AAAI/CCC Symposium on AI for Social Good

Invited Talk 2- AI for Social Sciences
Speakers: Eric Rice (USC School of Social Work) & Sharad Goel (Stanford University)

Amulya Yadav: All right. Welcome back everyone. This is, it's time for our second invited talk, and I am very happy to welcome Eric Rice and [Sharad Goel 00:00:18], as speakers of our invited talk. Eric is an associate professor at the University of Southern California, in the school of social work. He's the co-founding co-director of the Center for Artificial Intelligence in Society at University of Southern California. Eric received his PhD from Stanford University, and he's an expert in social network and community-based research. His primary focus is in youth experiencing homelessness. He's been closely working with homeless youth providers in Los Angeles, and many other communities across the country, to develop novel solutions to end youth homelessness.

Sharad, on the other hand, is an assistant professor at Stanford, in the department of [inaudible 00:00:59] science and engineering, with [inaudible 00:01:02] appointments in sociology and computer science. His primary research is [computation 00:01:08] social choice and emerging discipline at the intersection of computer science and social sciences. He is also particularly interested in applying modern computational and statistical techniques to understand and improve public policy. With that, I'll hand it over to them.

Eric Rice: Let's see. Does this work? Yes! Okay. Good morning. This is a little bit of an intimidating audience to speak to. A, we're in this small little intimate room, and B, I am not a computer scientist. This is the second time that I've been asked to come up and talk in front of a computer science audience, or maybe the third, in my entire life. Forgive me. The last time I took a computer science class it was back in 1992, and it was about using [pascal 00:01:59]. Just to give you a sense of how out of the loop I am.

What I have done with [Milland Tambey 00:02:05] is that we have created the Center for Artificial Intelligence in Society at USC, which is a joint venture between the Viterbi School of Engineering and the Suzanne Dworak-Peck School of Social Work. When I talk to people about this, people usually look at me like I'm some sort of crazy person that's told them that they should be pouring maple syrup on top of sardines or something like this.

It turns out, it's actually a little bit more like bacon-wrapped dates. It's really, really tasty, and once you start getting into it, you're going to not want to be able to put them down, but it's challenging. One of the challenges is, in part, learning to speak one another's language. When we talk about models and you talk about models, we actually mean something different. When you use the word qualitative data, we would call that categorical data. There's many other instances where I've noticed these language barriers that we've had to overcome.

Sometimes I've been joking with Milland recently that I feel like my role recently has become a broker of social problems, that I'm gathering together other social work scientists and other social scientists and other community partners who have problems that they might want to have novel solutions for, and then bringing them to him, and we're trying to figure out where the angle where artificial intelligence can enter.

I think that really is a good place for me to stop for just a second. Where is it that this can enter? Like [Molia 00:03:36] said at the beginning of the morning, the popular public and press is all worried about killer robots, and I guess Uber's driverless car ran off the road this morning in Arizona someplace. But no one's really talking about right now there are actually techniques that are really, really useful, that could help us solve real world social problems in the current era. Not in some sort of science fiction future.

When Sharad and I were talking about how we would split the time in this talk, we were really thinking that there's two major areas that AI can contribute. One is in the context in which there are social problems where there's a massive amount of data that's been amassed. Medical records is an example of that, where you can then do machine learning techniques that may give you better predictive tools that could be useful to practitioners.

This is also really relevant for the space of social welfare, where we have oftentimes child welfare data where we're trying to see which kids have abuse reports that have been placed against their families, and we're trying to understand what happens to those kids over time, and we're trying to make better predictions about who might need additional intervention.

But there's a whole other set of problems, and this is a lot more like the problems that Carla was talking about, which are much more about strategy. That's really what I'm going to talk about more today. What I wanted to talk a little bit about, for those of you who are not in social work, which is basically everyone except for my two students who showed up today, is that social work, relative to other social sciences, is somewhat unique. It's much more akin to public health, or to medicine, in that it is a interdisciplinary space wherein people are fundamentally concerned with intervention.

I was actually trained as a sociologist, just down on the other corner from this building that we're in right now, and sociology has a fundamental interest in understanding the world and observing it, and commenting on it, but not in intervening. That's actually a deliberate part of their mission, is to watch objectively. Whereas social work has the opposite approach, which is that there are an enormous number of social problems out in the world, and fundamentally the work within social work is oriented around trying to come up with solutions to those problems.

It would be things like fighting inequality, helping people to live to their full potential, and we really focus on marginalized and disadvantaged groups. Homeless youth, which is what I'm going to talk about a lot in the content of this talk this morning, is definite a part of that.

Social work, like engineering, has a set of grand challenges that our academy of social work has created, and when we've been talking, Milland and I, about what is it that this center is doing, so I think one of the questions that [Amolia 00:06:35] posed was, what is the space of these problems? What do we want to include as AI for social good, and what do we not want to include as AI for social good?

One could argue that some sort of better market prediction algorithm is for social good. We would argue that's not something that we're particularly interested in.

These are the sorts of problems, from a domestic standpoint, that we're really interested in. You can see end homelessness. That's one of the big 12 things that social work is trying to work on. Ensure healthy development for all youth. That is also a big part of what I'm going to talk about today. Reducing extreme economic inequality, achieving equal opportunity in justice. These are lofty goals, but if they weren't difficult challenges, wicked problems, as Amolia said, then we would have already had solutions for them. But we don't.

We're also really interested in, and this is the second time you get to see this slide this morning, the UN's sustainable development goals. We've been thinking that, although we are in the United States, and a lot of the focus of social work is domestic, it's not only domestic. There is also a whole set of problems which are international in scope, and you can see that there's some overlap. Notice no poverty is up here as number one, and that is one that you see reflected in about three of the grand challenges from social work.

Okay. Now I'm going to talk to you about how this actually plays out in the real world. This is the part that gets me so excited. This is a picture of homeless youth. These are two actual homeless youth who I have known. I met these guys a few years ago, actually, on the streets of Venice, and if you look at some videos of me online, you'll see these two pop up. When most people think about homeless youth, this is probably the picture that they have. There's a lot of stuff. They're sitting on a corner. They've got a dog with them.

There are 1.6 million kids between the ages of 13 and 24 who have this experience at least once during the year. One night during the year, in the United States, every year. In Los Angeles, we are unfortunately the per capita hub of homelessness in the United States. We have 47,000 people on the streets of LA every night, and 6,000 of them are these young people between the ages of 13 and 24, who are there without their parents.

One of the major challenges that these young people face is risk of HIV infection. There's a lot of dangers associated with street life, and a lot of the coping strategies of street life lead youth into unprotected sex and sometimes using needles for drug use, which both have high rates of HIV transmission. What you can see here is that in the general population of housed kids, between 15 and 24, about less than .5% have HIV, whereas, among homeless youth, Robin, my student just finished doing a study of LA homeless youth last summer, and just based on self report, it was 7% of the kids had HIV. And, the scary thing about that is that about 30% of them haven't had an HIV test in the last six months, so there's an enormous amount of the population that actually has no idea whether or not they have HIV.

Where does artificial intelligence enter this problem? When I first started talking to Milland, I was talking to him about the social networks of homeless youth. He and I are both really interested in social networks. I've been studying social networks since people ... When people used to ... There was no Facebook when I started doing social network research. I was doing social network research back in the mid 90s and people would ask me questions like, "What is this esoteric thing, a social network?" The world has definitely changed.

This is a face-to-face social network of homeless youth in Hollywood. There's about 250 of them, and these are the edges that we were able to collect between these young people who are accessing one particular drop in center. We get at this by doing very painstaking interviews where we collect a lot of information about who these young people are connected to, and then really by hand, try to figure out exactly who it is that the young people are talking about one another.

One of the issues with this is that sometimes this is very easy, because they'll be talking about some guy named Psycho P and there's only one guy named Psycho P. We definitely know that that's Psycho P. Sometimes there's three guys named Mike though, and they're all white, and none of them have particularly distinctive tattoos, they're all about 25 years old, and so what we can get to is something like this, which is a set of uncertain ties. This is adding into the same map a picture of people who may be connected to one another, but we're not entirely certain.

That was the map that I started talking to Milland about. What I was really interested in was with respect to HIV prevention was an intervention called a peer leader intervention. I know some of you have probably read the HEALER papers, so you may have some inkling of where this is all going. What I want to stress here is that within the context of social work and social science engagement in general, I've been struggling to think about how do we start these relationships that lead to really productive problem solving?

I think part of it is that within the context of both medicine and also in social work, we've developed an enormous number of treatments and interventions.

Oftentimes the question is, how do we deploy those things to scale? Or how do we deploy them most efficiently? Or how do we deploy them with a new population?

Within the context of homeless youth, one intervention that's been very much talked about, but not yet deployed, was that we should be talking about a peer leader intervention. Now, why should we be talking about a peer leader intervention? Well, homeless youth, about 40% of them have been in the foster care system, and have fallen through the cracks of that system, so they've been abused and then they've been failed by adult social workers. About 40% of them are LGBT kids, lesbian, gay, bisexual, transgender kids, who've been thrown out of the home, or run away, because of so much discrimination and violence. About 80% of the kids that you see on the street have experienced some form of traumatic violent sexual or physical abuse.

These are young people who do not trust adults, and they don't trust adults for good reason. Adults have failed them pretty profoundly. Yet, they take care of one another. This is part of their natural process. They are really amazingly resilient and young people who are very much concerned about fairness, justice, and community.

My thought, as a social work researcher, is how can I harness this toward something very specific that I know is a problem in this community, that this community may not know is a problem, which is their HIV risk.

What we want to do is we want to find some guy who's going to be our champion, who's going to go and talk to his friends about HIV prevention and the need to get HIV tested, and they're going to do that, and they're going to listen to this peer, not me, because they don't trust adults. What we can think about in this context is that we might be able to ... Just have a seat. We might be able to really intensively reach out to a small number of young people, we can break through the barriers of them not trusting adults, and we have limited resources in these homeless shelter situations as well, so these are the limiting constraints.

Now, when I've thought about these network problems in the past, I've usually thought about them in smaller networks. If you look at some of my earlier papers, when I first got out of graduate school and I was talking about HIV influence, you'll see network pictures of like five people. It's pretty easy to be like, "Oh, hey. This guy's going to bridge, that might be a really good person to pick." Or, "This person has a really dense mass of ties. That person might be a really good person to pick." But remember, my network is this big, and might be this big. Then, to make matters even worse, this is subject to change ... I have five minutes left? Okay. Subject to change really, really quickly. These relationship may break very soon over time. Fortunately I'm not going to talk about the actual algorithm. I will leave that to you guys.

What did we do? Well, Amolia created one of the two algorithms which we deployed. This is a description of it. Essentially what it does is it partitions the network, finds the right people to pick and then hands them off to us. What we did was we trained this small group of peer leaders in a three to three-and-a-half hour training, and then we followed up with them once a week to give them reinforcement around their work trying to disseminate messages about the importance of HIV testing.

We had a pilot study where we tried HEALER, where we tried a second pilot study about six months later in a different agency, where we tried the DOSIM algorithm that was developed by Brian, and then we also, about six months past the original HEALER trial, we did a second trial where we did just the most popular.

This is sort of the default public health approach to this, which is the idea that if you could pick the most popular people you'd have the most spread of your influence. This is sort of the best thinking in the public health field to compare it to.

This is what happened in terms of our information dissemination over six months. This is for the people who were not the peer leader, so in each case about 20% of the network was selected as peer leaders, and you can see that about 70 or so percent of the network got reached in both the HEALER and the DOSIM and in degree centrality it was far, far, far less.

Even more importantly, of the youth who needed to get an HIV test, remember I was telling you there's about 30% of the young people out there that haven't had an HIV test that we need to convert, 40% of the ones that got reached, about, in both of these situations got reached, whereas in degree centrality none of them actually, of those people that were reached, changed their behavior and started getting an HIV test.

What happened? Well, one of our thoughts about what happened is that essentially degree centrality picks the popular kids. It's sort of like the high school football team. Whereas the algorithms are thinking much more about specific pieces of the network. It's looking for cliques. It's kind of like, I don't know if you guys remember, The Breakfast Club, the movie from the 80s, where there's a bunch of different people from different cliques that come together for this detention. Well, we didn't put them in detention. We put them into a training session.

This is our thought process about what happens, and it matters because the social process of the intervention is actually impacted by having people that can let down their guard, and be engaged with us as adults, as opposed to bringing in their buddies from the high school football team, where everybody had to be cool with one another the entire time.

I want to end this with just one last ... Two brief little anecdotes, and it will only take me a minute or so to tell each one. This is David. He's one of the guys that we picked in the very first pilot. He is a guy with severe ADHD. He is kind of all over the place, to put it bluntly. When we came to the intervention, and the HEALER algorithm told us that he was one of the folks to work with, when he came into the intervention session, one of the social workers who I know well at the agency space kind of looked at me like this. Like a confused dog, like, "That guy?" We're like, "Yeah. That's what the algorithm said."

When we worked with him, it turned out that not only was he a great peer leader, you can see the impact that he and his cohort had, but it changed his life. No one had seen him before. A year out now, he's in housing, he has a job, he's thriving in ways that he never did before, and it's in part because this algorithm broke our own biases.

The agency saw him as difficult to work with. We didn't think he was going to be particularly easy to work with. He wasn't easy to work with. But giving him that chance was an amazing thing.

I'll tell you one other anecdote about a young man who's not shown here, who is this guy named Jake. He is a stoner kid, to put it bluntly. He pretty much never showed up to the agency sober. When we got his name pulled from the algorithm, we couldn't find him. I ended up having to go out onto the streets of Venice and try to look for him. I finally found him on the beach, smoking a joint with a couple of his buddies. When I came up to him, I said, "Hey, Jake, how's it going?" "Oh, it's good, man." Okay. "So, remember that program you signed up for where we have the computer algorithm that picks people to be peer leaders?" "Oh yeah. Yeah." "It picked you." "Oh, sweet!" "So, day after tomorrow, we're going to be doing this training. It's going to be at nine AM at the agency that's about a mile from here. Are you going to be able to come and do that?" "Oh, yeah, man, I'll totally be there." Of course I thought he would not show up.

But it turned out, that when I got there at 8:15 to start setting up, Jake was sitting on his skateboard with a cup of coffee in his hand, stone-cold sober. "Hey Eric! Can I help you get the stuff out of your car and help you set up?" And he turned out to be one of our best peer leaders. He was so smart, so engaged. He was really, the last time I saw him, he was actually sober. He would show up at the agency not high, and he repaired some of his relationships with his family. He has a child that he's kind of reconnecting with. Just amazing things. This happened over and over and over again, and I think it's in part because the algorithm only cared about what we told it to care about, which is how these kids were connected, and we didn't have our own filter with our own biases that were then getting in the way.

While there's all this fear sometimes about how algorithms are going to create these horrible biases because they're going to be prediction algorithms that go

sideways, I'd like to throw out there the notion that actually, sometimes, algorithms can help you break down biases, especially if you're very thoughtful about what you want to put in as the input.

With that, I think I'm just going to say thank you, because I'm probably over time.

Speaker 1: We have time for one [inaudible 00:21:38].

Eric Rice: From a social work and sociology standpoint, the difference between those people is they tend to be relatively newer to the network. They also tend to be of two types. Either less engaged in risk, and they have small networks where they're trying to stay away from people who are bad news, which is also what you see by all those isolates that are out there. Those people, it's not that they don't have any social ties at all, they just don't have social ties to the other kids that you see in these settings.

The other piece of it is that sometimes they're groups of young people that travel. These are actually really, really high risk networks with a lot of heroin and methamphetamine abuse. These kids move from city to city by hitchhiking and taking trains. Some of the networks like that, we definitely really need to penetrate, and some of them, it might be okay if we missed them, but because we don't know exactly what people's behaviors are when they're just coming into an agency, the algorithm is kind of agnostic to whether that little cluster might be something where we have to get to, or that we don't have to get to.

Sharad Goel: Hi. I want to continue talking about how AI can be used to address public policy questions. Particularly talked about a key decision in the criminal justice system, and [methodologically 00:23:06], I'll explain this in a minute, and methodologically the point is to show that not only can the sophisticated ML algorithms be used but in many circumstances very, very simple [hiristics 00:23:17] can do just as well, and I think this has broad implication for a lot of policy problems.

This is joint work with a bunch of people. John [Benn 00:23:24]. John, a PhD student at Stanford, Connor [inaudible 00:23:28] at the New York City district attorney's office, Rory [Schroff 00:23:31] at NYU, and Dan Goldstein at Microsoft Research New York.

The application here is judicial decisions, and particularly pretrial release decisions. Shortly after arrest in the US, within about 24 hours after a defendant gets arrested, the judge has to make this key decision of whether or not that defendant will be released pending trial, or whether or not they're going to be detained pending trial. This is a key decision point, because even brief amounts of detention can result in pretty significant harms to the individual, loss of job, disconnection from their family, all sorts of problems could result from even small amounts of detention. Remember, this is all done under the presumption

of innocence. They haven't been convicted of anything, they haven't gone to trial, so now the judge has to make this key decision of whether or not, while they're awaiting trial, if this defendant is detained or released. We call released on their own recognizance.

The goal is, of course from the court's view, the goal is to make sure the defendant appears at their trial, so they want to minimize flight risk, and at the same time they want to minimize this burden of bail on defendants. They have to weigh these two factors and try to make a reasonable decision.

The status quo right now, this is changing across the country, but in many jurisdictions the judge is using their intuition, and they have a lot of experience, so they are reading over the case files, they're talking to the defendant, they're talking to the attorneys, and they're trying to make an informed decision. At the same time, this is a perfect situation for machine prediction.

The machine, why this is a very good situation, first, almost all of the information that's available to the judge is available in structured, digitized form to us. We have these extensive case histories. We know the nature of the offense. We know a lot of information that's available to the judge. The second the outcome is particularly clear here. the idea is that you want to minimize flight risk, you want to know, does this person eventually show up to their court date when they're asked to appear? It's a very clear prediction problem. It's a great setting for using these types of machine learning approaches.

Here we just did the standard thing. We have an outcome. We have lots of features. We fit random forest. You could do whatever you want, but here random forest is kind of the state of the art black box ML model, and we based this on about 150,000 cases from a large prosecutor's office.

Then, for every ... Once you train this model, for every individual you have a risk score. How likely are they to appear or not appear if they're released? Then you just detain the ones that are, the defendants that are deemed riskiest, and the rest you release. It's a very simple strategy. We're just take all the information available to us, we predict this outcome, whether or not they will appear at trial, if they're released, and then we detain the ones that are riskiest.

Now we have for every individual risk score, so this gives rise to a sequence of policies, a series of policies. We have to determine what that threshold for releasing or detaining individuals is. Wherever I set that threshold, that's going to determine my policy, so I have to pick this one point. In this case, it's actually a little bit tricky to evaluate any given policy. For what I mean by evaluating a policy, I have to estimate two quantities for the policy. First, how many people will be detained under that policy, and that's pretty easy to do. If I give you my threshold, and I have the risk scores, the people that are detained is everyone above that risk threshold. That's straightforward.

The second is, how many people are actually going to show up if I carry out that policy. This is tricky, because we have a lot of information available to us, but we don't necessarily have everything that the judge has. You can imagine that the judge is really looking at the person. This is a face-to-face interaction, and there might be some ... They might be using something that we just don't have access to, and so this is a standard causal inference problem. You can think of this as omitted variable bias, or a selection issue that the decision, detain or release, is not fully captured by the available information to us.

I'm not going to go into details of how we address this. I'll leave it in the paper, but I just want to point out that this is a tricky problem that comes up in all of these types of policy decisions, is that we're making ... We're training on data that is collected in a very particular way, and there might be selection issues in the data that we have. If it's the case that the judge has access to information that we don't have, the decision detain or release might be related to the outcome, appear or not appear, in ways that aren't captured in the data.

Question in the back?

Audience: [inaudible 00:28:11]

Sharad Goel: Here I'm not going to talk about it. We actually have another paper that deals entirely with the ethical issues on this, that maybe I can discuss afterwards, but here I'm just going to lay out our basic strategy for doing it.

We have this. Now how well does this strategy do? Here are two dimensions I just talked about. Proportion released on their own recognizance, so this is the easy access to estimate, and then here is a portion who failed to appear. The status quo is indicated by this point right here, so about 69% of defendants currently, no algorithm, are being released, and there's about a 13% fail to appear rate. That's our status quo.

Now when we implement this algorithm, so when we look at what would happen if we were to make decisions based on this algorithm, we get this line, and each point on this line corresponds to a policy which is based on a different threshold.

If I release ... I can release a lot of people, setting a very high release threshold, so only detaining the riskiest people, and in that case I'm going to see many more people who fail to appear, because I'm releasing a lot more people, so I can choose any point on this line.

Now the one thing to note here is this point right there. What is this point? This point says that I'm going to have the exact same fail to appear rate as I have in the status quo, but the point is that this is significantly further on the X axis, meaning I'm releasing many more people. In numbers, we can release, if we use these machine predictions, we can release 45% more defendants, and at the

same time maintain that failure to appear rate. Many, many more people are released, and at the same time we don't see any decrease in this flight risk, which is by law what the prosecution or what the judge is trying to maintain.

This is just kind of, if we ... I would say a standard place to stop is we have this black box ML model. We can show that it dramatically improves performance. You release many more people while maintaining the same objective that was already being implemented. But the problem is that it's hard to roll this type of system out. We've seen a lot of talks today about this is not just a theoretical intervention, that we actually want to do this, and so what are the real barriers that exist when trying to roll out these policies, is that it's hard to go to a judge and say, "Well, my black box algorithm is saying that there is a 20% chance that this defendant will fail to appear if you release him, but I can't tell you exactly why. I trained some fancy computer science model. If I were to write this down, it would be a thousand trees, based on a hundred different features. There's no way you're going to be able to explain this to the defense attorney. The defendant won't be able to respond." But you can say, "Well, a computer told me to detain you, so this is why I'm detaining you."

No one's going to be happy about that. We want to move from these black box machine learning algorithms to something that's very simple, and particularly something that's interpretable. Not only simple to construct and simple to explain, but simple to understand for ourselves. How are we going to do this?

First, what do I mean by simple? Let me give you a strategy which I think is simple. Here is a linear scoring role. It's based on exactly two features. The defendant's age and the number of past court dates missed. These are both features that people in the criminal justice system regularly associate with flight risk, so they're pretty natural. How does this work? Well, we just add up two numbers. We're going to look at for example a 28-year-old defendant with one missed court date, so we look at the age column, so this is score of four. We look at one missed court date, a score of six, so we have four plus six equals ten, so now we have a risk score. This is their risk score. It's very easy to compute. It's transparent. This is our proxy for this machine learned, this pretty sophisticated random forest model that was based on all of the information, 100 different features, and it was complex, but at the same time we didn't exactly understand what it was doing.

Here we're going to the other extreme. We're taking something very, very simple. This is for a risk score. You can also represent it graphically in this two dimensional plot, where we're saying this is the region where we release people, if we set the threshold, for example, at 10.5. Remember there's always this extra parameter of where do you want to set the threshold. Here we can say let's set a threshold at 10.5. We're going to release all of these people and we're going to detain all of these people. Again, it's transparent. It's easy to explain to defendant why we came to a decision, and it seems to make a lot of sense.

Audience: Just to be clear, a 50 plus year old person going to have to miss eight court dates to have the same score as a 20 year old missing one date?

Sharad Goel: Yes. Here we don't ... Actually under this key, the threshold 10.5, you're always releasing the small number of people who are older than 50 who are defendants.

These decisions ... These types of very simple decision rules, I think they have three nice properties, what we call fast, frugal and clear. They're fast in that these decision roles can be memorized. They don't need a computer to compute it. For example, a random forest, even though it's easy in theory to apply these things, you could put this on your phone, but still it requires some sort of computing device. This type of decision we can just memorize or put on a note card, so it's very fast to do this. It's frugal in that it only requires very limited information. Remember, all of the case features, there's something like 100 different features that we're using about the defendant, about their criminal history, about the case. Here we're only using two features, so it's easy to, again, apply this. It's clear, and this is perhaps the most important, is that you can explain to the defendant why a decision was made, and you can override these decisions. You can understand why exactly you're doing that.

When it's this black box ML algorithm, it's very hard to understand why a decision was made, therefore it's hard to correct any type of error that you think the algorithm is making, because we don't really understand why it chose to recommend the action that it chose to recommend.

Now the question is, this is clearly simple, and there are certain advantages to that. But now the real question is how much performance are we losing when we go to something from one extreme, this complex random forest, to this really, really simple, linear scoring role with two features and a couple integer scoring weights? How much performance do we lose? Remember, this is the line that we had. Again, depending on where the threshold was, this is performance, and these are our simple rules. We're losing almost no performance by going from this state of the art black box ML method, to these linear scoring roles. We're losing almost no performance. This was quite surprising to us. We thought that certainly there would be some performance loss when we did this. There are still advantages of having a simple transparent rule, but we thought there would be some performance loss. In this case, there is essentially no performance loss going from these complex methods to these simple methods.

Audience: Interesting topic, how to go to interpret the message. Do you have a general approach, and are you going to-

Sharad Goel: Yeah.

Audience: Discuss that?

Sharad Goel: I'll talk about that in one slide.

Audience: Thank you.

Sharad Goel: Okay. First of all, why is it that we actually are able to improve on judicial decisions? Remember, these are experts. They have quite a lot of years of experience and they clearly are trying to make the right decision of who to detain and who to release, so why is it that either our simple rules or our machine learned optimized methods are outperforming the human experts?

There are a couple different reasons. Here's the performance. The empirical performance of all the judges in our data set. There are a couple things to note. First, our algorithmic decisions are beating essentially every judge in the data set. What's going on? First, judges are all over the place in where they set these thresholds. Some judges are releasing about 90% of defendants, and other judges are releasing only about 50% of defendants. Remember, these are roughly randomly assigned, so they're seeing the same distribution of cases, but judges still, for whatever reason, they have these different internal standards. Whenever you apply different standards, then you're going to lose efficiency.

The judge, a defendant comes in, they randomly are assigned to different judges, some of whom are harsh and some of whom are lenient, and because these different standards are being applied at random to the same set of defendants and distribution, then we lose significant efficiency, because sometimes, if you happen to get assigned to the lenient judge, high risk people are being released, and if you get assigned to the harsh judge, low risk people are being detained. That's clearly not an efficient solution in a policy sense.

The other is that, as we would expect, even if we account for the different standards that different judges are applying, they're not completely internally consistent. This is a hard optimization problem. It's hard to know ... They don't get a lot of feedback, so it's reasonable to say that they're not going to be as good as a statistical method in this case. These two factors, this judge to judge variation, and also this within judge variation, is what's giving us this type of significant performance increase.

Again, to see how this type of variation plays out, let me just show you that, again, this is our decision rule, the solid line, so we release down here, we detain down here, and the gray scale indicates the proportion that judges right now are currently releasing. The point is that if you don't have any prior history of failing to appear, most judges, or on average, are very likely to be released. But in this entire region here, there's not actually that much difference. Judges aren't strongly differentiating between these relatively high risk people. They're young, they have at least four prior FTA's, failure to appears, and someone for example who's relatively older, and also only has one prior. About half of them are being, in both of these cases, about half of defendants are being detained.

This type of relative, or lack of optimizing for the empirical risk, is again what is letting our very simple statistical approach outperform the judge in this case.

How are we doing this? Again, we're taking a very simple strategy for creating the simple rule. You can imagine that you can take a sophisticated strategy, for example you could fit a mixed integer program to solve for these types of simple rules, but we're going to take a very simple strategy, and the point is that we want to make sure that policy people can actually carry out, can build their own rules, with very limited statistical knowledge, and using off the shelf statistical software. What is this strategy?

First, you're just going to select a few features. Either ... We find that about two to five features works well in practice, so you're just going to select a few features, either because of domain knowledge, or you use something like forward [stop wise 00:39:55] regression, but any method that you want to select the features that you think are relevant here.

Then we're just going to regress the outcome on the predictors, again using nothing fancy here, we're just going to use logistic regression. You could try lasso. Any kind of standard regression. You could even use a linear probability model to fit these types of things. Then finally we're going to round the regression coefficients, and here we're just going to rescale first, before we round. When you fit this regression, you're going to get these complicated coefficients with like ten significant digits. Clearly for these types of applications you don't need that level of precision, so we're just going to rescale the coefficients to whatever scale that we think is reasonable, and then we're going to round to integer on that rescaled scale. This is all we're going to do. It's a very simple strategy to select, [regress 00:40:45] and round. It's about one line of code to implement this thing. It doesn't involve any detailed statistical knowledge. A little bit of domain knowledge, if you have it, but really there's nothing to this type of strategy.

We tried it in 22 UCI data sets. We compared this method, this select, regress and round method, to lasso, and on the average over these 22 UCI data sets, we find an average AUC of 87% on both data sets. Again, we're not using, here is our black box model, lasso, using all of the features. Not doing any kind of ... So the coefficient are relatively complicated. This is a standard ML approach, so we just have exactly, we have on average no performance loss doing this.

Now if we compare it, we also compare it to random forest, the same 22 data sets. This is a non-linear model, and it's even more complicated than lasso, and here on average we do see a little bit of a performance loss, as you might expect. Going from 87% AUC to 92% AUC, for random forest. We do see a little bit of performance loss, because some of these data sets, there is some inherent non-linearity. In our bail example that I talked about earlier, we didn't need to use these non-linear methods, but in some of these data sets we do.

Again, the difference is actually not that big. This very simple method that requires almost no statistical knowledge, but it is statistical, it is doing more than a human expert is doing, we can get significantly better performance than humans on real examples, and at the same time, compared to a benchmark of the best black box ML methods, we're getting pretty close, and in some cases it's almost indistinguishable.

Audience: Is it that you're running the simple rules on the outcome of the output of random forest, or the original data?

Sharad Goel: No. On the original data. We're using these black box models only as a benchmark. We never use that in the training process itself.

Wrapping up, Eric and I both talked about these ML approaches to policy problems, and I think there's this area of computational public policy, for lack of a better term, which is having increasing impact in how we make these types of decisions. We see this playing out in a lot of real world examples. There is this promise of greater efficiency, equity and transparency, in many examples, as has already been pointed out, I think there are many social, ethical, legal challenges that we've only starting to begin to understand, and I haven't discussed them in this particular talk, but I'm happy to discuss these during the Q&A session. Thanks.

Speaker 1: We have time for questions for both the speakers.

Audience: What I wanted to understand is that your work does a really great job of making the rules interpretable, but what I want to understand is how is it different than the actual black box models you are using, because what you are doing is you are taking the coefficients that the black box model is putting out, and the models that you're using, like random forest, are linear sort of scoring mechanisms. They are not doing anything non-linear with your data, and so when you arrive at the rules, and the coefficients you assign to them, your model comes up with simple interpretable rules, but it doesn't tell you why those coefficients, like why is age weighted so much more heavily in your data set than the missing [bail 00:44:20]? The numbers I'm trying to understand is ... The scores you are getting, for 47% versus 47%, that should be happening because you are taking the same scale of coefficients that the model has come up with and using them in the same way that the model is using.

Sharad Goel: There are two differences here. First, random forest is non-linear. This, in theory, should do ... Can, on some data sets, it can do much better. The second ... These are linear scoring rules here, but random forest is highly non-linear. This is one of the reasons that people use it. The second is that this is using exactly two features, age and your prior number of failed appearances. The random forest, in all our benchmark models, use all of the available information. This could be, in the bail example, this is about 100 different features. It's going

from a non-linear model that uses many, many features, to a linear model that uses only two features.

It's clear ... I think your point is valid that where are these numbers coming from. They're not clear. The point is that humans don't actually, can't actually write these things down. They are coming from a statistical method, but I would say this is much simpler than the random forest on everything, in the sense that I could put this on a note card, and I can give it to you, and I can inspect it, and I can say, "Okay. I understand that being younger is more associated with risk in this sense, and also clearly having a history of missing court dates is again more associated with risk."

The random forest, even on 1,000 trees, which is what we fit, we can't even write this down. We're going to have reams of paper, if I were to show you literally what the model is, versus putting it on a note card.

Audience: Just very quickly, how did you choose those two trees, rather than the other 98?

Sharad Goel: Do you mean the two features?

Audience: Yes.

Sharad Goel: We chose these ... There are two different ways you can do it. You can either do this automatically without any domain knowledge, and something like this is going to pop out. This is always going to show up. If you use something like forward stop wise regression, this is clearly indicative of your likelihood of fail to appear, and it's also pretty obvious that the best predictor of future flight risk is past history of failing to appear.

Then this one we chose from a number that we could have used, just because it seemed, in part because this is viewed as a non-controversial predictor. You could use things like housing instability. This is another reasonable predictor of flight risk. But here, talking to the domain experts, these were the two that seemed the most reasonable.

Audience: I had a question. If I understand correctly, this came out of just a standard logistic regression, maximum likelihood estimation, you scaled it, and then you rounded the coefficients. What's interesting to me here about this is this almost reverse of the rhetoric that I usually hear, because what I usually hear is, or what we've been thinking about is ... What you've presented in the end as the solution is the standard way that a social scientist would say, if you said, "Can you make me a predictive index?" I've done things like this before around housing. Do this. Then it's interesting that you're making the argument that that 5% difference maybe isn't enough to worry about the lack of interpretability, and I'm just wondering how you, as someone who's invested in computer science, how are you balancing the simplicity versus novelty of the

computational versus the applicability and how do you ... How are you wrestling with that?

Sharad Goel: One thing that's first, on the bail example, we actually see zero loss of performance. Here, this ... We're seeing no loss in performance, and so it's actually an easy choice. I think no one would favor the random forest over the simple rule for that particular application, which would be our motivating application, because there's no loss. As long as you believe that there is any value to the interpretability and the simplicity, then it's a clear choice.

On average, we are seeing data sets where there is a loss in performance, and then I think you have to make the decision. In terms of my personal preference, in some ways my main interest is in the application, and solving the problem. If that means inventing new methods, great. If it means using a hundred year old method, that's also great. I think that's ... Especially in this crowd, I think that's probably a standard view that we're all interested in solving problems, and to the extent that we have to solve hard problems, technical problems to solve the policy problems, we do it, but if we don't need to, then we don't do it.

Audience: Are you [inaudible 00:49:17] law enforcement in general to see how they agree with the results, or not so much the results, but the variables that you end up having out within your model?

Sharad Goel: The methods ... Some of these types of rules are already being deployed, independent of our work, are being deployed across the country. The federal system is mandating use of these types of rules. They're all ... This is relatively widespread now. The features that we're using past failure to appear and age are pretty standard. Usually the rules are more complicated, so they're using something like five or six different features, but these are pretty non-controversial in this world, to use these features.

Is that the-

Audience: They come up with new rules every time, right? And they're designed to address like corner cases and things that are just highly specialized based on their expertise. When you come up with new sets of rules, you have the potential for losing that type of nuance that might be inherent in the experts.

Sharad Goel: I think, yeah, I think that's ... I haven't talked about equilibrium effects. Even looking at this rule, you can say if you're 51 years old you'll actually never be detained, and maybe that's the right rule, but maybe it's actually going to change the equilibrium behavior, so you might not want to implement this. Maybe you want to have another edge case. We only went up to four. You could say, well, at five, this is where we're going to call it. I think you do have to think about it.

At the same time, I think the reality is that the rules that are being deployed are not actually as sophisticated as you might think. It's not that they've been developed over generations of people or experts going through lots and lots of these rules.

Audience: Let me just pace out a little bit more. One thing that we keep coming across is that the experts don't want to give up complete control, so what you end up with is, you try to build hybrid models, where you want to ... The expert, their [inaudible 00:51:26] is kin of like a [inaudible 00:51:27] in your system. Then you take the data and you turn this almost into like a ... Then you try to merge them together, and you accept the fact that it's not going to be a perfectly optimized solution, but you're still making them happy because they feel as though they've had some control.

Sharad Goel: Oh yeah. This I think is a great point. I probably should have stressed this a little bit earlier, is that these decision rules are not supposed to replace judges in this particular decision. They're only intended as aids, and by law I believe that actually has to be the case, that you can't simply say that, "Here is my machine output, and there's no chance of rebutting the decision." The human expert is going to look at this. Then they're going to decide whether or not this case warrants any sort of override of the machine recommendation.

One example is, I might have a sole caretaker of her children, who is deemed to be as risky as an individual without any children, and the judge might decide, their higher social cost to detaining the individual with children than detaining the individual without children, even though they're equally likely to be a flight risk.

In this case, the judge might decide, "I'm going to override the recommendation, knowing more factors about this particular case." Now, there is always the worry that the extent to which you let humans override the machine recommendations, there are other issues that are going to crop up. For example, there could be issues of bias that crop up. There could be issues of misunderstanding the risk, or misunderstanding the social trade offs in any of these decisions, so it's always a balancing act. But I think being as transparent as possible, and certainly making it clear that I would never advocate we just drop in this model and say, "This is how decisions are being made." in fact, that's one of the reasons that we aren't using these black box models, in that when they're clear and transparent, it's easier for humans to understand where they might be going wrong and where you might want to change the recommendation.

Audience: If they replaced age with [inaudible 00:53:29] prior FTA's with credit score, this would look almost identical to an underwriting model that you use in lending. What I was wondering is, when you crafted your model, did you see any additional ... Essentially cut points, what we see a lot when we're creating underwriting is, okay, but if this condition exists, ignore everything above in this

graph. I'm wondering, did your model provide any insights into what those additional rules would be?

Sharad Goel: Are you talking about comparing to a non-linear baseline? Is that the ... If you have a prediction score, or a risk score that's not simple linear combination of these two?

Audience: Generally ... You mentioned five or six potential variables, and that would be too difficult to represent on a two by two, but what you would see is that, if the person has an aggravated crime in their history, then add ten points on for whatever ...

Sharad Goel: Okay. Let's see here. This graph, which keep on referring to, shows that there's no increase in performance if we were to use a non-linear model, and we use all of the available information. We have about 100 different pieces of information available to us, and we feed this through random forest, which can give us all sorts of complicated non-linear decision rules. We actually, which is very surprising, is we don't see any increase in performance. Meaning that you could do these things, but in the end of the day, the simple two factor linear scoring rule is basically capturing all of the risk that's present in the data. It was surprising effect to us. We certainly didn't expect to see that, but I think that you can make these more complicated rules, but they don't seem to give you anything at the end of the day.

Speaker 1: Let's thank Eric and Sharad for the great talk that they've given.