Ethics of Big Data
What Happens Next - 07.29.2023

Larry Bernstein:
Welcome to What Happens Next. My name is Larry Bernstein. What Happens Next is a podcast which covers economics, education, and culture.

Today's Topic is Ethics of Big Data

Our speaker is Dick De Veaux. Dick is the C. Carlisle and Margaret Tippit Professor of Statistics at Williams College and the author of the college textbook Intro Stats.

I was Dick's student at Wharton in his Statistics 1 class in 1985. For those of you that know me personally, I have a very loud and distinctive laugh, and when Dick started to tell funny stories in class, I was laughing out of control. Dick said to me, "you, front row, center, now!" It turns out that Dick was also in the performing arts and was working on some new material.

I loved Dick's class because he effectively used storytelling to interest his students in statistics, while at the same time exploring the limitations of statistical analysis.

At Princeton, Dick won the Lifetime Achievement Award for Exceptional Dedication and Excellence in Teaching.

Today, I've asked Dick to discuss the problems with big data and the algorithms that we interact with each day in finance like with credit cards and mortgages as well as with crime prevention and parole boards. Let's begin the podcast with Dick's opening six-minute remarks.

Dick De Veaux:
We're going to talk about ethics and big data. And I want to start by saying what I'm not going to talk about, because the topic of ethics and data in general is huge. Anybody who goes on the web these days knows that there's now windows that pop up that say, "what do you think about these cookies? Do you want to accept them all and notice that if you accept them all, then there are 20 pages of stuff you've just committed to? Or do you want to reject them all or do you want to itemize these?" This is all because the privacy and security issues of data themselves is huge.

I live in Europe about a quarter of the time, and the rules in Europe are very different. In fact, they drive a lot of what's happening in the States. I'm going to plug one of my colleagues, Andreas Weigend, who has a book called Data for the People. He talks about how you should get control of your own data. That's really important for anything to do with your financial data, your health data. If you've ever been rejected for a loan or for a credit card you should be able to see the data that they base this on.

I want to talk about the ethics of doing data analysis. This is what the data scientists do. What are their responsibilities? Part of the problem is that algorithms are making decisions for us. Algorithms can be great for proposing solutions and for giving options for solutions. But I get very worried when the algorithm is making the decision without human intervention.

So where does some of these problems come from? Part of it is in the collection of the data. If you don't collect data that are representative of the population, you're going to get biased results. How are the data being collected these days? It is complete chaos. It's just collected all the time. We have no idea what most of these databases represent, and we're making inferences on them. The other problem is accuracy. How accurate are these data? I talk to some data scientists who are on the data collection and I ask them about data accuracy, data quality, and they say, “that's not my job. Who's got time?”

Another problem related to all of this is the transparency of the model. If it's a black box and it says, no, Larry doesn't get his mortgage. And then Larry says to me, “why?” I say, “I have no idea, but the algorithm said you're a bad risk, end of story.” Transparency is huge and unfortunately we're using them to make decisions about finance and we're using them even worse to make decisions about parole. Recidivism algorithms making decisions about people are black boxes.

Using data for purposes that they aren't intended. It's very tempting once you have a data set to explore it and use it and forget about whether somebody said it was okay to use their data for this or that. This happened to my wife actually at Mass General. The night before an operation, there was a guy who came by asking can we sign this consent form? And it turns out he wanted her to go through another MRI just for collection of data about his own research, which when I told the staff later, they were appalled. This guy had gotten in and was going around to people who were about to go into surgery and asking them to have another test so that he could collect data.

Larry Bernstein:
David Salsburg is the former chief statistician at Pfizer and the author of the book The Lady Tasting Tea: How Statistics Revolutionized Science in the 20th Century. He tells the story about experimental design. Ronald Fisher one of the giants of modern statistics got his first job working for a UK based industrial farm and he was tasked with determining which seeds to plant to maximize profits given local conditions. The farm has tons of data that was painstakingly collected over generations, and he literally throws it away because there was no design for the experiment. The farm did not use randomized plots using random choices of seeds. He believes

that the big data has no value. The staff is flabbergasted. Fisher's idea was that with statistics the experimental design is critical, and that just having a lot of data is meaningless.

Today we live in a world with both big data and enormous computing power. Given the change in the landscape, how crucial is experimental design?

Dick De Veaux:
Well, that's a great question. Experimental design is still important. And in fact, it's a heyday of experimental design in some sense because there are probably tens of thousands of experiments going on the web every day. The bank Capital One says they do 40,000 experiments a year. Some of them are very small questions about whether this font in a website is better than this. And they'll look at click-through rates and see which is better. And these are called A versus B comparisons. And that's basic experimental design. Unfortunately, they use mostly the simplest experimental designs possible. This just A versus B. And what Fisher brought in all sorts of other factors and made it made a whole industry of more complicated and efficient experimental designs.

Larry Bernstein:
When I took your intro Statistics class in 1985, there was no big data and there was limited computing power. You told us that statistics can help us solve real world problems with small amounts of data, but the experimental design has to be done properly otherwise its worthless. You also told us that the biggest risk was overfitting with too many parameters. Chat GPT use millions of parameters with seemingly no experimental design, yet the output looks pretty good, what is going on here?

Dick De Veaux:
Modern machining learning with non-parametric methods have hundreds of thousands, maybe millions of parameters, but they are regularizing the output. So, if you just fit the data at hand with these huge models, you'll fit it perfectly. You will fit last year's stock market perfectly, but it has no predictive ability. So, what can you do? When I was a beginning grad student in 1973, what people would say is, you can't even look at the data yet. You have to postulate the model that you want to fit. And it had very few parameters. Then as computation got more powerful and data got more plentiful, people started using highly parameterized models.

Here's what they do essentially, if you remember stepwise regression, variables come in and out of it. So, we start with this huge model, and if we fit it just to the data that we have, and we would fit the data perfectly. We filter some of those parameters down so that they don't overfit. And that's a huge part of training modern machine learning methods, is you use part of the data that the model hasn't seen to see how it's doing and you tune back and forth and try to get it so that it will do the best prediction on data that it hasn't seen and fit the data at hand reasonably well. And that's what's going on there.

Larry Bernstein:
On the final exam in your Statistics 1 class, you said that the baseball player Pete Rose was known as Mr. Clutch. You asked us to evaluate whether or not his nickname was valid? And you provided the students with Rose's batting average with runners on base and with the bases empty. And from that the student could evaluate whether the batting averages were close enough to be statistically the same?

I remember that question on my exam 38 years later. How do you find examples to make statistics interesting so that college students are fully engaged?

Dick De Veaux:
I'm always searching for something that cannot bore an 18- or 19-year-old to death. It's been a 40-year challenge and it remains one. As the population has changed, I use things like baseball a lot less than I used to.

Larry Bernstein:
Let's go back to the Pete Rose example. I think evaluating Pete Rose is more complicated in real-life than the data provided in the test question.

Dick De Veaux:
Now I'm thinking what we really should do is the experiment. So, what would that involve? That would have the pitcher randomly select pitches for both men not in scoring position and men in scoring position to make all the other factors fair, or to somehow at least add the factor of what kind of pitch he was getting and build a model that would take that into account.

Larry Bernstein:
With runners in scoring position, the infield players are in different positions and that could make their defense less effective.

Dick De Veaux:
Yeah, there's lots of reasons why it's not really comparable. And that's exactly the problem with a lot of A versus B tests these days, that people are assuming that all those other factors are comparable.

Larry Bernstein:
In your introductory remarks, you mentioned that you were concerned about algorithms that determine parole. Tell us about that.

Dick De Veaux:
The worst case that I know of is an algorithm that's used to judge whether someone should get parole. There's a great article that was published by ProPublica that just showed how biased this algorithm is called Compass. It's proprietary and based on 137 variables.

What happens is they put these data in and it's got the person's age and their prison record and all sorts of other demographic information. And then it comes out and says, should they be paroled or not? Or probability of recidivism within the two years, which they'll use to decide. There are plenty of cases showing two people, one of whom gets a low-risk score. One of them gets a high-risk score. And in all of the examples, the high-risk person is a person of color with a very limited criminal background. And the low-risk person is somebody who has 40 years of terrible crimes and is white. So, it's clearly this algorithm is racially biased.

My colleague Cynthia Rudin got the data and looked at it and used a decision tree, which is a very simple, transparent model that uses the data and does splits. The first split might be is the person 25 years old or older, yes, or no? And it keeps going down there. So you can follow it at the end and see what the decision was. What she found out was by using those data, she almost reproduced the same bad conclusions of the proprietary algorithm but hers were only based on a few variables. So first of all, didn't need 137.

Larry Bernstein:
There's a reason why parole boards starting using an algorithm. There was litigation that a parole board was being unfair on the basis of race. The judge likely concluded that these parole boards are systematically discriminatory. So, the parole board went out and acquired a third party algo. Now your point may be that what makes you think that that algo is going to be any better than that group of people.

Dick De Veaux:
That's exactly right. And that's the question though, is should we substitute a black box proprietary algorithm for a parole board? I mean, certainly there's a problem. And that was the great thing about Cynthia's analysis was to show that the problem was the quality of the data. ProPublica first of all, assumed that the algorithm was using race as one of the variables, but it didn't and it didn't have to. The decisions may not be racially equitable because of other variables that are maybe associated with race. But the point that I'm trying to make is that we can't see whether it's potentially discriminatory unless we can see what variables are important and unless we can see that the data are accurate. A black box is very dangerous for that reason. We need transparencies in these algorithms.

Larry Bernstein:
20 years ago, I got a call from the NORC to be included in a longitudinal study. I agreed. They started asking me questions about my income, but I didn't fit neatly into their framework because I am self-employed and earn my money trading securities with my proprietary capital. I was a true outlier from every perspective. My income is variable and some years I lose money. The questions just didn't apply to me, and I asked for a manager to get involved and they decided to drop me from the survey. What are the problems of a study when the outliers like me are removed?

Dick De Veaux:
It depends. Surprisingly <laugh>. What is the purpose of this analysis in the first place? If I'm trying to figure out the relationship between income and spending. Most Americans, not Larry's but normal people, then I might throw you out. If the purpose of it is to understand the whole population, then of course you're important.

This happened to me once at American Express years ago. I was analyzing one of their campaigns, which was a double mile's campaign. They have to pay for double miles for some airline. They want to know if it was worth it. They sent out 30,000 non double miles offers and 30,000 double miles offers and saw what the average spend was for each of those groups. And that was randomized. Nice experiment. And they came back and it was a hundred dollars difference. And they said, well, okay, now we know that it, on average it's a hundred dollars spend. We can figure out whether that's worth us buying the miles. Great. I thought we rushed this. Maybe we should look at the data a little bit. I did a box plot of the two groups, and there was an outlier, like you couldn't believe it.

There was one guy in the double miles group that had spent $3 million on his card in March. And I immediately called down, "that's a typo, right?" And they came back and said, "nope. That's what he did." So if you take $3 million and divide it by 30,000, you come up with a hundred dollars difference between the two groups. So. what should you do? Should you pay attention to that guy?

The funny thing was the discussion among the analysts. Somebody said, "well, maybe he was incented to spend that much because of the double miles." And I said, "actually, did you read the fine print? You only get it on the first 10,000 miles." I don't think when he went to Sotheby's and bought the Van Gogh, whatever he got for that. And it was pretty clear that we had to just set him aside.

Larry Bernstein:
One of your homework assignments that you gave us in Statistics 1 was to create our own regression to solve some hypothetical problem. And I decided to estimate the grade point average for the pledges in my fraternity.

I used a bunch of variables. Now there're only 25 pledges, so I am going to have some overfitting going on. First thing that I did was collect the data. I sat each pledge down and I'd asked him questions. My favorite question, I asked them was how many hours do you study? 1-5, 6-10, 11-15, etc. And one of the pledges asked if that was per semester, and I said no per week! Is there great confusion in answering survey questions?

Dick De Veaux:
I'm sure it's a concern. I don't think it's the largest problem in data. And you didn't have zero to five <laugh>.

And the other problem with that is if you ask people, for example, how much did they drink last week? People have no idea, what you need to do is ask them, "how much did you drink last night? And even though that seems less accurate because of human memory and what humans rationalize, et cetera, that's a better thing to do. Just asking people, "how much did you study last semester?" Can you imagine that one? <laugh>

Larry Bernstein:
There was a very famous longitudinal study at Harvard for the classes of 1941 and 1942 called the Grant Study which was funded by a department store magnate trying to figure out which managers to hire. Additional funding was provided in the 50s by Philip Morris and the questions started to include topics like smoking and drinking. George Vaillant ran the study and he found that when he asked the study participants if they were an alcoholic, the answer was always no. So, instead he asked how many Mondays do you miss work because you were on a bender all weekend. If the answer was more than two, then he labeled you an alcoholic.

Dick De Veaux:
That's something that psychometricians worry about all the time, and that's really a whole subspecialty. And what about asking a sensitive question about drug use that somebody may not want to truthfully answer? So, one thing you do is you randomize the question. So, you say if your social security number ends in this, answer this question, or if it ends in, in that, answer this question and you compensate for that later so that people realize that no one will know which question they answered.

Larry Bernstein:
In your opening remarks, you mentioned that algos might hurt people. Who are you worried about?

Dick De Veaux:
I'm worried about people that get discriminated against in every facet of life. And they're the ones that are most vulnerable to these decisions where they don't have control of their data. They need the analysis to be correct. Their life may depend on it. And if you're talking about financial decisions or health decisions or recidivism decisions, those are important. And the Larry Bernstein's, that's not a worry.

Larry Bernstein:
My brother had a roommate who worked in the mortgage department at a bank, and he brought home hundreds of mortgage applications, but he never got around to finishing the analysis. When my brother complained that he was making a mess of the apartment, he picked up the mortgage applications and tossed them in the garbage and announced that they were all rejected. Why do you think humans are better than algos?

Dick De Veaux:
Wow. <laugh>, wait, I'm appropriately outraged.

He rejected them all. So we could see what he did and that's bad. I don't trust people what I trust is a decision process that I can understand if I know what the criteria going in are. Now whether somebody follows that or whether they just go on a rampage and reject every cardholder for the next month, that has nothing to do with what I'm talking about. That's a behavioral problem.

Larry Bernstein:
What would you recommend to firms about how they should use algos in their decision-making process?

Dick De Veaux:
We want models that we can understand that are based on variables that we think are appropriate, and the data are good enough to make those decisions. I had some friends who did an analysis of judges across the country and they really wanted to find out whether judges were being fair in their sentencing. They got data from the judicial data set that's publicly available. And what they found were there were two judges who just were clearly giving harsher sentences to people of color. So they identified them and they publicly shamed them. <laugh> turns out that a Professor from Berkeley, a criminology professor, pointed out to them that there was a lot of missing data in this judicial database. And in fact, that these two judges that they had identified had very few numbers of cases compared to the other ones.

And it turns out that at least one of the judges was a person of color and they had been known by the Southeast Pennsylvania District Attorney's office as the two most advocates of people of color in the system. So, problems. One is missing data were causing bad information in the model. And the second thing was, what you probably remember from Stat 1 A was what the Berkeley professor did is he took 700 fair coins and flipped them a whole bunch of times. And two of them came out as having significantly too many heads.

Larry Bernstein:
You should expect outliers in a study, but that may not mean that the outlier is problematic.

Dick De Veaux:
So clearly there's something bad going on with those two pennies. Part of it was statistical naivete. They got a low P value for something and jumped to the conclusion. The second part was the missing data. And the third part was, did they do case control for the types of cases that these judges happened to have. You know, you can find out that the death rate of some inner city, state-of-the-art hospital in New York for certain diseases, much worse than some place in the middle of nowhere in the Midwest. And that's because the people that are coming to that hospital in New York are severe cases. So, you have to do some more sensitivity analysis than just look at averages. there's a great book called The Flaw of Averages.

Larry Bernstein:
By Sam Savage.

Dick De Veaux:
You know, averaging models can be really misleading. The problem today is it always has been. It's just we have more data and more models and more decisions.

Larry Bernstein:
Dan Willingham was speaking on our podcast in March where he discussed making learning easy. And I complained that I loved reading his books except for his textbooks which I found technical with too much jargon and it was boring. Dan said that is the way textbooks need to be written because we need to cover lots of material. Do you agree?

Dick De Veaux:
My editor called me up and she said, “I want you to write an Intro Stats textbook with Paul Velleman.”

I think one thing that comes across is voice. The other thing was some humor. Our book starts with, instead of introduction, it says, stats start here, footnote. And the footnote says, normally this would be called introduction, but nobody reads the introduction. So, we wanted you to read this. We didn't call it introduction and said, and besides, it's safe to put this here because nobody reads footnotes, <laugh>.

And you know what? I think it gets the students on our side. They realize we're humans. This isn't written by Chat GPT. It's not completely dry. Statistics might not be the course they're most looking forward to. But the other books all started with things like, in this chapter we'll be discussing the parameters of the hypothesis test.

We start with who wears seat belts, more men or women? Then we get to the methodology. I just listened to some of my teaching from last semester, and I realized some mistakes I made. Saying, now let's talk about how to graph a quantitative variable. <laugh> <laugh>, they look at me like I've just walked off the spaceship from another planet. I should have said, here's some data on incomes. What kind of graphic do we want to do for that? And then generalized it to that's a quantitative variable. It's easy to make this mistake even when you've been teaching for a million years because you're too close to the subject. I think what we try to do is get into the student's head.

Larry Bernstein:
How do you use stories in the classroom and in writing your textbooks?

Dick De Veaux:
Stories are the way that people remember facts. You are the prime example of that. You remember several stories that I told in Stat 1. No one remembers the formulas. They remember the stories. And that's why we put stories into the textbooks.

I'm not testing on the formulas anymore. Machines are really good at formulas. I'm testing on appropriateness of the analysis and all the things that we've been talking about. How good are the data? What do you do with the outliers? Is this test appropriate? Is this analysis appropriate? Is this analysis, ethical? Those are the larger things because where's programming going to go in the next 10 years with Chat GPT, I mean, routine programming, it's not going to be a skill worth having.

That's what I'm trying to teach.

Larry Bernstein:
When I was helping my daughter study for a geometry test in high school, I told her to try to do the most difficult questions in the chapter that were not assigned. She was very reluctant, but she did a proof for a trapezoid with diagonal lines. When I asked her how she performed on the

exam, she said she aced it and that the trapezoid proof was on the test. I remember taking your exams there was always a curve ball. How should students' study for your exams?

Dick De Veaux:
I made a mistake last semester of giving a practice midterm, because that's something a lot of my younger colleagues are doing. And the kids love that because then they feel like they're in great shape. They know what they're doing. But then you can't put the trapezoid on, can you?

If you put it on the practice test, fine, then you have to have another trapezoid on the real test. And so it's a vicious circle. What happened was I gave this practice exam that I borrowed from a colleague, and then I gave my own midterm. And I had a third of the class absolute panicked because there was a question where the information was there, but it wasn't in the way they had seen it before. They saw the curve ball for the first time, and they just stepped out, swung and missed.

We had a conversation about this afterwards, and it was really interesting. I loved this class. I had 60 kids. They were mostly freshmen, mostly wanted to be econ majors and were mostly varsity athletes. So I picked on this kid in the front row. I knew he was on the football team. I said, "how do you prepare for the game coming up?" He said, "we watch a lot of films." I said, "good. And if the team comes in and they do something different, it's not on the films. Do you go to the refs and say they cheat?"

I said, and, "let's talk about the other preparation, is that all you do all week, watch films?" He goes, "oh, no, man. We practice every day."

Larry Bernstein:
Thanks Dick, for joining us today.

If you missed last week's show, check it out. The topic was We Need More Fraternities!

Our speaker was Anthony Bradley who is a Fellow at the Acton Institute and a Professor of Interdisciplinary and Theological Studies at Kuyper College. He is also the author of the book Heroic Fraternities: How College Men Can Save Universities and America.

Anthony discussed why fraternities are so important to young college men, and why university administrators and others want to shut them down.

I now want to make a plug for next week's show with Michael Reid who is the author of the new book entitled Spain: The Trials and Triumphs of a Modern European Country.

Spain had an election last week, and I want to hear about the implications of a near tie in that election and what it means for Spain and Europe. The issues in Spain will be familiar to you: Too much immigration, abortion rights, and should Catalonia be independent.

You can find our previous episodes and transcripts on our website whathappensnextin6minutes.com. Please subscribe to our weekly emails and follow us on Apple Podcasts or Spotify.

Thank you for joining me, good-bye.