Racial Disparities in Police Stops Across the United States:
A Large Scale Analysis
April 11, 2022

Dean Morrisett

And finally, I want to introduce you to one of our faculty, whose research is dedicated to using technology in aid of social justice, by identifying and illuminating inequality. Emma Pierson is an assistant professor of computer science at the Jacobs Technion-Cornell Institute here at Cornell Tech and the Technion and a computer science field member at Cornell University. She holds a secondary joint appointment as an assistant professor of population health sciences at Weill Cornell Medical College.

Emma develops data science and machine learning methods to study inequality. Her work has been recognized by a number of best paper and talk awards, a Rhode Scholarship, a Hertz Fellowship, an MIT TR 35 under 35, a Forbes 30 under 30, she's got it all. And her research has been published in venues such as Nature and Nature Medicine, and she's also written op-eds for The New York Times, FiveThirtyEight, Wired, and various other publications.

Emma will be discussing her work on racial disparities in police stops across the United States. Welcome Emma.

Emma Pierson:

Thank you very much for the introduction. Can you hear me okay? Perfect. All right, I will share my slides then. Okay, great. Thank you so much for inviting me to give a talk today. It's really a pleasure to be here and I'm looking forward to discussing this work further with this audience. As mentioned, today, I'll be discussing a large scale analysis of racial disparities in police stops across the United States. And this work would not have been possible without a large team of excellent co-authors, it was quite a lot of work, as you'll see. What I'll be talking about today is based on two papers, which you can both find online published in the past couple of years.

So we're going to be talking about discrimination in police traffic stops today. Why is this something we care about? Well, it's one of the most common ways that Americans interact with the police, with tens of millions of Americans stopped every year. And if you've been paying attention to the news over the past few years, you'll notice that sort of high profile episodes of police violence often do begin with a traffic stop. And this is heightened concerns that traffic stops may be racially discriminatory.

And to be clear about what I mean by racial discrimination in this talk, this is when someone is treated more negatively specifically because of their race. So for example, they're stopped by police because they're black, they would not have been stopped had they been driving identically but they'd been white, or they're searched by police because they're black, they would not have been searched had they been behaving identically, but they'd been white.

So this is obviously very bad if this is happening in this large scale form of police interaction, but it is difficult to statistically test for, for a couple of reasons we'll discuss. And I want to mention at a high level is that, at some point, this talk may get a little bit more mathy, but even if you don't understand the math for a little bit, I think you'll still understand the main takeaways and of course, feel free to ask questions at the end.

Okay, so the first challenge we confronted when we embarked on this is that there was no unified data set tracking every stop that occurred in the United States. Policing data in the United States is highly decentralized, each department basically tracks data in its own format in its own system, there is no big database from which you can download everything.

So we set about creating such a database in collaboration with our journalist collaborators, journalists are awesome in many ways. And these journalists in particular were awesome, they did a huge amount of work, basically submitting data requests to more than 150 police departments over the course of five years. Now, when that data comes pouring in, it comes pouring in a million different formats, you get CDs, you get Excel files, you get PDF, like a million different things, and so it's impossible to do an analysis on data that is that heterogeneous. So the next big thing we had to do was to standardize the data, to put it into a common format. And there are many data formats, coding conventions, et cetera, it's thousands of hours of cleanup, tens of thousands of lines of code. But honestly, I don't think we should go into that.

I'm only going to tell you the good news, which is that that job is now done. And all the data and the code is available online, so you can download and analyze it yourself if this is a topic of interest to you, and I think there are many questions that remain to be explored in a data set that rich beyond the stories I'm going to be telling you today. This map on the left shows the states for which we have at least some data available. In total, the full data set encompasses 255 million stops from more than 56 city agencies and 33 state agencies.

In the main analysis I'll be focusing on today, that number is somewhat smaller, and this is true for a couple of reasons. The first is we filter for stops between 2011 and 2018, and this is to sort of impose some standardization in terms of the time period analyzed. Secondly, we analyze only stops of white, black, and Hispanic drivers because often the other race groups are quite small in the municipalities or areas examined. And then a third thing is that you have to filter for departments that even have the requisite data to conduct this sort of analysis at all. So for example, if a department does not track race, as some departments do not, in their stop data, you cannot analyze racial discrimination.

So these are the three questions we seek to analyze using this data set, which we've now put into standardized format. The first question is, do the police discriminate in whom they stop in the first place? Second question is, do they discriminate in whom they search after the stop? And then third, how can policy changes like specifically, the legalization of marijuana, affect the racial disparities that we observe? This is by no means a comprehensive list of questions. As mentioned, for example, we don't look at police use of force at all, which is why I think it's useful that the data is online to enable such further analysis.

Okay. So let's talk about this first question, are the police discriminating when they decide whom to stop? Now, one thing you might do just very naively is compare the number of stops per capita and break it down by race group. You might say, for example, white drivers are stopped once a year on average and black drivers are stopped one and a half times per year on average, and this is a disparity, but this on it own is not really enough to prove that the police are engaging in discrimination when they decide whom to stop because it's possible that some race groups drive more, or they might commit more violations when they drive, or they might drive in locations that have more police presence, et cetera. There are a million things that might differ that mean just comparing stops per capita is not enough.

So instead, what we're going to do is we're going to take a very old... well, it's not that old actually it's from 2006, but we're going to take an old and nice statistical idea called the veil of darkness test, which comes from a paper written by Grogger and Ridgeway in 2006. And the core idea here is we're going to make use of the fact that it is harder to tell someone's race when it's dark outside and therefore harder to discriminate on the basis of race. If you cannot see the driver's race, when they're driving, you cannot discriminate on that basis.

So here's sort of a cartoon illustrating this idea, okay? Imagine you compare two different stops that both occur at 6:00 PM, but one of them occurs 6:00 PM, June 14th when it's light outside, and one 6:00 PM, November 14th when it's dark outside. So in the left instance, the officer is able to observe the driver's race and therefore, it would be much easier to discriminate on the basis of race if indeed they were doing that than in the right instance when it is dark outside and they cannot tell the driver's race at all.

So we're not presupposing that racial discrimination is occurring. We're simply saying it is much harder to do that if it is dark and you cannot see the driver. It's not really clear that these two settings are the same, June 14th and November 14th, maybe driving patterns change from June 14th to November 14th. Really, what we would want is to have the two settings be exactly the same, except for the fact that one setting, it is dark, and the other setting, it is not, because then we can say it is solely the effect of darkness, and specifically that it makes it harder to racially profile, which is causing any differences we observe.

Now, June 14th and November 14th are quite far separated, but one thing we can do is we can take dates that are much closer together, but still differ in the level of darkness specifically by taking advantage of daylight savings time, you might think that daylight savings time is not good for anything, it turns out it is good for one very specific thing, which is studying its causal effects in this way, gives you sort of clean effects of what happens when you suddenly jerk the clock forward or back an hour.

So specifically what we're going to do is we're going to filter for data right near daylight savings time to get what's called a natural experiment. A natural experiment means the world is sort of naturally creating the conditions of an experiment. Here, the experiment is, what happens if you just made it dark at 6:00 PM, as opposed to light at 6:00 PM.

So we're going to filter for all stops around daylight savings time, those two daylight savings times boundaries, and then we're going to control, basically, for what time it is, the location where the stop occurs, and whether it's in the spring or the fall. And we're going to do a regression where the regression coefficient of interest is what is the effect of it being dark outside. So basically, when you control for everything else, how much more likely is the stopped driver to be black when it is dark than when it is light. And if stopped drivers are less likely to be black when it is dark, that is an indication that basically darkness is in some sense, protecting them, there is a veil of darkness effect indicative of racial discrimination. And that is in fact exactly what we find, the probability that the stopped driver is black falls after darkness suggesting that darkness has this protective effect and indicative of potential racial profiling.

For those who are sort of more graphically than numerically inclined, here is an illustration of the core idea with data that comes from a single state. So this effect is somewhat larger than the effect that we estimate on the data set as a whole, but it gives you the core idea. Here, what we're doing is we're plotting on the X axis, sort of the time relative to dusk, we remove stops right around dusk, because it's kind of unclear whether it's light or dark during that period, just as they did in the earlier paper. And you can see basically that in three different time windows, so at 7:00 PM, 7:15 PM and 7:30 PM. If you compare sort of just pre-dusk to just post-dusk, the proportion of stopped drivers who are black falls, that is sort of the graphical illustration of the veil of darkness effect. Okay, so that's our analysis of police stops. We leverage this natural experiment and this veil of darkness technique to provide evidence that the police are racially discriminating in whom they stop.

Now, let's talk about the second question. Are police racially discriminating in whom they search after a stop? So a little bit of context on police searches, just to make sure we're all on the same page. So after the police stop a driver, they can conduct a search in order to find contraband, contraband here being things like illegal drugs, weapons, et cetera, stuff you're not supposed to be carrying. And the purpose of a search is to find contraband, of course, so it's not just to take up the driver's time or for the police officer to make some quota or something like this.

And so since the purpose of a search is to find contraband, our definition of discrimination is are minority drivers searched when they are less likely to have contraband at a lower threshold of evidence. So for example, if the police are searching white drivers, only when they're 90% likely to have contraband, but searching black drivers when they're only 20% likely to have contraband, then that difference in thresholds is discrimination under our definition. Now, this of course is not all the bad... It's not comprehensive, it's not all the bad things police can do after stopping a driver, but it is an important form of discrimination, and that is what we examine in our analysis.

So how are we going to test for whether the police discriminate in whom they stop? And here again, we quickly run into tricky statistical issues. A first statistical test you might say is like, let's just look at how likely drivers are to be searched after a stop. And if we do this, as shown in right, we do indeed see big gaps by race. Minority drivers are more likely to be searched after a stop in both state stops and city stops.

We sort of run into the same issue we ran into before, higher search rates on their own do not prove that the police are being discriminatory. It is possible that some race groups are more likely to carry contraband. The purpose of a search is to find contraband, so if some groups are more likely to carry it, then the police may be more likely to search them simply in the course of normal police work.

So this problem has been recognized for a long time. And so what's been proposed and said is, well, don't look at the rate of the searches, look at the outcomes of those searches, specifically, look at how likely those searches are to find something, and we call that the hit rate. The intuition here is like, look if searches of white drivers are finding something 90% of the time, but searches of black drivers are finding something only 10% of the time. Then it suggests that the police are searching white drivers only if they're really, really likely to be carrying something, but they're searching black drivers more at random, on the basis of less evidence, so that's sort of evidence of this differential thresholds idea we were talking about before.

So if the hit rates, the rates of finding contraband vary by race, this is discrimination under our definition or under the outcome test. And when you look at the raw data, you again see some evidence of discrimination under this measurement technique. You see that the hit rates for black and Hispanic drivers are somewhat lower and both state stops and city stops than they are for white drivers suggesting discrimination.

But it turns out that the outcome test too suffers from a statistical flaw. And this is known as infra-marginality. And I'm going to illustrate this with a hypothetical example. This example is stylized, none of these numbers are real, these numbers are totally made up, but it sort of illustrates the idea of infra-marginality. Imagine there are two races of drivers, there are black drivers and white drivers, and among each race, there are two easily distinguishable groups. Maybe one of them is wearing hats, for example, there are those who are very likely to carry contraband, and those who are quite unlikely. Among the unlikely group, 5% carry contraband, regardless of their race, among the likely group, 50% of black drivers carry it, and 75% of white drivers carry it.

And importantly, imagine in this hypothetical world that the police are not being discriminatory, they search everyone who is more than 10% likely to carry contraband, so they apply the same threshold to both groups. So what are the hit rates going to be in this hypothetical world? Well, the police are going to apply this 10% threshold irrespective of race. And so they're going to search all the likely drivers. And they're going to end up with a hit rate of 50% for black drivers and 75% for white drivers. And so under the outcome test, that difference in hit rates is going to be interpreted as discrimination. But this is a misleading conclusion because by assumption, in this hypothetical example, the thresholds being applied are in fact the same. And it's worth pointing out that you can also get misleading results in the other direction where the outcome test fails to indicate discrimination even though in fact it is present.

So why is this happening? Why are we running into this misleading conclusion? Well, it's happening because the statistic we are measuring, the probability of carrying contraband conditional on being above the search threshold is not quite the same as what we actually care about, which is the threshold itself. Now, that threshold itself is hard to measure, it's not directly measurable from the data, you can't just take simple fractions. So what we're going to do is we're going to write down a Bayesian model to try to infer this threshold. This is where it gets slightly mathier, if you don't understand this bit, that's fine, you'll still understand the rest of the talk.

Okay, so what is a threshold test? Basically, what it does is it writes down a stylized model of a police stop. Stylized means, we're not necessarily attempting to capture every feature of the world, but we're trying to capture enough features of the world that we're able to sort of make conclusions that are supported by evidence. And the purpose of this model is to estimate the search thresholds that are consistent with the observed data, so the search rates and the hit rates we were talking about before. The thresholds themselves are not directly observable, so we're going to try to estimate them. And discrimination, as before, is if lower search thresholds are being applied in searches of minority drivers.

So how does the threshold test model a police stop? The threshold test assumes that when the officer stops someone, they estimate the probability, P, that person carries contraband. P is drawn from a risk distribution. So I've shown this at right, the risk distribution is this blue line, and the probability the person is carrying contraband is on the X axis and the fraction of drivers is on the right, Y axis.

So for example, if the police officer pulls over a bus driver, then P would hopefully be quite low, the person is driving kids around, hopefully, they aren't also ferrying weapons or whatever, so P would be quite low. In contrast, if the police officer pulls over someone and they're acting woozy, and they're drinking out of a bottle, it's possible that bottle is some innocuous substance, but it's definitely suspicious behavior, and so P would be more high.

If that P exceeds a threshold, the officer searches the person, and if they search them, they find contraband with probability P. So for example, the bus driver would be below the threshold, the officer would not conduct a search and would not find contraband. In contrast, the woozy acting driver would be above the threshold, the officer would search them and would have a 75% chance of finding contraband. The model allows the thresholds and the risk distributions to vary by race and location and discrimination, as before, is if lower thresholds are being applied in searches of minority drivers.

Now it turns out that it's a bit more complicated than this because when you actually try to fit this model, it turns out to be prohibitively slow. To deal with this issue. We had to create a new family of probability distributions, which makes the test run 100 times faster. I'm not going to tell you about the mathematical details behind that, but I will tell you that this is an instance where fancy math is actually quite useful pragmatically. Why? First of all, it means that you can actually fit the model at all on a data set of this size with hundreds of millions of traffic stops, so that's a very powerful thing. Perhaps even more importantly, it widens the scope of the test, such that non-academic practitioners can use it. So journalists, for example, or police departments, or anyone who doesn't have sort of a massive number of computers and a lot of PhD students who are willing to bang their heads against the wall for a month. So this is sort of an instance where the fancy math sort of directly translates into practical impact and we'll return to that point in a bit.

But rather than go into any more mathematical details, I'm just going to tell you the high level conclusions of fitting this model, which is, indeed, that the inferred thresholds are lower for black and Hispanic drivers than they are for white drivers. So here, what I'm plotting on the Y axis is the threshold estimated by that model, and you can see that the model estimates that minority drivers are subjected to sort of lower threshold. So the police require less evidence that they're carrying contraband in order to search them, so this, again, is indicative of discrimination.

So to summarize this search analysis that I've shown you today, I've shown you three things, I've shown you that search rates are higher for minority drivers, I've shown you that the hit rates, the rate of which searches find contraband are lower, and I've shown you that the thresholds are lower. And this is sort of a smoking gun indication, like a characteristic pattern for discriminatory searches. We've seen it as well in other data sets, for example, that we have analyzed. All three tests are pointing in the same direction here, suggesting discrimination against minorities, but the threshold test, it deals with the statistical flaws of the simpler test, so it's sort of reassuring that from a statistical standpoint, that all three tests are pointing in the same direction here. And I think it provides quite convincing evidence that there is racial discrimination in police searches.

I want to mention briefly that the same statistical methods here are much more broadly applicable besides policing, specifically, they can be applied in other data sets where you have both a binary, so a yes, no decision and a yes, no outcome. So in policing, the decision is, should the police officer search a driver and the outcome is, does that search actually find anything? But there are a whole bunch of other settings where you care about discrimination, where you have that same basic mathematical structure.

So for example, in medical testing, the decision might be, should the doctor test the patient, and the outcome would be, does the patient test positive. In the loan setting, the decision would be, should you grant someone a loan? And the outcome would be, do they repay the loan? So for example, in the medical testing situation, if you saw that minority patients were less likely to get tested and more likely to test positive when they do, that might be indicative of sort of bias in how much testing they're getting. They're getting tested only if they're more likely to have a disease. And in subsequent work, which I'm not going to talk about in detail here, I actually do take the techniques that we develop for policing and apply them to COVID testing, to show exactly that pattern with racial minorities being tested for COVID only when they are more likely to actually have COVID. So this general math is very widely applicable to a whole class of important decisions where you care about testing for discrimination.

Okay. Let's talk about the final question, which is about how policy changes, specifically the legalization of marijuana, affects the racial disparities we observe. So the first two points are very negative points. They're showing that there's a serious and sort of systemic large skill problem here consistent with many other sort of sources of evidence on this topic, but the third point is sort of, well what can we do about this from a policy standpoint?

So the technique we're going to apply here is called difference-in-differences. The question we're going to try and assess is what is the effect of legalization of marijuana on whether drivers are searched after a stop? And so what we're going to do is we're going to compare two differences, this is why this is called difference-in-differences. We're going to compare the change in search rates in two states where marijuana was legalized to the change in search rates in 12 states where it was not.

We take two changes, one in the legalization states and one in the non-legalization states, and then we compare those changes. So in the data, we're looking at, the two treatment states were Colorado and Washington is subsequent to sort of the time period, there's been wider legalization of marijuana, but this was the analysis we were able to do in the time period where we had data. And so what I'm plotting on the X axis is time and the Y axis is search rate. And basically, what you're seeing is... Oh, and sorry, the dotted vertical line is when the legalization of marijuana occurred. And you can see that there is like a dramatic fall in search rate basically for all race groups, post legalization, in these two treatment states, in these two states where marijuana was legalized.

So this is consistent with the idea that the legalization of marijuana resulted in a reduction in search rates, but it's possible, that it was just some other national trend that happened to occur in 2013. So this is the purpose of looking at the control states: Is it really the legalization of marijuana, or is it some other time trend, which is producing these apparent effects? Now these are the control states where marijuana was not legalized. And you can see that there isn't really any evidence of a consistent trend, some of them go down, some of them go up, some of them kind of stay flat, but there's certainly no evidence of some dramatic event occurring in 2013 that produced consistent national patterns.

If you want to sort of get actual numerical estimates, what you do is a regression where you try and basically actually estimate that difference-in-differences model, and basically what we find when we do that is that the legalization of marijuana produced a large drop in search rates for all race groups. It didn't, by the way, remove the racial disparities in search rates, but it did make them go down for all groups. And so, why does this matter? Because to the extent that searches are discriminatory, if they're occurring less frequently, it may sort of mitigate the discriminatory impact of those searches. So this is sort of an example of how policy change actually does result in change on driver's lives. Before moving to questions, I want to close by speaking briefly about the public policy impact of this work.

So as I mentioned, one of the advantages of the fast threshold test that we developed is that it's much easier for journalists and other non-academic practitioners to use. And in fact, a couple years ago, that was exactly what we saw, The Los Angeles Times was able to take our threshold test with some help from our team and use it to show that the LA Police Department was searching black and Latino drivers on the basis of less evidence.

And within a week, in response to the story that The Los Angeles Times published, the LAPD announced that they were going to dramatically change their random search policy in an effort to reduce this racial bias. So this is an instance where sort of the math translated quite directly into practical impact, and I think this is one of the reasons that it can be useful to work on these sort of discrimination questions, because it is very much a domain where the math is not isolated to the ivory tower, whether through collaborations with journalists, or with real world practitioners, or through expert testimony in federal cases, et cetera, you see many, many paths via which the sort of discrimination research exits academia and makes it into the real world.

Okay. Thanks very much for your attention. I'm looking forward to the subsequent discussion. I will say though, that if you have data sets, which you think might benefit from similar statistical analyses, I am a math nerd perennially in search of interesting and high impact questions, so my email is extremely open if you want to discuss any problems along this line.

Matt D'Amore:

Thanks Emma. Hi everyone, this is Matt D'Amore, I'm Associate Dean here at Cornell Tech. Please put your questions in the chat for Emma. Emma, I put your email in the chat, so now everybody has it.

Emma Pierson:

Thank you.

Matt D'Amore:

We got a couple questions in the chat and I have a few for you as well. First of all, thank you so much for that presentation, it was really enlightening. I really appreciated it and found it very thought provoking. I've also put down a link to The Nature Human Behavior article in the chat as well.

So the first question is, will the project be continuing to collect additional data and update it as we go forward? That's particularly of interest as more states, New York, for example, legalize marijuana, or is this a sort of a closed-end project?

Emma Pierson:

Yeah, that is a fantastic question. I personally am not continuing to do the data collection myself, but it is possible that the data collection will be continued under sort of the Open Policing Project, and I would suggest reaching out to either Sharad Goel or Cheryl Phillips, who are the two senior authors on the paper, to know if they have sort of independent data collection efforts going on. During the time we worked on the project, there were two massive data collection efforts, so sort of we did one and then another team revamped the whole data a second time, so it's a lot of work to sort of maintain the thing, but it's clearly invaluable as well as mentioned as additional states add.

Matt D'Amore:

Thank you. How did you decide to take on this particular project? How did the group decide that this was the data set, 100 million stops that they wanted to try to tackle?

Emma Pierson:

Yeah, that is a great question. So the original project began, I know with Cheryl, who was the senior journalist on this project, was engaging in this data collection effort and she and Sharad had a fortuitous interaction where this sort of became a collaboration between the journalist who had the data collection expertise and sort of the technical statistical side, which is often the way I think high impact stuff gets done with interdisciplinary work between technical and non-technical.

I think in terms of my own involvement, I was a first year PhD student who had not come to Stanford to work on policing, and I got looped into to run some of the preliminary analysis on the data just as it was sort of pouring in for the first time, and as a statistician, you sort of develop a spider sense, you feel it almost viscerally like, "Wow, there's a there there." And here, it was just very clear that something massive and very bad was apparent in the data that was being collected. Of course this was by no means the first time that someone had provided evidence of this. There's a huge amount of previous work on this. There's also the lived experiences of individual drivers that testifies to this quite forcefully. But I think sort of purely from a statistical standpoint, in this data, there was just sort of this like very visceral sense that this was something that you needed to put a lot of time into, I think.

Matt D'Amore:

Yeah, thank you. One of the questions from the chat: In looking at the data, did you see differences across gender, either compared against race or as a separate access from race?

Emma Pierson:

Yeah. That's a great question. There are definitely intersectional dynamics here, where, by intersectional, I mean, there's an interaction between gender and race, and gender does matter here. We looked at that, I remember, in preliminary analyses, we would make a graph stratified by gender and race. That was a long time ago, so I'm reluctant to speak specifically to them, but it is definitely the case that gender is an important covariant here and in criminal justice more broadly. And I think that's a fruitful direction for follow-up work.

Matt D'Amore:

One of the things that occurred to me as you were talking about the marijuana stops is the change in stop behavior suggests not just a change in threshold, but a different outlook of policing. I feel like there's something there, I'm not sure that it's statistical to talk about, it surprised me that with the change in marijuana regulation, that stops would necessarily go down, successful stops, sure, but stops, that surprised me. When you say success on your hits, I wonder if you can say a little bit more about that.

Emma Pierson:

Right. Yes, that's a good question. So to be clear about what's being plotted on the vertical axis there, it's the fraction of stops that result in a search, so it's not that the number of stops is going down. It's that how likely are you to be searched after stop that's falling.

Matt D'Amore:

Oh, I see.

Emma Pierson:

Yeah. And so, as to why that occurs, I mean, I think basically it's that legitimate reasons a police officer has to search you is now more restricted, it's like things that used to be crimes are no longer crimes. And so you can't search for them. And so plausibly sort of the fraction of instances in which you feel justified in conducting a search is just smaller.

Matt D'Amore:

Got it. So to a certain extent, that mean, sort of the threshold, it's looking at that, the threshold for conducting a search has to have gotten higher in those cases, because they must assume that the likelihood of finding something has gone down is that... I mean, statistics was a long time ago for me.

Emma Pierson:

Yeah. Well, I think it's that basically like the class of instances in which you could be defined as having gotten a hit, having had a successful hit is smaller, so it's like the very definition of what it means to conduct a successful search has gotten smaller. We also, in the paper, do analyze the change in thresholds. I don't offhand remember those results because it's sort of in the supplement to the paper. It was definitely not the case that legalizing marijuana magically made all the search thresholds equal, unfortunately. It is the case, however, that the fraction of stops that resulted in a search stop went down. So if you believe that searches are discriminatory, then reducing the number of instances in which this potentially discriminatory thing occurs is good.

Matt D'Amore:

So we're right on the cusp of marijuana legalization in New York, it's been recently done in New Jersey. What sorts of data could the city or the state look to, to see if this behavior... To see if New York state is following the trends that you've identified? We have folks from the court system on the line here, we've got folks from court research who are going to be talking in a little bit, so as we think about our data collection, what sort of data can we collect to make this analysis easier?

Emma Pierson:

Yeah. I mean, I think in the paper, at the end of the paper, we provide recommendations, this is not specific to marijuana legalization, but sort of broadly for what kind of data it's useful to collect. So in our case, bare minimum, it's useful to know, the race, the driver, it's useful to know whether there's a search, it's useful to know whether that search found something. Then there's like a wealth of additional helpful, contextual information like why was the search conducted, where, what time, et cetera, and what time of year and other things like this. And so I think I would refer you to sort of the place where we describe sort of best data collection practices, which is broadly useful to collect, that's not specific to marijuana.

In the case of the marijuana analysis, really, you need to you to know the race of the driver, whether a search was conducted. It's also useful to know, I think, the reason for the search, because in theory, if marijuana's legalized, people should no longer be searched. And I think in New York, certainly pedestrian stops are also of interest. So in this analysis, we look at traffic stops, but pedestrian stops are also very interesting as well.

Matt D'Amore:

Offline, maybe you and I could chat about what the data from New York looks like. And so we can chat about ways to think about that with this group.

I saw a question but I want to stick with the data question for a minute and then we'll get to the chat question.

Generally speaking, what can municipalities, and states, and legal aid organizations do to collect and expose data that might be useful for research like this? I'm thinking there are a lot of folks on this conference who are interested, for example, in housing, both assisting folks through eviction or through discriminatory housing issues, and how can we try to collect and expose data that might help in situations like that?

Emma Pierson:

Sorry, is this about discriminatory housing?

Matt D'Amore:

Yeah. One case that gets talked about a lot is both discriminatory rental behavior, but also eviction behavior. Rental behavior, I think might be really hard to track with this group because it doesn't necessarily get adjudicated, but evictions do. And so I wonder about whether there's a way to take your test and look at eviction data.

Emma Pierson:

That's fascinating. So evictions are not a topic I specifically studied, so I'm reluctant to kind of in a sense... The median, most natural thing, I think would be something which would allow a kind of benchmark test, by which, I mean, so there's not an obvious outcome in eviction, it's not clear what would make 'a successful eviction', a successful police search is one that binds contraband a successful eviction, not clear. So generally-

Matt D'Amore:

Yeah. We could look at a proceeding and whether or not it succeeds in evicting... Whether or not the grounds were-

Emma Pierson:

Upheld.

Matt D'Amore:

Upheld. So you could look at the claim and then look at whether or not the person actually was evicted.

Emma Pierson:

Yes. That's interesting, and we should talk about that more. You would have to assume I guess that the outcome of that proceeding is itself unbiased, which is a case you would have to... That seems like it could be plausibly influenced by say, the quality of representation... I'd be very happy to chat more about this topic, because I think... I teach a class called Data Science for Social Change and on Wednesday we're having a speaker from Matt Desmond's Eviction Lab coming in to speak to the class, so I think it's definitely something where I'd be very happy to discuss.

Oh, I did want to say though, I think a general class of statistical tests, sort of benchmark tests are basically like when you control for sort of the plausible explanatory factors, are people of group X more likely to be evicted, and then the question is, what are the reasonable explaining factors and do we actually have data on them? And that would be something you would need to talk to domain specific people about, but I think that would be sort of how you might make such a case.

Matt D'Amore:

Yeah. Like I said, court research folks are going to be presenting in a little bit, so be interested to see... We'll talk more offline about accessing their data.

A couple of questions from the chat, going back to the police stops, is there data on the police pretexts for a stop or to search, in other words, do they have to put down why they did the stop, and are certain justifications more likely to be suspect or pretextual.

Emma Pierson:

Yeah. That's a great question. In many jurisdictions we do have data on that and there's certainly been concern that, for example, like consent searches might be discriminatory. There's also been concern that... Okay, sorry. So there's data both on sort of why the stop is conducted and data on the search which is conducted. And there's been concern that both might be things you could study for racial bias. So for example, in the search context, consent searches have been looked at as potentially racially biased certainly on the stop side, this is also something people have looked at, for example, broken headlights, other things like sort of ticky-tac violations that are plausibly pretextual. And those are both you can examine with our data and in preliminary analysis we did, it did not make it into the final paper, but definitely worth taking a look at and possible to do with these data sets.

Matt D'Amore:

Yeah. A related question on data regarding data quality is the collection, and I guess this is going to be by the officers or somebody in the police station, is it logged automatically or manually? And if manually, was the possibility of bias in data collection considered, could it be biased towards not logging certain types of stops or searches? Is that something that either the data suggested, or that came up in the way you examine the data?

Emma Pierson:

Yeah, these are both great questions. So the way the data is logged is widely heterogeneous across departments. There are certainly numerous errors in the data. Some of them appear to be just benign but nonetheless... Sorry, they weren't done on purpose, but they are nonetheless insidious, like over densities of stops made exactly at midnight, they weren't something weird is going on with the logging there. There's also, however, evidence in some jurisdictions of less benign errors. So for example, a couple years back, the Texas Department of Public Safety got in trouble with some local journalists who provided convincing evidence basically, that drivers who were being recorded as white were not only not white, but were not even sort of plausibly white. So they would look at sort of their names, they would look at their pictures, and it was just very clear that there was sort of a systematic doctoring of the data in order to make it look like the per capita stop raters were less racially discriminatory than they were.

And lo and behold, after this was pointed out the problem fixed itself, so this was sort of more evidence that the department had been doing this deliberately. So in our analysis, in Texas, at least, I think we do sort of a sanity check correction procedure to make sure that the driver's name, which we do have access to, is consistent. So if you look at sort of the fraction of census people with that name, who have a given race or ethnicity, it's consistent with their recorded race. Problems like that, they do not describe the majority of states in the data set. There are tons of errors, but I think most of the errors are due to just people writing things, forms, and bad bookkeeping, et cetera. And just data collection is hard, it's not due to sort of deliberate malfeasance.

Matt D'Amore:

Got it. Got it. Sally has a question in the landlord-tenant space, but I'm not sure if it's a question or simply just a statement, Sally. Maybe to turn it into a question, Emma, in order to do this analysis, the dataset does need to include some observations on race, and that actually can be difficult. One of the challenges with eviction or court data, that may or may not be captured.

Emma Pierson:

Right, definitely agree, much better if race is included. I wonder if, in the absence of race, you can do something with location, you can argue racially disparate impact by this zip code, this zip code, or even finer-grained information. But I definitely agree, and particularly in driving, where people move around, yeah.

Matt D'Amore:

Yeah. Interesting about zip codes something to think about. Another question: Do you need individualized data? You're looking at aggregating the case-by-case data when you look at these, when you are performing your analysis. In other words, for the police stops, for example, you looked at individual 100 million individual stops to do your modeling.

Emma Pierson:

Yeah, it's a good question. Our data set does indeed contain individual stops. You can do some stuff with aggregated data. So for example, if you know search rates and hit rates by race and location aggregate, that's all you need to fit the threshold test. And similarly, in the eviction instance, I think if you did have eviction rates and eviction outcomes broken down by race and location, you might be able to do something with that. I really don't know without knowing more about the... But sometimes you can do some stuff with aggregated data.

Matt D'Amore:

And right now, the court and the state, there's the Open Data Act and a lot of work being done in the state to expose additional data and make it available, and folks, we're going to talk about that in the rapid fire in just a bit. So individual organizations, which don't necessarily... There's the court end, and the agencies, and the kinds of institutions that you looked at for our work, but we also have a number of legal aid organizations, with much smaller, much smaller data sets that might be interested both in applying this analysis either to their own caseload, possibly to check for implicit bias in their own work. Does the analysis that you're talking about work on smaller data sets that organizations might have available?

Emma Pierson:

You definitely don't need 100 million stops, you definitely need more than 10. So I think it's a case-by-case question, but again super happy to talk... Honestly, statistical testing for discrimination is, I think, one of the most intellectually interesting things I've ever worked on and then incredibly important, so very happy to chat about specific applications.

Matt D'Amore:

Yeah. When you asked, are there interesting data set out there… Could you give some examples, if there was a dataset you were interested in working on, what would it look like?

Emma Pierson:

I mean, the eviction problem is certainly one of interest, housing discrimination is certainly one of interest. My baby sister right now is a law student, and she's interested in these sorts of questions, and so I sort of go back and forth. She's like trying to explain to me, how do you make a prima facie case for discrimination.

I think my little sister has become much less annoying and more useful as she has aged. So I think all questions of discrimination and data sets to test them, I think are of a lot of interest to me. As I mentioned, these threshold tests are very widely applicable to cases where you have sort of a binary decision and a binary outcome. So loans are such a setting, medical testing in such a setting, to some extent, hiring is such a setting because you have the binary decision to hire and then the binary outcome if they are, so problems of that form, I think are very interesting too.

Matt D'Amore:

Yeah. We have contacts at the California Department of Fair Housing and Employment... or Employment and Housing DFEH, and I may mention this work to them as well. They're just beginning to expose some expert systems on their website to help people with intake and information and through that. But it's also worthwhile talking to them about what kinds of data they're collecting in this regard, because they're a clearinghouse for those kinds of claims in housing and in employment in California.

Emma Pierson:

Yeah. I'd be very happy to chat.

Matt D'Amore:

Yeah. I'll connect with them. Unless there are other questions from the chat, or Emma, any further comments. Well, thank you so much for your time. We really appreciate it.

Emma Pierson:

No. This was a pleasure. Thank you so much for inviting me, and again, very happy to chat more.

Matt D'Amore:

Thank you so much, Emma.