UNEDITED TRANSCRIPT

Becker Brown Bag: Learning From Data, Featuring Steve Levitt

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

1126 E. 59th St, Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

MICHAEL GREENSTONE: We're very excited to have Steve Levitt here today for Becker Brownbag. Now, I don't know if this will be part of Steve's speech, but one thing is clear from our organization of this event given that it sold out in about 11 minutes. Looks like we have a lot of great econ students here. Who knows what we should have learned from that? The price was too low, that's right.

[SCATTERED LAUGHTER]

So I want to thank all of you for joining us today. My name is Michael Greenstone, and I'm the Director of the Becker Friedman Institute. Many things you hear are not true. This is true. BFI is a collaboration of the 300 PhD economists on the campus here, and we really have two goals. One is to help foster the kind of research that Chicago has been historically associated with. It helps people understand the world in a new way, and Steve is about as good an example as exists.

And the second is to make sure that those ideas enter the broader marketplace, not just academia in a powerful way. And Steve is also a model for that. So in many respects, BFI is just an effort to try and emulate Steve.

I just wanted to mention a couple things. I think Steve is widely known. But he was awarded the John Bates Clark Medal, which is given to the most influential economists under the age of 40. He was named one of *Time Magazine's* 100 people who shape our world. Very famously, he is a co-author of *Freakonomics* and *SuperFreakonomics. Freakonomics* sold more than 4 million copies.

He also has a Freakonomics podcast which receives 10 million downloads per month. Again, maybe he does not charge enough for that. So we can ask him questions about that. But I just wanted to also say, one could go on and on about all of Steve's accomplishments but in many respects, I think Steve was really at the vanguard of teaching the economics profession, what you could do as people began to gain access to data and computers, and how data computers and economic theory could be used to understand the world better and lend often very surprising insights about the way the world operates.

So personally, I'm incredibly excited to listen to what Steve has to say. Today, I think he's going to be talking about AI and big data, so he's capturing some good buzz words. Hopefully he'll help us understand what they mean. And with that, please join me in welcoming Steve up

here.

[APPLAUSE]

STEVE LEVITT: Thank you, Michael. It's gratifying that so many of you are here, although I know really you came for the lunch, and you're all sorely disappointed that the caterer decided to boycott my presence. And instead, you're left with kind of something second rate. Michael talked about changing prices, so we did in essence change prices and raise the price for being here, but thank you for coming.

Before I start, I just want to mention something briefly. I'm trying to do something very different in the near future. I'm doing something non-academic where I'm trying to take *Freakonomics* kind of ideas, and instead of just writing papers, trying to actually go and change the world. And I'm looking for a handful of impossibly talented people to help me do that, so if you're the kind of person who thinks that a combination of *Freakonomics,* and like a startup culture, and doing good in the world is the kind of thing that might appeal to you, drop me an email or come talk to me afterwards.

OK, so back to the topic. So here's my motivation for what I'm talking about today. I have spent my life trying to understand data, and I've really done it through the lens of how economists think about data. And we have a set of tools, and I've learned from the really smart economists who developed those. And then over the last two or three years, more and more to the point where you literally can't turn around without hearing about modern data science, and AI, and these new tools and how they're going to revolutionize the world. Machine learning.

So I didn't know anything really about those set of tools, and I thought it would be interesting to learn about them. What was shocking to me when I began to study them just a little bit-- I'm definitely no expert on them-- is how completely and totally different the modern data science, machine learning approaches are to using data, relative to what economists have been doing for the last 20, 30, 50, 100 years.

And it creates a puzzle, because I know the economists are not stupid. The best economists who thought up how to use data were smart men and women who knew what they were doing and have generated enormous insight. But I also know that the computer scientists are probably even smarter than the economists, OK? I'll show you how different the approaches

are, but could they be this different because one group is just completely confused and doesn't know what they're doing? And if it's not that, how could it be that two totally different ways of thinking about data have arisen side by side, and they really have nothing to do with each other.

So that's the puzzle that I came to, honestly like a year ago, just wanting to understand. What I'll talk to you about today is my understanding since then of what I've come to understand. So in particular, what I'll do is spend probably too much time giving you a brief history of how economists think about data. And then I want to talk also about modern data science just to put the two into perspective.

And here are the questions I want to try to answer. Why are the approaches so different? Because they really are extremely different. I always put modern. I don't know what to call it. I call it modern data science. It's only machine learning [INAUDIBLE].

Is that going to make the economic approach irrelevant? Because if you just read newspaper reports, I think you would believe that that's true. My answer is going to be no, I don't think it will. And then I want to ask, well if you're an economist, what and how should you borrow from this new data science stuff? And what, if anything, can economic approaches contribute to modern data science? So that's my agenda for what I want to accomplish today.

Let's start with economic approaches to data. It starts with correlation. The most basic thing we have is correlation. I want to give you an example. When I was first starting out as a researcher many, many years ago, I was interested in the question of whether prisons reduce crime. So should we lock up more people, or should we lock up less people?

And so I went in the literature to try and understand what people knew, and I got a little bit nervous. I don't know if you've ever done research, and you start to see the titles of papers, and you think oh my god. People have already done what I've done. My work is going to be completely redundant.

So I stumbled onto a paper that the title was called, "A call for a moratorium on prison building." It was an empirical paper, and it sure seemed like these people had figured out that prisons did not do a good job of reducing crime. Otherwise, how could they have a title like "A call for a moratorium on prison building"?

So I went and I tracked down this paper. It's a 1976 paper from a journal called *The Prison*

Journal. And what they had done is, they had divided states into those states that had built lots of prisons between 1955 and 1975 and those set of six that hadn't built as many prisons between 1955 and '75. So there were the heavy construction states and the light construction states.

And it turned out that the prison capacity were they built a bunch of cells had gone up by 56% and had basically stayed the same in the other states. And then they looked at crime, and it turned out that crime had increased by an amazing 167% in the places where they built all the prisons and only by 145% in the other set of states. OK. And they looked at the data and said, well look. We built a bunch of prisons, crime went up more. That proves that if we want less crime, all we have to do is to stop building the prisons. If we took all the prisons away, think how amazing that would have an effect on crime.

This was published in a peer reviewed, academic journal. And what these guys were doing were using correlation to make judgments about the real world. But in general, we know that's not a good idea. But I still want to say the correlations are the single most important element of data that we have. And why is it? Because a correlation is the only thing that God, nature, the universe gives us for free. OK?

Correlations are everywhere. If you have two data series, you have an *x* and a *y,* it doesn't matter what they are. You could generate the correlation between those two, and those are right there for the taking. And that's incredibly valuable and useful because other than correlations, nature doesn't tend to give us things other than correlations.

The other reason that correlations are important is because I think it's important to be able to draw the distinction between things everybody can agree upon and things that you argue about because of your preferences, or your beliefs, or your techniques or whatever. And the beauty of a correlation is that there's no reason at all that you send out two people from completely different parts of the world or the political spectrum, and you tell them to compute the correlation between two variables, why they aren't just going to go and gather the same data, run the correlation, and get the same answer. So it's nice to have a starting point where everyone can agree, and then let divergence happen after that.

The problem with correlations is that they just aren't useful. They're almost never useful, OK? Because all they tell you about is-- what do they tell you? They tell you what has happened in the world. They tell you that it was cold, and it was dark. You know it rained, and the birds

weren't flying around. They tell you stuff like that, but they don't tell you anything about why. And for almost every question at least that economists care about, why is a critical piece, because economics is all about having theories about the world and testing those theories.

There's virtually no economic arguments that are simply descriptions without an interest in why two things are or are not linked together. So a real limitation of correlations is they don't tell us very much about the things that, ultimately, at least economists care about. OK.

So just a quick divergence. Why are correlations not very useful? It's because there are many ways to get to a correlation. So I'm going to use notation here where this arrow means causality. So you can have a correlation between x and y that happens because x causes y. Two variables x and y, x causes y. When we think about economic models, we usually write them down where we have an x variable and a y variable, and we want to show that x causes y or measure the extent to which it does. That will generate a correlation because if x causes y, then when x goes up, y goes up.

It can also be the case that y causes x. In your world, you think that x is causing y, but in reality, y is causing x. It's called reverse causality. It still will lead to a correlation, the same positive correlation between x and y that you get if x causes y, but y is causing x. So the inference about what happens if you want to go intervene in the world is very different.

But what I didn't say is, so why do we care? You might say, why do we care about the causal arrow? We care about the causal arrow if we care about public policy. Because in public policy, the idea is you have some variable you control. That's the next variable. That's how much you spend on education and public schools, or how much transfers you make to the poor, or whether or not you make marijuana legal. That's like an x variable.

And then the y variable are the outcomes, like are there lots of car crashes, or do wages go up, or do teenage girls get pregnant? Things like that. Those tend to be the y variable, OK? We don't directly control those y variables. We control the x variables. And so we care because the only lever we can push on is x. And if we push on x and it causes y, then y will change. But if we push on x, but it's really y causing x, then pushing on x doesn't do anything to y at all. It doesn't do us any good. OK.

You can also get the more complicated situation where there's some other variable that both causes x to change and causes y to change. And so then again, if you start pressing on x, x is not doing anything directly to y. You've got to push on the z, not push on the x. But if the x is

the thing you control, it doesn't do you any good. OK.

And then, most complicating is a case where x causes y, and y causes X. I think a good example of this is police and crime. So lots of reasons to think that police, which is like our x variable, we can control that. If you have more police, you're going to have less crime. But it's also true that when you have a lot of crime, you tend to hire lots of police. So y is also causing x, right? So we would not have people in blue uniforms standing on almost every corner in Hyde Park if we didn't have crime, right? The crime is causing those people to stand on the street corner. And hopefully, the people standing on that street corner is also causing less crime to happen.

But again, just a correlation in here will not help you understand exactly how much x is causing y and exactly how much y is causing x. The correlation is just a starting point for what comes next. And so in terms of public policy, you somehow have to have a way to start to think about inferring causality from what you see in naturally occurring data, which is just correlations.

OK, and we have a set of techniques that economists have developed. And this is really what I mean, in many ways, by the economic approach. It's a set of techniques economists have come up with to try to deal with the fact that the world only gives us correlation, but what we care about is causality.

The one that's kind of closest to my heart is what we call natural experiments, or what I actually prefer to call accidental experiments. I think it gets at the idea better. And the idea is super simple. The idea really goes back to randomized experiments. I will talk later about randomized experiments, but let me start by talking about randomized experiments.

So why is a randomized experiment so amazing and so powerful? What is a randomized experiment? You take a set of people, like this room, and we randomly draw out some of you to put into a treatment group, and the rest of you go in a control group. And then we give some treatment to one part of the group and no treatment to the other. We come back later, and we see what's happened to the outcome.

So maybe some of you, we give this new flu medicine that's been approved. So if you come down with the flu, we'll give you some flu medicine that's been approved yesterday by the FDA, the first one in 40 years. And if you're in the treatment group and you get sick, we give you that. We see how much better you get. And the other group, if you get the flu, we do nothing for you. And then we see how sick you get while you have it.

So the beauty of it is that because we've randomized you into treatment and control, if we didn't do anything to the treatment group and treated you exactly the same, when we came back later we would expect that on average, the people in the treatment group should look exactly like the people in the control group. And so all we need to do is compare the people in the treatment group to the people in the control group afterwards. We've now given the treatment group some treatment, and any difference we see, we think we can attribute to the treatment, the intervention.

So the real power of randomized experiments is that you expect that other than the treatment, the treatment and control group would have looked the same, OK? And that's the exact same logic that economists try to exploit in natural experiments or accidental experiments. We try to find two groups of people who we think would have been the same, except kind of by chance, one group got treated differently than the other group. OK?

But not with any real reason or logic, really by accident. And so the only difference between an accidental experiment and a randomized experiment is that, the experimenter doesn't get to choose who gets treated or not treated, or even how much they get treated or what. You just kind of snoop around in the data, in the archives, and you try and find some examples of this.

And so in general, what's interesting is the best accidental experiments arise out of the greatest stupidity. Because in general, we think in a sensible world, you should treat people who are the same in the same way. And that's the enemy of the accidental experiment. In the accidental experiment, you want to take two people who are truly identical, and because somebody blunders, you end up treating one very different than the other. That's kind of the logic of it. So how does that happen?

Examples are things like law changes. So you know the law change just happens to come in on a certain day. And some people are grandfathered in, or some people aren't. Some people live in a particular state, and others are just over the border in the next state that doesn't pass the law. Or the law says that if you are over the age of 65, this applies to you. If you're not over the age of 65, it doesn't apply to you. So you've got people near the edges.

Arbitrary rules, one example I've looked at is it used to be airlines used to have sales. And if the distance of the flight was 199 miles or less, then they would charge you $99. If the flight was between 200 and 999, they'd charge you some amount. And they had these really sharp

divisions, depending on whether the flight happened to be 999 miles or 1,001 miles. It didn't really make any sense business wise. It was just a rule they used, and that induced a bunch of variation in prices for two flights that were pretty much similar.

There's an example where you can look and see. And anywhere really where you see sharp discontinuities. Now let me give you an example in practice where we use this, and that's in Uber.

So Uber turns out to be a great example of a natural experiment. So you all know Uber. I don't know if you guys used Uber long enough ago where you remember surge pricing much more clearly. So it used to be, you would open up the app and it would literally tell you that this ride is going to cost you 2.5x what the ride usually costs, OK? And then a bunch of economists, including our own John List, went to Uber and told them that's a totally idiotic way to do it. Because when you make it that transparent to people, they're going to react badly. So now that's all hidden in the background. It still happens. You just can't see it anymore.

This on the x-axis is what Uber thinks is the exact, true price that they should be charging. And it goes from one all the way up here just to 2.4. It goes much higher, but the data gets sparse. And on the vertical axis, what we have is the share of people who, after they open Uber app, actually end up getting in a car and paying for a trip. So we've divided the data into really fine slices, OK? So anyone who is between 0.2 and 0.3. They never charge less than one, but there's lots of times when they think the right price would actually be less than their regular price.

So the bars here that are just plain white means that is not a bar where discontinuity has happened. So over these bars, there's no actual change in price. Wherever you see the white bars, there's no change in price occurring, even though the Uber model suggests that you're a little different. OK, so you can see there's no change in price as you move here, and there's no change in the purchase rate. So it's like super flat because all of these people kind of do the same thing.

And then where you see a red bar is the one, little, tiny sliver of data we have that's just before the discontinuity, where they suddenly raise price. And the yellow bar is the data just after the raised price. OK, so the people in the red bars and yellow bars are almost identical in the eyes and ears at Uber, but they face different prices.

What's so cool is when you go white bar, white bar, white bar, nothing happens. Then you go

red bar to yellow bar, and you see a big drop off. And you see again, white bar is pretty flat, drop off. White bar's pretty flat, drop off. OK The red bar to the yellow bar is showing you the response to price of people to these changes in price.

When you have that and build that data back into the usual way we picture for demand curve, this turns out to be this now legendary, in my own mind, demand curve that shows you what I think is the first real demand curve that we've ever seen of something you care about. But if you add up the area of the demand curve, that gives you consumer surplus. And by our estimates, the consumer surplus that was coming from Uber in the year of our data a couple of years ago was at almost $7 billion, which turns out to be huge relative to Uber's profits, which were probably -$2 billion dollars after the amount that's paid to the drivers, which is $3 billion or $4 billion.

So really, this turns out to be the single most important component, this consumer surplus. It's kind of not surprising in a way. People love Uber. People use it all the time. That's an indication that they're getting a lot of surplus out. But what's surprising maybe to us was that, the willingness to pay for Uber was something like three times the amount of the ride. So if you pay like $7 on average, if push comes to shove, consumers would have been willing to pay like 20 bucks for that same ride. And so there seems to be huge consumer surplus, which also is kind of odd if you think about it, because this is all consumer surplus in a world in which Lyft is also out there. So even given the other options, people seem to get tons and tons of consumer surplus from Uber.

Let's go past out now to talk about structural estimation. I don't want to talk much about it. It's a different approach, one that a lot of resources in economics are currently invested in. And what it tries to do, usually in the absence of really good accidental experiments or randomized experiments, it tries to use economics theory or maybe just functional form assumptions if you have to, to tell you given the data you have, how could you then come up with deep structural parameters that actually describe the causal mechanism in the world? And you could then extrapolate those to other settings.

It's a totally different discussion talking about how successful economists are in doing that, but it's a set of tools that economists use for doing that. In the nature of time, let's skip the third point. And then the last point is randomized experiments. OK, so economists were late to the game in terms of using randomized experiments. They've been used in agronomy and psychology for a long time, but only have they been really, really common in economics in the

last, say, 20 years.

But they've come to be a really powerful tool. There are randomized experiments in the lab economists use. But more and more, economists think about randomized field experiments, so experiments that are done out in the field that are creating insights about stuff. The one I want to talk about just super quickly was my own rather odd randomized experiment where I was interested in the question of whether or not people quit too much or too little. Whether they say, quit jobs or end relationships more or less than they should.

Economists have so much to say. We have all these models about how people should behave, but we don't really have a lot of evidence about how people do behave. So the thought was how could I figure that out? And it's not easy. So let's say I want to think about divorce. How would I figure out whether people stay married too long or too little? Let's say you take two sets of people, and they're right on the margin for getting divorced. One set does, and you see whether they're happier six months later than the people who don't get divorced. That would be the thought experiment.

How do you think about how you're going to do that in the real world? Because number one, I don't get to divorce and not divorce people. I couldn't really think of a good accidental experiment. But what I learned actually-- Dubner does his Freakonomics podcast. And after he does podcasts, 1,000 people write to me and say how they changed, how his podcast changed their life, because we share a joint Freakonomics email address. And it suddenly occurred to me, Dubner changes people's lives. Perfect! OK, I'm going to take advantage of the fact that people listen to Dubner, and I'm going to try to change their lives as well.

So what I did is I built the web page, and we advertised it. And it was called Freakonomics Experiments. You can go there. And we advertise, look, if you're having trouble making a decision in your life, come to our web page, and we'll help solve your problems. Also, we kind of pretended to try to solve people's problems by having them think differently and ask different questions.

But what we really wanted at the end of the day was for them to say, I'm still just as confused as I was before I got to your web page. And then we said, OK. We'll do you the ultimate favor. And we had this beautiful, virtual coin that would be tossed up in the air. If it came up heads, then you would get divorced, quit your job, get a tattoo, whatever. If it came up tails, you wouldn't.

And what was so interesting is that 25,000 people came to the website and flipped the coin. And even more interesting, they actually followed the coin toss. So here's another picture which is, for me, one of the most interesting, gratifying figures I've ever made in a paper. OK. So here before they flipped the coin, we asked them how likely they were to get divorced, quit their job, whatever. And they could give an answer anywhere from zero to 100.

Then we randomly assigned them, flipped the coin, and the people who got yes make the change, they're in the green line. And the vertical axis is how many people actually made the change. And if you got no don't make a change, then you were the orange line here. So you take people who said in the beginning, I am not going to change no matter what. They still flipped the coin, and it turns out that something like 20% of them who got heads ended up making the change. And only maybe 12% of people who didn't get heads made the change.

What's interesting is, across the entire scope of how likely people said they were to make the change, the green line is well above the orange line. The difference between these two lines, this vertical distance is how impactful our coin toss was on what people actually did in their lives. This is all assuming reporting, so there's a lot of issues floating around that we deal with in the academic paper, but I really believe this to be mostly true.

And what's interesting is that people kind of knew themselves. The people who said they were likely to make a change, they were much more likely to make a change, on average, than the people who didn't. But people were far too extreme in thinking they would for sure or wouldn't for sure make the change.

But then most interesting is we followed them up. And six months later, we asked them. Some questions say, should I go to this movie or that movie? But anything that was important, the people who got heads were happier six months later and were more satisfied with their choice and would do it again than the people who got tails. And because I can't think of any other reason why the people who randomly got this virtual coin to turn up heads are any different from the people who got it to drop tails, the only logical conclusion I can make is that the people who got heads were both more likely to change, and they're happier. So I think that the causal arrow goes from making a change to being happier, subject to a bunch of caveats you can read about in the paper. But I still think that it's actual, true result.

So what's good is it's changed the way I think about the world. And whenever anybody asks me any question about anything, I always give the same answer, which I tell them to quit. If

you ever are on the margin at all, you should change what you're doing because at least according to this evidence, people are way too reluctant to change. If you're not sure whether to do it, you'd think it would kind of be the same outcome whether you change or not. But it looks like people don't change or quit stuff nearly as much as they should.

I've talked too much about economics. But that's what economists do, more or less. They do things like randomized experiments, natural experiments, run regressions. That's the way we think about the world. Modern data science has a whole bunch of different names. So things like random forests, and cluster analysis, and deep learning. And it's not just that they have different names, but their techniques are completely and totally different than what economists do. And the easiest one for me to explain in a short amount of time is random forests.

OK, so let me just explain as an example of how they do stuff. The basic idea is that you have some outcome y, and you have two possible variables that can explain what happens in x1 and x2. And what you do is, you start by building what's called the decision tree. So a decision tree is the following. It says, I'm going to do a cut of my data. I'm going to cut my data into two pieces, and I'm going to choose the pieces I cut them into based on the cut that will maximize the difference between the two pieces of the data that are left. So I basically slice part of my data to give me now two data sets. And the characteristic of these two pieces is that they're really unlike the other one. That's the first step in my tree.

The next step in my tree is, I look at what's the next cut I can make? So now I have to two data sets. So I look at the cut in the first data set that will make the biggest difference between what's left in that data set, or I can cut the second data set. And I find the one that makes the biggest difference between those two. And I just keep cutting up until some point of stopping rule that I've defined about where to stop. That's what a decision tree is.

So here's an example where you start, and I'm not sure we can generate the data. I think we just generated fake data here. x1 runs from 0 to 1, and x2 runs from 0 to 1. I think their uniformly distributed between 0 and 1. And then we have some y variable. The y variable is equal either to 1 or zero. When it's equal to 1, it's a blue dot here. When it's equal to red, it's a zero. So the blue dots are 1 in the y variable, and the red dots are zeros for the y variable.

And so this is just our universe of data that we created. And when you start making the cuts, the best cut you can make in the data that explains the most is divide the data between cases where x2 is less than 0.3 versus where x2 is greater than 0.3, so that would be this line right

here. So the first cut of the decision tree split off this part of the data from this part of the data.

Then the next step is, now you've got where x1 is less than 0.88. So then you now cut this part from that part. And you just keep on cutting the data until you get to some you can't explain very much variation. That's a decision tree.

So to an economist, this is a somewhat bizarre way to think about cutting up data. It's a hierarchical type of thing. But what's most interesting is what you're left with at the end. I don't know if you can read these numbers, but you have this block here. And in this range, the mean value of y is 0.123. y is almost always zero, so almost all these dots are red. But look right above it to all the blue dots right there. So on the line right between crossing over from here to here, there's a huge change in your prediction, right? If you're just here, you're going to predict that it's going to be blue for sure. If you're just here, you're going to think that it's likely to be red. It's super nonlinear.

So the defining characteristic of this and really all of the new modern data science techniques is that they're enormously nonlinear in the sense that small changes in your variables can lead to really radical predictions in what's going to happen in the world. Whereas in general, almost all economic models have a kind of feeling of linearity built into them. OK?

So this is a single decision tree, but what a random forest is, is you take subsets of your data. You have a million observations, and you take say 100,000 observations at a time. You pull out 100,000 observations, you build a decision tree, you put those 100,000 observations back in. You build another decision tree, you pull another 100,000, build another tree. And you do a bunch of those until you've built a forest of decision trees.

And then what you do is, you let each decision tree gets to vote. So now you get a realization, out of sample of some particular sets of x1's and x2's, and then you let each decision tree have one vote about whether it thinks that y is going to be a 1 or a 0, and majority rules. So that's the way these models look.

So I didn't really explain to you the nuts and bolts of how we build econometric models, but I just gotta say it's nothing like this. There's zero familiarity to anyone who's done econometrics in what happens in this kind of a model.

OK. So then that's where you just say, well this is super weird that these exist. So now let me kind of give my sense of what modern data science, the core of it, is. The core of it's, it is

theory free. OK? It's just a way of cutting up the data, and it doesn't care why you're cutting up the data. Whereas econometrics at least pretends to care deeply about the whys that are underneath it. You at least pretend that you have a model in mind when you go and run the data, and that model is what you're testing.

Interestingly, it's focused almost all on correlation and pattern recognition. And pattern recognition is just another way of saying correlation, right? So these tools, empirically, turn out to be incredibly good at telling you whether it's a dog or a cat. Better than econometric models of trying to get the features of a picture and then tell you whether it's a dog or cat. These things figure out whether things are dogs or cats.

They're obviously black box-y, right? You plug stuff in, and even when you get the answers out, you often can't exactly tell why the model told you. It tells you it's a dog or a cat, but you don't really understand why the model thinks it's a dog or a cat. But the fact is when it comes to the kind of problem they've been applied to, they are really, really effective. When economists try to build a model that does what these models do, using our tools, we're not as good as these tools are at doing it.

So now back to the key questions quickly, because I do want to leave you a chance to do questions. Why are the two approaches so different is the first question. And the first one is, in part, because there's a different mindset of how economists think about the world and how computer scientists think about the world. But I don't want to oversell that because I think that's actually the least important reason why these things are so different.

The real reason they're different-- and once you see it, it makes it really obvious-- is that they were designed to do two totally different things. What economists try to do with data is we try to explain why something has happened in the past. We take a historical data set, we try to understand it, and we try to put the reasons and the causal arrows into it. When people use the data science approaches, they almost invariably are only interested about predicting the future.

So when Netflix wants to figure out what movie you are likely to like, they do not care at all about why you're going to like that movie, or what your background is, or what would happen if a completely different set of movies were released. All they care about is, can they put a movie in front of you that you're going to like so you keep on sticking with Netflix and not doing something else?

And so interestingly, it turns out that predicting the future in a static world is roughly the only interesting question that exists where correlation is good enough. As long as the world is the same tomorrow as it is today and I know that, who cares why today turned out the way it did today? Tomorrow's also going to, more or less, turn out the way today is, too.

And the premise of modern data science, essentially, is that what happened in the past is going to happen in the future, and so I don't need to understand why. So it's interesting. It was surprising to me. It took me a while to see that. Modern data science is the king of correlation, right? And so it is the king of predicting, as long as you're predicting the world where stuff's going to be the same.

The last thing I going to say, the other reason that they're different is that modern data science only ask questions when there's enough data to ask using modern data science approaches. It turns out modern data science is very greedy about needing data relative to old economic approaches. And so there are lots of questions you wouldn't even think to try to answer with modern techniques that economists are quite content to try and tackle. We don't need so much data because we impose more in the way of theory on the thing.

So will modern data science make the economic approach irrelevant? OK, so the obvious answer to that question is no. Because for the thing that econometrics does, which is explaining the past, modern data science has not prove itself to be particularly useful. But the fact is, if you go and think about it, mostly people don't care that much about what economists do, right?

So it turns out in business and in practice, predicting the future turns out to be a lot more important than explaining the past. And economists just haven't been in the business of explaining the future. Predicting the future is not our forte. It's not what we think our job is. And so I think we have lots of job security as economists because businesses don't care about understanding why the past is the way it is, so economists can keep on doing it.

So what should we borrow, though? It seems kind of sensible. There's all these new techniques. What is in it that economists can use? In those rare cases where economists are actually in the business of prediction, I think it's silly for economists not to use some data science tools because they are empirically almost always turned out in a horse race to do better than the econometric models, in doing pure prediction things when you have really thick data.

The other thing that I think is more holistic, but is the bigger value to economics, is that modern data science has had a spirit to it which is just that everything is data. Words are data, images are data. You can basically exploit anything and turn it into data. And that is a really smart and important idea that economists didn't really latch onto. Economists mostly thought that data was the stuff that the government produced and put into data tapes and books. Always, it was numbers. And always it was rectangular in the sense that the way data was structured was, there was a state, and it has a bunch of variables about it.

But anyway, it's super important that even stuff like voice and face, everything is data. And that's the thing. If you are a researcher or someone using data in business, your mantra should be, everything is data. It's just a matter of whether I'm clever enough and have the right tools to turn that thing into data that I can use to be effective.

There's a little stuff about how we can do things slightly better and binning-- I don't care about that. But the fourth thing that I think is potentially harder to see but interesting is that, if modern techniques get good enough-- if modern AI becomes better than humans at doing what humans do, then we don't need economists anymore, right? Because they'll be better. But more subtle is that these techniques might be really good at brute force looking for natural experiments, and then turning those up so that economists could then try to use the human element to try to determine what are their next experiments. Because what these techniques do a lot of is, they look for these highly nonlinear things.

So they might say, for some reason, patients who have these seven characteristics all at the same time tend to die a lot, even though if you only have six of them, you don't die any more than other patients. And then that might be something that would lead economists, or doctors in this case, to say, OK, can I think of a theory why those seven things together might be a sign of some underlying causal thing? So we might be able to use techniques in that way.

So what, if anything, can economic approaches contribute to modern science? I think this is actually a more important question, because I don't think people are asking this very much as opposed to a lot of people are asking in the other dimension. In a number of cases, I've been asked to build economic models in settings in which other people have built data science, modern data science models. And what's interesting is their models always beat my models in terms of prediction, but the correlation between my prediction and their predictions is actually really low. It's often 0.5 or 0.6.

And if you have two signals that are not very correlated, if you take a weighted average of those two, you could actually do better than either signal alone. Now you put more weight on the data science one than you do on the econometric one, but having multiple, relatively independent signals is really, really useful in prediction problems. So a couple times, we worked with companies and were able to convince them that some kind of old fashioned approach has really had real value for them, because it gave such different answers relative to the kinds of answers they were getting with newer approaches.

The other thing is that in a world that's changing, the world could change faster than data are generated that the modern techniques use to keep up. The model can't keep up with the changes, and that's another case where you can obviously see why you'd want to use the old techniques which are more anchored in causality. If you have a model anchored in causality, even if the world changes, you can often use theory to make a prediction about what will happen in the new world.

We're in a new world, right? So if people stop watching regular movies, and they only watch virtual reality, 3D movies overnight. Their old models aren't going to predict much of anything until they have enough new data on the virtual reality stuff, but economic models might tell you something about what the transition might offer.

As I said before, if there's enough data around, it's also true that if you're faced with no estimate at all because you don't have enough data to do modern techniques, you can fall back on the techniques economists now use.

My last point, which is almost unrelated, the most fundamental insight that's come to me over time is that, oftentimes when you talk to people, there's this idea that big data is kind of enough by itself. here Just having big data solves all your problems. And I have come to believe that that's not true at all. In fact, what I've come to believe is not only is big data not an answer by itself, but there's actually a complementarity between ideas and big data, not a substitute.

If you have big data and you have great ideas, you can do amazing things you couldn't have done either without ideas or without big data. But just having one or the other, ideas and no data or data and no ideas, I think actually leads you to a terrible place. I've worked with a lot of companies. I have not worked with a lot of companies that have a lot of ideas and no data.

I've worked with lots of companies that have a lot of data and no ideas, and I can tell you I've

never seen a good outcome from the use of data at the companies that simply don't have any ideas but have amazing data that they imagine should have answers to everything, but they literally cannot think of the question to ask. And so what they do is they stockpile terabytes and terabytes of difference of data, which becomes a bigger and bigger burden on them how to keep track of it. And they never do anything with it except draw incredibly terrible inferences which end up leading them to make bad decisions, worse than they would have made if they hadn't had it.

I think the future belongs to companies, and situations, and people who have big data. I think big data is incredibly valuable. And it will make those of you who also have great ideas much more productive as well. So let me stop there.

I only have just a few minutes for questions, but I'm happy to take them on whatever you see fit. We have microphones here if anybody has anything to say. So we can either let people go, or I can just keep on talking. Should I tell some stories?

[LAUGHTER]

Here's someone. Here's a brave soul.

AUDIENCE: Hi. Hello? Yes. I have a really brief question. The way you described the coin tossing experiment, I feel like it was counter-intuitive steps that you've taken from sort of sampling and the way you wanted to approach gathering data. I didn't know you can do that. And then the conclusion that you drew from that data were really non-intuitive. So how did you think about the process? Did you have a method? Did the experiment kind of have a life of its own as you've gathered data, and did you change some things about it as time went on?

STEVE LEVITT: The question is, so I was kind of surprised at what seemed to be logical jumps and fallacies in what you did in the coin tossing experiment, and did it emerge whole cloth or did it change along the way? In fact, here's how my thinking went. I've learned this from Michael, actually. So Michael was the first one who ever told me to do this. Whenever a student is presenting results, we'll say to them, if you could have exactly the data set you wanted, the perfect data set, what would it look like? You still ask that question, Michael?

So Michael asked that question. And I thought to myself, what is the perfect data if you want to understand about quitting your job and should you do it? And what I would want is, I would want people who are exactly on the margin for quitting their job. I want the people who wake

up every morning and say, my God, I hate my job. I want to quit my job. But they just said sometimes they quit, sometimes they don't. That's, in essence, a perfect data, because I don't want to fire a bunch of people who love their jobs and see if they're happy or not, because they're not on the margin. It's all about people on the margin, because whenever you're having trouble making a decision, you're on the margin.

So what I wanted to find was a pool of people who were so undecided about whether they wanted to get divorced, or quit, whatever, that they would come to my idiotic web site and flip my idiotic coin. So that is the definition of somebody who's on the margin. If a virtual coin can sway your decision, you are really, truly on the margin.

[LAUGHTER]

So for me really, it was the perfect data set. Now what was not perfect about it is that I couldn't really sit on the shoulder of these people or get inside their head and know how they really felt. So I had to rely on them telling me how they felt. I had to rely on them responding to my survey. Not everybody, when I contacted them six months later, wanted to come back and talk to me.

Now apparently, I dealt with that by being clever upfront. And when you said you were going to flip a coin, I also asked you to name a friend who you would also tell about the decision, who then I could go and talk to and ask. Did he really quit his job? Did he really? Is he happier, or is he lying to me? So it's good that I had these third parties who didn't have the same incentive to lie. Not like anyone had a very strong incentive to lie.

But I think one thing we've learned in experiments is that experimental subjects try to do what the experimenter wants. So when people were told to quit their jobs, I'm afraid they might tell me, I did quit my job, because they wanted to make me happy. And then if I asked them later were they happy, maybe they would tell me they were happy because they thought I wanted them to be happy. But the third parties who didn't even know who I was and why I was writing them, I thought, were much less likely to want to come out and say, oh yeah. Lie to me and say this guy was really happy, and this guy quit his job and stuff like that.

But in essence, this is one of the unusual papers I've written where from the very beginning, I knew exactly what I was going to do, exactly how to analyze data, and what I expected to see, as opposed to usually, I just get a pile of data and start thinking about it. And I kind of worry about the details of it. Let me take one more quick question, and then I'll let you guys go.

AUDIENCE: Thank you so much. I kind of wonder, can you elaborate more about your startup? What is your startup trying to do?

STEVE LEVITT: So I've been really lucky, right? So for an academic economist, I've had an incredible ability to talk to people, the opportunity to talk to people. So a book that's highly read, this podcast. I can get 25,000 people to get divorced or not get divorced.

[LAUGHTER]

But what's interesting is, I cannot point to any successful change in public policy, for better or for worse, that I think has happened because of my research. Most of the outcomes in the world, at least directly, would be exactly the same if I had never been here at all. It's a telling statement about the inability of at least my kind of research to have an impact. And in some sense, should I care about that? I'm not sure if I should care about that or not.

But I've gotten to an age where mostly, I think it would be fun to try to go and change the world and do stuff. In particular, my own view of the world is that when things are easy to do, people go and do them. And when you will be showered with accolades and love because you go out and do something that feels really good, then people go and do them. So I think a lot of the easy things that philanthropists can do have already been done.

I'll give you this. I talked to one very prominent philanthropist. He was talking about a problem, and I suggested a solution to that problem. He acknowledged that would probably would be a good solution. And he said, but the thing is, the only reason I do philanthropy is because I want people to like me. And if I did what you said, people wouldn't like me. So I'm not going to do it.

In a world in which that's the case, I think there's room for someone like me who doesn't really care whether people like me or not, to go out and try to make the world a better place by doing the kinds of things that other people like, who care maybe more about their reputation. Or maybe they want to run for president someday, or whatever, that they don't want to do those kind of things.

So essentially, what my startup wants to do is to find really smart people and to take ideas we have. I believe that there are amazing ideas out in the world, and sometimes good ideas are attached to people who want to spend their life pursuing those ideas. But more often than not,

it's like people like my dad. My dad's like a happy doctor who does his thing. But I'd say over my lifetime, my dad has told me five amazing ideas that I think could potentially have, in very small ways, some world changing. My dad's not going to go and quit being a doctor and go start something.

So I think we have the possibility to crowdsource really amazing ideas that other people don't have access to, and then glean the best of those and really go and put them into practice. But the problem is that, ideas alone don't win. What I've learned is that you can't go to a funder or employees and say, I have a great idea. You have to go with more. You have to go with something like a prototype. You have to go with an example of how it works. It has to be really simple.

And so what I want this organization to be able to do is to take ideas and turn them into the equivalent of a business plan, or a prototype, or a device that can then actually be there for people. People don't often have imagination, myself included, about what we are building. I thought, why would I want a cell phone? What could I possibly do with a cell phone? My life is perfect without a cell phone. Why would someone make a cell phone? 10 years later, I can't even imagine life without a cell phone. But someone had to put a cell phone in my hand, or I had to see friends use cell phones to see it.

So in essence, that's kind of what we want this startup to do. But we're not going to be massive implementers or manufacturers. We want to be able to sell, in an intellectual sense, our ideas to the kinds of people who would go and implement them, or would want to take them and do business. That's it. Thank you so much for your time.

[APPLAUSE]

MICHAEL GREENSTONE: That was a terrific speech by Steve. I just wanted to take one part that was my favorite part. Steve has this incredible manner of talking where, it's not a big deal, I didn't do, everyone's doing it. He said, that's what economists would do. And so I just want to underscore Steve. Among the many things he's accomplished, he got 25,000 people to flip a coin and change their life.

I think there's something larger to take from that, that I just want to underscore for all of you. It strikes at Steve's willingness to think about the world in a way that, the world doesn't have to be the way it is. And that, as he was saying, theory and ideas combined with data can often

allow you to alter the world or shape the world in a new way.

And so I think it was just a terrific speech. And the lesson I don't want to be missed is that I encourage all you to think about how not to be afraid about thinking about how the world could be different than it is and change it in some way. So thanks, Steve.

[APPLAUSE]