[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: So this is 14.31 or 14.310, Data Analysis for Social Scientists. I am Esther Duflo. I won't say any more for the moment because Sara will introduce all of us later.

So there is a lot of data out there that's almost too obvious to be said. There is data sources that are put together by various organizations. The World Bank is one of them. The Census is one of them. J-PAL, the lab that I head, is one of them. We have lots of data. Almost all of the journals now, academic journals, require or at least encourage people to post their data for public access.

There's, of course, lots and lots of data that exists on the web in the wild and can be harvested for use. There is data from Amazon, from Facebook, from LinkedIn, et cetera. There is lots and lots of data out there. And of course, there is always the option to go and collect your own by interviewing people. And it's actually becoming easier and easier to do that well.

So once we have all this data, is it actually useful? Is it actually data that can be harnessed to say something interesting about the world? Or is just like a bunch of data that journalists put out there in some way that is useful or not?

And of course, given what I do for a living, I would say that data is lots of good things, besides plentiful. First of all, it is often beautiful. I'll show you just one example. And we'll see a lot of other examples of beautiful data in this class, data that really lends itself very nicely to beautiful displays, network data.

So the data I'm just going to show you has been harvested, so to speak, by a researcher called Kimo Quaintance. He lives in Italy, and he interviewed some Somali people living in Italy, some of them very recently arriving, some of them settled in Italy for hundreds of years. And he asked them for permission to harvest their Facebook network data, something you can do with a tool called Netvizz, and then put them-- and then visualized them using another tool called Gephi.

And this is how it looks like. This is how it looks like when put together, the social network of these guys. So each round is what in social network parlance is referred to as a node. And this is just a person. Each little ray is a link between this person and their contacts on Facebook. So this is not some-- this is someone who is identified as a contact.

In the text that describes this picture, he mentions that there are actually many, many links, many more among the Somalis than among other people. And he says that in part because Somalis have exchanged Facebook, like each other on Facebook almost as a way of exchanging business cards. That it's actually more standard to exchange-- to befriend someone on Facebook than give your telephone number.

And the size of the bubbles is proportional to the importance of the person in the social network. So here you can immediately ask-- so there is already a lot of data here. It's not just a nice graph. You can see, for example, a person here on the top right to be quite an important person.

And here you might ask immediately, well, what does important mean? How do we define even someone who is important in a social network? What would you think about? Oh, yeah.

AUDIENCE: I'd think about the degree of the node.

ESTHER DUFLO: It could be the degree of the node, which in plain English would be?

AUDIENCE: The number of connections that they have.

ESTHER DUFLO: The number of connections they have. So it could be one way to do it, so how many friends someone has. Of course, it could be that someone has many friends, but those friends themselves don't have very many friends.

Or it could be that someone doesn't have all that many friends. For example, the guy over there in blue, in the bottom left corner, he's important, and yet, he doesn't seem to have all that many friends. So why could he be important in a network sense, despite the fact that he doesn't have many friends?

AUDIENCE: It could be that he's connected to more important hubs.

ESTHER DUFLO: Exactly. He could be important because he's friends with someone who is important. So maybe you're thinking that we are in danger of becoming quite circular over here. If I'm important because I have friends that are important, and of course, we are in a sense. But in a reasoned way.

If I say that I'm important because my friends are important, then it leads us to a somewhat recursive notion of importance in a network, which is very similar to how Facebook, how Google used to rank pages in the early history of Google. I don't think it does anything like that anymore. Now it's much more sophisticated, which is basically you were first on the Google rank if you were linked by pages that were themselves linked by many people.

And this is the same notion of importance in a network. It's referred to as an eigenvector centrality, which is basically this recursive notion of network, which corresponds to a nice mathematical object we can compute. And we see exactly this notion of recursivity.

So you can see that just in this beautiful picture, we actually have lots of data. Another thing we can infer from the picture is how different parts of the networks seem to be connected together. So there is-- which he visualized by saying, well, the light blue guys are in one corner, and then there are the purple guys.

Here's another [INAUDIBLE] zooming into that, into a part of the network, where he shows us what are the nature of the links that people are between with each other. So there are people who met because they're from the same hometown. There are people who took a class together. There are people who went to boarding school together. There are people who are working together.

And we can also visualize that with the color and order. The data also put the nodes somewhere in the network such that these things kind of are nicely separated. So that's one example of very beautiful data.

Data is not just beautiful, but properly used and coaxed into something, it's actually potentially insightful. It can tell us very nice and important stories. And I'll show you one of them, which also has a nice visualization, by the way, but of a very different type. And this is on pollution in China. And the question is whether pollution kills when there's too much of it.

And there's a lot of pollution in China, and in other places in the world, for example, in India. And when I say a lot, it's a lot, like the degree of the particular PM2.5 and PM10, which are these very, very fine particles that goes into your bloodstream, actually, when you breathe them, the concentration of those particulate matters are just absolutely off the charts in some Indian cities and some Chinese cities and some Pakistani cities compared to anything that we would consider dimly acceptable in the West.

And the question is, does it matter? And this is very, very nice work by Michael Greenstone and his co-authors, looking at one particular natural experiment in China. And this is the visualization in a geographic map.

The black line over here is a river. And the colors indicate the degree of PM10 everywhere. And it's a heatmap, which indicates how much-- what is the concentration of PM10 in micrograms per cubic meters in each of the pixels that we see on the data.

So another nice visualization of your data. And you can immediately see that this black line seems to be maybe relevant in some way. Why did they draw this black line? Because the policy that was given by the government was very different on the north and the south of the black line. On the north of the black line, heating was subsidized, and heating via coal-- coal heating was subsidized. And on the south, it wasn't.

So now we can see that, clearly, something seems to be going on on the north versus the south. But we might be worried, of course, if we were just looking at the entire north and say, well, it's quite red. There is a lot of pollution over there. And the South, it's quite green. You say, well, maybe there are other things that vary from north to south.

So maybe we wouldn't be entirely satisfied, but what we can do is zoom in. So we could zoom in geographically. But I'm now going to represent a different visualization of the data, where I'm zooming in in the two-dimensional plane of data, where I'm looking at the green north of the Huai River boundary. So negative means it's in the south, and then positive it is in the north. And we are going closer and closer and closer to the river.

And we are drawing just simple-- each of these bubbles is a simple average of the data around different points of the degree with the size of the circle in proportion to the population in each of these locations. So again, there's a lot in that graph. There's a lot of data represented very nicely in that graph. We have the river, which is the vertical line. We have the average concentration at each point, which is each of the bubbles. And then they also draw a nice line on each side.

You can see that they're not trying to impose any particular form. They're not trying to fit a straight line. They're just drawing the line that best fits these bubbles. We're going to show you how one does that. That's one of the things we want to show you. The only structure they impose, which is very reasonable given what they know about the environment, is that they want to draw a different line on the left side and on the right side. And then they want to see whether there is a jump exactly at 0.

And what you can see is that seems to be the case, where the line that starts from the south goes up as you get closer and closer to the boundary, which does make sense, because pollution travels through the air. So they probably would get some of the pollution from the north. And then there is a big jump at the north. And that's kind of a little bit more constant, maybe to the right.

So this jump you can have it. You have a nice little text box here that very nicely illustrates what they find. It's a jump of 4.1 microgram per square-- per cubic meter. And it's statistically significant. Again, we'll get to the detail of what that means. But that means that we think it's likely not to be zero. It's likely not a fluke that there is a difference here.

So once we know that for pollution, we can say, well, let's look at mortality. And then we can see whether we find something similar for mortality. And then you find something very similar for mortality, which is there is-- as you increase, as you get close to the river, you have the mortality increases. And then there is a big jump down. And then you go down.

So it's actually not mortality-- excuse me-- it's life expectancy. So life expectancy is higher on the south of the river compared to the north of the river. Another thing you see that is nice in this graph is unlike pollution, the mortality, other than the fact that there is this jump, the mortality doesn't seem to be widely different on both sides.

The thing is heat is actually a nice thing to have. So even though pollution probably kills some people, as we can see, it's probably-- it might be the case that heat also saves some people's life. So from this graph, you might not necessarily conclude that, oh, we better stop the subsidies for heat. The only thing you might be willing to conclude is that PM particulate matters are not good for you. And so that's a graph that I think that very nicely-- the little story very nicely encapsulates the type of insight we can get from data.

The next question you can ask before getting into this class is, well, does anybody care? You know, Greenstone and his friends write this beautiful paper. They might publish it, maybe even in *Science* if they're lucky. But will anybody read it? Will anybody pay attention to what they have to say? Will anybody be persuaded by data?

So that's where I have to maybe give a little disclaimer. A big part of my day job when I'm not teaching the like of you or doing research is to try to take my research to governments and to say, look, this is what we find. Maybe you should think about what this means for you.

And another part of my job is actually to try and work with government to generate and help them generate the data that will help them take better decisions, and other actors as well. So I sort of have a vested interest, at least for my own enjoyment of going to work every day, and thinking that data is powerful. But even with that caveat, besides this, I do think it's very powerful, because I have some pretty good experience with convincing people using good data.

I should immediately tell you that's not sufficient. In the new sort of new data-driven businesses, maybe data plays a very big role in making decisions. And I think Sara can talk more about that, because she knows e-commerce very well. And I think a lot of decisions are data-driven. In policy making, which is the area that I'm more familiar, data are a part of the stories, and largely-- occasionally not the whole story.

Nonetheless, let me give you an example where we can see the power of data and the power of very clear data. And this is an example, where, actually, I've been involved with, where we ended up changing-- we ended up presenting data and creating data, actually, in collaboration with the pollution control board in the government of Gujarat that actually led to changing the law.

So in case you think Gujarat is just a region in India, it is. It's a state in India. But India is a big country. So it's actually bigger than a fair number of small Scandinavian countries, for example.

And here is what we did. We worked with the Gujarat Pollution Control Board. Again, this is with Michael Greenstone, and also with Rohini Pande and Nick Ryan. And what we did is they got in touch with us, or we got in touch with them, actually, after finding out about a system they have to regulate polluting firms.

So what do polluting firms emit? Lots of horrible pollution, air pollution, water pollution, a bunch of quite noxious stuff. And they have a system that is quite unique within India, but we encounter in other walks of life as well to try and regulate the firms. In addition to the normal, I'm sending an inspector to the firm and looking how it's going, they also have a third-party private auditing system.

So what does it mean? It means that every firm which is particularly polluting, so take, for example, a firm that produce dyes, or a chemical industry, or something like that. So a firm that is in a polluting sector is required to hire an auditor. And what is an auditor? It's a private firm that has to have some minimum competencies in terms of engineering, and engineers, and scientists. It must visit the firm three times in a year, take some pollution measurement for air pollution and water pollution, and prepare a report that is sent to the firm and to the government.

And in principle, the information in that report is actionable in the sense that that's a substitute for an inspection. If the audit really signals something that's not right, the government can go and inspect further and punish, and all that. Now, it might sound familiar in the sense that this idea of auditing by private firms may look, for example, very much like what? Where do we have third-party auditing?

So that's the other thing with me. I don't really ask rhetorical questions. Or if I do, I really move very quickly. So when I ask a real question, actually, I appreciate an answer.

AUDIENCE: On the financial markets for credit ratings, we have Moody's S&P.

ESTHER DUFLO: Exactly. So credit rating is an example. Another one in these same financial markets? Yes.

AUDIENCE: For auditing firms.

ESTHER DUFLO: Yes, exactly. Corporate audits, very much the same idea. And you're almost too young to remember the 2008 financial crisis, and certainly too young to remember the Enron scandal. But if you're not, what is the problem that people have noted with this private auditing? Yes.

AUDIENCE: The Enron case, the auditor was paid for consulting work in addition to their auditing. So they had an incentive-- a perverse incentive to fudge the numbers to keep Enron going, even while fraud is ongoing.

ESTHER DUFLO: Exactly. So that's the-- there is a clear conflict of interest in a relationship. So in this case, it was particularly egregious because they had another side business. In principle, it's forbidden. At least since the Enron case, it's forbidden. But there is always this tension where the auditor has an interest in keeping the relationship going. And what best way to keep the relationship going is to keep your client happy. And in the case of pollution auditing, what's the best way to keep your client happy than to provide a clean bill of health, regardless of what you actually see in the data.

And by the way, as long as you're going to fudge the data, why bother going and collecting it? You can do the work much faster and much cheaper if you actually do it in your own home. So data becomes very plentiful if you can make it up. And the advantage is that the price of an audit becomes really, really cheap. So you can combine-- you can compete in prices by offering very, very cheap audits, which, in addition, will show the firm to be clean.

So what I'm telling you is not actually an exaggeration. It was very close to the situation when we arrived. And in fact, some of the auditing firms that were the good ones-- before the law, there were already consultants that were kind of really helping the firms, and then they moved into the auditing business. And then there was a huge industry of auditors that crept up to take advantage of that.

And in fact, it was known that the situation was so bad that the firms actually sued the government saying, look, you're not using the data, because it's so bad. So this entire thing is just a tax on us that is illegal. So they lost, because it was a bit too much.

But then the government was-- the first time we got in touch with the government was through their lawyer. And they were aware of the fact that there was a problem with this system. And so we had a chance with working with them to say, well, let's try and generate data that is going to give us an idea of whether we can reform the system in a way that makes a little bit more sense.

And what we propose is to say, look, the biggest problem is the conflict of interest, because there is a direct relationship between the auditor and the firm. So let's cut this, first of all. So let's say that the firm, instead of paying the auditor, is going to pay in a central pool. And the central pool is going to hire the auditor.

So whoever wants to be-- whoever auditor wants to be in the system can register, and then they'll be randomly assigned a firm, instead of-- so that severed the link. They are now responsible to the pool, which is administered by the government, instead of being responsible to the firm.

Now, they might still want to cheat, for example, because the firm might offer some side payments, et cetera. So on top of that, we're going to introduce some control on the auditors by introducing some random back checking, where the moment that they go-- that they register that they go, and then someone else-- like, in this instance, it was a university would send teams and take the same measurement within a few days so that you can see whether it's very different or not.

So what we did is that this system of back check, we introduced for all firms. And then the firms that were eligible for auditing, we randomly divided them into treatment firm and control firm. The treatment firm, we changed the system the way I just described, by having this-- the third part-- the auditor is paid by a pool, and not by the firm directly. And there is this-- and the back checking are actually used to monitor the monitors.

The control firms remained like the usual system. Why did we do that randomly? Because that way we know that it's not the best firms that select into the system or the worst firm that are forced to go into the system. We know that there is nothing different between the firm in the new system and the old system. And the data is provided by the audits themselves, and the back check, and the comparison between the two.

So here is something that you'll also learn how to do, which is an histogram. What's a histogram telling us? It's kind of a rough idea of what percentage of the data is in various bins by pollution. This is the control plot. And this is the audit data. And in red is the threshold. This is for one particular pollutant, which is SPM.

And so you can see that. And you can see that these firms are really doing well, because they are all not only compliant, but right at the limit of being compliant. So they are really targeting very well the thing. So that's the actual audit. And that's the back check. So you can see this is the same.

[LAUGHTER]

You can see, well, there was-- these issues that people were talking about. It's actually pretty real. So we better do-- there was some-- the data is useless. The data is entirely made up. And you can see, interestingly, there are some firms that are not polluting at all in reality that we don't see in the audit data.

That's the thing about-- if you're going to make up the data, you might as well just make it up yourself. And it's not even coming to your mind that they might actually be quite clean. But of course, the mass of the distribution comes from the right, unfortunately.

So this is the control plot. And here is a treatment plot. And you can see that in a treatment, the audit is quite different. The audit distribution is quite different. And it now looks very much like the back checks.

So you can-- something you're going to learn is you can reject that this distribution and this one are the same, that the data comes from the same distribution. You cannot reject these are so different. They are not exactly the same. You can still see a little bit of bunching. But they look reasonably similar.

So this is the data that we generated. And we shared with the government. And it took some time, but it was so clear that there was a problem-- number one, number two. That's not often that-- we're not often that lucky, but in this instance that the problem could be fixed that, basically, the Gujarat Pollution Control Board managed to work through the court to change the way the system was implemented such that the system is now changed to look very much like the one we checked. Yes.

AUDIENCE: I was just curious as to the timeline. When did you enter the project? Over what period of time did the effect actually occur?

ESTHER DUFLO: So the project was for two years. And the effect, you find in the first year. We find in both years. So we lasted-- we collected the data, and did the project for two years. And both years found similar results. So it's pretty immediate.

One thing that I didn't show you is that you might wonder whether this whole matters. And of course, what matters is not just the reported pollution, but the actual pollution. And we compared the firm's-- the actual pollution measured at an end line in the treatment and the control firm. And we find a reduction in pollution in the treatment group. We did that survey at the end of two years.

And then it took another couple of years to go from there to actually changing the law. There are many, many Ts to be dotted, and I to be-- no, the opposite-- whatever. You get the point. But we did that. I mean, our partners did that till that happened. Yeah.

AUDIENCE: Were the firms informed they were part of the survey, or were they just told, this is a government policy, and some of you are going to have [INAUDIBLE], and some of you are [INAUDIBLE]?

ESTHER DUFLO: So this was a-- so when there was actually the survey, at the very end, the only time the firms were surveyed by us was at the very end. And there they were informed that there was this survey. Before that, they received a government notice that, oh, by the way, you have a different system this year.

That raises a very interesting question that we're also going to talk about along the way, which is when you're collecting data, you're collecting data from humans. And human people or whatever, they have rights. And therefore, we try and protect them. But it turns out firms are not humans. So the way that firms are treated, although there was a human subject protection plan, you can collect and use data on firms, whatever they think, as long as you have some protection of the data, that you promise the data is not going to go in the wild world.

So that's the data. It can be powerful. Data can be insightful [INAUDIBLE]. The downside of data being powerful is that it can also be deceitful. And in particular, a lot of people know that data is powerful, and a lot of data are trying to-- a lot of people are trying to use data. We're all trying to use data to make a case. And sometimes you can mishandle the data. And it can make-- you can think that you have a case, or that you have a story, and really, it's not really the story you had in mind.

One example that is close to home is what explains autism. So autism is something that is deeply affecting to people when it hits them, and deeply scary for everybody else. And perhaps the result is one where there are a lot of stories circulating. One of them some time back which was actually published in *The Lancet* was the idea that maybe some agent in the measles vaccine was causing autism.

So that was supposed to be very careful measurement, et cetera. That's why it was published in *The Lancet.* But that data was fabricated. So of course, when data is fabricated, what can you do? That's one possibility.

But more recently, another story has surfaced without fabricated data, with absolutely real data, showing this very clear correlation. This data come from an MIT researcher called Stephanie Seneff showing a strong correlation between the cases of autism that we observe in the data and the use of a particular agent that is used to protect GMO plants. So this is called glyphosate. And you can see that in red, you have the glyphosate, and in yellow, you have the autism bar. And the correlation between the two is very, very strong.

So on the basis of that, you have many representations you can find on the web explaining that GMOs might be responsible for autism. And therefore, we better stop. And if it continues, maybe a huge fraction of kids will have autism in future years. And when you see this graph, what do you think? It might be happening.

AUDIENCE: Perhaps there are more children being diagnosed with autism these days than there were 10 years ago or 20 years ago.

ESTHER DUFLO: Yes. So the one thing that is very clear in this graph is both series are really strongly trending upwards. And maybe it's a diagnosis case, even if it isn't, just the fact that they're both trending upwards might give us pause.

AUDIENCE: Maybe it's not normalized to population growth.

ESTHER DUFLO: I don't think it is. You're right. It's a number of cases. I don't think population growth explains that, in the sense that if you did it a fraction, you would find the fraction keeps rising as well.

AUDIENCE: You can probably find lots of things that are trending upwards over the same time period. So just because they're both trending upward doesn't mean there's a causal link.

ESTHER DUFLO: Very much so. So I'll give you one of them, which is the organic food sales. So this is from-- actually, many people have observed that. This one I took from David Gorski's website, who is a doctor who has a nice blog on data and misusing of data. Maybe it's organic food sales that is causing autism prevalence.

It's a little bit tongue in cheek, obviously, because it doesn't have the nice-- like, this has a slight-- maybe the trend is like really nicely espousing themselves. And this is slightly larger scale. But you can see something pretty similar. So maybe organic food sales caused the problem.

You could even become a little bit more sophisticated if you wanted to prove that organic food sales is responsible for autism, because one place you could say, well, I'm going to compare, for example, Iowa and California. And you would notice that in Iowa, the organic food sales has increased much less than in California. And so have the number of autism cases. The growth in autism cases in California has actually been much faster than in Iowa. So maybe-- so that gives you one more piece of evidence that really organic food sales are responsible for the growth of autism.

So of course, this is slightly-- this is kind of a caricatural case. Any theories that strength trends strongly upwards is going to be correlated with any number of things that trend strongly upwards. So you could think, well, yes, whatever. It was kind of a cheap shot.

But there are two things to be-- that is worth remembering in that. One of-- the first one is that if you're very convinced about a particular causal story, you can look at the data in the way that is going to reinforce your belief in this very strong causal story. So when one looks at data, one need to use [INAUDIBLE] theory to know what to look at. But one also needs to let the data speak, and then in some sense, tie your hands to look at the data and not keep the part of the data that's good for you and the part of the data less good for you.

I wanted to make two points, but I'm forgetting what the second point is. So I'm going to skip the second point. So that one is a little bit trivial, but there might be some less obvious ones. And one that is sort of interesting, especially for-- you know, I'm a development economist. I care about-- I work a lot on developing countries. A very interesting graph-- by the way, this comes from-- this was done by us from R, which is something you're going to learn in this class, which plots this very, very strong correlation between enrollment in secondary school and log GDP per capita.

So clearly, where people are more educated, in countries where people are more educated, they are also richer.

AUDIENCE: Why does it go above 100%?

ESTHER DUFLO: Oh, because the enrollment rates, actually, are often-- enrollment rates are gross enrollment rates. So it's the number of kids in school divided by the number of kids of that age. So because some people stay in school longer than they should or come earlier, et cetera, you can-- if everybody goes to school, actually, the enrollment rate tends to be above 100. Very good observation.

And you can see a very strong correlation between 1 and the 2, and that correlation people have used to motivate policy. For example, build schools, and build support for education, spending on education, et cetera, also develop some theory of economic growth that relies on human capital.

And in particular, this correlation is much stronger than the similar correlation you could find at the individual level. So in general, one more year of education increases your earnings by about 7% to 8% in most countries. And this correlation is not the-- the units are not the same. So they need to be compared. But that would suggest a much, much steeper correlation between the two.

So that makes you to think, well, maybe the whole is better. It's more important, the sum of the part. When I am educated, after all, I'm teaching you, if I have a PhD from MIT, I teach you better than if I just had a high school degree. So my education benefits the rest of you as well. In the workplace, maybe when two people meet, they exchange ideas, and whoever has the best ideas can share them with another. So people keep meeting, and that creates this huge, what we call, externality of education. So that's one possible interpretation.

The other possible interpretation, though, is that that's actually misleading. And one way to look at it is to say, well, if that was the story, then when countries become more educated, they should also become richer. And you see that that's kind of hardly true. There is a slight correlation if you really squint, but it's much lower.

So when we look at difference in education and log per capita GDP growth, we don't see this difference anymore. So there doesn't seem to be-- maybe there is a causal story. Maybe there is no causal story. We were so convinced by the causal story here, and we have been for so many years that we maybe were not very critical in looking at this graph.

But now people are seeing that and saying, well, maybe there is no effect of education growth at all, or maybe something else is happening that goes in the other direction. So what could explain how we go from this graph to that graph? Like, what could explain that the first one is there and the second one is there?

AUDIENCE: 1% growth of a trillion dollar economy is much harder to do maybe than 12% growth to see than a million-dollar economy just on an absolute scale.

ESTHER DUFLO: That's a good point. But I don't think this is true. In practice-- not only it's a good point, but it's a point many, many people, very respected people, have made. So it's a very good point, which is this idea that they might be convergence. So the countries that are less educated also have lower GDP, and therefore they should catch up to the richer country.

So maybe if I took this graph and I added a base level of GDP, I could find the correlation back again. In practice, there is some convergence. So it might explain part of the change. That's a very good point. I don't think it would do all that much for you. But it was a good attempt. Yes.

AUDIENCE: Is it possible that's like the previous graph. It could be that the GDP explains the enrollment. So as in like vice versa, instead of enrollment explaining GDP. Because the countries that are richer, they can afford to have children not having to work. So it could be vice versa, which is why there's not as much correlation when it comes to growth per se.

ESTHER DUFLO: Exactly. So there could be a reverse. The causal story could be the other way. The story you are giving is a very good story. And we can think of other stories that would explain the reverse causality. Maybe you have one.

AUDIENCE: Other variables [INAUDIBLE].

ESTHER DUFLO: Exactly. So the third possible story is that there are hidden variables that explain both of these things. For example, give an example of what it could be.

AUDIENCE: Maybe [INAUDIBLE].

ESTHER DUFLO: The government supports education, and something else that makes them richer. For example, maybe governments that are able to have good schools are also able to control malaria, and are able to have a good functioning credit market, and all sorts of stuff like that. So that's the omitted variable story that could also explain the two things. So I think you're both-- all your three ideas are good. Yeah.

AUDIENCE: I actually have a question about this graph. [INAUDIBLE] y-axis. I'm not sure what GDP per capita growth means.

ESTHER DUFLO: So this is the log of GDP per capita in one particular year. This is take two years or two points, and take the ratio of the GDPs, and take the log of that. So for example, it says that between-- to get any of these points, you say, between 2000 and 2012, the DRC grew on average-- DRC Congo grew on average by about 2% a year. And they have about 50% of kids in primary school.

AUDIENCE: So then this actually might just [INAUDIBLE], but I think you could just see the improved economy, like I see China and India and [INAUDIBLE]. So it's just like a general [INAUDIBLE].

ESTHER DUFLO: Yes. So something else could be happening at the same time. The issue is this something else is also not pure noise. That's the thing that is important to remember.

AUDIENCE: I think it would probably interesting to look at the-- comparing GDP growth per capita to change in [INAUDIBLE].

ESTHER DUFLO: Yes. One could do that, too. And if one did that, one would find a slight positive slope. So you would have something that stops much less than that, but a bit more than that. And depending on the model you have in mind, you might want to do that. You might want to go not from here to here, which, after all, is a different story, but from here to difference in GDP. That is GDP growth versus growth in education.

So these were all very good points, and brought out the issue even on things that seem not at all obviously wrong. Correlation is not causality. It's important to give a-- to have a causal story when you look at your data. But a causal story is not causality either.

And when you read the newspaper, when you read-- but it's not just a newspaper. It's also a policy report. And it's also a lot of academic journals a causal story kind of plus some data that illustrates the causal story, it might be taken to be causality. And it's not always.

Sometimes even sophisticated data use might not indicate causality. For example, if we did your graph which I really should have put there on the slide, we would find this slide positive growth. So you can say, well, fine, the level of GDP to level of education, I'm definitely not going to take that seriously, because there are all these omitted variables and reverse causality.

But growth on growth? That seems like that all of these differences, fundamental differences between the countries would go away. And I can really take the growth on growth seriously. But even on the growth on growth, you could make a story that something else explains both the growth and the growth in education. So even though it's one more level of causality, someone can write you the equation that would really justify doing that, and be done. But a level of skepticism on whether we really believe that should still be applied.

So fortunately, there is actually a lot, a lot, a lot, and lots of data, data by the chock-full, actually. So one kind of hope that I've seen expressed in people is that there is so much data that we can become so much more sophisticated on our use of data, that we will not-- we will be able to control for everything. For example, when we look at not only we'll take growth on education, growth on GDP on growth on education, it will be able to put so many, so many things that explain everything-- the reverse causality and all the omitted variable. And what is left after we have done that will be the true story that we are after.

So I wrote this slide on a Saturday afternoon. And I had been looking for capoeira classes for my four-year-old. And she wanted to see a lot of videos. So we watched some videos. Then we went to the-- then I went to the Spurious Statistics website, which I'll show you in a minute, and downloaded a couple of graphs. Then I bought a doll house for the birthday of the younger one.

So there is a lot of data on me that was circulated in this one-- so you can see, first of all, that I have a bit of ADHD because I'm doing a lot of things at the same time. But secondly, so someone could say, well, this person is not very focused. So we can send some ADHD medication, or we can say, well, clearly, she has at least one kid in this age range that likes dollhouses and capoeira. So the kid must be about four to like both, to be on the older age of dollhouses.

And then immediately, *The New York Times* website starts proposing me ads for more toys, different kinds of toys that I might like. And actually, I received an email from a site that I had never registered for that is proposing me to sell something. So there is someone-- not someone, not a person, who is actually making constant use of the data that I generate, and try to get some predictive patterns of who I might be and what I might want next.

And this is done by what we call machine learning, where someone is trying to predict who am I. And I should say it doesn't matter at all for that particular use of the data whether it's causal or not. The fact that they propose me next another toy for the kid is not because they think it's a causal effect going from capoeira classes to a trip to Brazil or other toys for the kids. It doesn't matter to them. Just all that matters is that there is a good chance I might be interested.

And another type of question you can ask, we have a lot, a lot, a lot of genetic data on people. And there's a whole research agenda, which asks, for example, whether someone-- what are the genes in people that predict patients. That's also something that a little bit like the data that is here involves a giant fishing expedition, because we have absolutely no theory of telling us what it could be.

There are very structured ways to do that so you don't do that-- you have to do that in a specific way. You can see whether or not you agree with this use of the data. I won't give you any-- I won't give you my view at this moment, anyway. Maybe we'll do it later, or you can think of whether this is something that is interesting or not, or dangerous or not.

But what is clear is you want to be very careful of patterns you observe in the data, because forget about the causality interpretation. They might not be meaningful at all. Because once you search long enough, you're going to catch something.

And here is one that one could find if you look at data long enough is the total revenue by arcades and computer science doctorate awarded in the US. You can see a lovely correlation. And it goes-- it's more than just the time. It's the time, but it really follows very nicely. And having been at MIT long enough, I can understand this. I can build a causal story around it. Kids who like computer games become very good at-- are also people who are interested in science.

But here is another one which comes from the same website, which is the Spurious Statistics, which is the age of Miss America, and the murder by steam, hot vapors, and hot objects. Well, you can see that the correlation is even nicer. And there is a little bit more difficult to think that there is a causal story between the two. Here is another one-- the number of people who drowned falling into a pool and films Nicolas Cage appeared in.

So this website, Spurious Statistics, it actually gives you a little tool to build your own spurious statistics. So what does it do? It basically searches for time series that are nicely correlated. And it searches and searches and searches until it finds one, and then it finds one, and it is what it is.

So the thing is when you have enough data, some pattern will emerge. So you need to be very, very, very disciplined. So this is not disciplined at all on purpose. This is the search for something. But in reality, you will need to be very, very disciplined in the use of this considerable amount of data to extract something from it that is actually meaningful and not entirely due to chance. And we are going to introduce you to those techniques, how one does that, and one does that in a meaningful way.

So what do we need to learn? We need to learn-- so as I explained, unless we go to the data with a structured and disciplined way, the data is going to play tricks like this to us. We need to prevent it to do that to us. So we need to learn how to model the processes that might have generated our data. And in a sense, that almost requires no data to do that. This is a study of probability. This is how the data might-- how do we model those processes and how do we play with them? And so we're going to start with that.

Once we have those tools in hand, we can go back to the data and say, well, how do we summarize and describe the data? And how do we go from the data to the processes that might have generated it? So how do we go back, too? And this is statistics.

Once we have done that, we are familiar with manipulating data. We can start working on trying to uncover patterns between variables, how variables might be related to each other. So some of it is exploratory data analysis drawing plots. Econometrics, obviously, is going to help us. And machine learning, which is these techniques to avoid this, and in fact, to find patterns in a disciplined way.

Then we need to think about causality. There might be patterns with data that are very strong. For example, the relationship between GDP and education. I can tell you it's not a chance thing, but it's also probably not a causal thing. So how do we go-- how do we think of causality? What does it mean for something to affect something at the conceptual level? So I'll give you a causal framework.

How do you generate data that gives you-- that's going to give you a good causal framework idea? We're going to start by experiments. And then, well, suppose we don't have the ability to generate an experiment, how do we come close to that? And how do we assess how close we are to that benchmark in data that actually occurs naturally in the wild?

And of course, this is a practical class. So we are just trying to get the practice work in steps with everything. We are learning conceptually. So we are going to work with R in this class, not Data, partly because R is free, partly because it's very powerful, partly because it's the future. So it is better if we all learn R. So this is going to be a joint learning exercise.

We're going to learn about designing experiments. We're going to learn about finding data on the web in various ways. And then, very important, I think, is how to present our results in a compelling and truthful way. Beautiful graphs, such as the one I showed you, insightful tables, and good text is the things you need.

I'm going to stop talking now and leave the floor to Sara, who is going to introduce the lot of us a little bit more completely and talk to you about syllabus and logistics issue and take your questions. Thank you guys.

[APPLAUSE]

SARA ELLISON: So welcome. It's so great to see so many people here, and a few familiar faces. So I'm going to use the last 20 minutes or so to go over the syllabus that you have in front of you to sort of, as Esther said, answer-- fill in some of the specifics about how we're going to structure the class, what kinds of things we're going to require, how we're going to grade the class, answer any questions you have about that. And I'll also take a couple minutes to introduce myself and the other people who are sort of running this class for you.

I am Sara Ellison. And just by way of introduction, I'll just give you a couple facts about me. I have been at MIT for a long time. So I arrived here in 1988. I think there were about two restaurants in Kendall Square at the time. I came here to do my PhD in economics and have been here not quite continuously since, but almost continuously.

I've taught lots of different classes. I've taught undergraduates, PhD students, MBAs. I've taught statistics. I've taught industrial organization. I've taught microeconomics. My research is in the field of industrial organization. So I study how firms interact and what sort of markets look like. And I'm sure I'll have some examples drawn from my research peppering lectures as we go along.

I have three kids, all girls, and they all like math. So that's great. And actually, one of them, in fact, is an MIT student.

ESTHER DUFLO: I'm Esther Duflo. I've been at MIT almost as long. I came as a graduate student in 1995 and never left. I work on developing countries and the economic lives of the poor. I worked almost exclusively with data. I'm one of the directors of J-PAL, the Poverty Action Lab, where we try and set up-- we work with partners, NGOs, government firms, et cetera, all around the world, including North America now, to set up randomized control trials that are experiments to look at the impact of various interventions on people's lives.

One example of it was the Gujarat pollution trial that I showed you in a minute. I love data, and I'm very pleased to-- I love data, and I love teaching. So this is a great combination for me. And I'm very excited to try this out with you guys.

SARA ELLISON: And I should mention that Esther's superpower is that she can do so many different things. She is just-- and many at the same time. So you'll probably see evidence of that throughout the semester. She's amazingly productive and interested in lots of different areas.

I don't have much to say about course description. Obviously, Esther sort of gave you a very nice motivation into what things that we would like to cover this semester and what are some of our philosophies towards covering some of these topics. I do want to just point out that we have not listed-- I believe we didn't list any prerequisites in the course.

We will assume familiarity with basic algebra and calculus. It's not going to be a very calculus heavy course, but there will be some calculus. We're not going to assume any prior knowledge of probability and statistics. We're going to start from the foundations and build up from there.

Because this is kind of an experimental course, not only have Esther and I never taught a course like this before, but I think few people, if any, have. There is no good textbook that really covers everything that we're going to do. So there's no required text.

On the syllabus, I've listed a number of texts that if you are inclined, you can order. You might like to have some of them on hand. But I can't really say that any of them are highly recommended. I mean, I think they're all good texts, but they're not necessarily going to cover-- no one text is going to cover a majority of what we're doing this semester.

So the three listed in the beginning here will cover material that we're going to do mostly in the first half of the semester, and then the two books, the econometrics books, will cover some fraction of the material in the second half of the semester. All of them are excellent texts and could be good resources for the future. But you don't necessarily feel like you-- you shouldn't feel like you need to go out and buy them. We're finished 2 minutes early.

[APPLAUSE]