bar chart, when you have a graph, a time series, et cetera, not starting at zero is fine because nothing is supposed about the bar, about the height of the point. So the axis in a graph that looks like this, in a scatter plot, you can start the axis wherever you want because you're not assuming that the height really means something. But with the bar chart, you really do.

Because basically the surface is used to represent the height, so the height means something here. You're implying something from the height. Therefore, that's pretty bad because this graph, this one, if you look at the actual data on the graph, the people who finished no schools make about-- let's say people who finish one year of school make about a little over 500 whatever it is. $500,000 over a lifetime, and the people who finish with a 3.75 GPA make over almost $100,000.

So that means these guys make about twice as much, which is a lot. But if you look at the graph, they look like they are making four times as much. So that is very misleading. If you're going to present the graph in a way that is supposed to illustrate something, you're actually presenting the graph in a way that people will be confused. They are going to come out of this thinking that the people who graduate college with 3.75 make five times more or four times more than people who didn't finish, which is incorrect. Go ahead.

AUDIENCE: I totally agree. What if in the cases where you have large numbers, and we want to highlight the differences between one another where, let's say it's thousands, and then the difference between 1 and another is 100. Would you recommend [INAUDIBLE] from zero and then maybe zooming in to visualizations?

ESTHER DUFLO: So I read about that like too because I had exactly these questions. People seem to debate about this point. So I think the bottom line, which seems reasonable to me, is maybe in that case the bar chart is not the right tool to use. It's just not the right item to use. In that case, if you want to start them from high up and you want to emphasize difference at the margin, then you want a line or something like that because with a line, again, you're not-- the implicit assumption is not that the height really means something, per se.

So you can emphasize the differences. And you can do something like Catherine talks about if you're worried that people might get mistaken. You can even do something make a squiggly line to start at zero and start at 10,000 to further emphasize it. But that's not even needed because in a scatterplot or a line, the range of the graph is going-- people are going to pay attention to it. In a bar chart, they're not. So I think bar chart do need to start at zero.

I don't think that there is very many exceptions to that. There are exceptions to everything, but I couldn't think of one clear exception to that. In that case, you're just going to use something different. But here I think it would be-- the bar chart might be nice, you just need to start it at zero. People are going to get a real visual impression of the actual differences between people's earnings and proportions, which are interesting.

But then it needs to start at zero. Bar charts are meant to emphasize proportion differences. So if you mislead them by starting at zero, you-- so there is a great very many examples of Fox News and stuff like that over the internet making great use of the not starting the bar chart at zero to emphasize some point. Yeah?

AUDIENCE: When you want to do something like this that's relative to other categories, can you take the lowest one as a baseline and do a percentage? So take everything with respect to finished no school and say that basically the grads with 3.75 [INAUDIBLE] also have a bar chart.

ESTHER DUFLO: Yeah, you could do that. You could do that in bars. You could present that in bars relative to, and then things would start at zero if that's what you want to emphasize. So here's another one, another type of bar chart that's also very popular. Very popular Excel graphic. Again, those are all published. These were all published. I think these are a little bit gone away. So the tri dimensional bar-- the 3D has to be avoided, I think, unless it's needed. I mean, sometimes it is needed because you really have three dimension, but three dimension for the sake of it?

AUDIENCE: That means nothing.

ESTHER DUFLO: That means nothing. It's just aesthetic purpose. You are laughing. I've been to many seminars where people have presented little bar. It's just at some point, people thought it would look good. I mean, so I think our view of implicitly or explicitly we're all already influenced by having improved on the graph so maybe we think that that is particularly unappealing. But yeah. So the tri dimension is not spectacular. It's not really readable. The blue and blue are-- if you want to emphasize the contrast between blue and blue, the blue and blue are maybe not the best choice either. Yeah.

AUDIENCE: Oh, text placement. So for example, the percentages. At 60% on the far right, it's six.

ESTHER DUFLO: Yeah, it's six.

AUDIENCE: Just shows the point. The missing grid line at four. The fact that experience is not capitalized for young. I mean, there's a lot of things.

ESTHER DUFLO: Oh, there was another one in this one which I wanted to notice, which is these are more grid points than data points. I mean, again, the grid point is going to help you read them. But if you are more grid point than data point, you have too many grid points. So that is something that you can also take for. This one also has many grid points compared to the data point, and then it also repeats them just in case people haven't gotten them because they got confused by the 3D.

AUDIENCE: Why are they on the same grid line?

ESTHER DUFLO: Yeah. So that one I don't know what happened to it. That one disappeared. I think that one was an error. It shouldn't have been there. It's also because of the 3D the grid line has to go-- you have to go a bit deeper, so you have to have your inside grid. The labels, like young, old, they're in the graph. That's nice. But they are sort of hidden by the grid lines. Here I would say the grid lines don't do very much for you, especially since he is repeating the graphs anyway.

He's repeating the numbers anyway. The grid lines are supposed to help you read the number. Here is the last one. I call this type spaghetti charts. This one is much less-- so this is the initial DI worker by major cause of disability over different calendar year. What's good and bad about this one? So there is a title, though, so you can see it's like-- it says initial DI worker awards by major cause of disability. I think we can infer that 32% say circulatory problem. In 1975 10% was cancer.

AUDIENCE: I think it's actually kind of nice that there's different colors and different shapes and you can tell-- unlike the one with all the countries, you can kind of follow which data is where. But when I look at it, I see no trend whatsoever. It's just like, what is happening?

ESTHER DUFLO: Yeah.

AUDIENCE: The histogram where you have it adding up to 100%. I don't know if that's a histogram, but then within each bar you can color which disease-- because it's a percent anyway. That would have just been visually a lot easier to read the graphs.

ESTHER DUFLO: Yes, so that would be-- so one problem with this graph is that, I think, exactly what you're saying. And then they are trying to counter it with the colors. One problem is that there are many theories. That's why they call it spaghetti graph. There are many, many theories, so there are many theories are competing for our attention.

And at the end of the day, it's not clear what story it's supposed to tell. So the text is probably going to help you and say, oh, the circulatory has really gone down and musculoskeletal has gone really gone up, especially after 1996. Maybe you want to make the case. I don't know what the paper is trying to say, but maybe you want to make the case that after mental-- and then you see the mental disability going up a lot.

Back problem and mental disability are the big sources that are going up, which are more difficult to rigorously assess, let's say, and this has led to an increase in the caseload. Let's say that's the story. That's quite plausible, the story they want to say. There's so many things going on in this graph that it's not clear that it's making the story extremely explicitly.

AUDIENCE: What about the colors? I'm not sure green and red is the best selection, especially for colorblindness.

ESTHER DUFLO: Yes. So about 10% of people are colorblind in the sense of different types of different type of colorblindness, so red and green are difficult to separate. Yeah?

AUDIENCE: So I'm not sure if this a bad thing. If you want to order the legend, for example, in the same order that you see them-- say, on the leftmost, the topmost.

ESTHER DUFLO: That's a good point. Yes. Here there is no order whatsoever. So one might order them by the end. Something like that might help a little bit.

AUDIENCE: Are the colors and shapes [INAUDIBLE] as well?

ESTHER DUFLO: Sorry?

AUDIENCE: Colors and shapes are [INAUDIBLE]

ESTHER DUFLO: Yeah. I think the idea might be that it will look better in black and white, but this will not look good in black and white because the color will all be gone. Even though we have the shapes, it's going to become confusing. So in any case, since it will not be good in black and white, you maybe just-- the shapes become redundant. Using all of these different shapes for the tick marks contribute to the graph looking very busy as well.

OK. So these are examples. We had fun going through them. But honestly, I'm sure you take any of my graphs or presentation, you could do the same thing as well. So let's see what people have to say. So the big Pope of this field or the founding father of this field is this person Edward Tufte has four very nicely produced books and he gives a lot of lectures on how to present things.

A lot of the blogs, et cetera, essentially rehashed some of his basic principles and they are a little bit-- I mean, honestly, I enjoyed reading the books, but it's basically his views on things. So everything has to be taken with some grain of salt or a lot of grain of salt. It's not based on very recent science or anything. But he has a particular aesthetic about how to present things and you might agree, disagree, et cetera.

There are some things that he says are quite-- in general, I think there is a lot of common sense in there. But I just want to say that anybody who is thinking about this field has at least read this one book and then people have maybe moved on from this, or at least would be the same for you. You can read this one book and then feel free to move on from this and add your own thinking and the thinking of others. So his principles were reasonable. Show the data.

In particular, in a way that people can read it. And then maximize data ink ratio. So what this means is how does that happen? By erasing the non-data ink as much as possible. So this means grid lines, et cetera. So your professor's view on grid lines' probably coming from there. Grid lined, hash, et cetera. Erase redundant data. So for example, we've seen a lot of percentages in the different graphs. He would try to figure out such that you only write percentage once and it's good for the-- avoid what you call chart junk.

So all the moiré thing. He hates the hash-- these type of things. And what you call ducks, which are things to look good in graph but have no use. So in particular, this type of stuff. Three dimensional when it's not needed, et cetera. So his view-- and I think people might disagree on that-- is that you want to have as little non-data ink as possible and you want to have a lot of data ink. So his view is that graphs should be dense, very dense, has lots and lots of information.

These are different views. Right? These are different points. You could do the first thing, try to make the graph as economical as possible in terms of the non-data ink, and you could-- on the other hand, you could try not to put too much data in it. So his view would be that you remove the data ink and then add a lot of-- remove the non-data ink and add a lot of data ink. Make the graph very, very dense.

Something that actually makes a lot of sense, I didn't realize, but is the graphic tends to be horizontal. It's easier for us to think when graphic tends to be-- graphic, instead of being vertical, should tend to go to the horizontal. So the scale of graphic should-- in general, things should be spread out.

In looking at other things after that I got-- so all of these principles when applied leads him to some examples. Before I go to the examples, I want to say one other thing that is-- I think that you pointed out, a word that Leo said that is useful is conventional is also useful because-- and here it has to be a little bit field specific because the conventions are a little bit field specific. But people are coming at your graph with an expectation of what graph tends to look like.

So for example, with the bar chart, that's something that we are just hardwired to think of. We're going to want to compare the height of the bars if we see bars, therefore if you don't put them at zero, you're going to mislead the person. But this principle is more general than that, which is we are all coming at graph with conventions, and if the graphs are non-conventional, that's going to take effort to your reader. So the extreme version of that was, of course, the book example that we started with, which is highly non-conventional.

But in general, in looking at aesthetics versus information you have to think about whether it's going to look very odd for someone. So a lot of Tufte's principles-- and there, there is network externalities in the sense that if nobody adopts something, even if it looked great, if nobody has adopted it, you don't really want to be the first one to stick your head out because that will create an issue in other people doing it.

So it's like the QWERTY keyboard is inferior but since everybody uses it, the first person who introduces a better keyboard has no chance and we still are stuck with the QWERTY keyboard. That's a little bit the same with graph, and I think it's worth remembering when we look at these examples because not very many of them has actually caught on in this particular form and shape, and therefore, it's not clear that you should really use them. Yeah?

AUDIENCE: A lot of these points and things we've commented on in the past graphs, like aesthetic or clarity kind of things, but I mean, the bigger question is what type of graph should we be using. Does Tufte talk about that?

ESTHER DUFLO: Not really because it has become field specific, but I'm going to come to that. I'm going to go back to some of the empirical techniques we're looking at and then tell you, this is what you need to show for all of this. So let's look at what this leads him to. So here's one graph he likes for the scatter points. Take something like that. This is how we would plot it, plot one of the four.

Basically he has erased the axis because who needs the horizontal line? He has erased the vertical line. Who needs the vertical line? The points are nice and clearly labeled with the graph going between them. Here is your thing on grid lines. They are not going away, but he has only the two that people would think are particularly pertinent. So that would be one example of his minimal plot line. Again, for example with the vertical and putting the axis or not, conventionally we tend to put them.

Here I don't think it's a big deal, but you might want to think about it. Incidentally, this was done in R. So someone has taken the trouble to code all of his nice graph-- they all come from this book, in the last chapter of his book where he takes everything together to propose a graph. Someone has coded them and they're all on the web. So it's also good to-- a finger exercise for you with R.

In fact, this person has plotted them in ggplot, which is one of R graphic package but also in R base, which is just the base, and in lattice, which is another graphic package. So we're making the point that you can do pretty much anything you want in any of these three packages. So I have the reference for this person's website at the end. So that's one idea, minimum plot line. I'm not showing this saying that you have to go through them. I'm saying then the extreme application of the principles would lead him to these kind of ideas. And then we can revert back for type of graphs we would actually show.

AUDIENCE: [INAUDIBLE] --on top of the last data point on the right?

ESTHER DUFLO: 5%. So I don't know what that 5% is relevant, but there is something that is trying to be said in the text or something like that.

AUDIENCE: 5% drop.

ESTHER DUFLO: 5% drop between-- yeah, exactly. I don't know why it is relevant, why he wants to attract our attention for that. Maybe it's the biggest drop that we observe in the data from year to year.

AUDIENCE: Could be the 5%-- 5% of 100 is 20, so--

[INTERPOSING VOICES]

ESTHER DUFLO: The width of this-- the 5% is the width of this band and it's also the difference between the last two points. So that's telling us that the biggest drop in any two points is the last one, but a lot of the other points-- a lot of the other fluctuation since 1970s is between those two. So in that sense, it's nice that there is a lot of information in this graph in a way that, of course, needs to complement the text but that is readable.

That's a kind of a funny one. That's one that has really not caught up. It's about the scatter plot. He's saying, well, suppose you want to show a scatter plot for your data. Maybe you would want a regression line somewhere, but you show your data on miles of gas per gallon of fuel against the weight of the car.

OK, fine. You show your plot. So this is really about what should you do with the x-axis range. So here you just got rid of it completely. But here you're saying, well, when you have those axis, they are sort of wasted, so let's do something useful with them. So the first thing he proposes is to say let's indicate the range of the data. So the axes don't go to all the way to zero here or don't go to all the way to the origin.

They start at 1.5, which is a minimum value in the data set, and they finish at 5.4, which is the maximum value in the data set. Therefore, you're not wasted any non-data ink because you actually have information in there. Moreover, here you see it's a little bit elevated. And so this is actually the interquartile range, a little bit above, and this is the median and the mean.

You've not wasted any non-data ink. You have some representation of the-- the axis is used to present that you have 25%, 75%-- 25% of the data is there, 75% is on this end, et cetera. This has not caught on. This is a cute idea but I don't know if it's this type of stuff that would be particularly informative until it catches on. Here's another version of that, which is the dot dash, which also has not caught on which is a more extreme version of it.

Which is here you're representing more percentiles of the data as you go along. So again, you have the minimum here and the maximum here, and you present-- and the median here, and different percentiles at different points. So that's another one that I don't think has caught on, but the application of this principle. So you're not wasting any information. You have all the information-- you have a lot of information on the distribution of the count weight.

A lot of information of the distribution of the fuel, and then some information on how these two distributions relate to each other from the scatter plot. Here's one that I actually quite like. It's a redrafting of the box plot. So you're familiar with the box plot, which is usually like this and then like this and then like this. Where we represent the median and then the interquartile range and then sum the 95% confidence interval here.

So he says, well, why do we need the box? Kind of wasted. This is symmetrical so I don't get any more information from this side of it than this side of it, so let's redefine that better. We are going to put a point for the median. Then instead of putting-- we are raising the box entirely and we are now putting around from here to 95% we have it here. So it's sort of-- I like that one.

It's not caught on either, but it's nice, and it's readable, and it's clear so you can describe it. And then the advantage of that is you could have-- when you have very many of those box plots in a sequence they look a little bit busy, so here you can actually have-- you can show several. So this shows the number of station. This is not a time series. This shows the number of station reporting the magnitude of Fiji earthquake for different values. Yeah?

AUDIENCE: How come the last two dots don't have any lines attached to them?

ESTHER DUFLO: These one? Yeah, he's not representing that data, I think. He's just representing the 95% confidence interval.

AUDIENCE: Like I don't even know where the-- because there's only a dot for the 6.1 and 6.0.

ESTHER DUFLO: Why are they not-- I think it's because there is only one data point.

AUDIENCE: Oh.

ESTHER DUFLO: Yeah. In that. So this is average over the years so there's only one data point in that year. And that's one that I think is quite nice. Also answering your bar chart. Your bar chart-- your grid line example. A clean way of having grid lines in a bar chart is to put them inside-- just make them inside the bars. So this is the average score on a negative emotion trait from about 4,000 participants.

You also have a good description of the sources of the data, et cetera. Then you have the labels right below, not on the side with a different legend. This grid's very nicely in black and white of course. You have something that doesn't like axes so the axes are not there, but the numbers are from one to three on the x-axis, and you have the thin grid line to help you through the bars.

That's the vertical version. And the horizontal version, which can also be useful, depending on what you're trying to represent. Again, with the idea that graphs are better going this way than that way. If you have some long lines, here you have a difference between a very small number and a very large number. Maybe it will look better horizontally.

So that's the big highlight of Tufte's book in terms of what, at the end of the day, practically is proposed. Some more chart rules from the Flowingdata person. Few. Not few rules. That's his name is Few. Bar charts must start at zero. So this field loves to talk about the pie chart. So Tufte hates the pie charts because we are not very good at comparing surfaces and angles so it's very difficult to actually compare slices of a pie.

So there is all debates about whether you should have pie charts or not. I think in scientific article we just don't use them. They look cheap. So that's why I didn't discuss more, but if you have to have them, you should not have too many slices. There's a very nice example that floats around the internet in Wikipedia on the number of something about countries where they have one pie chart per country-- one slice per country. That's really too many. You can't really see the difference.

Show the data. So that means through the data in as uncluttered a way as possible. It means that, for example, if you have too many points cluttered together, you can't see them, that's not helpful. So you have to think about, should I aggregate? Should I bin the data to see it more easily? Should I run, say, a nonparametric regression line instead of showing all the points because that's what summarize the data effectively? Should I maybe separate the data into different graphs?

So this is a small multiple idea that we had at the first graph. They didn't have all the lines together. They had different lines. And then the chart must be self explanatory. That I could not repeat enough. 9 out of 10 graphs are not self explanatory that I see in journals. So you have to explain the encoding what means what. You have to label the axis, include the units, include the sources, and circle-- this one is circle size by area not diameter. If you have round to indicate populations, it has to be sized by areas not diameters because diameters is misleading.

OK. Let's apply those principles to the graph we saw. So this combination of these ideas and these ideas, and some common sense in general. So this is the original line chart. That is redrafting. So here I decided that the grid lines are still helpful. The zero line is still different because the zero line as a threshold doesn't go up and down. We want to know so we keep that.

Contrary to what you said, we have actually increased a little bit the range because we want to be able to show all of the data. We've kept the scale the same because we want to compare across charts. That means there is a lot of white space. He likes the grid line so he kept them, but they are a little bit paler. And then the thickest line is actually the line that shows the data you want to show.

The title is now there at the top. That is not necessarily always a choice that you have when you actually publish, but in presentations it makes more sense to have the title at the top than at the bottom where we put it frequently because the title is actually-- because we tend to read from top to bottom, left to right. It does make sense to have the title at the top. What else?

AO has become adult only, no cash, welfare cash assistance, SSI cash assistance. We remove caseload because caseload is in the title. We explain that it's percentage change so we don't need to write all of the percentages. And you've got the same data but readable. OK? Do you like this one? I think it's still too many grid lines for my taste, but otherwise it's--

AUDIENCE: If I just look at that side, I look at [INAUDIBLE]

ESTHER DUFLO: Yeah, the grid lines, they sort of mix up the mix up the two graphs. So either we put a bit more white space or maybe we remove some of the grid lines. It's not clear we needed negative grid lines. Anyway, maybe we could have done a Tufte thing where we just put the relevant grid lines. And you can note something that is going to be useful in a lot of cases when you want to present coefficient in data analysis, in scientific papers you typically want the standard error as well.

But you don't want them with the same importance because you want people to emphasize. So here we have the standard error a little thinner. Alternatively, you could use dash, dash, dash, or something. Here's a very uncluttered graph find, and this is what it becomes. So this is nothing very tricky. But basically you remove all the points that he doesn't want to pay us to pay attention to.

There are four points that are-- five points that are mentioned in the paper as being above the regression line for a reason or another. China, Thailand, Costa Rica, Malaysia, Philippines, so he puts them in. He doesn't put the three letter codes. He writes them out fully. And now we can see how the data we get from the graph, that generally the number of machine exported goes up with education, but there are some outliers.

China, Thailand, Costa Rica, Malaysia, Philippines are those outliers that we want to discuss something about. They are not outliers, but they are above the regression line anyways. So it's emphasized with the different color for the graph and the labels. So now we have the data in a clearer way. This is our friend, the graph, the moiré graph. Here's how he moved it. We start at zero. So now we see that it's not-- we see the doubling and not the multiplied by five.

Because there is quite a width in this graph, it is quite a difference between no school. We need the zeros, and these guys do spend a lot of money. It would be a very high graph, so he moved it. He moved it sideways. And the moiré has disappeared. In fact, the color was doing nothing to this graph because once it's properly labeled you can see it very clearly. The color is quite strong.

The grid line, I think, helps a lot here, so we definitely would want to keep them because it does help to see that people are above 5,000, et cetera. So this is a case where I think I would keep the grid lines. Maybe you don't need to put all of them following the Tufte example of insisting on some of the grid line. I don't think it would be shocking to have only some of them. But they are nice and not very thick, so they are not encumbering our reading.

This one mainly needed some flattening out. That's what it looks like. I think it has still probably too many tick marks on the side because I think we-- I think the rule that there shouldn't be more tick marks than data point is a good one. Here we have probably too many, especially given that he is adding the things anyway too. But now, so it's flattened out because it was doing nothing in particular. The legend is nice.

That's something that, instead of putting in the kind of default Excel box, the legend is now indicated on the first graph. It's more directly visible, and the color helps-- the color is quite contrasted so we can move from there to there easily. And the zero line, which is relevant here, because we compare everything to zero, whether the young makes more money than-- whether the change in weekly wage has been positive or negative, that's kind of relevant, so the zero line is emphasized and we see whether the bars are going down or up from there.

I think there's one very nice idea of the spaghetti charts. It's the idea of there is just too much-- it's not that any choices was bad. They sort of did what they could with what they were given, but there are just too many lines. Here is how we craft it. So it's kind of a radical choice because it's not a Tufte choice because you can see that he kind of replicated the same data four times. You would see where it's very redundant, there is a lot of redundant data in here.

So just to say that these are guidelines, but that's not sufficient because actually it's the same graph four times, but with emphasis on what you want to see. And so now you can really see that the circulatory as the source of caseload, circulatory reasons, being able to work has really gone down, and mental, musculoskeletal have gone up, have taken the bulk of it, with cancer also increasing slightly. If that's the first graph in the paper and it's all about it's very difficult to assess people from mental disability, we've learned something.

So one of the rules that the FTC keeps revising, keeps editing and revising, that's sort of a rule of everything. I think every writing, but certainly scientific writing, you need to edit and then edit again and then leave it for a week and then edit again. And of course, that applies to the graph. This is how you're going to learn how to get better. I think it's a good idea to go over this list a little bit between the Tufte's principle and the other principle.

I showed you basically had no clutter, show the data, don't be misleading, self-explanatory as, did I tick these boxes? Do I tick these boxes? And to finish by giving you some examples coming from economics. That's a very recent paper that data has a huge amount of play in blogs and articles, et cetera, partly because what it has to say is very important, but also partly because all the stories can be told and pictured.

This is a work by David Cutler, Raj Chetty, and a bunch of other people on life expectancy at 40 by income. And the reason why they can do it, just the background of reason why they can do it is because the IRS-- it turns out when you file your taxes you have to tell the IRS whether you're dead, because when you're dead you still need to file your taxes once but then after you're dead you don't need to do it anymore.

So that's actually a pretty useful box in the IRS form that tells you whether this person is dead or not. So they use that because they have access to all of the IRS MicroProfile. They use that to look in different places in the country and then different income level, what is people life expectancies. So this is the geographic variation.

So here we're using the map. So there is also a lot of debates on maps, whether maps are useful, not useful. A thing that is very good to know is that there are maps in R. So if you want to, you can-- if you have data that geographic by county or by state, et cetera, and you want to put them in R, you can.

People sort of disagree on whether maps are good or not because big territories are given a lot of importance visually, but then they might not be that important in practice. There are also maps that-- particular maps of the US where each state is actually a same square, so they're all the same size. Those have become popular.

In 538 you see a lot of those maps of the US where things are geographically oriented so we are learning something from the fact that states are the East Coast states or the South states, but nothing is said about the size. Schwabisch on his blog has as a routine to do that in Excel, actually. Nice little maps of the US with square representing the states. Anyway, this is the graph here, and you can see pretty clearly that people in the Eastern Coast states live longer than-- in fact, on both coasts lives longer than people right in the middle. Yeah?

AUDIENCE: I have two questions. First, so life expectancy at age 40, is that estimated once they hit 40? What does it mean? I know what life expectancy is, but at age 40?

ESTHER DUFLO: At age 40 is exactly moving forward. Once you reach 40, moving forward.

AUDIENCE: How would you get that from knowing what [INAUDIBLE] take the expectation?

ESTHER DUFLO: You move up. So you see someone at 40, you're looking for him when he was 41, 42, 43, et cetera.

AUDIENCE: I just like looking at Florida because [INAUDIBLE]. The expectancy there is really high.

ESTHER DUFLO: So this is people who were in Florida at age 40 and looking them up, finding them up. It's a bit more complicated than that. I don't particularly want to go into the detail of what they're doing, but that's the idea is you want to-- the question you're asking yourself is if I observe someone at 40 the next year, do I still see him at 41, et cetera, and what's your probability of still being there and then recalculating the age from there. So that's one graph they are doing. Another graph they are doing is by income decile.

So they're looking for people-- there are so many people that for each decile they can-- actually, centile, they still have enough people to look at them and say, what's the life expectancy? That's, of course, the big headline graph which is that rich people live longer. But what is striking in the graph is how much. So that's the graph as it got published in JAMA.

So you see that inside the graph they also put in some actually textual information to put more richness to the data. But the key point is that someone who had the 1%, bottom 1%, their life expectancy at 40 is 78.8 years. Someone in the top 1%, their life expectancy at 40 is 88.9. So here, in a sense, it's like we want to-- so it doesn't start at-- it doesn't start at zero, but this is fine because this is dots. You want to emphasize the big vertical differences. We knew that rich people live longer.

Maybe we didn't know it was type of linear in percentile and we may not have known that it was 10 years. Interestingly, here you were asking. So choices matter because it's linear in percentile but not in income, because the person-- the difference in income at different percentile keep growing. In fact, if you draw it in income, it's much flatter. It flattens out. But in percentile it looks linear. So that's a choice of visualization that is pretty important as well.

AUDIENCE: What's a hint that [INAUDIBLE]?

ESTHER DUFLO: Between one or the other?

AUDIENCE: Like just outright values and percentages.

ESTHER DUFLO: It depends what you want to emphasize. I think here they probably felt that the linearity in percentile is quite striking and cool, so that's what they want to show. This is how the New York Times ran their data article, that graph. So you see there is much more data and information in this one, and this one is just more aesthetically pleasing, maybe. But we are lacking a lot of the data that we might want to know. There is no confidence interval anymore.

Maybe it looks more accessible. And then this is sort of combining the two. That's another interesting choice that you might want to have is to say, well, we had the geographic information that I first showed you and we have the income information. How do the two map together? And here you have to make some choices, because you're not going to be able to combine this entire map county by county and this income variation.

So here you have to make a choice and say, what should I show that is going to illustrate my case? So they pick-- how many? Four cities. They pick four cities, New York, San Francisco, Dallas, and Detroit to make the point. And what is the key point is that the big differences that we see territorially between places where people live longer and places where people live less long, it's actually really, for the poorest people that we are seeing at the top, the bottom 10% of the population, basically. And after that, for the rich it doesn't really matter where you live anymore.

So here, again, it's like there is no hard and fast rule that tells you you have to present, how many lines you have to present, but it's showing four lines that are first very ventilated and then will get together as kind of a way of illustrating this point that, for the poor, the quality of-- maybe for the poor there might be a difference between the quality of the infrastructure, your public health, hospital. We don't know what it that is underlying that, and the rich are able to-- as you become richer, you're able to compensate with whatever-- maybe move for better hospitals or something like that, or find a better health care within the system.

So here you have to-- because this is a choice, it's not a substitute for the actual analysis that you would want to do something. Even at the purely descriptive level you would want to do something more systematic. But it's going to be-- this is really the point of conveying a story and illustrating it quite differently from actually stating the point. That's kind of an economical way of representing the combination of the geographic data and the income data that is then going to be represented more systematically in the regression.

I didn't get to what I promised to you I would do, but you can look at it in the slides, which is for all of the methods that I talked about, RCT, difference in difference, regression discontinuity, and IV, I showed you example of what you'd like to show in all of these cases. I was pretty convinced that I would not have enough material to for this entire lecture, so I went too slowly. So we are done teaching this class, so thank you very much.

[APPLAUSE]

It's been a pretty great ride, at least for Sarah and I. So we hope you liked it and don't forget to file evaluation, and good luck to all of you in what comes in the future.