Webinar Series

Data Cultures in Higher Education: Emergent practices, professionalism and the challenge of social justice.

Organizers

United Nations Educational, Scientific and Cultural Organization

UNESCO Chair Education and Technology for Social Change

EDUL@B

UOC-EPCE

Universitat Oberta de Catalunya
Faculty of Psychology and Education Sciences

A hierarchy of limitations in machine learning

Data biases and the social science.

Dr. Momin M. Malik

Berkman Klein Center
for Internet & Society
Harvard University

29 September – 16-17hs GMT+2

Registration: http://symposium.uoc.edu/54681/

Series Coordinator
Juliana E. Raffaghelli

Translation

A hierarchy of limitations in Machine Learning. Biases in Social Data

M: I'll be going through the topics that Juliana had also put in her blog post that I believe she shared out before the seven hour long with the link and those were some of the questions that we went over and back and forth through conversation we had over email over a few months. And the first question Juliana had asked is how does my background lead me to my current work and what is this hierarchy that I am talking about.

As previously shown I have a background going across a lot of disciplines and I started moving more from humanities and social sciences to computer science because I had this background in history of science, science and technologies, statistics, and I was really dissatisfied with the state of discourse that I saw around, back and then was Big Data. As I learned more, I think machine learning is the most important thing to think about, but of course it takes press under lots of labels: artificial intelligence, deep learning, and all these are related to data science. And so I went into computer science trying to get more of the background to engage on a more granular level with some of the material critically, and were both externally and internally. And so I would say what the paper that would be the topic of this talk is really about is: does modelling really work and, as all discussed shortly, machine learning really is just another type of modelling, almost identical to statistical modelling but for very different purposes and with a different style. And the answer is, as with many things in social science, maybe, sometimes. If it's done right and we get lucky. Often we don't even have the ability to tell whether it worked or not and that also makes it challenging. And I'll be breaking down some of the claims that happen around machine learning and interrogate them critically and show how both internally and externally they are probably exaggerated for the most part, especially when they are made in press releases, made by businesses selling products, but even academic papers I think don't do a very good job in understanding the limitations and in framing things in a responsible way.

The thought that Juliana had in her blogpost comes from a paper and it's the sort of tree of methodological approaches and this is the structure of my paper and I go down different branches of this and say what are the trade-offs? Why do some people advocate going down this branch vs that branch? And I don't have the depth to know what are the other branches, other types of these things. Under 'simulation' I know there is agentless modelling and there's system dynamics, equations, there's partial differential equations and other types. So you can choose different approaches to modelling things even within these qualitative, certainly there's many different approaches, content analysis or grounded theory from ethnography that I again don't have a very good grasp of the range, so I left those as dotted lines. Where this came about and maybe it'll explain more what I am trying to do with this is seeing a lot of literature about these binaries. Juliana shared with me some of the research she had done in the 2000s looking to the paradigm wars, and this is something I hadn't been heard about but I do know some of the products of it, these methodological syntheses or items to say how do we combine quantitative research and qualitative research. This work is, Juliana tells me at the end of the paradigm wars, and this is from 1988, and this tries to say what is quantitative research good for, what is qualitative research good for, how do you do either of them properly, and how do you combine them. And these are debates that still go on today but are completely unknown within the majority of computer science and machine learning literature

that I read, so I do want to bring that in. But this is one branch and I think a lot of us haven't studied this in great depth, have a sense that experiment vs observation is another big debate in the quantitative literature, especially in psychology, econometrics, economics more generally that can you say anything without experiments, without manipulating the world. That's on the one hand where people take causality to be by definition manipulation or intervention and causality being what really matters, on the other hand people saying a lot of experiments in psychology has no ecological validity, the way people perform and behave in these very artificial settings doesn't generalise at all. You've made the conclusion that doesn't apply in anyway to what happens in society and in the world, that's the flip-side of observation and experimentation. A good review of all of this is in a collective volume from 2014, it's specifically about field experiments but it goes into that division, but I think a big breakthrough came when I came across this article from 2010 by Galit Shmueli. She says there is a difference between explanation and prediction and the failure to understand this difference is a major pedagogical failure within statistics and also a major misunderstanding around machine learning, and, specifically at the end of the article, she gives a mathematical example of how a model that predicts well, as prediction is defined in statistics and machine learning, can actually not capture the causal process, so some sense of false model, one that reflects this toy model worse can do better at this task of prediction. And, even though that is a very artificial example put in mathematical terms, this decoupling this idea of whatever predicts will explain was really helpful for me to think about the difference between the methodologies that are often approached in statistics versus what machine learning is trying to do. And since 2000/2001 there has been a lot written about how machine learning is converging with statistics, how early machine learning was a set of more ad hoc metrics but gradually it was put into a statistical footing, using things like probability theory, concentration of measure, estimation theory and all these other things from statistics. It's diverse a little bit around deep learning but the basic idea is you use data to reverse engineers some underlying function or boundary or thing to imagine exists in the world. Whether you do that for the purpose of explaining things for understanding or merely making predictions is the major difference between what statistics and machine learning try to do. Of course, either can do either task and they do them slightly differently but that's the big difference between them, even though they both will use linear regression, logistic reversion, kind of the first thing they will try, exact same mathematics, exact same software fermentations and a lot of same datasets as well, but a difference in the style and output.

So some other things that I've come across that helped me construct this hierarchy: simulation is, compared in specifically agent-based, simulation is compared as an alternative to statistical modelling, so if you have data, you probably are not going to be very well served by simulation because it's hard to put those data into simulation. You can just initialise and then just compare outputs, but it's difficult to use it directly versus if you can't manipulate a system whether for ethical or logistical reasons, maybe statistics is not going to give you what you really care about. So simulation, advocates say that simulation will give you the causality that you might care about. Or in epidemiology, and I've renamed this category, I first called it mathematical modelling but then I found out in epidemiology they distinguish data modelling from mechanistic modelling. And mechanistic modelling I think has gotten more attention under Covid-19, things like the SIR model, susceptible, infected, recovered or removed. These are sets of differential equations that are difficult to fit to data, but give a conceptual understanding and allow for arguments about the dynamics and how epidemics work. And that's not to say that they can't be compared but the primary value of these equations and these derivations is not to fit to data or to explain the shapes we see in data, but rather as set of first principles to understand what we might expect to see in data or how we might try to manipulate the world to change data that we end up seeing and so I renamed this mechanistic after epidemiology.

That gives this big branch in tree and implicitly, when we choose a discipline, when we choose a methodology, we are on some branch of this tree, and we've implicitly made a set of decisions, whether that's in a particular project or even in our training, about what we've prioritised, what we can do, we cannot do and each of these has trade-offs. So I talked about some of them between for example simulation or equations, and using data or modelling data, and earlier, I think we maybe can get into discussion later on, there were these really acronymic debates about quantitative vs qualitative in current research. And I again found this article that really helped me make this link between machine learning and statistics as kind of a branch in this tree. I would say mainstream machine learning is about following this sequence of branches and I would define it as a kind of working realist definition as extrapolating from correlations to anticipate outputs of a static system. I would say that the learning in machine learning is completely a metaphor now, maybe early on it attempted to capture some processes of learning, but it bares very little resemblance to even in cognitive learning theory and almost no resemblance to social culture learning theory for example, so it's purely a metaphor for improving with more inputs. And it's really about using correlations and, as I talked about more in my paper as a lot of social science literature talks about, correlations use the mean, the adverts, some sort of central tendency and that privileges the mass at the expense of outliers or people who don't fit into whatever is average, and that can have powerful normative consequences as well. And so machine learning is building on the same set of machinery that you've committed to appear when you're going with probability based modelling. And lastly, I'll get more into this later as well, but it doesn't really anticipate the ways in which a system can change. If there is variability that has been observed in the system already, machine learning can be able to account for that but it fails miserably when something completely new appears in the system, so it's very much reactive, it's not making these larger connections that statistics tries to do in terms of finding causal relationships that can be then generalised to unseen systems, and certainly imagining entirely different ways of thinking, of knowing, of being, of doing that would be further in these branches. And even inquiry as a choice of branch, you don't need to approach the world through inquiry, it can be through speculation, through practice, through a whole bunch of things, so even this is a branch of an even larger tree.

The next thing that - I will put up the questions - Juliana had asked about is things about responsibility for quantification and what are some of the drawbacks of quantification. And these are again things that were well known in the paradigm wars from what I've gathered and understood from some of the literature that's trickled down from that, and of course a lot of these scientific technology studies, history of science, has contributed to. And this is the idea that quantification is privileging one set of meanings over all others. Theodore Porter in a wonderful 2012 piece talks about thinning description, description is of course the anthropological idea but it contrasts this to what post World War II behaviourists who took thinning out meaning to only be what is observable, only what we can directly see, being the only thing that matters as kind of the price worth paying, sacrifice a few meaning was a price worth paying for these objective results, or proper objective methodology. And of course, solidifying one set of meanings indeed does likely get to all the results of what you can build, but the problem is that you contribute to using one set of meanings over every other possibility and then you forget or you lock that in and you can't go back and nothing in quantification can ever undo that. Exploratory data analysis maybe find some discrepancies in the data, at most you can go back to the quantification process and understand what happened in it, but quantification itself can never fundamentally reverse that process of quantification of choosing one set of meanings over any other, and I'll go to some examples of this later. Another problem, and this I think can directly contribute to some of those paradigm wars literature, is the difference between what is measurable, what is observable, and what we actually care about. And there's a version of a law, called the Goodhart's Law/Campell's Law from 1975, they were published about the same time, and this is, different versions of it are: if you optimise to a measure, that measure will become bad or will no

longer measure what was originally measuring. Any statistical regularity that is used will cease to be useful. And what it's really talking about is this gap between the thing we care about, in social science and psychometrics we call it 'the underlying construct', and what is operationalised, what we're actually able to observe and measure. And some examples of this, a recent paper that got a lot of attention, published by Obermeyer et al. talked about how the healthcare system was discriminating against black patients because it was optimising the healthcare costs which were measurable rather than the actual target of health. And because historically less money has been spent on black patients because of systemic discrimination and racism, the system was just using those same patterns, optimising along the same patterns. In education, I think we all know, the grades are poor proxy for ultimately what we care about which is learning or maybe preparedness for whatever that might be, and learning is infinitely multifaceted and subjective and contingent and, crystallising it into one set of meanings which are how well you can view on some standardised assessment or some standardised grading of essays, is it putting poor proxy. Maybe there's no better proxy or anything that can be measured, but when we do commit in quantification and measurement, that's the sacrifice we make for being able to get grades for millions of students, to be able to aggregate those, to be able to look at trends across time and space. Citations in academia, we are judged by our H-index, so our impact, what journals we are published in and their impact factor. Ultimately, those are based on things like citations which are a poor way, I think we all understand, of assessing in that. Some things that are heavily cited have no impact, some things that are never cited have extremely high influence or they change people's perception of the world, of theory, of methodology, and those aren't things that can be easily quantified, or can be quantified citations as what ends are being used? And lastly, both arrests and convictions are a poor proxy for some underlined construct that we can imagine of crime. Juliana asked me for examples of that measurement in machine learning, and I think, there's some examples of that measurement and also just some examples of assuming that what is available is what we care about, assuming that grades are learning or that citations are impactful.

There's a press release and this was earlier this year and I think March or April from Harrisburg University in Pennsylvania. They claimed that a forthcoming paper can predict criminality with no racial bias. There was a group of scholars under the Coalition for Critical Technology who wrote a letter to Springer saying on the face of it, this is implausible, if not impossible and it makes categorial errors. One of the biggest things is that there is no such thing as criminality, let alone one that can be predicted automatically. Criminality is based on what people decide is criminal or not, so for example in the United States, an example of seeing that is wage debt, employers not paying their employees is not actually criminalised or is not enforced, and people can get away with that easily versus things like loitering, somebody just decided that that is criminal, loitering is when you are just in a place or wait outside a place and are not doing anything specific, that is criminalised, and other behaviours are criminalised. But even given sets of codes or given sets of laws, we never actually have statistics about crime, what we have are the proxies of arrests and convictions. In the US we're certainly seeing the unavoidability of how these are really really biased measures: black people, Latinx people, and other minorities are arrested at far far far higher rates. Similarly, convictions, people who even just have darker skin are given longer and more severe convictions for similar crimes, and certainly using the quantitative evidence here. But the most important thing is just people's experience of being brutalised, being profiled, leads to the system of criminalisation and criminality, and that's not something that is captured in faces. And to think so, I mean, the letter talks about physiognomy and these old practices of measuring skulls around eugenics and it's a version of that, that there's no causal connection between this thing that they want to measure criminality and what is measurable in another way, faces, and yet they find a correlation and assume that because there is a direct link rather than examining the metrics of things that are happening. And this is, this paper wasn't actually ever published, it was withdrawn before it was sought, and I signed

onto the letter, but I think it's in itself unremarkable. But this is the kind of assumptions and measurements that happen in machine learning or the use of proxies without realising that they are proxies and that's a major danger, and this I think it's the case that could be very harmful going forward and so I thought it was important to intervene but this is not the first study to do this, this won't be the last study to do this, both in this domain and many many others. And so that's the kind of problem of measurement.

One important thing that might have been talking about is that machine learning is matching whatever is measurable and not the underlying meanings. And an example I like to use is around image recognition and deep neural networks which are the most impressive things to come out of machine learning. There, what the systems do is correlate patterns in pixels, and what's amazing is that they can automatically extract different patterns they correlate to human given labels, and so that's the kind of connection they make. But in social science we would say there's an underlying construct of catness that both produces humans labelling these spectres as those of cats and teams certain patterns of pixels, but of course catness is not one thing. In order to make human labels we have to decide what counts as cat for this purpose and what doesn't. And I give an example, is this a cat or not? Well, somebody on Mechanical Turk might immediately say 'yes it's a cat'. Sure, it looks like a cat, but if we cared about phylogenetics, it's a kind go evolutionary biology, then this is a false saber-toothed tiger which is a marsupial and not at all related to felines, or that order. Conversely, we would say well does this picture contain catness? And again, if somebody on Mechanical Turk was clicking through it, they would say 'no', but if we cared about is there something of this (unintelligible) in the picture, well there is this cat, this size, in the picture and so we would say 'yes, there is a cat'. And so we can decide what we care about capturing in catness, and there is no one thing that we can do for that, we have to decide on that. And if we ignore that, then we'll get edge cases, we'll get outliers that would be misclassified and that would be labelled as 'a cat' or 'not a cat' in ways that maybe detrimental. And again, this is a probably harmless example, but it's showing the kind of importance of considering what the underlying construct is and not just focusing on the immediately amazing performance of having billions of images that can be labelled as having cats or not having cats such as using Google image search.

I won't go too much into the things that have come out of disciplines like psychometrics, about how you create measures and how you validate them, but there's construct validity, there's face validity, criterion and related validity, internal and external validity, all of which are by and large unknown in machine learning. And some other papers, in addition to mine, are trying to bring this into machine learning, but in machine learning is really only looking at external validity, and ignoring all of these other things, and sometimes that's ok, sometimes it's good enough to ignore everything else. But other times, when we want to make claims about the world, external validity alone is not good enough. We can achieve external validity while failing at all these other things.

So those are some examples of measurement which matters, another thing that I, going a little bit into my paper, is this idea, performativity of course is much older, and do the (unintelligible) , but in science and technology studies Donald MacKenzie and others wrote this book *Do Economists Make Markets?* Not markets exist and then economists study them, and specifically Karen Healy has a very good summary of this, is that 'the performativity thesis is that economics produces a body of formal models and transposable techniques that, when carried out into the world, reorganises the phenomena that the models purport to describe…'. And one way of understanding this is that whether or not the models were true before or not, when people start using them to think about the world and approach the world they in effect become true or they make themselves true. And this is hard to find very clearcut precise examples of, but in a more conceptual sense if we look at models

sort of being self-fulfilling purposes, we can see examples where this might be happening. One example that I had in a work of mine is around Facebook's people you may know recommendation system that assumes that triadic closure exists, and this was also the example Karen Healy gave and my paper goes into more quantitative detail about. This model assumes that friends, our friends, know each other, and in social network analysis we call this triadic closure or transitivity, and it built the system based on that. If you are a facebook user, you'll see people you may know you have 38 friends in common, you have 115 friends in common. What happened when the system was introduced in Facebook is law beholds the connections people were making were friend of a friend connections. Was this really using the underlying principle of triadic closure or which kind of existing latent sense in Facebook or did it just give people an easy button to click for (unintelligible) the behaviour that Facebook assumes it was using or seeing? And so the manipulation of systems with models, with these machine learning implementations of models that we usually call algorithms, are these changing the world in a way that would affect what we think is happening? And again, that's very subtle, and something I struggled to apply, but it's a very important insight to think about this framing about how are models affecting the world, rather than just reflecting the world.

Another major theme, and I am working on a stand-alone piece as well, is the problem of this word 'prediction'. And I love this quote from Daniel Gayo-Avello, he's talking about predicting election results with Twitter in 2012, he says: 'These predictions are not predictions at all. I have not found a single paper predicting a future result. All of them claim that a prediction could have been made; i.e. they are post-hoc analysis and, needless to say, negative results are hard to find.' There's a couple of things packed into this quote: one is a critique of publication bias that negative results don't get published, which applies to machine learning as much as anything else. The other thing is that prediction in statistics, going back I found cases maybe even as far back as Carl Peterson in the 20s and 30s, is used as a synonym for correlation, and this makes a certain kind of sense that the best way to extrapolate, to look at the future, is to use correlation that you've seen in the past, but the way is used certainly now across statistics and machine learning gives the impression that there is something much more profound going on than what is actually going on. People would proclaim that they've predicted elections on Twitter when in fact they say looking retrospectively I was able to get a result that correlates two things. Do these generalise to the future? We don't know and we have strong reasons to believe that no, they would not. And so I would also encourage that paper, this version is on archive, has this marked here title, has also published a version in a (unintelligible) magazine.

And as a way of explaining that more carefully, there's this great example from the New England Journal of Medicine, it's a joke paper, it says chocolate consumption leads to winning Nobel prizes, and indeed, if you look at the plot, the more chocolate consumption a country has, the more Nobel Laureates it has per population. In this article, Messerli jokes that, well, writes that there's proteins in chocolate that would stimulate brain function, so obviously if countries want to win more Nobel prizes they need to purchase mass quantities of chocolate and feed them to their people. And the joke of course was that correlation has not causation and that we can invent plausible causal stories to explain correlations that can be completely inverse. But the important thing here is that machine learning would use correlations like these to make predictions, it will make retrodictions, it will make descriptive statements about (and statistics does this too, this is not unique to machine learning), descriptive statements about a correlation and simply call that a prediction and say, I predict that Switzerland will have, if we extract from this linear trend based on the chocolate consumption, we would predict that Switzerland would be up here, we're pretending that was held out from a sample or truly out of sample when in fact we just do that for retrospective illustration purposes. And I can get more into cross-validation and how validation works in machine learning but the idea is that you

can fit different models again and again and again until you get good results, or you can have papers that people try to write again and again and again until you get good results.

Prediction, which is correlation, is not explanation, which is causation. So a graphical model, and this is something that comes out of machine learning but is very seldom used for actually modelling causality, is something like national resources lead to both science funding and chocolate consumption. The science funding leads to Nobel prizes and so we observe a correlation, going backwards to this tree, between these two things, but we can't actually intervene on chocolate consumption and expect to wine up increasing Nobel prizes. In more simple terms, we can ask where does this correlation break down, do past patterns fail? For example, the small European countries that have been very successful in the past, how are they going to be compared against other places now, India, China, Brazil, Russia, which are on this plot. There's a lot based on the history of Nobel prizes, which maybe doesn't look like the future. We can also say Nobel prizes supposedly are awarded on an underlying construct of merit. Who decides that? How is that measured? And that's also a very subjective process that we can easily imagine biased, we certainly know about the bias against women that has existed in Nobel prizes, and what other biases still exist in terms of country background, race, and other personal characteristics.

I also want to point out that this is not an obvious use of the word 'predict'. There's this great philosophy book from 1998 called *Predicting the Future*, it points out that trend projection and curve fitting are reasonable ways of making predictions, but they are rather rudimentary. And Rescher is trying to say that things like scientific laws or analogies are much more robust and sophisticated in scientific ways of making predictions and I think what people tend to not realise and, within machine learning and certainly hearing about machine learning results, is that the predictions of machine learning are extrapolations based on projecting forward, based on curve fitting. And there are critics within machine learning as well who are very strongly against curve fitting but that's what the vast vast majority of machine learning does. It takes some set of data and it fits a curve to it or a decision boundary or something. Does that generalise? Maybe, maybe not. If the world changes, probably not.

A good example of this is Google Flu Trends, and here cross-validation was done properly, holding out data as a way of testing the model and, as you introduce the word data to the model, indeed it does very well at keeping up at the trends. So what this was Google flu trends using Google search data to try to beat the Centers for Disease Control and Prevention numbers about incidents of the flu in the United States that has to manually compile reports of how many flu patients they see and they send them to the US government centrally. And so Google was trying to use the search terms people used to do this process much faster and get projections of the flu incidents before the CDC can compile the official numbers laboriously. It is properly done in terms of the modelling and in terms of the validation, unfortunately there were things in the world that the model just couldn't pick up because it was based on correlations, because it hadn't observed that variability in the past. So for example, I think, yes, so the CDC numbers of the flue, here there was a second wave of the flu that Google flu trends completely messed in 2009. Later on, the creators went back, they recalibrated it, it did well, but then there was another season where completely overestimated the amount of flu. And a paper from David Lazer and colleagues called 'The parable of Google Flu Trends', he rises that it was half winter detector, half flu detector, because it picked up on non-causal, sometimes he calls it the 'spurious correlations' between winter search terms and flu, it wasn't able to shift when there was something different in the world that no longer matched that, and I will return to this example as well.

Another example that Juliana asked is why after all of this knowledge that correlations and causations, about measurements, does the world keep moving more and more to being committed to correlations? I think one thing is certainly that machine learning has some impressive successes, and there is really good work I think now in the history of science about the histories of artificial intelligence. I recommend an article from Matthew Jones, 'How we became instrumentalists (again)?' about Positivism after WWII, about how machine learning filled a need, especially in the military, and also businesses, to take these large data bases that they had and do something with it. Does that thing work? Is it meaningful? Is it accurate? Well, by the time you've made a decision on it, maybe you don't have the time to test it properly. In statistics, data mining for a long time was a kind of insult that 'oh that is just data mining', when in fact data mining filled very big need for businesses which was to take these massive data bases and do something with them. And another reason I, or another really great example I'd like to want to bring out, this is from Dan Bouk, *Creditworthy.* There was a black legislator in the 19th century in Massachusetts, he was born into slavery and then after the end, or after the emancipation, moved to Massachusetts, worked as a janitor, but served also as a state legislator. He introduced the bill to ban an insurance company practice of charging Black people more. There was a local lawyer who critiqued this bill saying that well if you just look at the numbers the death rate for quoting coloured people was much higher than the death rate for whites. So it's not discrimination, it's just a matter of business, right, it's not motivated by bad will or anything bad, and in fact, if you exclude this practice, you're probably going to cause terrible economic collapse, you'd cut Black people off completely from being able to access insurance.

This argument in 1870, or I forget, maybe was a little bit later, was rejected and his allies noted that these statistics, these empirical realities that this lawyer was bringing up, were asking the wrong question. It wasn't about who was poor insurance risk but about the potential for equality, and the idea being that if you offer Black people insurance rates that are worse than white people, they're further entrenching this inequality, you are constraining the potential for equality. This was a fight that in the 1800 was one but as people like Kelly Horn, historians, one of my colleagues, (unintelligible name) and others are picking up was sort of refought and lost in the 1970s around actuarial fairness. Insurance companies instead, if it predicts, it's fair, and they won that argument, and that's become more and more entrenched in the world. And I think we can indeed reference the central mass over the outliers, over minority groups and that would probably make more money, would probably lead to more optimal outcomes as defined quantitatively, but of course the consequences of that are much harder to quantify, much harder to see if we're only looking at statistics and numbers and that's maybe why we still move more and more towards a hope in Positivism, this hope that what is measurable will give us everything, is everything that we care about and need. And so that's something that I think whether in the use of machine learning or even just in the extension of what are used to predict scores or insurance rates we need to think critically about and be willing to reject and not accept this argument that 'oh it's just the empirical reality, oh it's just optimal, oh it's just fair, it's just efficient', when in fact these notions of fairness, and even efficiency are made on the basis of a very specific set of values that we maybe should reject.

Going back to Google Flu Trends, I mentioned that the model was constructed properly. One important thing about machine learning is that it only has past data on which to validate. A typical way of doing that, since it's no longer guided by theory, is to hold out a portion of data. If you have a time series, you should do this in temporal blocks which they do correctly. Use a first portion of data to do your curve fitting and then test on the remaining portion of data and see how well you would have done if indeed there has been this new data. In my paper I talk about all the ways this

can break down, and one of these ways, as Google Flu Trends showed, is if there is unexpected variability in the world. If there is some underlying causal process that has only acted one way up until now, but now acts another way and will completely throw off your machine learning performance.

There's an example that I'd like to give: a paper was originally published in 2002, it found about 70 genes that correlated with developing breast cancer. And of course, they did the machine learning properly, so this is very optimal when holding out a certain portion of data, specifically they used leave-one-out-cross-validation (LOOCV), training on the other sets, asking on this held out thing, but does it actually generalise to the world?

It was only in 2016 that, again I'm showing the branch, instead of going down this branch, or rather k-fold, leave one out as an example of k-fold cross-validation, another paper went on this other branch doing two out-of-sample testing and experiment.

This is published in the New England Journal of Medicine in 2016, there they compared a traditional diagnosis versus the diagnosis given by the machine learning system, and if the machine learning system said the risk was high, the clinical risk was high, people were put into chemo-therapy, no question. Similarly, if both tests showed low risk, people were not put into chemo-therapy. The question was when they disagree, when the clinical risk is low but the gene expression risk is high, what do you do? And same with the other side, the doctor says to treat and the model says don't, you have this question about these two squares of the box. So they did an experiment, people falling into these two boxes were randomised into chemo-therapy and not chemo-therapy and what they found is for people in this box chemo-therapy actually led to the worst outcomes. In this box, chemo-therapy led to similar outcomes and, since it's a very painful and long process, if it doesn't affect your outcomes, your preference is not to give people chemo-therapy. And so that was able to then give a sort of decision criteria about how to use the machine learning system, machine learning alone would have made things worse. So if you just use this row of using the things that were high, but now with this experiment we can see there's a secondary diagnosis, it can catch false positives and avoids unhelpful chemo-therapy but only as a secondary diagnosis. And this, we don't understand what in these several gene expressions is causing breast cancer, the initial experiment was only done with I think 64 people, or maybe, certainly less than 200 people, which is not a very good sample size. So we want to repeat this with the larger more representative sample. This test also was only done in Europe, across several countries, but still it wasn't global, so there's a lot of ways in which the sample was not representative, all the same considering it still applies around sampling. But this is a good approach of doing experimental testing of a machine learning system as a whole to see how it interacts with the world.

And I want to then conclude about imagining the future of machine learning and social research. An example, see, Juliana brought up an article by Buckingham Shum that gives two examples: one is the Open University in the UK, they have this project Open University Analytics, and a similar project in Georgia State University. And I would say, before accepting the claims that a machine learning system works and predictive analytics work, we should have real world and holistic testing and that might be, for example, in Georgia State University, all the money that was spent on this analytics platform could have been put into additional teachers as well, additional guidance counsellors and human support. And one big thing that critics point out is that Georgia State University along with this predictive system about who is going to drop out, who is going to fail, also hired a lot of new guidance counsellors. So was the causal impact really just the guidance counsellors, what if they

had spent the money just doing more bad instead. So there are a lot of things happening that are hard to tease apart. One thing that I like about the Open University Analytics Project is that they in fact did qualitative assessment, and I think ultimately that might be the best way to judge these predictive systems, not to do experiments, not to measure how well they do, but to do qualitative research and see how they impact people's, in this case learning and teaching and interventions. And as I talked about in my paper, as a lot of literature talks about, experiments have endless contingency as well. It's really hard to definitively say that an experiment does or does not generalise, does or does not give us what we want to know, even if the results are significant. Do the results from UK transfer over to the US? Do the results from a university serving primarily first immigration students generalise to a really prestigious university serving mostly lead students? So there are all these questions and there's something called experimental regress that you can always challenge every little bit in the experiment that you never quite know, you have to have a whole set of theories about what's valid, what works that are completely before the quantification process. For labelling, I think, I guess some examples in my paper about the importance of bringing in the main knowledge of doing qualitative coding for machine learning imitation and labelling, developing a codebook saying what should we classify as what category when the human labels are doing it before putting those human labels into a machine that can, into a model that can try to scale them off. And I think rejecting new and even existing forms of governance by correlations and that would also mean thinking hard about what's a better alternative. One good thing about metrics is that they can level the plainfield, whereas personal discretion is a recipe for authoritarianism, for discrimination, for further entrenching elites, but metrics can do the same things if gaming the system is the best way to win, and people with resources know how to game the system. We should think about alternatives to fairly distributing research resources rather than citation based in disease and metrics. We should think about how we can have ways of assessing students in institutionalised standardised way that aren't grades, and that's going to be a very hard process and in some cases metrics might be the least valid solution, but they are not ever the only solution. And as long as we rely on them, as long as we further entrench them, whether that's with machine learning or with statistics, we are going to run into the same problems again and again. And thank you for your time.

Traducción Inglés-Castellano

Una jerarquía de limitaciones en Machine Learning. Sesgos en los datos sociales

M: Voy a retomar los temas que Juliana también había puesto en su entrada de blog que creo que compartió antes trámite el enlace, y esas fueron algunas de las preguntas que repasamos una y otra vez a lo largo de la conversación que tuvimos por correo electrónico durante los meses pasados. La primera pregunta que me había hecho Juliana es cómo mi experiencia me ha conducido a mi trabajo actual y cuál es la jerarquía de la que estoy hablando.

Como se mostró anteriormente, tengo experiencia en muchas disciplinas y comencé de las humanidades y las ciencias sociales para pasar a la informática porque tenía también experiencia en historia de la ciencia, ciencia y tecnología, estadística, y estaba realmente insatisfecho con el estado del discurso que veía alrededor, en ese entonces se hablaba de los Big Data. A medida que aprendí más, comencé a entender que el machine learning sería el aspecto más importante para considerar, pero, por supuesto, este se encuentra bajo muchas etiquetas: inteligencia artificial, deep learning y todos estos están relacionados con la ciencia de datos. Así que me interesé a la informática tratando de entender el trasfondo para tratar en detalle y críticamente parte del material, y lo hice tanto externamente como internamente. Entonces, les comentaría el tema del artículo que será al centro de la charla de hoy: ¿el modelado realmente funciona? Y, como se discutió en breve, ¿el machine learning en realidad es sólo otro tipo de modelado, casi idéntico al modelado estadístico pero para propósitos diferentes y con un estilo diferente? La respuesta es, como ocurre con muchas cosas en las ciencias sociales, quizás, a veces. Si se hace bien y tenemos suerte. A menudo, ni siquiera tenemos la capacidad de saber si funcionó o no y eso también se convierte en un desafío. También desglosaré algunas de las afirmaciones que suceden en torno al machine learning y las interrogaré críticamente y mostraré que, tanto internamente como externamente, probablemente son exageradas en su mayor parte, especialmente cuando están hechas en comunicados de prensa, hechas por empresas que venden productos, pero creo que incluso los artículos académicos no hacen un buen trabajo para comprender las limitaciones y para enmarcar las cosas de manera responsable.

Lo que Juliana escribió en su entrada de blog se basa en un artículo y es una especie de árbol de enfoques metodológicos y esta es la estructura de mi artículo: voy por diferentes ramas de esto y me planteo ¿cuáles son las ventajas y las desventajas? ¿Por qué algunas personas prefieren bajar de esta rama en cambio que de esa? Yo no tengo la profundidad para saber cuáles son los otros tipos de ramas. Bajo 'simulación', sé que hay modelado sin agente y sé que hay dinámicas de sistemas, ecuaciones, hay también ecuaciones diferenciales parciales y otros tipos. Entonces se pueden elegir diferentes enfoques para modelar cosas incluso dentro de estos aspectos cualitativos, ciertamente hay muchos enfoques diferentes, análisis de contenido o teoría fundamentada de la etnografía, de los cuales nuevamente no tengo una buena comprensión del rango, así que los dejé como líneas punteadas. ¿De dónde surgió esto? Quizás se explique mejor lo que estoy tratando de decir a través de la discusión de la literatura. Juliana compartió conmigo algunas de las investigaciones que había realizado en los años 2000 sobre la 'guerra de paradigmas', y esto es algo de lo que no había oído hablar, pero conozco algunos de sus productos, o sea estas síntesis metodológicas o elementos para entender cómo combinar la investigación cuantitativa y la investigación cualitativa. Juliana me comentó que este trabajo se realizó al final de la guerra de paradigmas, y esto es de 1988 y trata de explicar para qué sirve la investigación cuantitativa, para

qué sirve la investigación cualitativa, cómo se hacen bien ambos y cómo se combinan. Estos son debates que aún continúan hoy, pero que son completamente desconocidos dentro de la mayoría de la literatura sobre informática y machine learning que leí, entonces los quiero incluir. Pero esta es una rama y creo que muchos de nosotros no hemos estudiado esto en profundidad, no tenemos idea de que experimento versus observación es otro gran debate en la literatura cuantitativa, especialmente en psicología, econometría, economía en general, que se puede decir cualquier cosa sin experimentos, sin manipular el mundo. Eso es, por un lado, donde las personas consideran que la causalidad es, por definición, manipulación o intervención, y que la causalidad es lo que realmente importa, por otro lado, la gente dice que muchos experimentos en psicología no tienen validez ecológica, la forma en que las personas se desempeñan y se comportan en estos contextos artificiales no generaliza en absoluto. Se llega a la conclusión que no se aplica de ninguna manera a lo que sucede en la sociedad y en el mundo, esa es la otra cara de la observación y la experimentación. Una buena revisión de todo esto se encuentra en un volumen colectivo de 2014, aunque trate específicamente de experimentos de campo, entra en esa división. Creo que se produjo un gran avance con este artículo de 2010 de Galit Shmueli. Ella dice que hay una diferencia entre la explicación y la predicción y que no entender esta diferencia es un gran fracaso pedagógico dentro de las estadísticas y también un gran malentendido en torno al machine learning y, específicamente al final del artículo, da un ejemplo matemático de cómo un modelo que predice bien, como se define la predicción en las estadísticas y en el machine learning, en realidad no puede capturar el proceso causal, por lo que se crea una sensación de modelo falso, uno que refleja peor este modelo de juguete pero performa mejor en la tarea de predicción. Y, aunque ese es un ejemplo muy artificial puesto en términos matemáticos, el desacoplamiento de esta idea de que cualquier predicción explicará fue realmente útil para mí para pensar en la diferencia entre las metodologías que a menudo se abordan en estadística y lo que el machine learning está tratando de hacer. Desde 2000/2001 se ha escrito mucho sobre cómo el machine learning está convergiendo con las estadísticas, cómo el machine learning en principio era más un conjunto de métricas ad hoc, pero gradualmente se colocó en una base estadística, utilizando elementos como la teoría de la probabilidad, la concentración de medida, la teoría de la estimación y todas estos otros aspectos propios de la estadística. Es un poco distinto en torno al deep learning, pero la idea de fondo es que se usan datos para realizar con la ingeniería inversa alguna función subyacente o límite o cosa para imaginar que existe en el mundo. Ya sea que se haga con el propósito de explicar las cosas para la comprensión o simplemente para hacer predicciones es la principal diferencia entre lo que intentan hacer las estadísticas y el machine learning. Por supuesto, ambos pueden hacer cualquiera de las tareas y las hacen de manera ligeramente diferente, pero esa es la gran diferencia entre ellos, aunque ambos usarán regresión lineal, regresión logística, o exactamente las mismas matemáticas, exactamente las mismas fermentaciones de software y mismos datasets también, pero con una diferencia en el estilo y el resultado.

La simulación es una de las cosas que me ayudó a construir esta jerarquía. En comparación con la que está específicamente basada en agentes, la simulación es una alternativa al modelado estadístico, por lo que sí tienen datos, probablemente no se verán muy bien servidos por la simulación porque es difícil poner esos datos en simulación. Pueden simplemente inicializar y luego comparar los resultados, pero es difícil usarlo directamente si no pueden manipular un sistema, ya sea por razones éticas o logísticas, tal vez las estadísticas no les brinden lo que realmente les importa. Entonces, los promotores de la simulación dicen que la simulación les dará la causalidad que podría interesarles. O en epidemiología, y he cambiado el nombre de esta categoría, primero lo llamé modelado matemático, pero luego descubrí que en la epidemiología que distinguen el modelado de datos del modelado mecanicista. Y creo que el modelado mecanicista ha recibido más atención con el Covid-19, sobre todo con el modelo SIR, población susceptible, infectada, recuperada (o fallecida). Estos son conjuntos de ecuaciones diferenciales que son difíciles de

ajustar a los datos, pero brindan una comprensión conceptual y permiten argumentos sobre la dinámica y cómo funcionan las epidemias. Y eso no quiere decir que no se puedan comparar, pero el valor principal de estas ecuaciones y estas derivaciones no es ajustarse a los datos o explicar las formas que vemos en los datos, sino más bien como un conjunto de primeros principios para comprender lo que podríamos ver en los datos o cómo podríamos tratar de manipular el mundo para cambiar los datos que terminamos viendo, así que cambié el nombre de esta categoría en mecanicista según la epidemiología.

Así se forma esta gran rama en forma de árbol e implícitamente, cuando elegimos una disciplina, cuando elegimos una metodología, estamos en alguna rama de este árbol, y hemos tomado implícitamente una serie de decisiones, ya sea en un proyecto en particular o incluso en nuestra formación, sobre lo que hemos priorizado, sobre lo que podemos hacer, lo que no podemos hacer y cada uno de estos tiene ventajas y desventajas. Hablé sobre algunos de ellos entre, por ejemplo, simulación o ecuaciones, y el uso de datos o datos de modelado, y antes, creo que tal vez podamos comentarlo más adelante, existían estos debates sobre la investigación cuantitativa versus la investigación cualitativa. Encontré este artículo que realmente me ayudó a establecer este vínculo entre el machine learning y las estadísticas como una especie de rama en este árbol. Yo diría que el machine learning convencional trata de seguir esta secuencia de ramas y lo describiría como una especie de definición práctica que indica la extrapolación de correlaciones para anticipar los resultados de un sistema estático. Diría que el aprendizaje en el machine learning es completamente una metáfora ahora, tal vez al principio intentó capturar algunos procesos de aprendizaje, pero, por ejemplo, se parece muy poco incluso a la teoría del aprendizaje cognitivo y casi no se parece a la teoría del aprendizaje de la cultura social, por lo que es puramente una metáfora para mejorar con más inputs. Realmente se trata de usar correlaciones y, como hablé más en mi artículo y como habla mucha literatura de ciencias sociales, las correlaciones usan la media, los anuncios, un tipo de tendencia central y que privilegia la masa a expensas de valores atípicos o personas que no encajan en todo lo que es común y eso también puede tener poderosas consecuencias normativas. Por lo tanto, el machine learning se basa en el mismo conjunto de maquinaria que aparece cuando se utiliza el modelado basado en probabilidades. Por último, también hablaré de esto más adelante, el machine learning en realidad no anticipa las formas en que un sistema puede cambiar. Si ya se ha observado una variabilidad en el sistema, el machine learning puede ser capaz de dar cuenta de eso, pero falla miserablemente cuando aparece algo completamente nuevo en el sistema, por lo que es muy reactivo, no hace esas conexiones más amplias que intentan hacer las estadísticas en términos de encontrar relaciones causales que luego puedan generalizarse a sistemas invisibles y, ciertamente, imaginar formas completamente diferentes de pensar, saber, ser y hacer que estén más allá en estas ramas. Incluso la indagación como una elección de rama, no es necesario que se acerquen al mundo a través de la indagación, puede ser a través de la especulación, a través de la práctica, a través de un montón de cosas, así que incluso esta es una rama de un árbol aún más grande.

Lo siguiente que me preguntó Juliana es sobre la responsabilidad de la cuantificación y cuáles son algunos de los inconvenientes de la cuantificación. Y estas son nuevamente cosas que eran bien conocidas en la guerra de paradigmas por lo que he recopilado y entendido de parte de la literatura que llegó de ahí, y por supuesto, muchos de estos estudios de tecnología científica, historia de la ciencia, han también dado su contribución. La idea es que la cuantificación privilegia una serie de significados sobre todos los demás. Theodore Porter en un gran artículo del 2012 habla sobre el proceso de reducción descriptiva. La descripción es, por supuesto, la idea antropológica, pero contrasta con lo que los conductistas posteriores a la Segunda Guerra Mundial consideraron sobre el significado de la reducción, o sea equivalente sólo a lo que es observable, a lo que podemos ver directamente, siendo lo único que importa, el único precio que vale la pena pagar sacrificando

algunos significados, es un precio que vale la pena pagar por estos resultados objetivos, o por una metodología objetiva adecuada. Y, por supuesto, la solidificación de una serie de significados probablemente llegue a todos los resultados de lo que se puede construir, pero el problema es que se contribuye a usar una serie de significados sobre cualquier otra posibilidad y luego se olvida o se guarda ahí y no se puede volver atrás y nada en la cuantificación puede deshacer eso. El análisis exploratorio de datos puede encontrar algunas discrepancias en los datos, como máximo puede volver al proceso de cuantificación y comprender lo que sucedió en él, pero la cuantificación en sí misma nunca puede revertir fundamentalmente ese proceso de cuantificación de elegir una serie de significados sobre cualquier otro, y haré algunos ejemplos de esto más adelante. Otro problema, y esto creo que puede contribuir directamente a la literatura sobre la guerra de paradigmas, es la diferencia entre lo que es medible, lo que es observable y lo que realmente nos importa. Hay una versión de una ley, llamada Ley de Goodhart / Ley de Campell de 1975 (se publicaron aproximadamente al mismo tiempo diferentes versiones de la misma ley): si se optimiza a una medida, esa medida se volverá mala o ya no medirá lo que originalmente estaba midiendo, cualquier regularidad estadística que se utilice o ?? dejará de ser útil. Y de lo que realmente se está hablando es de esta brecha entre lo que nos importa, en ciencias sociales y psicometría lo llamamos "el constructo subyacente", y lo que está operacionalizado, lo que realmente podemos observar y medir. Encontramos algunos ejemplos sobre esto en un artículo reciente que llamó mucho la atención publicado por Obermeyer et al. y habla sobre cómo el sistema de salud discriminaba a los pacientes negros porque estaba optimizando los costos de atención médica que eran medibles en lugar del objetivo real de la salud. Dado que históricamente se ha gastado menos dinero en pacientes negros debido a la discriminación sistémica y el racismo, el sistema simplemente estaba usando esos mismos patrones, optimizando los mismos patrones. En educación, creo que todos lo sabemos, las calificaciones son un indicador poco fiable de lo que en última instancia nos importa, que es el aprendizaje o tal vez la preparación para lo que sea, y el aprendizaje es infinitamente multifacético, subjetivo y contingente. Cristalizarlo en una serie de significados, como qué tan bien se puede rendir en alguna evaluación estandarizada o en una clasificación estandarizada de ensayos, quiere decir crear un indicador malo. Tal vez no hay un mejor indicador o algo que se pueda medir, pero cuando nos comprometemos en cuantificación y medición, ese es el sacrificio que tenemos que hacer para poder obtener calificaciones para millones de estudiantes, para poder sumarlas, para poder observar las tendencias en el tiempo y el espacio. Citaciones en la academia, somos juzgados por nuestro índice H, o sea, nuestro impacto, en qué revistas publicamos y el factor de impacto. En última instancia, las revistas se basan en cosas como las citaciones que son una manera poco eficiente, creo que todos lo entendemos, de evaluar eso. Algunas cosas que se citan en gran medida no tienen impacto, algunas cosas que nunca se citan tienen una influencia extremadamente alta o cambian la percepción de la gente del mundo, de la teoría, de la metodología, y esas no son cosas que puedan cuantificarse fácilmente, o de las que se puedan identificar los fines para los que se están utilizando. Por último, tanto los arrestos como las condenas son un indicador poco fiable para el constructo subyacente que podemos imaginar del crimen. Juliana me pidió ejemplos de esa medición en el machine learning, y creo que hay algunos ejemplos de esa medición y también algunos ejemplos de asumir que lo que está disponible es lo que nos importa, suponiendo que las calificaciones sean aprendizaje o que las citaciones sean impacto.

Un comunicado de prensa, a principios de este año, creo que en marzo o abril, de la Universidad de Harrisburg en Pensilvania, afirmó que un artículo de próxima aparición podría predecir la criminalidad sin prejuicios raciales. Hubo un grupo de académicos de la Coalición para la Tecnología Crítica que le escribió una carta a Springer diciendo a primera vista que esto es inverosímil, si no imposible, y que comete errores categóricos. Una de las cosas más importantes es que no existe la criminalidad, y mucho menos una que pueda predecirse automáticamente. La criminalidad se basa en lo que la gente decide que es criminal o no, así que, por ejemplo, en los Estados Unidos, no se criminaliza la deuda salarial por la que los empleadores no pagan a sus empleados y la gente puede salirse fácilmente con la suya. En cambio, cosas como holgazanear,

alguien simplemente decidió que eso es criminal, holgazanear es cuando estás en un lugar o esperas afuera de un lugar y no estás haciendo nada específico, eso está criminalizado, y otros comportamientos están criminalizados. Pero incluso conjuntos de códigos o de leyes, en realidad nunca tenemos estadísticas sobre la delincuencia, lo que tenemos son los indicadores de arrestos y condenas. En los EE. UU., Ciertamente estamos viendo la inevitabilidad de cómo estas son medidas realmente sesgadas: los negros, los latinos y otras minorías son arrestados a tasas mucho más altas. Del mismo modo, las condenas, las personas que incluso tienen la piel más oscura reciben condenas más largas y más severas por delitos similares, y ciertamente utilizando la evidencia cuantitativa aquí. Pero lo más importante es simplemente la experiencia de las personas de ser brutalizadas, de ser perfiladas, que conduce al sistema de criminalización y criminalidad, y eso no es algo que se captura a través de la fisionomía. La carta de hecho habla sobre la fisonomía y estas viejas prácticas de análisis eugenésico del cráneo y la propuesta del artículo es una versión de eso, pero no hay una conexión causal entre lo que quieren medir, la criminalidad y lo que se puede medir de otra manera, o sea las caras. Sin embargo, encuentran una correlación y la asumen porque hay un vínculo directo en lugar de examinar las métricas de las cosas que están sucediendo. Este artículo en realidad nunca se publicó, fue retirado antes de que se publicara, y yo mismo firmé la carta. Creo que el artículo en sí no tiene nada de especial, pero este es el tipo de suposiciones y medidas que ocurren en el machine learning o el uso de indicadores sin darse cuenta de que son indicadores y ese es un peligro importante. Creo que el caso de este artículo podría ser muy dañino en el futuro, así que pensé que era importante intervenir, pero este no es el primer estudio en hacer esto, este no será el último estudio en hacerlo, tanto en este dominio como en muchos otros. Y ese es el tipo de problema de medición.

Una cosa importante de la que hablar es que el machine learning hace coincidir lo que es medible y no los significados subyacentes. Un ejemplo que me gusta usar es el del reconocimiento de imágenes y las redes neuronales profundas, que son las cosas más impresionantes que surgen del machine learning. Lo que hacen los sistemas es correlacionar patrones en píxeles, y lo sorprendente es que pueden extraer automáticamente diferentes patrones que correlacionan con etiquetas dadas por humanos, y ese es el tipo de conexión que establecen. Pero en las ciencias sociales diríamos que hay un constructo subyacente de la característica gatuna que indica a los humanos que etiquetan estos espectros como los de los gatos, y además combina ciertos patrones de píxeles, pero por supuesto que la característica gatuna no es una cosa sola. Para hacer etiquetas humanas, tenemos que decidir qué cuenta como gato para este propósito y qué no. Les doy un ejemplo, ¿esto es un gato o no? Bueno, alguien en Mechanical Turk podría decir inmediatamente 'sí, es un gato'. Claro, parece un gato, pero, si nos interesa la filogenética (es una especie de biología evolutiva), entonces este es un ‘falso dientes de sable’ que es un marsupial y no está relacionado en absoluto con los felinos, o ese orden. Por el contrario, ¿esta imagen contiene dicha característica gatuna? Y de nuevo, si alguien en Mechanical Turk estuviera pasando la foto, diría "no", pero en realidad hay un gato de este tamaño en la imagen y entonces diríamos 'sí, hay un gato'. Y entonces podemos decidir qué nos importa capturar en la categoría de gato, y no hay nada que podamos hacer, lo tenemos que decidir nosotros. Y si ignoramos eso, obtendremos casos extremos, obtendremos valores atípicos que se clasificarían erróneamente y que se etiquetarían como "gato" o "no gato" de formas que pueden ser perjudiciales. Este es un ejemplo probablemente inofensivo, pero muestra la importancia de considerar el constructo subyacente y no solo enfocarse en el increíble e inmediato resultado de tener miles de millones de imágenes que pueden etiquetarse con ‘gatos’ o ‘no gatos’, como usar la búsqueda de imágenes de Google.

No me adentraré demasiado en disciplinas como la psicometría, en cómo se crean las medidas y cómo las validamos, pero hay validez de constructo, validez aparente, validez de criterio y relacionada, validez interna y externa, que en general son desconocidas en el machine learning.

Otros artículos, además del mío, están tratando de aportar esto al machine learning donde en realidad solo se mira la validez externa y se ignoran todas estas otras cosas, y a veces está bien, a veces está bien ignorar todo lo demás. Pero otras veces, cuando queremos hacer afirmaciones sobre el mundo, la validez externa por sí sola no es suficiente. Podemos lograr la validez externa mientras fallamos en todas estas otras cosas.

Así que esos son algunos ejemplos de medición. Otra cosa de la que quiero hablar, adentrándome un poco en mi artículo, es la idea de la performatividad que, por supuesto, es mucho más antigua, pero en los estudios de ciencia y tecnología, Donald MacKenzie y otros escribieron el libro *¿Los economistas crean mercados?* Y no los mercados existen y luego los economistas los estudian. En particular, Karen Healy escribió un muy buen resumen de esto que dice 'la tesis de la performatividad es que la economía produce un cuerpo de modelos formales y técnicas transponibles que, cuando se llevan a cabo en el mundo, reorganizan los fenómenos que los modelos pretenden describir ... '. Y una forma de entender esto es que, independientemente de que los modelos fueran verdaderos antes o no, cuando la gente comienza a usarlos para pensar en el mundo y acercarse al mundo, en efecto se vuelven verdaderos o se vuelven realidad. Es difícil de encontrar ejemplos precisos y claros de esto, pero en un sentido más conceptual, si miramos los modelos como propósitos autocumplidos, podemos ver ejemplos donde esto podría estar sucediendo. Un ejemplo que tuve en un trabajo mío es acerca del sistema de sugerencias de amigos de Facebook que asume que existe la clausura triádica, y este también fue el ejemplo que dio Karen Healy y mi artículo entra más en el detalle cuantitativo. Este modelo asume que los amigos, nuestros amigos, se conocen, y en el análisis de redes sociales lo llamamos clausura triádica o transitividad, y Facebook construyó el sistema en base a eso. Si son usuarios de Facebook, verán entre las sugerencias personas con las que tienen 38 amigos en común, 115 amigos en común. Lo que sucedió cuando se introdujo el sistema en Facebook es que la ley contempla que las conexiones que la gente hacía eran amigos de amigos. ¿Facebook estaba realmente utilizando el principio subyacente de la clausura triádica? ¿O qué tipo de instrumento latente existe en Facebook? ¿O simplemente se le dio a la gente un botón fácil para hacer clic según el comportamiento que Facebook asume que estaba teniendo al usar o ver algo? Entonces, la manipulación de sistemas con estas implementaciones de modelos de machine learning, que solemos llamar algoritmos, ¿están cambiando el mundo de una manera que afectaría lo que creemos que está sucediendo? Eso es muy sutil y algo que me costó aplicar, pero es una idea muy importante en la que pensar sobre cómo los modelos afectan al mundo, en lugar de simplemente reflejar el mundo.

Otro tema importante, sobre el cual también estoy trabajando en un artículo a parte, es el problema de la palabra "predicción". Me encanta esta citación de Daniel Gayo-Avello que habla de predecir los resultados de las elecciones con Twitter en 2012, dice: "Estas predicciones no son predicciones en absoluto. No he encontrado un solo artículo que prediga un resultado futuro. Todos afirman que se podría haber hecho una predicción; es decir, son análisis post-hoc y, huelga decirlo, es difícil encontrar resultados negativos.' Hay un par de cosas que sobresalen en esta citación: una es una crítica del sesgo de publicación que los resultados negativos no aparecen, que se aplica al machine learning tanto como cualquier otra cosa. La otra cosa es que la predicción en las estadísticas, sobre la cuan encontré casos incluso desde Carl Peterson en los años 20 y 30, se usa como sinónimo de correlación, y esto tiene cierto sentido que la mejor manera de extrapolar, de mirar al futuro, es usar la correlación que se ha visto en el pasado, pero la forma en que se usa ciertamente ahora en las estadísticas y en el machine learning da la impresión de que está sucediendo algo mucho más profundo de lo que realmente está pasando. La gente proclamaría que han predicho elecciones en Twitter cuando, de hecho, dicen que mirando retrospectivamente se pude obtener un resultado que correlaciona dos cosas. ¿Esto podría generalizar para el futuro? No lo sabemos y tenemos fuertes

razones para creer que no, no lo haría. Les recomiendo este artículo: esta versión está en archivo, tiene este título señalado aquí, y hay también una versión en revista.

Para explicarlo más detalladamente está este gran ejemplo del New England Journal of Medicine, es un artículo en broma que dice que el consumo de chocolate conduce a ganar premios Nobel y, de hecho, si miran el gráfico, más consumo de chocolate tiene un país, más premios Nobel tiene por población. En este artículo, Messerli escribe que hay proteínas en el chocolate que estimulan las funciones cerebrales, así que, obviamente, si los países quieren ganar más premios Nobel, necesitan comprar cantidades masivas de chocolate para la población. Y la broma, por supuesto, fue que la correlación no tiene causalidad y que podemos inventar historias causales plausibles para explicar correlaciones que pueden ser completamente inversas. Pero lo importante aquí es que el machine learning usa correlaciones como estas para hacer predicciones, hace regresiones, hace declaraciones descriptivas (y las estadísticas también lo hacen, esto no es exclusivo del machine learning) sobre una correlación y simplemente la individua como predicción. Entonces, si extraemos de esta tendencia lineal basada en el consumo de chocolate, prediciendo que Suiza está aquí arriba, estamos fingiendo que se mantuvo fuera de una muestra cuando en realidad solo lo hacemos con fines de ilustración retrospectiva. Y puedo adentrarme más en la validación cruzada y en cómo funciona la validación en el machine learning, pero la idea es que se pueden ajustar diferentes modelos una y otra vez hasta que se obtienen buenos resultados, o hay artículos que la gente intenta escribir una y otra vez hasta obtener buenos resultados.

La predicción, que es correlación, no es explicación, que es causalidad. Entonces, un modelo gráfico (y esto es algo que surge del machine learning pero que raramente se usa para modelar la causalidad) se refiere a los recursos nacionales que conducen tanto al financiamiento científico como al consumo de chocolate. La financiación de la ciencia conduce a premios Nobel y, por lo tanto, observamos una correlación, volviendo hacia atrás por el árbol, entre estas dos cosas, pero en realidad no podemos intervenir en el consumo de chocolate y esperar ganar cada vez más premios Nobel. En términos más simples, podemos preguntarnos dónde se rompe esta correlación, ¿fallan los patrones pasados? Por ejemplo, los pequeños países europeos que han tenido mucho éxito en el pasado, ¿cómo se van a comparar ahora con otros lugares como India, China, Brasil, Rusia, que están en este gráfico? Hay muchas cosas basadas en la historia de los premios Nobel que tal vez no se parezcan al futuro. También podemos decir que los premios Nobel supuestamente se otorgan sobre la base de un constructo subyacente de mérito. ¿Quién decide eso? ¿Cómo se mide eso? Y ese es también un proceso muy subjetivo que podemos imaginar que esté fácilmente sesgado, ciertamente sabemos sobre el sesgo contra las mujeres que ha existido en los premios Nobel, y ¿Qué otros sesgos existen todavía en términos de información general del país, raza y otras características personales?

También quiero enfatizar que este no es un uso obvio de la palabra "predecir". Hay un gran libro de filosofía del 1998, *Predicting the Future*, que afirma que la proyección de tendencias y el ajuste de curvas son formas razonables de hacer predicciones, pero son bastante rudimentarias. Lo que Rescher está tratando de decir es que cosas como las leyes científicas o las analogías son mucho más sólidas y sofisticadas en las formas científicas de hacer predicciones. Creo que lo que la gente tiende a no darse cuenta viendo los resultados del machine learning es que las predicciones son extrapolaciones basadas en la proyección hacia el futuro, basada en el ajuste de curvas. También hay críticos del machine learning que están muy en contra del ajuste de curvas, pero eso es lo que hace la gran mayoría del machine learning. Se necesita un dataset y se ajusta a una curva o un límite de decisión o algo así. ¿Eso generaliza? Tal vez, tal vez no. Si el mundo cambia, probablemente no.

Un buen ejemplo de esto son los Google Flu Trends, y aquí la validación cruzada se realizó correctamente, reteniendo los datos para probar el modelo y, a medida que se introduce la palabra datos en el modelo, funciona muy bien para mantenerse al día en las tendencias. Lo que hacen los Google Flu Trends es usar los datos de búsqueda de Google para tratar de anticipar las cifras de los Centros para el Control y la Prevención de Enfermedades sobre casos de gripe en los Estados Unidos que tienen que compilar manualmente informes de cuántos pacientes con gripe ven y los envían al gobierno de EE.UU. de forma centralizada. Entonces, Google estaba tratando de usar los términos de búsqueda que la gente usaba para hacer este proceso mucho más rápido y obtener proyecciones de los casos de gripe antes de que los CDC pudieran compilar laboriosamente las cifras oficiales. Está hecho correctamente en términos de modelado y en términos de validación, desafortunadamente hubo cosas en el mundo que el modelo simplemente no pudo captar porque estaba basado en correlaciones, porque no había observado esa variabilidad en el pasado. Estos son los números de la gripe reportados por los CDC, aquí hubo una segunda ola de gripe que los Google Flu Trends distorsionó por completo en 2009. Más tarde, los creadores volvieron, lo re-calibraron, fue bien, pero luego hubo otra temporada en la que se sobreestimó por completo la cantidad de casos de gripe. Un artículo de David Lazer y sus colegas titulado 'The parable of Google Flu Trends' dice que Google Flu Trends era por una mitad detector de invierno, por otra mitad detector de gripe, porque detectaba de manera no causal las que a veces él llama 'correlaciones espúreas' entre términos de búsqueda de invierno y gripe y no lograba a cambiar cuando había algo diferente en el mundo que ya no coincidía con eso. Volveré de nuevo a este ejemplo más tarde.

Otro ejemplo que me pidió Juliana se refiere a las motivaciones que siguen llevando el mundo a comprometerse cada vez más con las correlaciones, no obstante se conozca mucho más profundamente el funcionamiento de las correlaciones, las causales y las mediciones. Creo que una motivación es sin duda que el machine learning tiene algunos éxitos impresionantes, y creo que ahora hay un trabajo realmente significativo en la historia de la ciencia sobre las historias de la inteligencia artificial. Recomiendo un artículo de Matthew Jones, 'How we became instrumentalists (again)?', sobre el positivismo después de la Segunda Guerra Mundial, sobre cómo el machine learning llenó una necesidad, especialmente en el ejército y también en las empresas, de tomar estas grandes bases de datos que tenían y hacer algo con ellas. ¿Eso funciona? ¿Es significativo? ¿Es exacto? Bueno, para cuando hayan tomado una decisión al respecto, tal vez no tengan tiempo para probarlo correctamente. En estadística, la minería de datos durante mucho tiempo fue una especie de insulto "oh, eso es solo minería de datos", cuando de hecho la minería de datos llenó una gran necesidad de las empresas, que era tomar estas bases de datos masivas y hacer algo con ellas. Y otra razón por la que yo, u otro gran ejemplo que me gustaría destacar, es de Dan Bouk, *Creditworthy*. Hubo un legislador negro en el siglo XIX en Massachusetts, nació en la esclavitud y, después de la emancipación, se mudó a Massachusetts, trabajó como conserje, pero también se desempeñó como legislador estatal. Presentó el proyecto de ley para prohibir la práctica de una compañía de seguros de cobrar más a los negros. Hubo un abogado local que criticó este proyecto de ley diciendo que si solo se miraban las cifras, la tasa de mortalidad por citar a personas de color era mucho más alta que la tasa de mortalidad de los blancos. O sea, no es discriminación, sino solo una cuestión de negocios que no está motivada por mala voluntad ni nada malo, y de hecho, si se excluyera esta práctica, probablemente se causaría un terrible colapso económico, quitarías a la población negra el derecho de acceso al seguro.

Este argumento en 1870, o tal vez fue un poco más tarde, fue rechazado y sus aliados notaron que las estadísticas, las realidades empíricas que este abogado estaba planteando, estaban haciendo la pregunta equivocada. No se trataba de quién tenía un alto riesgo de seguro, sino del potencial

de igualdad, y la idea es que si ofreces a los negros tarifas de seguro peores que a los blancos, estás afianzando aún más esta desigualdad, estás limitando el potencial de igualdad. Esta fue una lucha que empezó en el 1800, pero personas como Kelly Horn, historiadores, uno de mis colegas, (ininteligible) y otros recuperaron esa lucha en torno a la equidad actuarial, la combatieron y la perdieron de nuevo en la década de 1970. En cambio, las compañías de seguros, si predecían, entonces era justo, y ganaron ese argumento, y eso se ha afianzado cada vez más en el mundo. Creo que de hecho podemos hacer referencia a la masa central sobre los valores atípicos, sobre los grupos minoritarios y eso probablemente generaría más dinero, probablemente conduciría a resultados más óptimos según se define cuantitativamente, pero por supuesto las consecuencias de eso son mucho más difíciles de cuantificar, mucho más difícil de ver si solo estamos mirando estadísticas y números. Tal vez por eso todavía nos movemos cada vez más hacia una esperanza en el positivismo, esta esperanza de que lo que es medible nos dará todo, es todo lo que nos importa y necesitamos. Eso es algo que creo que, ya sea en el uso del machine learning o incluso simplemente en la extensión de lo que se usa para predecir puntuaciones o tasas de seguros, debemos pensar críticamente y estar dispuestos a rechazar y no aceptar este argumento de que 'oh, es solo la realidad empírica, oh, es simplemente óptima, oh, es justa, es simplemente eficiente', cuando en realidad estas nociones de equidad, e incluso de eficiencia, se basan en un conjunto de valores muy específicos que quizás deberíamos rechazar.

Volviendo a los Google Flu Trends, como he dicho antes, el modelo se construyó correctamente. Una cosa importante sobre el machine learning es que solo tiene datos anteriores para validar. Una forma típica de hacerlo, dado que ya no se rige por la teoría, es retener una parte de los datos. Si tienen una serie temporal, deben hacer esto en bloques temporales porque así lo harán correctamente. Usen una primera porción de datos para hacer el ajuste de curva y luego prueben en la porción restante de datos y vean qué tan bien lo habrían hecho si hubieran tenido estos nuevos datos. En mi artículo hablo de todas las formas en las que esto puede fallar, y una de estas formas, como se demostró con Google Flu Trends, es si hay una variabilidad inesperada en el mundo. Si hay algún proceso causal subyacente que solo ha actuado de una manera hasta ahora, pero ahora actúa de otra manera, afectará por completo el rendimiento del machine learning.

Hay otro ejemplo que me gustaría hacer: hubo un artículo que se publicó originalmente en 2002 y encontró alrededor de 70 genes que se correlacionaron con el desarrollo de cáncer de mama. Por supuesto, usaron el machine learning correctamente, y esto es óptimo cuando se retiene una cierta parte de los datos. Específicamente usaron la validación cruzada dejando uno fuera (leave-one-out-cross-validation; LOOCV), entrenando los otros conjuntos de datos, preguntando sobre lo retenido, pero ¿esto generaliza realmente al mundo?

De nuevo estoy mostrando las ramas del árbol, en lugar de bajar por esta rama, usando la validación cruzada dejando uno fuera como ejemplo de validación cruzada de K iteraciones, otro artículo del 2016 pasó a esta otra rama haciendo la prueba y el experimento de dos fuera de muestra.

Esto se publicó en el New England Journal of Medicine en el 2016, donde compararon un diagnóstico tradicional con el diagnóstico dado por el sistema de machine learning, y si el sistema de machine learning decía que el riesgo clínico era alto, se ponían personas en quimioterapia sin lugar a dudas. Del mismo modo, si ambas pruebas mostraron un riesgo bajo, las personas no recibieron quimioterapia. La pregunta se presentaba cuando los dos diagnósticos no estaban de acuerdo. Cuando el riesgo clínico es bajo pero el riesgo de expresión génica es alto, ¿qué haces? Y lo mismo con el otro lado, el médico dice que hay que tratar y el modelo dice que no, entonces la

pregunta recae sobre estas dos partes del cuadrado. Así que hicieron un experimento, las personas que encajaban en estas dos partes fueron asignadas al azar a quimioterapia y no quimioterapia, y lo que encontraron es que para las personas en esta parte, la quimioterapia en realidad condujo a los peores resultados. En la otra parte del grupo de validación, en cambio, la quimioterapia produjo resultados similares y, dado que se trata de un proceso largo y muy doloroso, si no afecta los resultados, la preferencia es no administrar quimioterapia a las personas. Este experimento luego pudo dar una especie de criterio de decisión sobre cómo usar el sistema de machine learning sin usarlo de manera exclusiva porque hubiera empeorado las cosas. Entonces, con este experimento podemos ver que hay un diagnóstico secundario que puede detectar falsos positivos y evitar la quimioterapia inútil, pero solo como un diagnóstico secundario. No entendemos qué está causando el cáncer de mama en estas diversas expresiones genéticas, el experimento inicial solo se hizo con alrededor de 64 personas, ciertamente menos de 200 personas, lo que no es un tamaño de muestra muy bueno. Entonces lo ideal sería repetir esto con una muestra más grande y representativa. También, esta prueba se realizó solo en Europa, en varios países, pero aún así no fue global, por lo que hay muchas formas en las que la muestra no era representativa. Pero este es un buen enfoque para realizar pruebas experimentales de un sistema de machine learning en su totalidad para ver cómo interactúa con el mundo.

Me gustaría cerrar la presentación imaginando el futuro del machine learning y de la investigación social. Juliana mencionó un artículo de Buckingham Shum que da dos ejemplos: uno es el de la Open University en el Reino Unido que tiene el proyecto Open University Analytics, y el otro trata de un proyecto similar en la Universidad Estatal de Georgia. Antes de aceptar las afirmaciones de que un sistema de machine learning funciona y el análisis predictivo funciona, deberíamos tener pruebas holísticas y del mundo real. Por ejemplo, en la Universidad Estatal de Georgia, todo el dinero que se gastó en esta plataforma de analíticas, también podría haberse gastado en maestros adicionales, consejeros adicionales y apoyo humano. Y una gran cosa que señalan los críticos es que la Universidad Estatal de Georgia, junto con este sistema de predicción sobre quién abandonará y quién fracasará, también contrató a muchos consejeros nuevos. Entonces, el impacto causal fue realmente solo con respecto a los consejeros, ¿y si hubieran gastado el dinero simplemente causando peores consecuencias? Así que están sucediendo muchas cosas que son difíciles de separar. Una cosa que me gusta del proyecto de la Open University Analytics es que, de hecho, hicieron una evaluación cualitativa, y creo que, en última instancia, esa podría ser la mejor manera de juzgar estos sistemas predictivos, no haciendo experimentos, no midiendo qué tan bien lo hacen, sino haciendo investigación cualitativa y viendo cómo impactan a las personas, en este caso el aprendizaje y la enseñanza y las intervenciones. Como mencioné en mi artículo y como menciona mucha literatura, los experimentos también tienen una contingencia infinita. Es muy difícil decir definitivamente que un experimento generaliza o no, nos da o no lo que queremos saber, incluso si los resultados son significativos. ¿Los resultados del Reino Unido se transfieren a EE. UU.? ¿Los resultados de una universidad que atiende principalmente a los estudiantes de primera inmigración se generalizan a una universidad realmente prestigiosa que atiende principalmente a estudiantes líderes? Entonces están todas estas preguntas y hay algo llamado regresión experimental, siempre se puede desafiar cada parte del experimento que nunca se conoce del todo. Se debe tener una serie completa de teorías sobre qué es válido, qué funciona y esto sucede completamente antes del proceso de cuantificación. Con respecto al etiquetado, mencioné algunos ejemplos de (ininteligible) en mi artículo sobre la importancia de incorporar el conocimiento principal de hacer codificación cualitativa para la imitación y el etiquetado de machine learning, desarrollando un libro de códigos que diga qué deberíamos clasificar, como por ejemplo qué categoría usar cuando las etiquetas humanas lo están haciendo antes de poner esas etiquetas humanas en un modelo que puede intentar redimensionarlas. Creo que rechazar formas nuevas e incluso existentes de gobernanza mediante correlaciones también significaría pensar mucho en cuál es una mejor

alternativa. Una cosa buena acerca de las métricas es que pueden nivelar el campo, mientras que la discreción personal es una receta para el autoritarismo, la discriminación, para afianzar aún más las élites, pero las métricas pueden hacer lo mismo si engañar al sistema es la mejor manera de ganar, y la gente con recursos sabe cómo engañar al sistema. Deberíamos pensar en alternativas a la distribución justa de los recursos de investigación en lugar de citaciones basadas en métricas. Deberíamos pensar en cómo podemos tener formas de evaluar a los estudiantes de manera institucionalizada y estandarizada que no sean calificaciones, y ese será un proceso muy difícil y, en algunos casos, las métricas pueden ser la solución menos válida, pero no siempre son la única solución. Mientras dependamos de ellas, siempre y cuando las afiancemos aún más, ya sea con el machine learning o con las estadísticas, vamos a encontrarnos con los mismos problemas una y otra vez. Gracias por tu tiempo.

Webinar Series “Fair Data Cultures in Higher Education”
Transcription and Translation: Sofia Morandini