[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: And today, we are going to cover a lot of stuff, but a little bit lighter than the previous two lectures. So it should be fine. I don't exactly know how far I'm going to go. So I'll continue tomorrow whatever I haven't covered today. It's going to be basically these three things, where can we find data, how can we start looking at it to give us a sense of what distribution the data might have been generated by, and then since we've gone through-- there's all sorts of things we can do once we know probability.

And one of the things we can do is extract much more from the data than what the raw data would be. And I'll give you an example of that, very powerful example of that. So that's what we're going to do today and maybe some of tomorrow. And then tomorrow and next week, we'll go back to probability, talk about marginal and conditional distribution and then function of random variables. Then we'll go back to some examples again later.

OK, so that's the plan for today. Where can we find actual data? Where can-- where does it live in the real world? And in particular, how can you get hold of it yourself? So there are basically three ways. There are probably more, but three sort of broad things I want to cover today.

OK, so where I can find data? So one possibility is that actually, a lot of data floating by around. Most of it is actually much, much, much more data than was there even when Sarah and I were doing our PhD or our undergraduate degrees. So one first step would be to see what does exist out there. Then possibly what exists out there is not what you want. In particular, when you run experiments, it's often the case, although not even always the case, but it is often the case that the data that you might need to evaluate the effect of your experiment is not actually-- it doesn't actually exist because you need the specific data that will be needed for your experiment.

In that case, you might need to collect your own data. And then something which is a bit in between, which I'll actually cover first, is there is actually a lot of data that's in the internet, that the internet keeps generating for us. It is not necessarily there in a smooth, already prepared for us database, but it can be gettable with not too many steps. This can be combined potentially with experiments. Or this can be combined with surveys.

So let's start with the more traditional way of going and looking for existing data libraries. What I'm going to give you today is really just a very-- like a tip of the iceberg of the data sources that might be available and how to find them because if I had started listing data sources, I think we could have been there till the end of the semester looking at web page URLs on the screen. And it would have gotten somewhat boring and maybe not all that useful.

In fact, I saw doing some little searching, I saw that there are now databases of databases. So maybe we could have done a joint project of doing a database of the database of the database, et cetera. And eventually, it would have been boring. So I just putting the most popular sources of data. And then keeping in mind that eventually if you want to do something for which you might need existing sources of data, you might just be able to-- once you know the type of thing that you are looking for, you'll be able to search for data more efficiently with what exactly what it is you're looking for.

A great resources, particularly for MIT students, is the MIT Libraries. But even for other in the wide world, you should feel free to go in this web page is the library. In particular, there is a librarian in the social science library here, Catherine McNeil, who really knows a lot about data, and is always willing to help out in finding data, even buying data. Yes, the MIT library might consider buying data that you might need, as long as they think that it might be a data purchase that would benefit other people in the future.

So if you consider buying data about cupcake production in Kalamazoo, maybe they'll be like, well, maybe not. But if it's something that actually would be useful to have, for example, in my experience, MIT Libraries are often quite willing to purchase GIS data, which they add to their GIS library, and many people use later.

But certainly, she can help you in finding data that are already there, in particular, finding the data on the topic that you are interested in and show you how to look for the data in the MIT Libraries and others. So that's the first big resources to go to. They have a web page that is updated regularly. And then there is actually a real person to talk to, which is quite useful. If there is an interest, she is always delighted to do a session. So we could invite her during a recitation or something. So maybe we'll poll you on that in the future.

Now popular sources of data are data.gov. The executive branch of the government generates a bunch of data that are just put on the public domain, partly for transparency reason, partly because they like to have people looking at it. So that's a source of data on anything and everything from tariffs to protection of nature in national parks, and other things like that.

Another very popular source of data is the IPUMS website. The IPUMS has the census, the censuses, at least not the-- very old censuses are actually there in their entirety. More recent censuses, we have extracts available, 1% extract of censuses of the population. And the census ask a few questions. And it's on a very large number of people.

It's freely available, very easy to work with. The IPUMS website also hosts a bunch of other data sources that we use very frequently. Some of them you're going to see today. For example, the current population survey, which is a quarterly survey of the US population available on a fairly large sample state by state that has more information than you might-- than you might think you might want to need on things like wages, hours, type of jobs people have, and stuff like that, and many more of these public service-- public survey data are there.

For example, the American Time Use survey, very interesting survey that asks how many hours people spend doing various things in the day. If you've seen several recent articles in the press, not so recent, about a few weeks ago, about how men and women share or don't share equally the responsibility in the household, this is American Time Use survey data. So these are just a few examples.

AUDIENCE: Time Use Survey data also tells us that Esther and I worked much harder when we were in school than you guys work. So there has actually been a sort of a series of articles about how the-- the amount of time that students spend studying in college has been going down pretty precipitously. Maybe that's not true of MIT students. But in general, this has been documented.

ESTHER DUFLO: I actually don't think it's going to documenting it for me, because I was a student in France. But I can tell you I worked harder than you did when I was in college. There is actually-- probably an international time use survey to be found somewhere, but probably not on that. Although it is-- it might even be linked.

So you would you might be able to compare the work habit of French students aged 17 to 19 to American students aged 17 to 19. And it would be interesting potentially. What would be very interesting in France would be to show that it's totally bimodal, with 90% of the students doing exactly nothing and 10% of the students doing working, pulling the mean to a reasonable number.

A few years ago, the IPUMS project started repeating the same effort of anonymizing, and cleaning up, and putting online the US Census for other countries. And this is the International IPUMS project. So on the International IPUMS Project, you have censuses from lots and lots of countries. So you might think that it's very easy to-- very difficult to get data from Kenya, think again. For example, the IPUMS International has censuses from Kenya. Yes?

AUDIENCE: What are regions where that IPUMS doesn't have that information?

ESTHER DUFLO: It has-- there are more places it doesn't have the information from then places it does have the information from. Basically, it started, I want to say, about 10 years ago, putting them on. And what they're doing is they're buying the censuses from the statistical offices and working with the statistical offices to clean them up and put them in and doing that progressively. So I don't know how many places they've reached.

So there are more places that don't have the censuses online then places that do. But there are a fair amount. That means there is lots and lots and lots of very rich microdata that can be used in different countries. So these are two interesting data sources.

Another very rich sources to find data on anything and everything is this ICPSR website at the University of Michigan. So this is really the first big dataverse. I'm going to be talking about dataverse today. What's a dataverse? Basically a dump where people put databases. Usually researchers put data sets that they have collected, or assembled, or cleaned, or something like that. And then they put them. And ICPSR organization has done a wonderful job working on how data needs to be anonymized, made secure for the people who have given their information, so your names, address, et cetera doesn't show up, and documented such that other people can use it.

So the ICPSR data comes with very good documentation, dictionaries, et cetera that's reasonably uniform from data set to data set. Which means that once you understand how to read a code book for one data set in the ICPSR dataverse it would be useful for another collection as well.

AUDIENCE: As more and more data becomes available about the same set of individuals, there's also some studies where people are able to de-identify [INAUDIBLE]

ESTHER DUFLO: Re-identified, yes.

AUDIENCE: --something one is also seeing when it comes to some of these data sets?

ESTHER DUFLO: I wouldn't say in those because this is very much a pre big data world, where you have a data set of a few variables on a few people over here and a few variables on a few people over here. But I think you're exactly-- you're asking exactly the right question. So there are databases that are publicly available, but publicly available without identifying information.

And identifying information, you have to ask yourself, what's identifying information. You could say, well, very many people who live in Boston. So the fact that I live in Boston shouldn't be identifying information. But once you cross that with-- suppose there is a data set-- there is a very useful data set called-- on aging, which followed a cohort of older people. They started at 45. They were not so old yet. But they are now older. They have followed them for a long time, maybe about 20 years, people who were, when they first started, 40 and older.

When you access the public release of this database, which I think you can get from ICPSR, by the way, you cannot get the state where people live. And you must say, come on, how am I going to identify someone from the state where I live in? And it's exactly for the reason you're talking about, well, maybe not just from the state, but from the state and the fact that they have cancer, and the fact that they have two kids, and the fact that they do this and then do that, it might be conceivably possible to cross that with other data set that make their way in internet, and figure out that this person is Mr.-- one Mr. Smith who lives in such cities.

And someone can start calling him and say, hey, you want some new magic bubble gum cancer treatment. So for this reason, actually, the requirement in term of the type of information you can leave on this public data set have become stricter and stricter, which doesn't mean you can never access the state. In this aging data set, for example, but to access the state in the aging data set, you're going to need to first apply to MIT for a human subject clearance, where basically, you're going to commit not to share the data with anybody and to use it on an encrypted disk. And then once you have that, you can call-- email these guys with a request saying, please give me the state. It's going to take some time. But you're exactly right.

So ICPSR, I think, is kind of the grandmother of this idea that researchers should be sharing the data that they collect, or they accumulate, and they clean up. Then the baby of that is a Harvard MIT Data Center has a Dataverse which resides at Harvard, but is open to the MIT community, in fact, is pretty open to everybody who wants to get data from it, which does a little bit the same thing works on help with people anonymizing their data, storing it securely, and distributing it to people. Even-- Yes?

AUDIENCE: Does J-PAL also publish some of its data sets?

ESTHER DUFLO: Yes, I'm coming to that in exactly one minute. In fact, I had it on this slide because it lives in the Harvard Dataverse. But I thought I will put it on the international data sources. And then Amazon has a dataverse. You would think if there is something to do, they would have done it. So in fact, I did not know till I prepared this lecture that Amazon has a dataverse.

But then I thought, by first principle, they must have a dataverse. So I looked for one. And in fact, sure enough, there it is. It has a lot of public data, actually, publicly available data. So you can just register on the dataverse and get it. So there is also some source of data there.

The other interesting sources of data, so if you're interested in household surveys, but not in the US, so if you're interested in the US based household survey, the IPUMS place is a good-- page is a good place to start. If you're interested in international household survey data, there is a lot of it. First of all, there are the demographic and health surveys. There to answer your question, apropos the demographic and health surveys, there are very few developing countries in the world, which does not have a demographic and health survey.

So what people have done is either using them individually or combining them, they have the same questionnaires in all countries. So they are very easy to combine. As their name indicate, they have mostly have data on demographic and health issues. For example, they don't have anything on income. But they have assets. By assets, I mean, tables, and refrigerators, and the cycles, and this type of stuff.

And from the assets, you can construct a measure of how wealthy a household is because it turns out how much stuff people have is quite correlated with how wealthy they are. And then there is lots and lots and lots of data on maternal health, and child health, and a little bit of data on education, and stuff like that. So the DHS is a very rich source of data. They are freely available. You just need to register to say what you are planning to do with them.

This is individual level data. Again, you have geographic information. And you can-- that's already there on the website. And with one more level of permission, you can actually get GPS coordinates of the location of the villages. Yes?

AUDIENCE: I was wondering so when you're using a public data set, you have to be concerned about how about accuracy in the data set, and then how do you go about figuring out what to trust or not to trust?

ESTHER DUFLO: So a good thing, if you've never collected data, is that you assume that the data set you get from the internet is pretty good data. And once you start collecting and cleaning data, you will never assume that ever again. So obviously, the data sets that are out there have issues. And their quality differ from data set to data set.

Data sets like the demographic and health survey or the data that I'm going to talk about in a minute, the Living Standard Measurement Survey from the World Bank or the Rand public data set are all pretty good quality data sets. So when you collect your own data, when people collect their own data, there are a number of steps that they can take to ensure that the data is not fake, that the data is not too noisy, and that the data is informative of what we are trying to get.

And when you get data from these collections, DHS, LSMS, the Rand data set, or the US equivalent of those, you can be sure that those steps have been taken. Now, there are sort of technical steps, which is, for example, when you send investigators to the field, it would be very nice for them to sit under a tree and enter the data.

And some of them do that almost surely. So a technical step that you need to do is that when you do a survey, you back check, which is you're sending another team of people who go behind the first team to collect the same data from the same household and cross-check. So this is sort of you know that you have to back check some amount of your data. And people do do that, certainly in those big data sets.

There are also-- so you can be sure when you're using these sources of data, and when you're using data sets from reputable organizations like J-PAL, for example, that's something like that would have happened. Now another issue, which is an issue that is true both-- whatever the source of data you're trying to collect is that you're asking questions which may or may not be the right question to get at the variable that you're looking fro.

For example, time use data might have a tremendous amount of recall error, where people just don't really remember. And possibly the recall error has changed over time. Now that people have calendars, they remember exactly what they are doing. And before they didn't, I'm making that up, I don't know if that's the case.

And then one might be worried about despite all your best effort, the quality of the data is not perfect just because it's a hard question to ask. So this is a conceptual issue that you can never be sure that anyone has exactly answered the right question it is your job as the user of the data to decide whether the data is actually getting at what you want it to get at. Sorry, there was a question over there in this neighborhood?

So the demographic and health surveys have this kind of uniform data on a lot of countries, usually several rounds. I'll give you one example of that data is very bad, for one thing that I know is immunization in India. Maybe in other places it has great data on immunization, but immunization, when I started working on immunization, which is something I have done a lot of work with over the future, one big issue is people don't have their immunization card necessarily.

So if you ask people is your child fully immunized, which means, the question you're really asking is did they get the five shots that constitute full immunization, they will typically tell you something. So about 30% of people say yes or said yes in India in 2005. You might think that's not very much. And in fact, that is not very much.

But when we looked at the same-- when we looked at the same data asking much more detailed questions about each shot, because it turns out, for example, a BCG leaves a mark most of the time. Other shots get administered in ties. Kids in India get shot because they are sick, adults too. Every time you go to the doctor in India, you get a shot no matter what.

So people might get confused that they have gotten five shots. But these were not immunization. So you can ask question's a bit better, et cetera, so when we did a bit better for a small place in Rajasthan, we found out that the full immunization rate in our sample was 1%. So like, wow, that's not the same. So what we did to try and understand the discrepancy with the DHS data is that we ask the exact same DHS questionnaire for our sample for not the same people because we had just gone over this very long survey, so clearly they remembered, but a set of comparable people.

And we got the same 30%. So we realized that it's not that they had done their job wrongly. Is that the way that the questions are asked, it just didn't work for this population, when you would have asked very differently. So this is something that when you use the data that's already there, you have to take it for what it is with some amount of-- some amount of questions.

That is why, in a lot of cases, and I'm going to that in a minute, when it's possible to use administrative data or to use data that's not self-reported about what people have done, that's really nice. For example, if you're trying to collect prices, then-- or to have-- to get questions on prices, you could either ask people how much they paid for something, or you could try and go and find the price. And you would probably have much more reliable data if you go and find the price.

So the World Bank has a bunch of data, again, on anything and everything, practices of business, everything you're interested. And in particular, they have conducted for a very long time-- Angus Deaton was involved in setting it up. Angus Deaton got the Nobel Prize this year. If you remember, this living standard measurement surveys, these are very rich household surveys on people in developing countries. They have them for lots and lots of countries.

In some countries, you even have panels or repeated cross-sections. So you can see how things change over time. And most of them are freely available. And they have data. And they have this very long questionnaire. Usually they stay in a household from 2 to 4 hours. So they ask them every-- from every type of questions, lots and lots of questions on education, on what people consume, and how much they are-- how they earn their money, et cetera, et cetera. So they are very rich description of the households there.

Rand has a bunch of data sets as well, including again three very nice household panel surveys, one for Indonesia, which they are at least at round 4, one for Malaysia, and one for Mexico. And the one for Indonesia and the one for Mexico is are panels with excellent follow-up surveys. So you have something like 98% of households in Indonesia that are followed every five years now for about 20 years.

So there is a lot to look at in terms of how-- even if they've moved, they look for them wherever they've moved. So very rich, very good data. Bit complicated to use because just-- the questionnaires are about that fat. But once you get into it, totally worth it. Yes?

AUDIENCE: So for developing countries, things change a lot faster. So how frequently is the status [INAUDIBLE].

ESTHER DUFLO: So they are not-- the data sets, they are collected once for a set of households. And unless you are in the case of panel data, where you come back to the same household regularly, they're not updated. They're just put there. They might not be the data set you go to to have your daily update of the macroeconomic condition of a particular country.

But they might be the data set to go to if you're trying to understand a particular behavior that people might have in poor countries. So for example, if you're interested in healthcare behavior, what do people do when their sons are sick versus their daughters are sick, you might not care that the data set is five years old. It might not even be true today. But at least if it was true five years ago. You're learning something about a phenomenon that is useful to understand how people behave and how they function. Yes?

AUDIENCE: So you mentioned earlier about looking for data sets that have back checks on the surveys that they do. Is that what would be comprised of a follow-up survey? Or are there--

ESTHER DUFLO: No, so a follow-up survey and a back checks are different things. So a back checks, you never get the data on the back checks. What you get is the final product after the back check has been done and the data has been reconciled. It's the outcome of that process.

But what the back checks are is that you send your team of investigators, they go to the village, they collect the data, they come back. In the past, they used to fill up the data on paper. And now they put it on tablet. So that makes this process much easier. Regardless how the data comes through, someone has it, either on paper or on their computer.

In parallel, you send another team that goes to the same household a couple of weeks later. And they ask a subset of the questionnaire. First question they will ask has anyone visited you. So if the answer is no, that kind of helps. Second question they will ask is, in particular, the questions that are at risk of being forged are the questions that are useful for-- that have a filter.

So for example, has anyone be sick in a household? If the answer is yes, you have a five page questionnaire whether you went to the doctor, and how much you had to wait. And so the interviewer might feel like-- I would kind of shave a good 20 minutes on the survey if the answer to this question is no.

So you can go to this kind of tempting question and ask them again. And then if things are roughly the same, things change over time. And there is a recall question. So you don't accept-- you don't expect 100% matching between one and the other. But you can, once you look at the data, you can decide that, in fact, that person has been visited and roughly, the job has been done. Then you're happy.

If it's not been done, you go back to your interviewer, and you confront them, and then you fire them, and all sorts of stuff like that. So if your incentives are nicely in place, then you don't have this problem. And therefore, the data of the first collection is reliable. And that's what you're going to put on the internet.

AUDIENCE: So just to be clear, a back check is taking a subset of the initial questions and then re-surveying the same people, as compared to a follow-up survey, which is it with comparable people? Or is it--

ESTHER DUFLO: Sometimes. So you have both. You have cases where you have what we call repeated cross-sections, where you have a survey of a sample of people, and then a couple of years later, precisely because things change over time, et cetera, you go back and you resample another set of people, not necessarily the same, but from the same-- you take your sample in the same way from the same community. So you have comparable information from two communities. But the people cannot be linked.

Another situation is-- so the demographic and health surveys are like that. They are never panels. They never go back to the same people. But they have repeated cross-section on many countries. For example, India is at the fourth one. So you have you can say how things-- you can see how things evolve over time.

And maybe then you can-- if you want, you can aggregate the data at the level of the district, or the state, and construct a panel of regions. But you cannot construct a panel of individuals or individual households. And then there is the cases like the Indonesia family life surveys or the J-PAL data set that I go to in a minute are usually panels, which means that the same exact household were interviewed several time.

So then you attempt to go back to the same people and ask them the questions. Sometimes you don't find them. But the objective is to go to the same people. And I clearly don't teach fast enough. You have the objective of going to the same people to ask them the questions again. So there, you can see how particular household, how the same--

It is the same people, usually the same questions. Although when you go back again, typically, you want to add some stuff that you realize you forgot last time, et cetera. So sometimes the questionnaire change a little bit. Or you realize that a particular module, nobody-- people were staring at you blankly when you were trying to ask these questions. So you just get rid of it.

So there might be some changes in one way to the other. But typically, there will be at least a big chunk of it that is the same question that is asked from the same people. And the data sets are presented to you in the way that you don't know who the people are, but you can link them. They have a number, 12573, in particular, village number 120, in particular district. So you can link them from wave to wave. So the LSMS, the Rand databases are like that. They are all panels. Some of the LSMS are panels. And some are not.

Another source of data is replication data from researchers. So more and more individual people or group of researchers are cognizant of the way that it's important to make data as available as possible for other people to use it for replicating their work, et cetera. So J-PAL does have a bunch of data available.

We put them-- we have our own little collection in the Harvard data dataverse. The address is there. And IPA, which is like our sister NGO, they also put-- they also have a data set. So these are data sets that are coming from specific project. Typically randomized control trials. We'll discuss later what a randomized control trial is.

They are usually panels, simply because most of randomized control trials start with a baseline survey, which is the situation before you're doing your intervention. Then your intervention takes place, and then you have follow-ups. So they are usually panels, although not always. But that could be used to do-- to study any other questions.

What is a bit sad about these databases in my view is that they are not used enough because they are very rich because when you go and see people, you might as well ask them a lot of questions. So we have this very, very big data set, very rich data sets, but then you run-- you write one paper that's typically the effect of what you are-- the intervention that you are looking for. So you use five variables from that super rich data set. And it really could be used to do all sorts of other things. Yeah?

AUDIENCE: So when you say panels, they are a group of researchers or a group of communities that you're serving a panel?

ESTHER DUFLO: A panel is a data set which has the feature that the same person, in this context, the panel could be five people on a day discussing the important issues of the day. But in the context of this class, a panel is a data set that follows the same unit over time. So it's not the panel of doctors. A panel is a set of-- is a data set that follow the same unit over time.

The unit could be households. In that case, you have a household panel. The unit could be a farm. You have farm panel. The unit could be a community, a town, a state, a state panel, a country. But it's a unit that is followed over time.

AUDIENCE: And the unit is the preceding adjective to panel? Say like--

ESTHER DUFLO: Household panel, exactly. Correct, household panel, the unit is a household, exactly right. So from a cross-- from repeated cross-section of people like the DHS has, the DHS is not structured as a panel because it interviewed, let's say, 5,000 people in the country. But it does not interview the same people. So it's not a household panel.

It's not a village panel either because it's not the same villages either. But you could aggregate it at the level, say, of the state within a country, take the averages for each state within the country in each year, and you would obtain a state panel, exactly. So from repeated cross-section, usually, when aggregating enough, you can construct a panel at a larger level of aggregation. Does that make sense?

So in a lot of cases, the J-PAL and IPA data sets are household panel or farm panel. It's not always true. But it's often the case, lots and lots and lots of data. The American Economic Association journals and pseudoscience, by the way, requires posting of any data used for research unless you can not post it for some reason because it was confidential data that you're not allowed to post.

And in recent years, it's been actually pretty good about requiring that people actually comply with that. So there is a lot of data on the AEA website. You simply go to the American Economic Association website and go to the journal. And there is lots of data on this, that, and the other. It's not all done in a way that's very intuitive. But you can look for it because people-- researchers being like anybody else are very sensitive to incentives.

And therefore, most people put the data up when they have to. And it ended up posted in the AEA website. In particular, if you're interested in your paper and thinking, oh, I wish I had the data, before emailing the researchers, you should check to see whether it's not on the data set, for example.

So these are like-- until a few years ago, this might have been the end of-- these are the data that exist in the world. But of course, the internet has led to the explosion of data sets that actually are readily available. And not just data set that you can harvest yourself.

The FiveThirtyEight website, Nate Silver's website, has lots and lots of data on polls, obviously, since that's kind of their bread and butter, but on many other things because they also have lots of data on many other things. And they have a GitHub site where they have, basically any story that they are doing, they have the data that's posted.

So you can go on their website and see whether you're interested in the data. For example, on anything, now that they belong to ESPN, there is also lots and lots of data on sports, for example. Students from this class are very usefully pointed us to the Yahoo Data Dump, that's apparently the biggest data set that exists at the moment. It's something like 13 terabytes of data, on one particular thing, obviously, Yahoo being pretty much essentially a news website, what they have is exactly this, like sample of anonymous user interaction on the newsfeed of several Yahoo properties.

So basically, they have behavior of people on the site. That's lots and lots and lots of behavior. And that's available. You can get it if you have space to store it somewhere. Of course, using that data requires thinking hard about how you can even manage so much data once it's available. But they give some example of what they did with it.

Of course, with that much data, it's mainly machine learning type application. For example, trying to predict patterns someone who did A, B, and C, what is-- what they're more likely to do next, that type of stuff.

There are some sites that-- so these are kind of one off stuff. But there are some sites that are specializing in aggregating data sources. So you can go to them and see. For people who are interested in sports, there is a project by UK researchers called Open Source Sports, which has lots and lots of data on lots and lots of sports.

And then on the NBA specifically, I'm pointing it out because that's the data we are going to use today, NBA Savant-- NBA Savant is collecting lots of data on NBA. There are also data d sites that specialize in keeping track of what happened in the internet in the past, storing it up. In particular, there is a site called Wayback Machine, whose objective is to try to keep record of lots and lots of stuff that happened and that have gone.

So for example, if you're interested in the history of, say, the prices on one particular website moving backwards, you might not-- it's not there anymore. But they might have it. I suspect they are charging for it, though. I haven't gone-- I haven't searched long enough on it to let you know.

AUDIENCE: So a lot of it's free, at least. They may charge for part of it.

ESTHER DUFLO: So that's in terms of anything that-- we are going to talk about web scraping in a minute, which is mostly forward looking, which is could set up a scraper and collect data from now for a while, but if you wanted to look backwards, then the Wayback Machine has it. So kind of generating data set for you from the internet.

And as I was saying initially, there is much, much more. It was just like a tiny flavor of what is already available with minimal or no effort in order to get your hand of it and that you can work with. MIT has a ton of data set, a lot of geographical data, for example, that lives in the Rotch Library. And the Rotch Library, they are very helpful in terms of telling you what they have and also teaching you how to use it, ArcGIS and the like.

The Barton Catalog, if you go on the Barton Catalog and you search for data on cupcakes, CD only, and a year, you will find whatever they have on cupcakes for those years. And you might not find very much. But then you can ask them to purchase the collection and see what they say. And of course, Google is always a resource.

But after all this effort, what if this is not what you were looking for? So sometimes, what you're finding is available, but not free. So your library might be able to purchase it. Or you might need to get an agreement from the place. Sometimes it is free or people would share it with you, actually, for-- the people who are collecting it, which could be a branch of the government, or it could be a survey organization, or it could be a firm will actually share the data with you if they think that there is something in it for them or if it's their job.

But the access is restricted. For example, for confidentiality reason or for example, because that's data that these people collect and then sell. In some cases, researchers use data for project that the company that gathered them actually sell them. But they are willing to give them to researchers.

So if you're interested in using administrative data that requires various kind of permission, J-PAL North America has put a very useful set of resources on their website that tells you how to get access to various kinds of administrative data, how to set up the permission and all of that. So you can follow their tutorial there, lots of resources.

Sometimes the entity that owns the data, for example, suppose you want lots and lots of data from Facebook, they will not just-- it's not there on the internet. They might still be interested in sharing it with you if it's part of a research project, a prospective research project along with an experiment or a retrospective if you're interested in looking at, say, we have-- looking at the history of what people declare on their profile as for their political affiliation over time, I'm making this up.

Then they will have a lot of requirements in terms of where you are allowed to work physically, whether you're allowed to-- what you are allowed-- whether they want to review thing you write before you can post it and the like. But it's more or less organized depending on the company. If it's a very small company, somewhere here in Cambridge, you can just go and talk to them.

Typically, the steps is you're going to have to comply with the partner's requirement for data security. And you have to go through human subject approval here at MIT to make sure that it's OK for you to use this data and the condition in which you're using it are OK. In particular, there is no risk to the subject, et cetera, et cetera.

We don't like people to work with data that has not been anonymized unless it is in a room that is not connected to the internet and all sorts of stuff. So you need to comply with the requirement of the partner and the requirement of MIT. But there's a lot of data that's one step further. It is there. It's not immediately available. You might have to pay for it. Or you might have to convince someone to give it for you that would otherwise want to charge you for it. But it is gettable. And sometimes, it's really not there. So you have to go harvest it yourself.

AUDIENCE: Are you going to say anything about freedom of Information Act?

ESTHER DUFLO: Oh yes, no. So I should have. That's kind of a way of harvesting data. So imagine a third bullet here, Freedom of Information Act. Some data that is not in the internet today really should be there because it belongs to the public. For example, Angela, I think, used Freedom of Information Act to get some of the data on her-- for her thesis. So maybe you can just say it and I'll repeat it for the camera.

AUDIENCE: Data from the Drug Enforcement Administration by Freedom of Information Act request. I do not recommend for this class trying to do that because they are extremely slow and do not like to give away their data for probably all these reasons. And for whatever reason, some government agencies don't have a standardized way to request data from them. It's not like they have a way to engage with researchers. They just say we don't care about you're a researcher from MIT or Joe Schmo who wants information on his files from the Drug Enforcement Administration because he thinks that they've been spying on him or whatever.

So basically, they just put everyone through the same process. And in order to do that, you have to write them a letter following a certain legal format and say I'm requesting this under the Freedom of Information Act, and then you have to go back and forth and appeal a lot. And it's a very exciting process.

AUDIENCE: So just a tiny bit of background, the Freedom of Information Act, I can't remember when it was passed, but it was basically a law that said that government collected data, unless it's confidential, unless there's some reason that the government needs to keep it confidential, should be available to any member of the public who requests it. And so Angela was talking about a cumbersome process to sometimes get the government to comply with this. But in theory, this goes for the federal government, state governments, local governments.

ESTHER DUFLO: Police, it is generally, as a note of background, a lot of what I'm telling you today, I'm not recommending you're doing for this class. Trying to get a data sharing agreement with Facebook is also not happening in this class. It's going to take a little longer. But it is useful to have all of the sources.

Other countries also have the equivalent of Freedom of Information Act known to people who like these as FOIA, for example, India has a Freedom of Information Act equivalent. And you can request data. And they give it. So more and more data becomes publicly available anywhere over time, governments tend to be more transparent than they have been. But what is not there and you want, you might be able to get hold of.

AUDIENCE: What about things like websites like LinkedIn, or Twitter, or--

ESTHER DUFLO: Yes, I'm coming to Twitter.

AUDIENCE: Have their API public, but then they stopped that for a while?

ESTHER DUFLO: I'm coming to Twitter in a bit, yeah.

AUDIENCE: I have a quick question. So the-- it seems that there's more confidential information than non-confidential information on the government and all of the different levels. Is there like a website or someplace that shows what is publicly available?

ESTHER DUFLO: Data.gov has a lot of stuff. So that's a good place to start. And then typically, you will-- the way you arrive at the desire to do a Freedom of Information Act request is that you are thinking about your research project or the project you want to do, could be for a company or a senior project, and you're thinking, oh, I really need some data on the number of arrests that have been made for possession of regulated prescription drugs.

And then you're like, OK, let me see whether the Drug Enforcement Administration has it, or like no, OK, then probably I will need to-- so you arrive this way, usually. At this type of data, you arrive at, you don't start from the universe of the data that could possibly exist, and see what is there, and what you need to get from-- but you're thinking of the data you need, and then try and find it from available sources, and determine that the rest should be accessible, this way or that way.

AUDIENCE: I just had a quick anecdote with regard to the difficulty of FOIA. There's a great story about how somebody wanted but some budgetary data from the state of California. And while the governor didn't refuse, because he can't, he sent back a note saying that the cost to the state government to actually compile the data they wanted would be more costly to the government than if they didn't send it altogether because they didn't have the infrastructure to do so. So that would probably be another try to get other sources as opposed to going to the official.

ESTHER DUFLO: I mean, or you could fight back, saying that you must be joking. Can't get tax-- you can't get budget data.

AUDIENCE: So what's the limit to this Freedom Act? I mean, for example, if it's-- let's say Facebook.

ESTHER DUFLO: Oh Freedom of Information Act is for public data. It's only public data-- Facebook can keep the--

AUDIENCE: Public locations like a company. Let's say they have--

ESTHER DUFLO: It's not public data. Freedom of Information Act applies to public data, government-collected data.

AUDIENCE: It could be demographics. It could be--

ESTHER DUFLO: Yeah, but Freedom of Information Act is government-collected data. So the government does not-- for obvious reason, government does not regulate what companies must make available. Now, companies have their own self-regulation and so-- and some of them do put data out. But they do whatever it is they want.

So anything that is not government-collected data, if it is not there, then you are entirely reliant on the goodwill of the company in question.

AUDIENCE: --in mind because, I mean, the government would have my address, would have my passport.

AUDIENCE: So there are exceptions to FOIA that include identifiable data. For example, I can't get certain data.

AUDIENCE: But there is about there is research that just came out that said with three little pieces of information that are not your credit card, that is not your name, that's not your address, you can easily recognize people from a data set. So this is very argumentative.

AUDIENCE: So it is extremely-- so whether or not something is re-identifiable and whether you will re-identify it is sort of up to the government agency. And if they don't want to give it to you, then they will make an argument that, oh, this is-- I can't get-- I can't get a zip code level data on pharmacies because it might be re-identifiable. But I can get state level data. It's basically coming down to argument.

ESTHER DUFLO: And this goes back to the point that you were making earlier, I think this is the-- unfortunately, many more things become re-identifiable over time because we have much more information from other sources than we used to have. So this is in permanent flux. So another answer they can give you is other than it's going to cost us too much money to generate that information is to say, well, there is no way we can give you that data in a way that protects the people on whom it is collected. So sorry. So that could happen. What is clear is that their duty to even answer you is only for government-collected data.

So the internet has a lot of data. I'm going to very brief-- I'm going to scrape the surface of web scraping, partly because I assume that there is at least more collective knowledge of it in this class than I have, and probably even more individual knowledge in any of you. So I'm going to stay at a very general-- very general level and give you-- my original intent was to have more on collecting data from social networks, that's your Twitter question. But I'll have just a little thing how it's different, why it's different.

So what's web scraping? Again, I'm sure it's more news to you than it is to you. More news to me than it is to you. So I'm going to try to stay at the level where I'm still comfortable. Basically, you could-- there are different ways, different things you could want to do, one of three things.

You could pull data from one page. You could want to pull data from one page. You might want to crawl an entire website for data. You might have data that is a set of forms running in the background and that you want to access those forms. Or you might want to do any of the above in an ongoing fashion, for example, because you're interested in a time series, the same-- you want to construct, for example, a panel, in which case, you need repeated observation over time.

AUDIENCE: [INAUDIBLE] crawling the web is very time consuming and requires a lot of computational power. So I tried at some point in my research to crawl Google Scholar and there's also limitations from the website itself, where Google would kick me out after I've done [INAUDIBLE] a day. So I was only allowed to collect data for 500 people a day. So in order to get sufficient data set, I have to do this for six months, and it was still not as big as I had hoped for.

ESTHER DUFLO: None of that is-- this is type of stuff where I'm going to give you the general principle. And whenever you get into the detail, you're going to find some snags, which you can then bravely vanquish. It's like the FOIA between the general principle that any government-collected data that's not identifiable you can get, and Angela's experience, or your experience in the background, there is a number of roadblocks in the middle. But hey, if it was too easy, then someone else would do this job.

So let me give you an example. Sarah and Glenn, her husband and colleague, wanted to write a paper on what the internet did to the price of used books. So what they want to do at the heart of it is to compare the price of the same used books in-store and online. So for that, they need a panel of prices of the same used books, the same used books title, both in physical location and on the web over a period of time.

So physical location, you have to show up to the physical location. So Sarah goes to the physical location, write down the prices. But for the web, they're actually on the site, which is abebooks. That's how it looks like. One of the title that they wanted is *The Frugal Gourmet* by Jeff Smith.

So if you wanted the price of *The Frugal Gourmet* by Jeff Smith, you'd have to enter it in the form here. And then it gives you back this web page, inspected the web page on the bottom to start seeing what we're looking for. But what you would see if you-- before you inspect it, is just the image.

And what you really want from that image is the title, maybe-- you could use the title or the ISBN number. You want the date. And you want the price. And there are probably several prices on the same day. So you might want to collect all prices for the same day, I assume. And you want to do that maybe several times to get your panel, maybe not every day. I think they did it at several-- so that's what we have.

And that's what, we want a nice table that we can import in R with the name of the title or the ISBN number for lots of titles. The date, and the price or prices for that date. So how do we do that? So that project is a little less ambitious than maybe your Google Scholar project. So some websites, for example, Twitter, will not really let you do that without trying to figure out what it is you're doing.

So they have their own application program interface, which basically export the data for you in exchange of you registering with it. And then you get the data sort of on their terms. But it's publicly available-- it's public data. So you can-- it's visible. So in principle, you could crawl it. But they will give you-- they have an application that is specifically developed to exchange data.

Twitter has one. And Facebook has one. And in fact, many other website has one. But for these big social network website, using their API is the way to go. I didn't know about the Twitter recent things. So typically, you would use those ones to sign an agreement. So you can get-- even get authorization from individual people to follow them, et cetera, to harvest the data from those sites.

So each will have a specific way to proceed. The advantage is that generally it looks like-- once you learn for that particular website, you can reuse. So for example, if you remember the data you saw in the first lecture about the social networks of the Sierra Leone people, they had been extracted from Facebook was using Facebook own API, which is called Netvisa or something like that.

I can tell you more about that later if you want. Or we can-- or we can follow up as needed if people want to work on that. But a lot don't. And even when they do, they don't necessarily-- they keep updating the website and don't update the API that go with it or something like that.

So the website is usually better maintained than the API. So you will typically want to go to the website. So the website, in a sense, becomes your API. That's the-- the web page is what's there in a much more reliable way. So you set up-- you have to set up your own little software that is going to extract what you need.

And there are many sets of tools and an infinite number of tutorials on the internet that will help you extract the information you're looking for on a page. Ah yes, really, you can do web scraping in R. For simple table, actually, it's really easy. There is something called XML Library and a command called read HTML table. And I tried it out before putting it. If you have a table, it's actually really easy.

If what's in the background is a table in the HTML file, you write read HTML, and poof, you get your table. So if that's a table you want, actually, why not? There is a nice video that explains how to do it. Something that I have not invested in terms of figuring out whether it worked, so I'm putting this with a disclaimer, is something called Rvest, which is basically copy and paste of Beautiful Soup for-- I'm going to Beautiful Soup in a minute for R.

So in principle, it's supposed to be doing the same thing. And maybe it's not good yet. Maybe it is already good. And you can try using that. Certainly, if you have easy project, like extracting table, the read HTML will work fine. And this might work better for easy project.

A more conventional way is to use Python. There is some entry costs because you have to learn Python. But the advantage is that the internet is full of tutorials. And because it's more conventional, there is more tutorial for using Python than for using R.

You need Python, which is free and pip, which is going to be used to request all the library that you need. Both of them are free. And then you will need to use this famous Beautiful Soup. So Beautiful Soup is basically a set of tools in Python that are already pre-programmed to say if you are looking for-- that very easily tells him, please find all of those things that looks like prices on the data set.

So you don't have to yourself analyze the text and figure out that there is a class that's actually a price, et cetera. So a lot of this is already-- most of the things that you might want to do are already-- they already have a little comment in Beautiful Soup. So they become-- it becomes relatively easy to use, even if you are not a champion programmer, which I am, I'm not.

So back to the abebooks data website. I've inspected the page. And I'm looking for prices. So I've clicked on the price here to inspect the element. And you can see that there is a class. Price is a class. So all the prices will be identified by this keyword. So it's going to be really easy for Beautiful Soup to search for the prices.

So you will basically instruct them to search for anything that's a price. It's nicely indicated by the thing. So it's not going to be too difficult for you to do that. And then you're going to do it again and again with the caveat that maybe after some time you're going to be kicked out by the website and stuff like that. Again, Angela used a lot of web scraping for dissertation. She's like a living repository of various techniques. So she'll do a bit more in recitation if people are interested on how to get there.

So this is using data that's in the internet for a fleeting period of time and you harvest before it disappears. And then finally, you might decide that you want to collect your own data. You might think, oh, well, it's not feasible. But it's not-- it's not that infeasible. You can collect data from the internet by using survey tools, such as Survey Monkey, for example.

You can install apps on willing participants that will track their movement or other things you're interested in. For example, if you're interested in knowing how a sample of people are commuting or are circulating within MIT, there are free apps that people can download on their phone that will track people's movement and download them and give them to you.

And you even communicate them back to them saying please, today, don't take the T to come to MIT, but take your bike. And then you can see whether people respond to this kind of entreaty or not. That's not a very interesting experiment. But that would be one example of an experiment where you would be collecting your own data. Of course, you need to get approval from the people where you are installing the app, et cetera. But it wouldn't be that difficult to do.

Or you can sit in the Science Center and administer some questionnaires. And of course, if you have more money, you can organize a data collection team to collect whatever it is you'd like. And then you enter into these kind of issues of how to get good quality data and stuff like that. If you want to go this one step, you probably need a little bit of advice. J-PAL is going to put on the internet over the years a data collection course that is going to go over all sorts of methods to collect data.

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: For all of them, for absolutely all of them. This is the steps, steps for collecting your own data. You need to obtain the funding you may need. It might be zero if it's installing an app on your friend's data set. You need to prepare a data collection and data management plan.

So the data management plan, you have examples on the library of what you need to do. But what's very important is how will you keep the data safe? Because you will have individual level data, maybe not highly sensitive, maybe highly sensitive, depending on what you collect data on. But whatever it is, if it's answers on the internet, if it's surveys you collect in the Science Center, you will have individual level potentially identifiable data. So you need to say how it's going to be kept.

In particular, you will need to encrypt it and stuff like that. So you need to develop that plan. You need to decide whether it's going to be shared eventually. And if so, how. Once you've done that, you take that and your proposal for the project. And you obtain human subject approval from the IRB, from the-- so at MIT, it's the Committee on Use of the Human as Experimental Subject. In any other institution, we'll have an IRB.

The IRB is going to check that you're not doing something that's unethical. And also that you're protecting people's-- the safety of people's data. Then you pilot your data collection instruments. You see whether they snags, it works, et cetera. And when everything is in place, you implement. So this is kind of the broad steps of collecting your own data. Yes?

AUDIENCE: So in some scenarios, it says resource limited requirements. You might want to use paper collection methods translate that [INAUDIBLE] that goes into data collection? Are there packages readily available--

ESTHER DUFLO: For writing questionnaires?

AUDIENCE: Well, for translating a stack of paper. [INAUDIBLE]

ESTHER DUFLO: No, what you're doing is that you're typing it up.

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Not good ones. So at J-PAL, we've experimented with a lot of that, as you can imagine. So we used to do everything on paper and enter everything. So basically, you create a software that is-- looks very much like your survey. And the data entry people enter it from that database. And you have to enter it twice because of data entry errors. And then reconcile the two, it's a huge pain.

So we tried to do exactly that. There were, for a while, a company that offered to scan survey. It never really worked too well . And we've switched to a digital data collection almost 100%. So even in resource poor setting, it is cheaper to do digital data collection. Tablets are now cheap enough that it's just not worth the hassle.

AUDIENCE: When you take that out [INAUDIBLE] you create your own application for [INAUDIBLE].

ESTHER DUFLO: So there are applications that help you write your own application. So there are software that are now very well done that you use to create your own application for your own survey. So you can-- it's quite easy to do. So for us, often, we have reasonably complicated surveys. So we have to write-- we check-- but even for a very complicated survey, it takes just a couple of weeks to get the survey very nicely. So it's something that is as easy for someone to use or even easier than a paper form, very intuitive, and people can type on it, et cetera. You can do-- there are all sorts of resources now to create those survey collection forms. Angela?

AUDIENCE: Oh, I was just going to say that if-- I have done it where there's a PDF reader package in Python that works pretty well. So if you have OCR papers that have regular tables, it can work pretty well. It depends a lot on the shape of the data inside of the document. So if it's a pretty regular table, you can extract the table using Python and extract the data from there. I think it depends on--

ESTHER DUFLO: How it looks like.

AUDIENCE: You could easily write a code, if you can imagine writing-- if you had the table in front of you, you could easily write code to generate the data you wanted from it, then you could do it. And if not, then it's probably better to get it out.

ESTHER DUFLO: But the bottom line is this has been leapfrogged. If you're going to collect your own data today, you typically will collect it even via phone or smartphone.

AUDIENCE: I have a question about the protection of subjects. How are firms treated differently than individuals in data protection? And does that vary based on the size of the firm?

ESTHER DUFLO: That's an excellent question. They are treated very differently because the protection of human subject as a subject of research, we're going to go through-- when we, at the time where we discuss randomized control trial and A/B testing, we will go back to that. And I'll do-- I'll do a whole-- not a whole lecture, but I'll do a whole little module of a lecture on human subject protection.

So I'll just give you the bottom line now which is research-- research is the production of generalizable knowledge. Whenever you want to use subject in order to produce generalizable knowledge, this is research. You have one set of regulations that applies to you that are international agreements on how humans-- how research should be-- sort of the burden of research should be equally distributed across people.

The research should be beneficial. There should be a principle of beneficence for the most part. And people-- and people should have the possibility to consent or not to research. All of these things, justice, information, and freedom, and beneficence are the three guiding principles for research subject.

Firms, there are regulations that just have to do with the fact that they can not-- there are very different regulations in the sense that when they collect data on you, you sign a release that is very long and you never read. And then they can do whatever they want that is compatible with that release.

So if someone wrote in a data release, I think I might be wrong on that, or tell me if I'm wrong on that, but if someone wrote in the data release that, by the way, your data is going to be put freely on the internet with all of your identifying information, do you agree? And you agree in principle, you're bound by this agreement. Or in practice, someone would go back and say like, sue you for obfuscation. So there are some limits to this principle.

But there is no overall regulation like there is in research. So firms can do much more than researchers, which has led to the big debate on Facebook works. Because Facebook works, for a while, was not covered by research, by the same protection of human subject as research because they were doing it for themselves. And when it became a problem, in a sense, is when we find out about it from the public because it is shown as a research project. In principle, they can experiment with you all they want.

We are going much even slower than I anticipated. I knew that we would be slower than I had planned. But I have 10 minutes to start looking at data once we have it. So once you have some data, I want now to have a what Sarah talked about, and I talked about it, they come back together, and say we'll have some data, let's look at it. What distribution might have it come from?

So we can start-- what I want to do now is start plotting some data. So how I can-- now that we have some data, we can look at it. We can start by plotting histograms. And we'll be plotting kernel density then estimates of CDFs, conditional distribution functions, and finally plotting bivariate distributions, which can be very useful when the data comes in this form.

So what's a histogram? As a definition, a histogram is a-- formally, it's a function that counts the number of observations that fits into each bin that you've specified. So if n is the total number of observations that you have in your data and k the number of bins, the histogram meets the definition that n as a sample size is the sum of the total sample, the sum of the observation in each bin.

So that's formally a histogram is that. In practice, you typically plot it. So you bin the data, and you count the number of observations in each bin. And then you plot it-- you plot it by plotting a rectangle that's proportional to the number of such cases. You can also choose to divide it by the total number of observations to your data set to obtain the density instead of obtain the-- obtain the number.

So you could either-- so that's an example of a histogram that comes from data that J-PAL data set from an experiment we did in Bihar. And this is female height. So you can see that we've decided we have a certain number of bins. And a number tells us in this case, the density, so the fraction of observation that fits into this bin. The height of the-- the height of each bar is the fraction of observation that fit into each bin.

So this was generated in R. So this is how one could have written it. We get the data in. We summarize it to make sure these are all numbers and stuff like that. Then we keep only the female by taking a subset. And then OK, let's try and do the histogram. PDF, it's telling us that the output is going to be this PDF file. And then we want to plot a histogram.

And if we do that, well, it's not that nice. It's kind of pretty ugly, even. By default, you're not going to have the number of bins that you like, the color is ugly, et cetera. Maybe there is some person whose height is zero that is kind of screwing us up. So let's try to do a little better.

So first, I'm not going to have the zero people because it's supposed to be adult height. So presumably, that's a problem. I'm going to plot the height between a meter 20 and 2 meters, which would be quite large already. I'm going to say how many bins I want. I'm going to given x-axis and the y-axis. And finally, I'm going to use the blue color that we use later. And I'm getting my histogram.

What happens when you use less or fewer bins, you could play with a larger number of bins. You can see that you get more or less precision-- more or less fineness in term of the variables you have. Even with lots of bins, sort of little bit jaggery.

So one next step you might want to take is to say, well, let's try and see whether we can estimate the shape of the distribution in a more continuous way. So instead of doing a histogram, really, intuitively, what I would like to do is to draw a line that goes around the histogram, sort of espouses it nicely to draw a continuous function.

So the kernel density is a nonparametric way to do that. It's a way to estimate the probability density function for a random variable. It's a straightforward extension of the histogram. At each point, instead of drawing a vertical line over this bin, I'm still going to take a bin. But I'm going to take an average of all-- instead of taking the number of observations here, I'm going to take-- I'm going to count the observation in that place.

But I'm not going to-- I'm going to count them over the bin by putting more weights. It's going to be a weighted sum of the numbers, where I'm going to put more weight to the points that are closest to my point. So what I'm going to do is to say-- for each place at which I'm interested in valuing my histogram, and I'm going to do a lot of points to make it nice looking, I'm going to draw a little function, could be any form of function.

But a nice function would be a bell-shaped function. I'm going to draw, suppose I want to estimate the histogram here, I'm going to draw an interval, a symmetric interval around that's going to be my bandwidth. I'm going to draw a little function that's going to be my kernel. And I'm going to count-- I'm going to take the weighted average of indicator function of how many observations I have exactly here.

So instead of the histogram would be a bar over the bandwidth. And this is going to be something that is going to look smoother. So that's what this function over here tells us. So what function do we use for the kernel? We could even use something that is flat, where I'm going to take the weighted average of the bandwidth. Or we could use something that is-- but if we use something that is bell-shaped, it's going to be-- it's going to give us something that is smoother and more continuous.

So the most common one is something called Epanechnikov, which is one of these bell-shaped function. In practice, of course, R does that for us free of charge. What we need to specify is the size of the bandwidth. More on that later. That's the only big choice we need to do is the size of the bandwidth.

And then what kernel we want to use, what function we want to use for the weight, we want to use-- that's in practice, not so much-- not so important. Unless you use a uniform, it's going to-- all of the bell-shaped one are going to look pretty much the same. Epanechnikov is what is often used as default. And you can go with that. The bandwidth is going to matter a lot.

Let me show you what comes out of this. This is the default bandwidth. NRDO is in-- this is the bandwidth choice is over here, BW is bandwidth. That's the default bandwidth for Stata. For R. I'll tell you how they pick it in a moment. And the Epanechnikov kernel. This is what we're getting.

If we pick different bandwidth, we would get different curves. And you can see that the-- picking a bandwidth that is very large, so the one is the-- picking-- the one is the one that's at the top. And then I have enlarged the bandwidth as we go along. So picking a bandwidth that is very large, the data is trying to find-- the function is trying to find data here, when it really doesn't have any because the bandwidth is too large compared to the optimal one.

So the choice of the bandwidth, I'll go back to-- I'll restart next time exactly where I am. But the choice of the bandwidth is dictated by a tradeoff between two things. If your bandwidth is too small, your function is going to look very, very jaggery, jagged. If your bandwidth is too large, you might miss some important features of the data that would have been picked by a larger-- by a smaller bandwidth. And I'll show you an example of that next time.

The optimal bandwidth chooses-- try to weigh in these two problems. So it's trying to weigh in the variance, the smaller, the bandwidth, the bigger the variance, and the possible bias, which means that there is an important feature of the data that you really want to get at that you're missing. So with a too large bandwidth, you're going to-- you risk getting bias. So you don't want too much variance. You don't want too much bias. The optimal bandwidth optimally chooses between the two.

There are any number of recipe to do that. And the one that is optimized by default in R seems like one of the very conventional ones. I don't think there are-- there are many. And I don't think there are some that are widely preferable to others. So the default one in R is one of them.

That should do for most application. But then sometimes, you might want to try with a bigger and smaller bandwidth to see what happens. I'm going to pick up exactly-- pick up from here, exactly here, tomorrow. In particular, we'll discuss more about this application. We'll see other kernels and stuff like that.