Transcript from the
Dean's Symposium on Social Science Innovations
May 3rd, 2021
Division of Social Science
Faculty of Arts & Sciences
Harvard University

[00:00:11.72] LARRY BOBO: I think I'll go ahead and start, Blake, if we can do that.

[00:00:25.21] BLAKE: You're all set.

[00:00:26.54] LARRY BOBO: Right. Good afternoon, everyone, and welcome to the Dean's Symposium on Social Science Innovations. My name is Larry Bobo. And as your divisional dean of social science, I am delighted to inaugurate and host this convening.

[00:00:41.87] Before turning to the substance of this afternoon's symposium itself, I wanted to open with a word of thanks. Thanks to all the faculty, graduate students, and staff who helped get us very successfully through the long and arduous demands of the COVID-19 pandemic era. It has not been easy researching and teaching remotely.

[00:01:04.94] It has been hard to go without the sort of human contact that makes a liberal arts college experience while embedded in the leading research university, the vibrant and stimulating environment that attracted us all to this very special perch in the academy. But the effort, dedication, and creativity you have all brought to fulfilling our shared mission of research, teaching, and advancing knowledge has truly been impressive and wonderful.

[00:01:32.70] It was a sense that out there in our community, there was a hunger for more substantial occasions for serious intellectual colloquy across our disciplinary lines that inspired me to launch this symposium. I hope this becomes a more regular vehicle through which we convene and engage with one another on important topics that touch upon the agendas, the activities, needs, challenges, and most exciting new contributions across the many fields represented in the social science division.

[00:02:05.39] In some respects, this gathering could be characterized as a session on big data opportunities, contributions, challenges, and limitations. It impress me as an arena in which several of our departments are actively engaged in cutting edge scholarship and development. There are other such arenas we will tackle in the coming academic year. More about this later. But this topic lent itself to bringing together work now taking place in economics, government, sociology, thinking from the history of science and psychology as well.

[00:02:41.72] I am particularly pleased that Elizabeth Phelps, the Pershing Square Professor of Human Neuroscience, agreed to chair this important session. Liz's own research addresses how

emotional states affect our lives and how even very seemingly subtle variations in these emotions can alter our thought and action. And she uses a very diverse arsenal of methods of data collection from behavioral studies and physiological measurements to microbiome assays, functional magnetic resonance imaging to computational modeling and more in her research.

[00:03:19.28] It's my pleasure at this point to turn the session over to Liz who will explain the rules of the road for posing questions as well as introduce each of our speakers and keep us all on track. Liz, the floor is yours.

[00:03:33.66] ELIZABETH PHELPS: Thank you, Dean Bobo. So welcome to the symposium. I'm honored to be able to serve as your chair and moderator. Just a few rules of the road before we get started. We have three speakers talking about their use of big data in the social sciences from different perspectives and one discussant. I will introduce each speaker in turn.

[00:03:54.29] This will be followed by an opportunity for you to ask questions. So we're asking that you use the Q&A function on Zoom to enter your questions. And then I will moderate that discussion at the end of today's symposium. If you have a particular speaker you want to address your question, please indicate who that is. Or let us know if it's a question for the entire panel.

[00:04:21.86] So I want to get us going. Without further ado, I'm going to introduce our first speaker. Our first speaker is Raj Chetty. He's the William A. Ackman professor of Public Economics and the director of Opportunity Insights which uses big data to understand how we can give children from disadvantaged backgrounds better chances of succeeding. Dr. Chetty is an economist whose research combines empirical evidence and economic theory to help design more effective government policies. The title of his talk today is leveraging administrative data to improve equality of opportunity. So, Dr. Chetty, take it away.

[00:04:59.90] RAJ CHETTY: Thanks so much, Professor Phelps. Thanks, Dean Bobo, for organizing the symposium and for including me. It's a pleasure to be a part of this distinguished panel, and thank you all for joining. So I'm going to share my screen here. Hopefully you're all seeing some slides.

[00:05:12.98] So I'm going to talk about how administrative data, in particular data from government records, has I think really opened up tremendous possibilities in social science research. And I'm going to focus specifically on an example that our team has been studying over the past several years, which is on equality of opportunity and how to give, as Professor Phelps was saying, kids from disadvantaged backgrounds better chances of succeeding.

[00:05:38.74] And so I'm going to talk about how modern data is allowing us to really understand that question in a very granular way. But I'm going to start at a much bigger picture level by talking about the American dream, which is, of course, a complex, multifaceted concept that can mean different things to different people.

[00:05:56.18] But I want to distill it to a single statistic that some colleagues and I set about to measure in a paper a few years ago, which is a traditional way that people have conceptualized

the American dream-- the idea that this is a country where through hard work, any child should have the chance of rising up and achieving a higher standard of living than their parents did.

[00:06:16.46] And so what we set out to do is assess the extent to which America actually lives up to that aspiration both today and historically. And we measured a very simple statistic-- what fraction of children go on to earn more than their parents did, measuring both parents' and kids' incomes in their mid 30s, and adjusting for inflation. And you can see the data here plotted by the year in which the child was born. You can see that kids born in the middle of the last century was a virtual guarantee that you were going to achieve the American dream. 92% of children born in 1940, we estimate went on to earn more than their parents did.

[00:06:54.51] But if you look at what has happened over time, you see a dramatic fading of the American dream such that for children born in the middle of the 1980s who are turning 30 around now, it's become a 50-50 shot, essentially a coin flip, as to whether you're going to do better than your parents.

[00:07:11.21] And so motivated by this broad trend in our research group, we and many other social scientists are very interested in understanding what is driving this trend and ultimately what we might be able to do to reverse it. Now, traditionally, when social scientists were faced with a problem like this, we had relatively limited data to rely on.

[00:07:29.69] So if you look at the data points plotted on the slide, for instance, and think about what might have changed in the US economy society over the past half century, there are numerous things that you might think of, from fundamental changes in industrial structure, global competition, to increasing segregation, civil rights movement-- lots of things that could have played a role. And so it becomes very difficult to disentangle the role of specific factors when you're forced to rely on broad, aggregate comparisons like this over time or if you're to look across countries, which is another approach that social scientists have traditionally used.

[00:08:05.44] Big data, and in particular, the type of data I'm going to focus on, administrative data from government sources, has really in the past 10 years opened up tremendous possibilities and being able to study questions like this in a much more precise way by unpacking these national trends. And so just to tell you a bit about how we approach that in our research group, the Opportunity Insights, what we're doing is using big data broadly to study how to increase upward mobility in various ways. And the starting point for a lot of our work and what I'm going to show you today is the construction of a longitudinal data set-- that is a data set that allows you to follow people over time-- which we construct from anonymized tax and census records covering all Americans over the past 30 years.

[00:08:51.71] So this is really an incredible data set in terms of scope and scale. Think of every person in America has a 30 row block in this data set spanning 30 years, giving us information about their incomes, of course, which you'd seen tax returns, but linkages between parents and children, where people live, information about where they're employed. So very detailed information in an 8 billion row data set covering essentially everybody who's been in the United States over the past 30 years.

[00:09:23.23] Now that data which our team set up for analysis about 10 years ago and now is used by something like 100 researchers around the US, really, I think, opens up lots of different possibilities for questions one can analyze. What I'm going to show you today is how those data have allowed us to analyze local differences in upward mobility in a way that I think can really shed light on the drivers of the fading of the American dream, and what we might be able to do going forward to increase upward mobility in the US.

[00:09:55.21] So to do that, I'm going turn to showing you the data in this map here, which shows you the geography of upward mobility in the United States. I'm going to start by describing how we construct this map and then discuss what I think we learn from it. So what we did to construct this map is took data on 20 million kids and their parents, essentially all kids in the United States born in the early 1980s.

[00:10:20.41] And we mapped them back to the area in which they grew up, dividing the US here into 740 different metro and rural areas. And in each of those areas we construct a simple measure of upward mobility. We ask what is the average income at age 35 for kids who grew up in low income families. That is families at the 25th percentile of the national income distribution. That's families earning about $27,000 a year.

[00:10:47.68] The map is colored. So that blue green colors represent areas with higher levels of upward mobility and red colors represent areas with lower levels of upward mobility. If you start by looking at the scale in the lower right side of the screen, you can see that there's an enormous spectrum in terms of rates of upward mobility in the United States in parts of the center of the country. For instance, places like Dubuque, Iowa, kids who are starting out and families making about $27,000 a year themselves, one generation later, are earning more than $45,000 a year. So a substantial amount of upward mobility across a single generation.

[00:11:24.13] In contrast, if you look at much of the Southeast or a place like Charlotte, North Carolina, for example, you see that kids starting out and families at the same level of income are ending up actually with lower average incomes one generation later. And so to the question of whether the US is really a land of opportunity, you can see that even today in the current generation, there are some places where kids truly have great chances of rising up. There are other places where, today, kids' chances of escaping poverty unfortunately remain quite low.

[00:11:58.40] We did not have access to this sort of information 10 years ago. This is only feasible because of having the enormous sample sizes provided by government data, in this case from tax records that allow us to essentially look at everybody in the United States and track them longitudinally over a 30-year period, which, if you can imagine, trying to survey people on the scale over 30 years would be an incredibly challenging activity. And so it's really these data that open up an analysis like this.

[00:12:26.55] Now, when you look at a map like this, naturally from a scientific perspective, we are interested in understanding what is driving these differences in upward mobility that we're seeing across areas. And I think that question is also of great interest from a policy perspective because if we can figure out what's going on in the blue green colored parts of the map, perhaps

we can figure out how to implement solutions in the red and orange colored parts of map that would lead to higher levels of upward mobility.

[00:12:53.08] Now you can see the broad regional variation in upward mobility in this map for yourself. Higher levels in much of the Midwest, lower levels in the Southeast and the industrial Midwest-- places like Cincinnati, Detroit, and so forth. What we've been doing in our research group is conducting a series of studies to essentially understand what is driving this variation and what can we learn from a policy perspective going forward to improve upward mobility.

[00:13:20.42] So what I'm going to do here is now, in the interest of time, just give you a brief snapshot of what I think we've learned from that type of analysis. And so to do that, to go further, what we did after constructing this map at the broad national level is to zoom in to more of a local level. And to show you that data, I'm going to toggle over to this website-- opportunityatlas.org-- which is a freely available website that you can visit yourself.

[00:13:49.13] And the way this website works is, you can enter in any address, very much like a Google map, and literally zoom in to look at levels of upward mobility not just at the National level-- here, I'm showing you the same map that I started out with. But now I'm going to enter an address in New York, and we can literally zoom in and look at the same statistics that I've just been sharing with you-- census tract by census tract in New York City.

[00:14:15.13] And the first thing I'd like you to see here is-- if you look at the spectrum of colors on your screen, you can see that you can go from the darkest red colors to the deepest blue colors within New York City by just going two miles down the street, right? And so a different way to say that is you can go from Alabama to Iowa in terms of rates of upward mobility just within New York or really within any city.

[00:14:36.97] And so what that shows us-- being able to zoom into this even more granular level-- is that the roots of upward mobility are not about differences across states or differences across cities, but rather differences across city blocks where we're seeing kids have incredibly different opportunities. The particular address that I entered here, 530 Sutter Avenue, is one of the largest public housing projects in New York called the Van Dyke Houses. And I'm just going to zoom in there a little bit further. And what you can do with these data is look at the data not just in aggregate but separately by race and ethnicity or by gender.

[00:15:11.69] And so if I click on Black Americans here to focus specifically on the outcomes of Black kids growing up in these neighborhoods, you see a striking pattern in this case, which is that kids growing up in this Brownsville area of Brooklyn, north of a street called Dumont Avenue, really unfortunately have very poor prospects of rising up. They have average incomes of only $18,000 a year in adulthood for kids who grew up in the Van Dyke houses.

[00:15:36.56] But what's striking and interesting about this particular example is if you just go across the street south of Dumont Avenue, you see that kids growing up there have average incomes of $29,000 a year. So if you just cross the street, you have quite different prospects. And it turns out just south of Dumont Avenue, there's a mixed income housing development that was developed by a congregation about 30 years ago. And our sense is that's part of the reason why

you're living in a different environment with less concentrated poverty, you see very different outcomes for kids.

[00:16:05.61] So as you can imagine, this very granular variation gives us great precision now in trying to understand what drives these differences in upward mobility and what we might be able to do about it. So going back to the slides using that data, a number of scholars have exploited this data to try to study what is it about neighborhood environments that are leading to such different outcomes for children.

[00:16:31.89] And so just to pick one example, here, I'm taking a paper by our colleague Rob Sampson in the sociology department who builds upon this data merged with some very nice data he's collected in Chicago to really show how the toxicity of neighborhoods defined in certain ways in terms of violence and incarceration is very highly predictive of differences in outcomes for children.

[00:16:55.86] So when we're able to put out data like this in a way that respects privacy-- and I know Gary King is going to talk more about this in a second. We use a tool developed by our colleagues in the computer science department called differential privacy to release all of the data that I've been showing you in a private manner that doesn't compromise any one individual's data.

[00:17:15.75] What that enables is scholars like this, like Rob Sampson and his graduate student, Robert Manduca, were able to do a downstream analysis that then really digs in and figures out what's driving these differences across places. And I think that opens up tremendous possibilities in the field. It's not just established scholars like Rob Sampson who are able to do that type of work.

[00:17:36.93] One of the things that I had not envisioned when starting to do this work is big data I think also makes this kind of information much more accessible to our students. So I teach introductory economics class called Economics 50 which many freshmen and sophomores take at Harvard. And the first assignment in that class is to take the Opportunity Atlas, look up the neighborhood where you grew up, and then talk about and try to understand how your prospects might have been very different if you had grown up in a different part of your city and try to understand why that might be the case.

[00:18:08.81] And just to illustrate the creative things that our students are able to do with this data, here I'm showing you a slide from a project by a student, Naomi Vickers, from this class, who's from Richmond, Virginia. And she noticed a very striking pattern which is if you look at maps of redlining from the 1930s-- so some of you might know that this was essentially credit scoring that the government did that was based on race, among various other factors, that ended up classifying some neighborhoods as being less creditworthy than others. She shows that those classifications persist to the present day in terms of the economic opportunities of children, showing the incredible lasting legacy of, I think, some of the institutional factors that led to segregation in the United States.

[00:18:51.62] And so this kind of work by many scholars, including our own team, has led to an understanding of what some of the characteristics are of neighborhoods that tend to promote upward mobility. Here, I just summarize them briefly in the interest of time. They tend to be places that have lower levels of poverty, more stable family structures-- so they tend to statistically be places that have more two parent families. They tend to be places, as you might expect intuitively, with better quality schools. And they tend to be areas with higher levels of social capital. So, places where someone else might help you out even if you're not doing well.

[00:19:28.91] Now, social capital-- just focus on that factor for a second. That's a complex factor that can be very difficult to measure in practice. And that's a limitation of the type of government data I'm talking about where we're not able to measure something like social capital super precisely. This sets up, I think, some of the issues that Gary is going to talk about in his talk where I think private sector data, for instance, from companies with social networks, have tremendous potential in enhancing our understanding of those types of factors.

[00:19:58.61] So what I want to do in the last couple of minutes-- I've shown you how these types of data, I think, can really enhance our understanding from a research and scientific perspective, in terms of what's driving these differences in upward mobility across areas. I also want to note that for those who are in this literature, you might know that a lot of what I'm showing you here is correlational. We and others have done a lot of work looking at families that move across neighborhoods either voluntarily or as part of randomized experiments where some families are given housing vouchers to move to higher opportunity neighborhoods.

[00:20:30.71] And from that analysis, we now have a pretty clear consensus in the literature that a lot of the variation we're seeing is driven by the causal effects of place rather than different types of people living in different places. And so all of that, I think, sets us up well for what I wanted to cover in the last couple of minutes here, which is how I think we can take this type of research and really have an impact on policy to try to change the way we do things in the United States to increase opportunity for everyone.

[00:20:59.36] And so in light of the type of evidence I've been showing you here, in our group, we basically organize our efforts to translate research into policy into three broad areas, which I think follow logically from this perspective that the roots of opportunity are very local, and they're heavily determined by where kids grow up.

[00:21:18.01] So one approach is to try to reduce segregation. If you see that there's a higher opportunity area down the street in your neighborhood, you might ask, well, can we help more low income families move to those high opportunity areas? And so I'll talk a bit about how that is one approach that I think is potentially scalable and something very tangible that we can do.

[00:21:39.04] Of course, we can't possibly move everyone to a higher opportunity area. And so it's also important to think about how we bring opportunity to people living in lower opportunity places at present. And so that's the second area of place based investments. And then finally, once kids turn 18, of course, the key touchpoint for most kids is not where they're living, not their household, but rather where they're going to college, for many children. And so I think there's a big role that institution of institution of higher education, including Harvard, can play in

amplifying the impacts of colleges on upward mobility. And that's another area where we're also focusing.

[00:22:17.89] So what I'm going to do here is just briefly show you, focusing on that first bucket of reducing segregation, how there can be a very direct translation from these sorts of data to impacts on policy. And then I'll conclude. So to give you a very concrete sense of that, I'm going to show you another snapshot from the Opportunity Atlas here from the city of Seattle where you see the same mix of red and blue color. Some high opportunity neighborhoods, some low opportunity neighborhoods.

[00:22:46.33] But what we've done here is superimposed in the bright green dots that you see, the most common locations for families receiving housing vouchers, that is, rental assistance from the government. Currently, some of you might know that we spend about $45 billion per year in the United States on affordable housing programs that are intended to give families access to better neighborhoods. But surprisingly, when you look at this map, you see that most of the families that are receiving these vouchers, which are worth about $1,500 a month, are still living in low opportunity areas. Despite getting that assistance, these dots are clustered in the red and orange colored parts of Seattle rather than the blue green colors, right?

[00:23:25.48] So when we saw this data, we teamed up with the Seattle and King County Housing Authorities to ask, "why is that the case?" And is it possible for us to maybe help more families move to high opportunity places by assisting them in the housing search process? So what we did is designed a randomized trial to help families that are receiving housing vouchers move to high opportunity neighborhoods if they want to.

[00:23:49.42] And specifically, what we did in that trial is gave half of the families randomly selected some additional services when they came in to apply for a voucher. Think of it as essentially assistance in the search process, connecting you with a landlord, a little bit of financial support, helping you identify properties and high opportunity neighborhoods that might work for you, et cetera.

[00:24:10.55] And so the goal here was to test are families not moving to these high opportunity areas because they have some preference not to do so, because they're maybe close to their family or close to their job? Or are there these small barriers that are preventing families from moving to neighborhoods where we know their kids would do much better, and they're just not able to get there for some reason?

[00:24:30.29] And so here are the results of that trial, just very simple. It's a randomized experiment. In the control group, only 14% of families use their vouchers to move to high opportunity areas. That jumped to 54% in the treatment group. And we estimate that the families in the treatment group that move to these higher opportunity places, their kids will go on to earn, on average over their lifetimes, about $200,000 more. So a small intervention, as you can see here, has a dramatic effect on the effectiveness of this existing government program.

[00:25:03.01] Now I want to flag here that another, I think, limitation of this pure big data sort of approach which is-- we might be interested then in understanding why was this program so

effective, what are the barriers that are preventing families from moving to higher opportunity areas. And this is where I'm anticipating what Mario is going to talk about-- And I've learned a lot from type of work Mario and others have done-- ethnographic work.

[00:25:28.31] We, as a complement to our study, work with sociologists to interview 200 of families that participated in this experiment and tried to identify some of the mechanisms that were preventing them from moving to these higher opportunity places. And you can see for yourself some of the remarks. They highlight things like the emotional and psychological support that came from the counselor who was helping them find housing, the connections with landlords, the short term financial assistance. Basically this customized support package really made the government assistance much more effective.

[00:26:00.98] So why does this kind of work matter? Last thing I'm going to show you. So in light of this research, there was... with bipartisan support a bill passed in Congress a couple of years ago to expand what we did in Seattle to cities across the country; they allocated $65 million to do similar demonstrations in cities across the United States. And most recently, there's now been a $5 billion per year expansion proposed of the housing voucher program again with bipartisan support in light of this work that I think has the potential to have an impact on thousands of kids and millions of kids across the United States in terms of where they're growing up and what their opportunities are.

[00:26:41.38] So just to conclude, I want to emphasize that I've focused here on one particular approach that I think stems from these data that tackle some of these challenges facing our society in terms of inequality and opportunity, reducing segregation. But by no means is that the only approach we should take. I think it's very important to think about other complementary strategies. And we're pursuing similar work in other spaces along the lines of what I've been showing you here, again, facilitated by the availability of modern administrative data.

[00:27:09.56] So let me end by just sharing what I hope will be two broad themes that you take away from this. I think the availability of modern data has really transformed our ability to study critical questions of our time. And I hope some of you will be interested in connecting with our team. You can read more on this website or email us if you're interested in using some of this data yourself.

[00:27:28.87] And then more substantively, for those who are listening and are interested in these issues, I hope one message that comes from this is if you take that initial slide of the fading American dream, that can kind of seem like a depressing figure and a daunting fact. How can you have an impact on such a big thing as any one person? But I hope some of these data I've shared today will show you that we can have a big impact in our own neighborhoods, in our own universities, in our own institutions to create better opportunities for all. Thank you so much.

[00:28:01.24] ELIZABETH PHELPS: Great. Thank you so much, Dr. Chetty. That was so interesting. I'm going to be following up with you myself at some point about that. So our next speaker is Gary King. Oh, sorry. Before I introduce Dr. King, I just want to say, if you have a question that's specifically for Dr. Chetty, go ahead and enter it at any time in the Q&A. I'm not

going to be reading any of these questions until the end, but you can start to enter questions as they come up.

[00:28:30.00] All right, so now our next speaker is Dr. Gary King. He is the Albert J. Weatherhead III University Professor and Director of the Institute for Quantitative Social Science. Dr. King is a political scientist who develops and applies empirical methods in many areas of social science, focusing on innovations that span the range from statistical theory to practical application and topics from fairness in legislative redistricting to improving social security forecasts.

[00:29:01.47] The title of his talk today is 'new leverage from partnering with private enterprises generating important sources of big data'. So Dr. King--

[00:29:11.61] GARY KING: Thanks, Elizabeth. And thanks, Raj, for an incredible talk on the importance of data from government. I'm going to talk about the importance of data from industry to complement what you talked about. But to do that, I thought I would start with something that really pissed me off just to give you a feeling.

[00:29:34.29] So in 1995, I got my copy of Science Magazine which I subscribe to, have subscribed to since then. And they had an article, and this was the article in 1995. They did a perfectly great survey of 60 scientists and social scientists and others and asked them-- those at the frontier-- they considered it the frontier-- and they asked them what they see for the future of science. OK, so what did 60 scientists and social scientists forecast in 1995? There were physical and natural scientists and there were social scientists.

[00:30:06.84] The physical and natural scientists talked about breathtaking discoveries and inventions and engineering marvels and problem solved, and it was really inspiring. The social scientists-- there were a smaller number of them. But every single one of them mentioned zero discovery, zero invention, zero engineering marvels, no problem solved, no public policies that we were going to contribute to-- nothing. What did they talk about? They talked about the fact that we studied this. And over the next 25 years, we're going to study that, right, which is a very nice thing. It's very nice that we study different things. But that's different than solving problems.

[00:30:44.34] So fortunately, the social scientists in 1995 were wrong. Since then we have made spectacular progress much due to new data sources like some of the great work that Raj just described. Where do we get the data from this is the real question. Where do we get the data about the social world that we study?

[00:31:07.11] Well, not long ago, we had access to almost all the data about those people and groups and societies because we either created it inside the university, we obtained it from governments, or we purchased it from willing industry partners. And we sort of had-- we didn't have everything. But we had lots and lots of stuff.

[00:31:27.45] Today, we have more data than ever before. And that has created these incredible innovations, this incredible progress. But we have a smaller fraction of data in the world about the subject that we study, than ever before, because most of it now is locked up inside industry

and other organizations. They have done a great job at collecting information about the people and groups and societies that we study, and they have it all locked up. So we have no choice, if we're going to do our jobs, but to find ways of working with industry that has done an incredible job of collecting this information.

[00:32:06.73] So how do we do that? And my main message is-- it's not by compromising with them or balancing our interests and their interests because they have no necessary means or reason to give us data at all. I once had a graduate student call up a company and said-- and they said, a naive graduate student who said, do you have any research data? And the company looked around and said no, and they hung up.

[00:32:32.10] No, of course they have research data, right? Not only do they have data on their customers, they have data from their finance system and their HR system and their air conditioning system and their roads and ground systems. Pretty much everything these days, every system produces huge amounts of data. And it's just-- the company doesn't think about it as research data. They think about it as they're doing their business. We think of it as a research data, right?

[00:32:57.36] Also, their goal is not to get an article published in the American Journal of Political Science. Their goal is whatever their business goal is which is just a perfectly legitimate thing. And we have to find ways not of compromising with them, not of balancing, but two things-- educating ourselves about them in detailed probably qualitative ways like Mario will tell us, really understanding them. Not only understanding the company, but understanding the person you're talking to, and that person's boss that that person is paying attention to, and the people that work for that person. And then also, spending time to educate them about our interests because the people in those companies-- they're basically our former students. They would like to create public good just like we would. And so we need to explain to them what we need to do to be able to create that public good.

[00:33:50.34] So first is education and the second is creativity. Yes, we have questions. And then we would like to go get the data to answer the questions. But it is incredibly important to be flexible about the questions you would ask. In school, from kindergarten up to the beginning of graduate school, we are taught that the teacher asks a question, and you don't change that. It is a fixed, immutable, unchangeable question, and you do your best to answer it. And that's what school's about.

[00:34:19.38] But actually, starting in graduate school through a professional academic career, most of the time, we adjust not only the data sets we collect, but the questions we ask. Raj's examples we're really terrific examples. I'm sure that a lot of those questions we would never have thought to ask, we would never have imagined we could ask and get a serious answer for it. So you have to focus on creating the new questions, not only looking for the data to ask the questions that we want.

[00:34:46.62] So education and creativity-- I'll come back to this. I'm going to give three examples of these points. First is-- three research examples. First is the effect of the news media, OK? So we did a study of the effect of the news media. We wanted to randomize, respecting

their constraints. Now lots of studies of the news media had been done. But it's very difficult to figure out the effect of the news media because for the most part, the media organizations are businesses which are following people. If they don't, they go out of business. And we want to know the effect of the media on people. And so there's a huge endogeneity problem. And as we all know, the way to deal with that is to randomize.

[00:35:29.79] So how do you randomize in the media? Well, there's the constraints of the journalists and the constraints of the social scientists. The journalists-- they absolutely have to have total control over what's published and when because when we're negotiating with them, as soon as they use the word journalistic integrity, that means like we have to start all over, right? You have to respect their needs. And their needs are to have total control over what's published and what.

[00:35:53.58] What does the scientists need in order to randomize, in order to assign treatment to a group that we choose, in other words, to tell the organizations what to publish and when. That's right. So we also need total control over what's published and when. So do the journalists. So all we have to do is come up with a solution here with no compromise.

[00:36:17.76] So how do you do that? Well, there's lots of ways. You can't do it the way that you intended. We have to do it in a different way. We found 48 media outlets. These are some of them-- small media outlets-- that were willing to talk to us over about five years. And us includes Ariel White, who was a graduate student and now is at MIT, and Ben Schneer who is post-graduate student and is now at the Harvard Kennedy School.

[00:36:41.97] And we talked to these organizations and the people in the organizations and their supporters and their readers and their subscribers for a long time in lots of forums. We went to the conferences. We did interviews. We really tried to understand their perspectives. And then we came up with some solutions. We made it possible to have total control over what's published and when for both the journalists and the social scientists.

[00:37:07.98] How do we do that? Well, there are lots of design features of the experiment that we never would have put in initially. But after we understood them, we could put in. So the journalists completely choose the content. But then we approved the content that they choose for the experiment. If we reject it, they can still publish. It just won't be part of the experiment. So we get total control, and they get total control. And there were lots of design features like that. Everybody gets total control. There's no reason for balancing, no reason for compromise if you're creative about the design.

[00:37:39.61] We also developed a number of novel statistical methods because we knew that their patience once we started the experiment, was going to be the main bottleneck. And so we used these-- we've developed these new statistical methods to reduce the number of observations needed-- matched pair designs and sequential randomization.

[00:37:59.82] Our results, just to give you a feel for it-- three small media outlets publishing about some very specific topic-- like Uber drivers in Philadelphia and what they think about automated or driverless cars. They have an effect on the National conversation not only about

what Uber drivers in Philadelphia are talking about, but about the broad area of the economy and jobs or in other areas. That has a big ... an enormous-- quite an enormous effect on the National conversation, which changes the agenda, which does have very clear effects on public policy. So we would not have been able to do this without the detailed understanding across these sectors.

[00:38:40.65] Let me give you a second example. So Facebook has tons of data. They have more than two billion users. Imagine going to get an NSF grant where you say you're going to collect data on two billion human beings through personal interviews. It ain't going to happen, OK? So I made many trips to Facebook, trying to convince them to make data available. On one particular trip, I went there. Didn't really work. No big deal. I come back.

[00:39:08.43] And I'm in my hotel room packing. And I get an email from them. And they say, hey, Gary, what do we do about this? And "this" was Cambridge Analytica, which was the biggest scandal in social media up until that point. And it was about an academic sharing data inappropriately with private industry. It was the worst timed lobby event probably in the history of the world so far as I know. So you don't want to do-- I just went home and I thought, this is over.

[00:39:38.46] But they called me three days later, and they said, hey, could you do a study of the 2016 election and tell everybody we didn't do anything wrong or we didn't change the outcome or maybe if we did something wrong, we'll... tell us what it is and we'll fix it right away? Losing $100 billion in market cap sort of focuses the mind. But I said, I'd love to do this. I'd love to do the study of the 2016 election. But I need two things. And I think you're only going to give me one. But I got to have two.

[00:40:07.02] First thing is I need complete access to all the data and people and processes in the platform and everything just like you give employees. And they said-- Mark Zuckerberg said, OK. Well, he said, what's the second? I said, the second thing is no prepublication approval. When I am done, I will publish. And you will not have approval of a publication.

[00:40:26.97] He said, no, no, we can't give you that. We never give employees that. We never give anybody that. I said, OK, well, I'm not going to do the study. He said, no, no, I want you to do the study. I said, well, OK, give me those two things. And he said, no, we can't give you those two things. I said, well, OK, then I'm not going to do the study. And so we went back and forth like that a few times.

[00:40:44.91] And finally, I realized, wait a second there's a solution here. There's a creative solution so that everybody can get everything that they want. Instead of doing a study of the 2016 election, I'm going to change the question. So here's the way we're going to do it, and this worked. So I said, we're going to have outside academics. They're going to send in proposals. They are not going to have any veto by the company. They can publish it whenever they want. However, somebody approves the proposals and it's not going to be Facebook.

[00:41:12.97] [PHONE RINGING]

[00:41:14.93] Sorry. Somebody should mute themselves. [LAUGHS] So it's not going to be Facebook that approves it. Instead, we set up a trusted third party, a commission of senior distinguished academics, who take one for the team, agree not to publish ever on the basis of the data, but they get all the information inside Facebook. They know what's going on inside Facebook. So they won't agree to a perfectly meritorious proposal that is on the same subject about something that Facebook is being sued about that is not public, things like that.

[00:41:46.83] So Facebook agreed that the trusted third party makes all the final decisions. And then once they make the decisions, they give the academics access. The academics can publish with no company approval. And if Facebook reneges, then this trusted third party set up at an organization we created now at Harvard called Social Science One would report publicly.

[00:42:07.95] So this completely solved the problem. No balancing, no compromise, legal agreement signed. There were public announcements. Mark Zuckerberg talked about this when he testified before Congress. Facebook assigned 30 people to this project. We got eight ideologically diverse foundations to give us more than $10 million to get out to academics who were also [INAUDIBLE] data access.

[00:42:33.54] There was just one problem. The particular implementation plan that Facebook had was-- well, illegal. No, they didn't know this. So it wasn't their fault. But the regulators said, hold on a second. You can't do this. So what we thought was going to take two months actually took two years. And it became a question-- how do we protect user privacy way better than the plan was and also protect social science analysis-- which Raj referenced earlier.

[00:43:03.82] So our solutions without balancing were new methods of data sharing, which is the differential privacy that Raj mentioned. So we had to modify this concept of differential privacy where instead of trying to deidentify data, which doesn't actually work even though we've been doing it for decades, it adds specially calibrated random noise to the data. And secondly, we develop new statistical methods to deal with the fact that the noise creates biases in the data.

[00:43:32.22] The result of this was Facebook has now released a data set with 17 trillion numbers in it. That's a T for trillion. All on the effects of social media on elections and democracy. We, in the first round, have given around 100 researchers access. Now that's almost 200. From this, we've built some institutions. This organization called Social Science One now has a consortium of social science research centers like the IQSS like organizations that grew up. And we helped to create around at other universities. We now meet regularly.

[00:44:07.47] And they call us or we call them, and we say, how did you ever convince this company to do that? Or just as difficult-- just as difficult-- somebody will call us up and say, how did you ever convince your general counsel's office to do that thing, right? And actually, figuring out how to do that is really hard because these kinds of problems-- we've never had these kinds of problems before. So we have a partnership with Microsoft. We have an open DP, open differential privacy initiative among a number of other things.

[00:44:38.97] OK, last example. Reverse engineering censorship in China. So one way of really understanding a company is to create some technology, license it to the company through

standard corporate procedures, and help found the company, which I did. And then you stay involved, and you get to understand them. So I sort of knew what they were.

[00:44:59.93] Our goal-- my two graduate students, Jen Pan, now at Stanford, and Molly Roberts, now at UCSD-- was solely to obtain data on Chinese social media posts so we could do a methodological study of automated text analysis in Chinese. That was our goal. But in the course of doing this, we stumbled upon the fact that we had access-- Brandwatch, the company that I founded. They didn't know-- had access to all of the posts, the social media posts, before the Chinese government decided which ones to censor and take off the web. And we thought, holy cow! Let's forget that study on automated text analysis in Chinese.

[00:45:38.70] And so we completely changed the question and now looked at censorship. So we had a big pile of censored posts and a big pile of uncensored posts. We had to figure out what the point was. So now, just to give you a feel for it, everybody knows the goal of censorship, right? It's to stop criticism and protest about the state, the leaders, and their policies. The first thing we learned was this was completely wrong. The difference in these between the two piles of censored and uncensored posts was basically the same.

[00:46:06.94] So we tried a lot of things. It took us a long time. But we finally figured out, with millions and millions of censored posts and millions of millions of uncensored posts, how they differed, how did they differ. We posed the question in a different way. We said, wait a second. It could be that their goal is to stop criticism of themselves in the state, or it could be that their goal is to stop collective action. And when we ask the question this way, it became crystal clear. The first was wrong, the second was right.

[00:46:33.99] Let me make this clear to you. So if you write into Chinese social media or at least one we did the study, if you wrote on Chinese social media, if you write that the leaders of this town are all stealing money, they have the money in overseas bank accounts, here's how much. And by the way, they all have mistresses and here are their names, that will not be censored. But if you say, and let's go protest. That will be censored.

[00:46:58.29] In fact, if you say, the leaders of this other town are doing such a great job, let's have a rally in their favor. That will be censored also. They don't care what you think of them. They are a bunch of dictators. What should you think of them? They just don't care. But if you have the ability to move people, they will stop you. And that's what we discovered. It's collective action. It's not criticism.

[00:47:19.59] And the implications of this were pretty cool. We can predict and have predicted which officials, which local officials were in trouble and we're likely to be replaced. We can easily predict the arrest of some dissidents before they're arrested, peace treaties before they're signed, scandals before they emerge even if we don't even know what the scandal is, disagreements between central leaders and local leaders.

[00:47:46.28] I'll leave you with one more substantive thought which is it's not only the leaders of China that are OK with criticism but are not OK with collective action, it is probably also the leaders of companies, it is probably also university leaders. Isn't that true, Professor Bobo?

[LAUGHS] Pretty much anybody that has to deal with a lot of people-- criticism is OK. But if they get together, it's actually a big problem.

[00:48:12.85] OK, so let's remember-- to work with industry and others, it's not about compromise, it's not about balance. It's about education, educating us about them in a very detailed way and them about us. And it's about creativity-- asking new questions, enabling the data to suggest new questions to us which will help us come up with unexpected solutions. So I look forward to your questions and to the next talk.

[00:48:42.98] ELIZABETH PHELPS: Thank you so much. I'm going to come back to this in the question period. But I'm kind of wondering what social scientists would say in 2021 about the future of their field. We'll get back to that. All right, the last speaker today is Mario Small. He's the Grafstein Family Professor in the Department of Sociology and a visiting professor at the Harvard Business School.

[00:49:06.83] Dr. Small's research is currently focused on the relationship between networks and decision-making, including investigating the ability of large scale data to answer critical questions about urban inequality and the role of qualitative inquiry in cumulative social science. The title of his talk today is 'why big data analysts need ethnographic and other qualitative inputs'. Dr. Small--

[00:49:30.31] MARIO SMALL: Thank you for that opportunity, and thank you, Larry and the rest of the speakers, for this. I'm going to share the screen right now. Just going to make sure that worked. Could I get a thumbs up to-- awesome. Thank you very much.

[00:49:42.75] So what I'm going to talk about is both a work in progress that both combines government data and also private data and in both cases to tell us why we can't quite make sense of them unless we also have access to field data. And so it's actually a talk that's very consistent and actually dovetails quite nicely with what you heard from Raj and Gary.

[00:50:07.73] So the argument is actually pretty straightforward. I'm going to give you right away so that we can start from the same page. What I'm arguing is that we need qualitative research in this space, not for illustration, which is a common idea out there. Also not for what I often hear people describe as richness-- sort of to make findings richer. Not for texture is another kind of phrase that you use. And if you could see my hands, you would see me waving them in a way that sort of is consistent with the typical invocation of the word texture. And also not for relatability in the narrative. Not so that people can sort of connect better to what you're saying.

[00:50:46.40] But I'm going to argue is actually instead that you need it for science, meaning that for some questions, it's actually impossible to do actual science on the basis of big data without using field work. And this is both before and after doing the work.

[00:51:00.29] And I'm going to make this case with an example. And the example is a study that I'm doing right now-- a study of access to banking in minority neighborhoods. Actually, minority neighborhoods... they're actually built very much on some of the work that Raj has done. What

I'm going to show is that both how to conceive of questions, which big data to get, and how to explain results all require some aspect of field work to understand this broader question.

[00:51:26.95] OK, so the basic context is access to banking. And the basic idea is as we just saw. And I think at this point, there's no social scientists who would disagree with this core idea-- neighborhood conditions matter. They matter for inequality-- excuse me-- for mobility, for life chances, for a whole bunch of outcomes that all of these papers down here have documented many times over.

[00:51:49.66] The question, I think, at this point-- and Raj very much spoke to this-- is how. What is it that living in disadvantaged neighborhoods does to you that makes your life difficult? What makes them difficult places to live? What I'm going to focus on is one answer, which is that these neighborhoods provide limited access to resources. And the particular resource I'm interested in is access to banking services-- brick and mortar banking services. Why would anybody care about brick and mortar banking services in the internet age?

[00:52:18.82] It turns out brick and mortar persists over essentially last 10 years before COVID. About 17,000 new branches were actually opened in the United States, even though the overall number of branches closed. And it turns out it has to do with a whole bunch of reasons, including persistent consumer preference, the need for banks to sell certain kinds of products that can only happen in person, and also the need for certain kinds of activities for legal reasons to happen online. You have to show your ID for certain kinds of things, including closing your account.

[00:52:51.47] Second-- it turns out that proximity matters. So there's some evidence of this. It's correlational, but it's still pretty interesting. Turns out that even today, having no bank-- and here's a citation for those who are interested. Having no bank within three or five miles of your home residence increases the probability that you do not have a bank account. And if you're curious, today, about 7% of families have no bank account. And the proportions are higher both among low income populations and also among African-Americans. And this association of proximity to having no account is actually the strongest for people who are low income.

[00:53:24.10] OK, so there's a standard way of answering this question, and that practice is the way I answered this question back in 2006 in a paper with Monica McDermott, which is basically to count the number of banks in a zip code. Banks that have no zip codes, you call them a bank desert or a banking deserts. And then you compare whether neighborhoods that are, for example, more minority or more poor are more likely to be banking deserts.

[00:53:46.96] Let's just do a version of this. This is Upper Manhattan. This is Central Park. And the idea is you-- this is literally zip codes, and you would look at zip codes and you say, oh, these are banking deserts, and this is not a banking desert. And you would do this on a national scale.

[00:54:01.36] I've done quite a bit of field work in Boston, in New York, in Chicago, some in Philadelphia. And I found a couple of things. Number one-- even though it actually is quite convenient to use zip codes because we have the data from the Federal government to answer this question, people do not actually live in zip codes. And this actually sounds like a pretty obvious point, but it turns out it's actually pretty consequential for the understanding of banking.

[00:54:25.03] I'll give you an example. This is another map, again, of upper Manhattan. And this is Central Harlem right here. And this is literally where I lived for four or five years when I was an assistant professor at Princeton and when I was doing research in New York. And shortly before I moved to this neighborhood-- you see this star right here-- Harlem used to be what's called a food desert. There were no supermarkets selling fresh foods or organic foods and so on. So right before I moved in, this supermarket opened up here. And headlines were everywhere-- Harlem no longer a food desert.

[00:54:59.56] Well, in the five years I lived actually only one zip code away from this supermarket, I never went there, not even one time. And the reason is that it took one hour to get there even though I was quite close. Why? Because to get there, I'd have to take the 1 or the 9 or the ABC or D trains, taken down to a-- excuse me 125th Street, which is the heart of Harlem. And then waited a 125th Street right here-- I hope you can see my cursor-- for the bus to come in. And take the bus across and get off at the supermarket. It would take an hour.

[00:55:30.55] In contrast, because the A train, which is an express train, stopped right here, I could literally be all the way down here to Columbus Circle in 15 minutes and actually all the way down in Chinatown in about 25 minutes. It was literally faster to get all the way down to Chinatown than to come here. So accessibility was actually very much affected in terms of my lived experience by that layout, special layout of the neighborhood and by the infrastructure of the transportation grid. And so those two things matter to how we understand the lived experience of accessibility.

[00:56:03.22] The second thing I found-- and I saw this especially in Chicago-- is that it's not as if people living in places that are often called banking deserts have no options-- they do-- it's that those options are often not only financial institutions. You've seen these, and these are in cities, suburbs, rural parts of the country. These are in New York. They're all called check cashing stores. Sometimes are called payday lenders, although payday lenders and check cashers are different things. Cash advance stores, cash for you-- and these are places that-- these are generally called alternative financial institutions where you can either cash a check or retrieve a loan at high rates.

[00:56:37.91] Now the thing about AFIs that's important for this talk is that they typically charge very high fees under strict terms and as a result have high default rates. So to give you an example, a typical sort of loan that you can get is a $400 loan for which you would pay in two weeks $470. So that $70 is the interest. Annualized, it's about a 450% APR, which is part of what many people call these predatory institutions.

[00:57:06.35] But the interesting part of this is that many, many people, at the end of those two weeks, still can't pay the $470. And so they just roll over into a new loan for another two weeks under very strict terms. This happens actually quite a bit. States have started regulating these and it is the reason they're called predatory.

[00:57:24.92] So the way to think about this issue is not that if I live in a poor neighborhood I have no options, it's-- number one-- that what my options are depend literally on the layout of my space, my physical space, and the accessibility of modes of transportation. And second-- that

the undesirable options might actually be closer to me, meaning faster to get to than the desirable option-- that it might actually be easier to get to an AFI than it is to get to a bank.

[00:57:51.18] So that completely changed how we approach this question. So the question I asked is really what is the probability that the nearest alternative financial institution is closer to get to or faster to get to than the nearest bank and whether that probability is higher in minority neighborhoods. So that's the question we asked. We did this for the 19 largest cities or 18 largest cities plus Boston. So we do some checking of the data. We did it for every block in the city. So we literally calculated for every block in the city how many minutes it would take to walk or drive or take public transit to the nearest AFI and also to the nearest bank. And then we compared the numbers.

[00:58:33.85] I won't go a lot into the data but I'm going to note it because this is a talk on data and big datam and these data are big. With the Google Maps database, which it turns out at this point is the best data source there is including administrative data from the Federal government. And that was not the case 10 years ago. We got literally the street infrastructure, including direction, speed limits, and so on from open street maps which is publicly available to nonprofit companies, essentially, organizations. We got the public transit schedules from the federal government, and we got the neighborhood characteristics, of course, from the federal government.

[00:59:06.73] OK, just really quickly what we did-- so here's how the algorithm works. We picked a block-- so for those of you who are not familiar with these data, the country is divided in the states. The states are divided into counties. Counties are divided into census tracts. Census tracts are divided into block groups. And block groups are divided into blocks. Blocks are exactly what you think they are. They're a block. And so this is a census block, and a set of blocks constitutes a block group.

[00:59:30.49] So we pick the centroid of every block in the city. We figured out where-- this is not 10. But we figured out where the 10 closest institutions... say the 10 closest banks were, as the crow flies, right? Just literally drawing a straight line on a map. And so you can see in here this would be the closest one. Except you don't actually get to a bank the way the crow flies, right? If you're driving, for example, here, you'd have to drive this way, turn right, and turn left, and then you get to the institution. So that's what we did.

[00:59:59.15] But now imagine that in this particular city, there were 3 one way streets going this way. Well, if that were the case, you'd have to come this way, find some other place to turn around and get all the way here. It might actually be the case that this is not the fastest to get to. Or imagine there's a big wall or a big cliff or a tunnel or something that's actually obstructing so you'd have to go around it in order to get to it. In that case, this actually turns out to be the fastest to get to. So that's what we calculated.

[01:00:23.16] So in this case, the block groups datum was four minutes. We averaged it out for all of the blocks and-- sorry. The blocks datum was four minutes. We averaged it out for all the blocks and the block groups. In this hypothetical case, it's 4 and 1/2 minutes. And we did it for

all of the block groups in all of the 19 cities-- about 22,000 block groups, OK? So what do we find?

[01:00:42.33] So first, turns out there are far more banks than AFIs out there. And so banks are typically faster to get to on average. You're curious, by foot, on average, in the 19 largest cities in the country, about the 19 largest cities in the country, it takes about 13 minutes to walk, about 12 minutes to take public transit, and about 2 and 1/2 minutes to drive. With AFIs, it takes longer-- 24, 21, and about 5.

[01:01:09.32] Now the question, of course, is whether these differences remain or what these differences look like when you compare neighborhoods of different racial composition. So here's what you see. This is just a rough first cut. If you look at neighborhoods that are predominantly Black, meaning 50% or more Black-- and here, neighborhoods means block groups. So predominantly Black block groups, predominantly Latino/ Latina, predominantly Asian, predominantly white.

[01:01:32.21] What you're seeing here is that for predominately white block groups, it's only 10% of the time that the AFI is actually more accessible than the bank. Accessible here means faster to get to. Very similar figure for Asian block groups. So for Latino block groups, more than 30% of the time, regardless of mode of transportation, the AFI is faster to get to. And for predominantly Black block groups, it's more than 40% of the time.

[01:01:57.08] Now, obviously, these differences don't reflect differences in poverty rates, income, and employment, et cetera. So we accounted for this. So let's see what happens if we account for poverty in the neighborhood, population density, portion foreign born, unemployment rates, college education rates, homeownership, housing density, the vacancy rate, commercial density-- basically everything we could think of that we could measure at the block group level. What are those differences look like? I'm going to show you the results.

[01:02:21.68] I'm going to walk you slowly through the first of these, which is just for travel by foot. And you will very quickly see the pattern. This is the predicted probabilities-- these are-- in economics you'd call them margins, in sociology-- I don't know in political science. In sociology, it's predicted probabilities. For a neighborhood that is at the mean of all of the characteristics-- this is just the grand mean-- except it's either 10% or 30% or 50% or 70% or 90% white or Black or Latino and either low poverty in the bottom panel or high poverty in the top panel. Low poverty means 10% poor, high poverty means 50% poor, OK? So we're looking at extremes.

[01:03:00.62] What you're seeing here is for low poverty neighborhoods, as a proportion white increases, the probability that the AFI-- by the way, excuse me. The dotted lines are unadjusted figures. The solid lines are adjusted figures with confidence intervals. And what you see is that as the proportion white increases, the probability that the AFI is closer goes down dramatically. This is regardless of whether the neighborhood is high poverty, low poverty.

[01:03:25.88] But as the proportion Black increases, the probability of the AFI is-- this is on, again, just the probability-- actually goes up. Same for proportion Latino, although you can see that for proportion Latino, once you adjust a whole bunch of characteristics the effects flatten.

This is for travel by foot. Notice also that the top panel and the bottom panel look, actually, quite similar. Poverty doesn't really matter that much once you take race into account. Or put differently, race matters a lot more than class.

[01:03:52.00] Here's for travel by public transit. It's a very similar story. Here's for travel by car. Turns out cars kind of level inequality a little bit, certainly with respect to proportion Black. But again, the basic story looks very similar.

[01:04:06.14] OK, so we have found these results. It took us forever to put-- years to compile these algorithms and run these data, et cetera. We're pretty excited. So we sent these results to kind of a general science journal-- that's lowercase 'science' journal. And the editor who was not a sociologist or an economist or a political scientist said, well, this is obvious. Minorities just prefer payday lenders more than whites. We said, OK.

[01:04:32.03] So we did the following-- what happens if we compare the opposite ends? What happens... we know that the people who would want a two week loan, regardless of the rate, are more likely to be at the bottom of the extreme distribution. And we also know that the people who-- that if you went to college and if you own your home, it's pretty hard to avoid having to go to a bank. You needed a mortgage to get the house most likely, and you likely needed a college loan.

[01:04:56.73] So what happens if we compare neighborhoods that are overwhelmingly white but also poor, very poor, highly unemployed, very uneducated, and composed primarily of renters? But then on the other side, we compare to highly minority neighborhoods that are really affluent-- 10% poor, very low unemployment, very educated, very homeowner. Let's see what that figures look like. This is what you're seeing.

[01:05:20.63] On the left here are essentially the destitute white neighborhoods. And the right here are the affluent neighborhoods. White in pink, majority Black in blue, and majority Latino Latina in green. And what you're seeing here is that the probability that the AFI is faster to get to is actually still higher in the affluent, highly educated, et cetera, homeowner minority neighborhoods than in the white ones. And in the poor destitute white ones. This is with travel by foot or travel by public transit. And actually, it's a little more complex for Latinos that travel by car as you can infer.

[01:05:54.56] Obviously, these confidence intervals are quite large because there aren't that many highly educated, highly affluent, et cetera, et cetera African-American neighborhoods. So you want to take this with a grain of salt. But it's actually still quite striking that even at these extremes, you don't find what you would think you would find. OK, so so far so good, right? I'm just going to show you two more things and then we'll wrap up.

[01:06:13.05] And so far, AFIs are more accessible as the proportion of minority increases and unlikely to just be a function of preferences, I think. But maybe you're like me and still skeptical. OK, fine. What if minorities actually prefer AFIs over banks? You have to kind of go beyond the big data. And here's what we did. So we just asked them, see what we find.

[01:06:32.13] So we launched a new survey. And I'm going to show you just one set of results. This is nationally representative data-- 3,000 respondents. And we asked them, would you prefer-- I'll show you the wording in a second-- banks versus check-cashing places, and then separately, banks versus payday lenders. The question is, literally, is there a racial difference.

[01:06:49.83] And so here's the question. There was two separate questions. One is, we want to learn about your preferences. Suppose you had a $100 check that you needed to cash in person. Where would you cash it? We told them examples of what we meant by a bank and by a check cashing establishment. These are the biggest banks in the country. These are the biggest check cashing establishments. And of course, the options were randomized, et cetera.

[01:07:11.93] Separately, we asked them, OK, now suppose you needed to borrow $500, here are your options. Where would you go-- payday lender or the bank? That's it. So we just asked them. And these are what economists call stated preferences. I understand that Samuelson and others believe that stated preferences are worthless, that instead we should believe revealed preferences. I think that there's something to the idea of focusing on revealed preferences. But I believe that stated preferences are important as well.

[01:07:35.78] OK, so here's what we found. For all racial groups, the average person preferred banks over AFIs, OK, across the board. But there were differences in the rate at which they did so. And here is the interesting part-- OK, so here's-- and I'm sorry this is not that pretty. We've had some COVID related difficulties at home. So I didn't get a chance to make this pretty. But you'll get the picture. OK, here's what happens when you judge-- this is a simple regression. All you're adjusting for is race. What you're seeing here is that African-American respondents are less likely than whites to prefer banks over check cashing places, OK?

[01:08:14.33] After you adjust for demographic characteristics, the obvious things, you make that affect much smaller. After you adjust for a set of variables that account for how comfortable people feel with banks-- and you might recall that last year, a lot of people were reporting-- for example, African-Americans were reporting racist experiences with banks. People believing their checks were not theirs, et cetera. Turns out that makes the effect much smaller and no longer statistically significant. After you adjust for the feelings about check cashing places-- which, I guess, are more positive than we thought-- again it gets smaller, network effects get smaller, and state fixed effects making it even smaller as well. Similar story with respect to Latinos, except that demographic characteristics account for it as well.

[01:08:53.18] OK, so far so good. So there is a base preference. But once you account for demographics, it goes away. Now what happens when you compare banks to payday lenders? Here's the answer-- same story, a little stronger effect. Again, it gets smaller. Again, it gets even smaller. But it doesn't go away. And again, it gets even smaller, and it doesn't go away. Network effects, again, it gets smaller, but it doesn't go away.

[01:09:16.08] This is just an indicator for whether you know somebody who has a payday loan, has taken a payday loan, in your network. And state fixed effects-- we can't make the effect go away. This is ongoing. But so far, we don't know what's going on. And that's the key. We don't know what's going on. And so we need to go out and ask.

[01:09:32.77] So what we're doing now is we're doing-- you guessed it-- more field work. We're going back, and we're doing in-depth interviews with people across all 50 states divided into states where payday lending is heavily restricted versus not to figure out how people are making borrowing decisions and what seems to be playing a role in this preference. Is it beliefs, is it not understanding, is it availability, is it et cetera?

[01:09:53.41] But then separately, we're, again, asking managers how we locate. I mean, there's a supply side aspect to this question that, again, the data are not enough to tell us. So we're going to interview managers and we're going to ask them about these decisions, how much of it is data based, how much of it is intuition, guessing. And then we're going to bring them our data and say, well, how would you interpret these findings, to see what they say.

[01:10:13.27] So in sum, why qualitative research? Because I hope as I've showed you-- it's important not just the question we ask-- literally how you phrase the broad question of how neighborhoods matter, but also the particular data needed to answer it. It's not just that data became available. They've been available for the last only about eight years or so, 10 years or so. But kind of what is the particular way you should think about the data given what you find in the field. Without it, I think data collection can often be driven by either convenience-- the data were just there-- or priors. And we know that our priors are affected by a whole bunch of things other than proximity to lived experiences.

[01:10:47.53] And finally, the data are going to inform us still working on this how we interpret the location results, how we interpret the survey results. And I think without these, rather than science, what we have are just provocative findings. And I think for this reason, qualitative research is actually not just nice to have, but actually important to the actual scientific enterprise.

[01:11:07.54] OK, thank you very much. I'll also thank Armin Akhavan, Mo Torres and Ryan Wang who helped a lot on the first project. Thank you.

[01:11:19.09] ELIZABETH PHELPS: Thank you so much, Dr. Small. Fascinating work. I'm now going to turn it over to our discussant to start our discussion. And then we'll open it up for questions, which I'll be monitoring through the Q&A function. Our discussant is Peter Galison. He is the Joseph Pellegrino University Professor and Director of the Collection of Historical Scientific Instruments. Dr. Galison is a physicist and historian of science. His writing and films explore the complex interaction between the three principal subcultures of physics-- experimentation, instrumentation, and theory and the embedding of physics in the wider world. Dr. Galison--

[01:11:59.25] PETER GALISON: Well, these are wonderful papers. And I have some thoughts that I thought I'd bring to help us launch our discussion.

[01:12:17.08] The first thing that strikes me as an outsider to quantitative social science research is that the statistics here occupy a kind of mesoscopic scale, that is to say, they are between what Maxwell used to refer to as the molar and the molecular. They're not biographical, they're not individual case studies. On the other hand, they're also not at the scale of nation, states, or cities

as a whole. But instead, they're looking at local neighborhoods which are immensely important and give us a new kind of realism in how we understand what's going on.

[01:13:02.13] And that struck me across the board with all three studies, that is to say, there's a use of statistics here to try to get at a kind of reality on the ground that becomes invisible when we look at too wide a scale or too narrow a scale. If we focus entirely on a particular family, we may gain insights. If we focus at the city, at the scale of a New York or a Boston, we lose the dynamics of these local neighborhoods and considerations that are so important. And I think that that's true for Gary King's work as well in ways that I want to suggest now.

[01:13:41.67] Just to summarize briefly what I took away from these papers, in Raj's analysis, he wants to say, look at these individual neighborhoods, these opportunity neighborhoods, or these squelching opportunity neighborhoods. And we can actually look through data diachronically at states diachronically. But also, we can look at what happens when one intervenes and offers the possibility of helping people move into another neighborhood where their opportunities may be greater.

[01:14:16.41] Or conversely, he says, let's not look at college in general. We have lots of statistics on the economics of what the results of going to college for one's fortunes and wages. But he says, look at the particular ways that immersion in particular colleges functions in moving bottom 20 to the upper 20 percentile.

[01:14:39.55] And I think that this is also true in Gary King's work. He wants to dig down and silo the siloers, so to speak, in industry. He wants to be able to look at the media effects not in general, not in referring to just the bubbles of particular partisan media structures, but to look in particular and enough fine grained detail, although statistically, to be able to look at the way that-- at the effects. And to look at censorship, in his study of Chinese censorship in particular, not as one of opposition or dissent being squelched, but of the threat really that's felt by organizations. Gary's interested in particular in looking at new methods by which the otherwise closed data empires of industry might become accessible to study and has proposed this third party adjudicator entity that might help make that possible.

[01:15:41.16] Mario's work focuses, again, on the ground and locally, not asking after financial institutions in general, but at the different kinds of financial institutions and at something as pragmatic as how long does it take to get there. One of my favorite moments in the history of the discipline of history is when Fernand Braudel calculated how long it took letters to get across the Mediterranean and along with letters came armies and policing of different sorts. And that, he says, explains much of the rise of the modern world.

[01:16:15.56] So this kind of time to institution from different blocks and blocks of blocks become important and form the empirical basis for his refutation of the notion that these are choices-- preferences-- but instead involve much more deliberate structural aspects of the way our financial institutions in the so-called alternative financial institutions are structured and laid out.

[01:16:46.60] So here are some-- here's some questions that I have for you. And maybe some of them will prove interesting in what follows. The first is that some of the results that Raj presents are important because they put on an empirical basis things that we might have suspected already. And other times, I found some of them quite surprising, including the map that he opens with of the United States where, for instance, you see a tremendous amount of blue in his color coding system starting from North Dakota and South Dakota down to Oklahoma and West Texas where there seems to be a certain amount of economic mobility which might surprise us politically in our analysis since so much of the accounting of a rise of conservative politics is predicated on the notion of people who are prevented from or will find themselves in a downward spiral from generation to generation. As I thought, it was just really an interesting thing that this micro analysis then aggregates into this map of the country as a whole, which I think is regionally very informative.

[01:18:05.84] And the second is a question of open upward mobility versus access on the universities. Here, interestingly, Raj shows that at the elite or more wealthy universities, there's a larger boost in the number of people that go from... the percentage that go from a low percentile to a high percentile in income. But of course, there are fewer of them there who start out in the lower percentile.

[01:18:36.28] And my question was really just does this added delta that the elite universities see at the left hand side of that chart result from resources that they have to do a certain kind of training? Is it because the wealthier schools are establishing networks of wealthier possibilities for the students that graduate and so they're attached to those networks after they graduate? Or is it because the elite universities are selecting for qualities in the first place for those relatively few numbers of people from the lower quintile, and therefore they come with other cultural capital that allows them to move through the system in a different way?

[01:19:25.88] To Gary, I am entirely sympathetic to the idea of wanting to avoid the split the difference compromise. I've faced similar sorts of problems in a very different way in research that I've done on national security secrecy where government secrecy is actually easier to access than industry. And indeed, that's why many of those three letter agencies outsource some of their work to industry because it's free from the prying eye of FOIA or other forms of inquiry. So I thought that was very interesting.

[01:20:07.67] So the stick here in establishing a third party adjudicator, as I understand it, Gary, from your intervention is that if Facebook, for instance, or another industry were to renege, then it becomes possible to announce that this had been reneged upon, and that that would presumably be embarrassing and forms a kind of countervailing force.

[01:20:31.86] So first question is that really a stick, do they really worry about being embarrassed in this way? But more to the point-- and for me, much more important is what's the carrot? I mean, in my research on, say, nuclear weapons history or its development or other forms of national security issues related to the physical sciences, the government agency will have a balancing act to perform-- is releasing this information, which they don't really want to do-- I mean, their potential embarrassment, all the same thing reasons why industry might not want to release it.

[01:21:09.59] On the positive side, they want to conform and please Congress and conform to the FOIA request structures and so on. Well, my experience with industry is there's only one hand here. It's like, does this help me? No. End of discussion. So my question is why-- I still, at the end of your talk, didn't quite understand why an industry would just not say no. And you would say, well, I have this compromise, and they'll say, no. I mean, I don't understand why that discussion went on when you were packing your bags in your hotel room, why they didn't persist. So that was my question.

[01:21:47.54] And secondly, on censorship, again, this seems to me very recognizable from a different domain where-- once I was meeting with somebody, who had been a leader in the Tiananmen Square, in a public cafe in Beijing. And this person was saying to me-- making various criticisms. And I said well, aren't you worried? I mean, what about-- the waiter speaks English. I saw him speaking English not long ago. And this person said, I'm not worried at all. That's not the problem.

[01:22:24.72] Whereas as when I had spent-- I, as a student, spent quite a bit of time in Berlin. And during East Berlin days, if you said something in a cafe, and a waiter understood what you were saying, you were in deep trouble. So that's a radical change in the kind of Soviet block notion of censorship and intelligence gathering and surveillance than in the Chinese government treatment.

[01:22:49.86] So my question is has the Chinese government correctly assessed that the real challenge is-- the real threat is non-state structures, or do you think it's just a choice that could have gone the other way had they chosen a East Berlin, East Germany style form of surveillance and control?

[01:23:15.15] My question to Mario-- my first question to Mario is-- I found that the analysis absolutely riveting. Your qualitative analysis explores the widespread and selective predatory structures of these alternative financial institutions, and they complement your quantitative analysis rather hand in glove. Can you say a little bit more about what's going on beyond this? I know you are still intending to do the kind of micro ethnological work of interviewing bank managers and so on.

[01:23:48.17] But I wonder whether there are other more external structural aspects. Like are the AFIs less scrutinized when they're in minority communities than they would be if they were in a majority community? Are the enforcement different if they are using unscrupulous or marginally legal techniques of loans? Would that be less enforced there than, say, in other neighborhood? So do they have less information inside these minority communities than other communities might have? Are they cut off from the availability of knowing what might be legal, borderline legal or not?

[01:24:33.78] And then finally, are these AFIs really free floating micro organizations or do they belong to, are they broader structures? Are there big companies that run these AFIs in franchise like structures? And if they do, do these large companies themselves have connections to the-- I mean, whether the banking or other aspects of the financial structures? The picture that one has

naively, which I suspect may be wrong, is that these are small-scale, mom-and-pop loan organizations. But are they? How do they really work in their fiscal wiring diagrams?

[01:25:22.05] So to all of you, thank you for these absolutely terrific papers. I know these are only glimpses at your much larger and longer research programs. But it's been a great pleasure to have gotten at least a glimpse inside what you're doing. Thank you.

[01:25:37.46] ELIZABETH PHELPS: Thank you, Dr. Galison. So I'm going to kick it off by letting the speakers address the questions posed by Dr. Galison. So Dr. Chetty, we'll start with you.

[01:25:46.99] RAJ CHETTY: Sure, yeah. Well, thank you so much, Professor Galison, for the very thoughtful remarks. And thank you, everyone, for the very complementary presentations. So I'll just pick up on the two points you raised on why the map looks more blue in the rural Midwest and the point about the elite colleges.

[01:26:02.50] So on the first one, I, too, was struck with that. So just to remind everyone here, we're seeing the highest levels of upward mobility in much of the rural Midwest. From an economist perspective, that's often surprising because one of the hypotheses that economists tend to have in their minds is that maybe these differences we're seeing across places are due to differences in the types of jobs that are available or the industrial structure of an area.

[01:26:25.57] So for instance, the tech sector in the Bay Area, many think it might lead to pathways to upward mobility naturally and so forth. And so when you look at a place like North Dakota, South Dakota, Iowa, if you set aside the fracking kind of boom that did have an impact on North Dakota, in general, those are not the places where you would think of as having vibrant labor markets.

[01:26:46.59] And so what's striking to me about those places is what is happening is that kids who grow up in a place like rural Iowa, they're moving to a place like Chicago or New York City when we're measuring their incomes as adults. And that, on the theme of this event, is one of the powers of big data, administrative data. You can follow people no matter where they go in the US.

[01:27:08.58] And so what's striking about a place like Iowa is despite sort of a brain drain phenomenon where a lot of the most successful folks are moving out to a different place, we still find evidence, both in our data and to the extent there's historical data, generation after generation, some of those areas have incredibly high rates of upward mobility.

[01:27:28.30] And so the question, of course, is why. As you noted, there isn't a clear alignment with politics. Lots of people have taken our data and tried to correlate in various ways with measures of partisan support one way or the other, and you don't really find a super systematic pattern. What I will say in terms of what seems to explain it is if you think about the four main factors I put up that seem to predict these differences in upward mobility across areas.

[01:27:52.08] The Midwest, the rural Midwest in particular, has a lot of those features of having relatively high rates of social capital, high levels of integration, Iowa is thought to have some of the best schools, historically. So it seems to line up with those patterns, and it's that kind of data that's driving those correlations. But I think the fact that we're having this conversation just more broadly illustrates the power of these data to be able to look at an example like that and exchange ideas about what might be going on.

[01:28:20.79] Briefly, let me talk about a similar point in the context of colleges that you noted. As I think you exactly rightly pointed out, traditionally in economics and education and related disciplines, there's a literature on the return to education, one more year of education. But of course, just thinking of education as one more year of education and not thinking about what is going on in that year doesn't make a whole lot of sense when you dig into it.

[01:28:44.46] And so I didn't spend time on this in the presentation here. I didn't actually show the slide. But the slide that Professor Galison was referring to is one where we put out statistics for every college in America on colleges' contributions to upward mobility, analogous to neighborhoods. And there, you have to think about, as you were pointing out, both how well low income kids at a place like Harvard do. The answer is extremely well.

[01:29:09.27] I think a very encouraging fact is kids from low income families and high income families who attend Harvard have very similar outcomes after college and have, as you might expect, tremendous prospects. But the challenge is that a lot of our nation's best universities have very small numbers of low income kids despite the fact that we have spent an enormous amount of effort expanding financial aid, doing outreach, lots of things over the past 10 years to try to make our institutions more accessible.

[01:29:36.28] And so I think you were asking, why is it that we're seeing these great outcomes for lower income kids at places like this. And that is a question that we and others are spending time studying at the moment. It's basically a question of causality. Is it the case that if we were to take a child and give that child a seat at Harvard versus whatever other university they would have attended that their life trajectory would be different. Or is a lot of this about selection and who's getting in and-- these are talented kids who would have done well anyway.

[01:30:05.07] And I don't think we have a definitive sense yet, but my instinct is that there is a quite significant causal effect. Some of it might be related to the networks you get connected to. I hope some of it is due to what you learn from professors at Harvard. But whatever it is, there does seem to be a treatment effect. And that's why I think it's very important for us at institutions like this to think hard about what the barriers are, that despite the fact that costs are now very low in a financial sense to attend Harvard, we still don't have that many low income kids on campus. Thank you.

[01:30:41.81] ELIZABETH PHELPS: Thank you. Dr. King, do you want to address--

[01:30:45.57] GARY KING: Sure. Thanks, Elizabeth. And thanks, Peter, for the great comments. Just to combine them, your questions-- or one way to combine them is why do why

do these companies listen to you and why does China not just stop you, right? Roughly speaking. And I think the answer-- we certainly ask those questions, especially the second case.

[01:31:08.03] But I think the answer really is that they're just more sophisticated than it seems at first glance. There's more to a company or a government than it seems at first glance. Our first analysis of a company is, well, what's a company? They are legally required to maximize shareholder value. That's it. If they do anything else, they get fired. On the other hand, they also have to hire employees, which are basically our former students. And employees like very much to be creating some public good.

[01:31:39.40] One of the early scandals before Cambridge Analytica that Facebook got into was from data scientists at Facebook publishing things in academic work. And they saw what trouble they got into. And they said-- they literally said to their data scientists, OK, no more publishing. What is this publishing thing anyway? It doesn't benefit us. We're not making any money from it. No more publishing. And the data scientists said, oh, OK. We'll just all leave. And then, so that just wasn't an option, right?

[01:32:09.97] So what's in the long term interest, what's in the interests of their employees, what's in the interest of their stockholders, what's in the interest of people who may want to join the platform. It's not only Facebook. Every company has very similar kinds of issues. A company is really a collection-- is really a "they" rather than an "it". The 48 media organizations where we wanted to experiment on them-- they really wanted to know whether they were having an impact. They've devoted their life's work to having an impact. If they really weren't having an impact, they really wanted to know. And so we figured out those things.

[01:32:47.38] Similarly with censorship in China. Sure, we absolutely paid attention. And my graduate students and I conducted an entire additional study that we did not publish. The purpose of which was just solely to figure out how to keep my students and staff and volunteers-- lots of people volunteered for this project-- safe. At one point, we employed 80% of Harvard's Chinese speaking undergraduates. So we absolutely had to keep everybody safe.

[01:33:17.50] But at the end of the day, we figured out the different kinds of interests of the different people in China. And in fact, I wrote and published three articles about this. And each one, on a separate trip to China, I gave a full presentation of our results. And we just figured out how you do it, how you speak to people. And so far anyway, it was OK. And everybody seems to be OK. [LAUGHS]

[01:33:45.21] ELIZABETH PHELPS: Dr. Small, you want to address the question--

[01:33:47.85] MARIO SMALL: Yeah. No, thank you. Thanks for those comments. My mind is running from all of the ideas I just got. I'll just answer them very quickly. The second one is easy. The question is whether the payday lenders and check cashing places are all mom and pops, or they're larger organizations. Within our first cut on this, it's actually more heterogeneity than we thought. So it's certainly not all mom and pop shops. But they're actually quite a few mid-sized entities at least on the first cut as indicated by their names.

[01:34:19.65] So we know from their names or we infer from their names how big they are. And it seems like there's actually quite a bit of a range. There's a possibility that a conglomerate owns a whole bunch of different entities under different names. And we haven't explored that yet. But I think you're completely right. That's kind of the next step in the systemic part of this.

[01:34:37.98] The other part of the question, again going back to the system, is what is happening on a kind of a larger level affecting this. And one of the things I've started learning about this-- and I should say, some of you know that I'm actually a neighborhood researcher, not a banking researcher. So everything about this new-- is that banks play a bigger role on people's decision to turn to AFIs than we actually even realized.

[01:34:59.52] So we sort of conceived the choice set as the nearness of one versus the nearness of the other as kind of a simple, elegant way, we thought, to approach the question. It turns out there's actually more reality to that than we figured. There's a nice paper by Emily Williams. She's an economist at the Harvard Business School. And she's looked at a practice that banks have had in which they stack debt from largest to smallest.

[01:35:23.89] So let's say I have $500 in my account. And I go to the supermarket in the morning, and I debit $100. And then I wait until the end of the day to write my check for $1,000 for my rent because I know that tomorrow, I'm going to get my paycheck. And that way, I don't carry an overdraft fees. Well, it turns out the banks know that rather than charging the amount as they come, if they stack them, if they put the $1,000 check, or whatever it is for the rent check, for the highest one, and the $100 one later, they're more likely to get an overdraft fee, and they're likely to get more overdraft fees from that same person. And so they'll do this on purpose.

[01:36:08.16] And there have been some class action suits in multiple states. And in different states at different times, this practice has been outlawed. And so you can imagine how you can use that variation over time, across space, to see the impact. It's a beautiful paper. What she's found is that immediately after this practice has been banned in a state, on the part of banks, the proportion of people going to AFIs and AFI business goes down, meaning a lot of people are essentially going to AFIs because they didn't want to get those overdraft fees from the banks.

[01:36:40.12] So that tells me, going back to your question, that the large, systemic picture has to be not just about the decisions of AFIs and how they're lobbying, but more broadly how people are acting financially-- [INAUDIBLE] sort of banking services from all of the multiple options they have. And that's actually something that I haven't done yet. But again, I think it will require some combination of talking to politicians. We're not going to be able get the answers just by downloading it from a data set from somewhere. We're going to have to talk to people, we're going to have to get a sense into the law, we're going to have to get a sense of these lawsuits and sort of put together the system from multiple methods.

[01:37:16.44] ELIZABETH PHELPS: Thank you. All right, so I'm going to go now to some of your questions. And I'll combine a few for each of you. I'm going to start with some questions for Dr. Chetty. So one was do you believe the strategy would effectively work in developing countries such as Brazil and India? And I'm assuming that's a strategy of assisting people with housing. The other one was when you mentioned the difficulty-- and that was from Carlos Millo.

[01:37:41.90] And ?Hafong? Lee asked the question, when you mentioned the difficulty of measuring social capital, what did you do that-- proxies? Could you share some ideas of measuring the social capital?

[01:37:53.96] RAJ CHETTY: Yeah, both great questions. So on the developing country question, we just speak to that a bit more broadly beyond the housing issues. So we actually have some graduate students, former graduate students at Harvard-- Paul Novosad and Sam Asher who are now at Dartmouth and Johns Hopkins-- who are leading a great effort, as an example, in India to do a lot of very similar work on the geography of mobility in India.

[01:38:15.65] Now the natural challenge, of course, is in developing countries, you don't have the type of tax records, census data that you have in the US that pretty much covers the entire population. You have to be more creative in terms of putting together other types of data sources from mobile phones to certain information from the census to other types of private sector data sources and so forth. And they and others are making incredible efforts in that space. And I think there's tremendous potential to do this type of work in developing countries. And a lot of my colleagues in the economics department are trying to think in that space as well, those focused on developing countries.

[01:38:50.67] So I think the big picture answer is absolutely yes. But the hurdles are even higher at this point than in the US. Remember, even in the United States or in European countries, 10 years ago, we were not at a point of being able to do this work. But given the way technology is changing, as Gary was saying, everything's leaving a digital trail. I'm quite optimistic that we'll be able to investigate analogous issues in developing countries.

[01:39:15.26] And some ways, I think the stakes are even bigger. What I see from the studies that have been done in developing countries is things like the neighborhood in which you grow up, moving to a different place-- they're even more dramatic consequences than what we see in the US because of the tremendous variance and opportunities in some of those areas.

[01:39:33.23] On the second question on social capital-- great question. As many of you might know, social capital is a term that has been discussed in sociology and other fields for many decades. People here have written extensively about it. Our colleague, Bob Putnam, of course popularized the concept greatly in the public discussion. But pinning down exactly what it means, how we measure it in a precise way, and so forth has been difficult.

[01:39:58.39] So people use proxies like participation in civic organizations or religious organizations or Bob Putnam's famous book, Bowling Alone, the extent to which people are bowling together. And it turns out all of those proxies, as rough as they are, do still correlate very strongly with our measures of upward mobility. But we wanted to go further. And actually, what we're doing now-- one of our main projects at the moment at Opportunity Insights is trying to measure social capital much more systematically using social network data, by looking at who's friends with whom and so forth.

[01:40:33.99] Turning to private sector data sources very much along the lines of what Gary was discussing. And our hope is when we put out those measures, we will have much more precision

in saying, it's exactly this type of social capital that seems to matter for economic mobility or for health and have a more refined discussion about how you get more of that type of social capital or what the determinants are, why it's less available in certain areas and so forth.

[01:40:59.09] ELIZABETH PHELPS: Great, thank you. There's several questions for Dr. King. So I'm going to just throw three at you because they're somewhat related. So first, Katie Giles ask the question-- just to touch on it. You might have missed it. But was Facebook funding the research? Or was it an outside sponsor? And if they were funding it, how did that impact the relationship with the study and communications about it?

[01:41:22.01] And then Carlos Millo ask the question-- sort of he says, big data is really the new oil, right? So companies like Facebook, almost impossible to beat, as they have a huge amount of personal data to be able to better understand their users. And he asks the question, what have users start to open online their personal data but protected by differential privacy to any other companies in order to break this data monopoly created by big tech companies? Do you think such initiative would be possible?

[01:41:50.76] And finally, Scott Wallander asked, are there roles that industry partners can play in partnerships with you all that go beyond simply data providers? And this, I think, actually is something that other people could speak to as well.

[01:42:02.37] GARY KING: Great, thanks. Thanks for the questions. To Katie, we took no funding from Facebook to do the study that I described. The funding came from eight ideologically diverse foundations that contributed quite a lot of money. We went one more step because we were a little concerned about individual researchers taking money from a foundation that had a particular ideological slant. And so we convinced the foundations for the first time to pool all their money, put them in one pot. And we had the SSRC as a nonpartisan financial agent watching the money basically.

[01:42:38.99] And then the decisions were not made by Facebook about who would get the money, they were not made by the SSRC, they were not made by the foundations in particular. They were made by Social Science One, this organization that was composed solely of academics. So that's how we did that.

[01:42:58.10] For Carlos, yes, absolutely, differential privacy could be used to solve the problem of individual ownership, right? If you have data, you may be torn between contributing your data to the public good, but violating your privacy. But with differential privacy, if you do it the right way, it is possible so that you could contribute data without your privacy being violated at all.

[01:43:26.30] The mathematical principle behind differential privacy is if I have a data set, and I ask you for my data, and I run an analysis on my data set without your data, I should get essentially the same results as if I run my analysis with your data contributed to my data set-- no, your personal information contributed to my data set.

[01:43:45.74] So that means that it ought to be possible to get public good from the data and to convince people to contribute data because if you're going to get the same answer, essentially the

same answer either way, that means your privacy can't be violated, but we might actually learn something to help not only you, but everybody else.

[01:44:04.26] And finally, for Scott-- absolutely, the role of industry is not merely to just give us stuff, because then, as Peter mentioned, we wouldn't get any stuff. It has to benefit them as well. So whether it benefits them by looking good, that's OK with us, whether it benefits them because they actually get to look good in front of their employees and make their employees' jobs feel more important because they're actually contributing to public good, or because some of these companies have huge highly intrusive sources of data on billions of people, and if companies that are essentially monopolies don't eventually figure out how to provide public good, the government takes away that monopoly.

[01:44:53.84] AT&T figured out that with Bell Labs, IBM figured out that with IBM Research, Microsoft figured that out with Microsoft Research. We think the rest of the companies will figure that out as well. And we hope to be there to help them not only figure it out, but provide data for public good.

[01:45:14.02] ELIZABETH PHELPS: Thank you. It's just a follow up. This is for Raj. Are there constraints on the government data that people have-- the researchers have to pay attention to get access?

[01:45:25.19] RAJ CHETTY: Yeah, absolutely. And so I can speak a bit more to that. So the way it works is the government now has sort of a call for proposals. Part of this came out of some of this early work where there was something passed called the Evidence Act that now tries to make more data from the government available to researchers. But of course, that data by law is never going to be publicly available, right? We wouldn't want everyone's tax return information to be publicly available, obviously.

[01:45:51.29] And so there's a long process through which one gets access to these data. It's often a time consuming process. So one of the challenges our graduate students face in our field is a big hurdle cost in terms of spending a year or two getting access to these types of data sets. And I think as social scientists, we need to figure out structures that can make it easier, especially for junior scholars, to get going on this sort of research more quickly given how critical it is to do cutting edge work.

[01:46:20.09] But the formal structure is that you submit a proposal in one of these calls for proposals. If it's selected, then you go through various security clearance procedures. And then you're able to access the data on site essentially at government agencies-- the Treasury, the Census Bureau. All results are then reviewed for disclosure purposes, sometimes put through a differential privacy protocol along the lines of what we were discussing earlier and then can be released to the public. So it's quite tightly controlled. It's not as easy as downloading a survey data set and working with it. But I think we're making strides in making it more accessible.

[01:47:00.04] ELIZABETH PHELPS: Great. Thank you. So Dr. Small, so for Sarah Avila said, amazing research. Thank you for sharing. I'm cracking up with Mario Small on his comment

about not having enough time to make the table pretty. We can all relate. The best is the unimpressed expression of the other presenters. That's for everybody.

[01:47:18.21] I had a question for you. So you talked about different data sources. You said Google Map is becoming better than all the other data sources used. I'm just wondering why. What is it about maps relative to other data sources you use?

[01:47:30.18] MARIO SMALL: That's a great question. And actually, what's interesting about this is if you asked me this question about 15 years ago, I would say that Google Maps database is the worst data set you could use maybe, yeah, 15, 18 years ago. And we have some evidence of that.

[01:47:44.40] So Stacey Lindow, back in the early 2000s-- she's a researcher at the University of Chicago-- sent a team of researchers, including residents of the South side of Chicago, out to the South side of Chicago to do an asset mapping. They literally recorded, went block by block and recorded every single establishments they saw. And then they compared the results to the Google results. And there were something close to a 50% mismatch in both directions. So about half of what Google said was there was not actually there. And about half of what they recorded was not on Google. It was terrible. Just absolutely terrible. A couple of things happened since then.

[01:48:22.30] So first-- so Google, the way it works now and part of the reason why it's so good for this particular question, is first, they take all of the federally available administrative data that you could use. And those data are pretty good for-- they are great for certain things like, for example, the project that Raj is doing, there is no better source of tax data than the federal government. But for other things, as I'll show you, they're actually not necessarily optimal. So they take, for example, the records of all establishments they have.

[01:48:51.09] Now for many establishments, they'll have records that are updated once a year, some once every three years, some even longer than that. As we know, for a lot of establishments, there's turnover over the course of the year. And so a lot of the data are going to be missing those things. Second, a lot of establishments don't actually do the grade up-- your business goes out of business, you just shut down. You don't necessarily keep up with all of the records expected to the federal government.

[01:49:15.36] Google, what it does is-- the second thing it does is you've seen the Google Maps car, the Street View. You know, the car that goes out with a camera. So what you may not know is that the Street View car has gone through multiple streets in all of the cities multiple times over the years. So for example, I typed in my address, and I can go back to, I think, 2005 or something like that. There are six or seven different images. I know my old car, I can see the trees get bigger and so on in the place I currently live.

[01:49:48.21] And so what Google does is it takes all of that image data. Remember, the image can look up and can take snapshots of awnings and signs. And because they have petabytes and petabytes of data and extremely powerful computers, they use essentially those data to improve the data that they got from the federal government. So for example, they'll know at this point

with extreme accuracy whether BBQ's check cashing location is an actual check casher as opposed to a barbecue restaurant, right, which was not the case many years ago.

[01:50:21.30] And the last thing they do is they gameify their data access. So they have something called-- there's two things they have-- sort of two aspects of it. One is sort of if you're a business, there's something called-- I think it's actually called Google Business-- where you can go up and you can see how Google records your data and improve it. And so you can say, hey, you guys made a mistake in these ways.

[01:50:42.48] Now if you do that, that doesn't mean that the data are ultimately updated. What happens is now the second thing, which is their Google Local Guides come into play. Google Local Guides are just people who go on to Google and get points, literally stars, for improving the data for saying, you guys are wrong, that restaurant closed last week or taking a photo of it and uploading it or reviewing it or a whole bunch of things. And you get stars for doing-- the more you do, the more stars you get. It's like Wikipedia or something.

[01:51:15.76] But then also-- it's not just that. Also, if you do really, really well, you get early access to a whole bunch of Google products when they come in. So you see certain products before other people do. So a lot of people spend a lot of-- they have essentially an army of people taking the data that were first downloaded from private-- and actually not just public sources, but also private ones, then refined with the year's worth of Google Street View data and then improved by the business owners and then of course refined by the people themselves.

[01:51:45.55] And so what you have is a constantly updated iterative process where the data are actually much closer than the lived experience on the person on the ground because it's what the cars saw improving on what we downloaded from the federal government corrected by what people are saying than you would if you just downloaded the data from the Federal government because you would have essentially administrative data that are not necessarily as close to the question.

[01:52:08.92] So I actually think for the question we're trying to get at-- it's kind of funny because I literally avoided Google for many years because I thought they were so bad. In fact, the first version of this project, we use Bing data from Microsoft Bing. You might not know that Microsoft also has a search engine. It's called Bing, and it's not a successful. But they also have a database.

[01:52:28.56] And we first did, uh, New York, Chicago-- actually, we just did Manhattan, Chicago and one or two other cities with Bing. And then when we heard about Google getting better, we actually, for about five or six cities, looked at the Google data and looked at the Bing data. And actually, Google had way more establishments than Bing across the board. And that is part of why we're doing Boston.

[01:52:53.22] Then we went to Boston, and we looked at the neighborhoods that we knew in Boston and the establishments that were in the Google data but not in the Bing data. And we did not find any errors in the Google data. Actually, Bing was just missing a whole bunch of things

because it didn't have all of these elements. And so, yeah, at this point, I think Google is as good as it's going to get for this particular question.

[01:53:13.96] ELIZABETH PHELPS: Awesome. Thank you. I had no idea. OK, I'm going to take the chair's prerogative and just throw out last question because we only got about six minutes left. Dr. King, you cracked me up in the beginning of your talk when you talked about how social scientists didn't project sort of interesting or big discoveries in the 20 years coming up.

[01:53:38.73] So I'm going to give the speakers an opportunity to correct that error and ask you, in the next 20 years, each of you to comment on what kinds of developments do you see coming in social sciences due to the use of big data sort of increasing. So Dr. Chetty, I know this is like-- I hate being asked these big picture questions. But here I go.

[01:54:03.09] RAJ CHETTY: I'm worried that this is being recorded, in 20 years from now, we're going to have a session looking back on what I say here. But I would say one broad thing, and I think it's a theme that's been echoed throughout this event today, which is I think of this data sort of permitting a personalization of social science. So traditionally, in economics and the other social sciences, I think there is a tendency to sort of one size fits all answers.

[01:54:30.02] So take an example of a question I'm often asked-- are charter schools good or bad? Are they better than public schools? Well, turns out the answer is it really depends upon which specific charter school you're talking about. What the potential alternative public school is and the specifics matter, right? So if I can make an analogy to medicine, if you were trying to give someone a treatment, the first thing you would try to do is have a diagnosis of exactly what is ailing the patient.

[01:54:53.82] And similarly, I think in the context of the economy or society, the types of issues that might be of limiting opportunity in one city or one region of the country might be different from the types of things limiting opportunity in another neighborhood, another place down the road as we were seeing in some of those maps I was showing.

[01:55:10.59] And so my hope is we'll be able to get to a version of social science 20 years from now where the response in terms of scientific understanding, in terms of policy, is not this is the general thing we figured out that one needs to do, but here's exactly what seems to be going wrong in this situation, here's a treatment that has shown some record of success. We can do further pilots and that's how I think we can have a more scientific advance in our field. And I hope as a result, our stock of knowledge and ability to provide informed policy advice will be much greater 20 years from now.

[01:55:47.29] ELIZABETH PHELPS: So better policy advice across the board, right? OK. Dr. King, do you want to--

[01:55:53.94] GARY KING: I think more studies like Raj and Mario, we will see them in the next 25 years. And if 25 years from now, we are all here and there wasn't spectacular progress

relative to now, then I'll really be pissed off, OK? So I think, actually this time, plus or minus several decades, is equivalent to when they first handed out microscopes to microbiologists.

[01:56:23.31] I mean, it was very nice of us to be studying things 100 years ago. But if you don't have the information about the subject of your study, you're not going to make that much progress. We, for the first time, actually have enough information to be able to solve or ameliorate some of the major challenges that affect you in society. And I think we're beginning to really tackle them one by one. And we will absolutely change public policy.

[01:56:52.32] In the last decade, governments are actually quite good now about letting you evaluate their public policies. Holy cow! Who would think that a politician would allow that? In fact, in many governments, they require it. All the public health interventions that we've seen in the last year up to but not including the vaccine were basically social science interventions. And trying to get people to take the vaccine-- that's another behavioral social science issue.

[01:57:22.18] So I think, actually, the future is incredibly bright. It's really exciting. I can't wait to see all the kinds of things that we're going to learn.

[01:57:29.31] ELIZABETH PHELPS: OK. Well, Dr. Small, you get the last word for today.

[01:57:32.55] MARIO SMALL: Yeah, so I actually agree with both of those statements but I'd add a qualifier to both. I think the future is very, very bright with respect to what we can learn. And I think our abilities are going to-- especially at the level of granularity, and you used the word Raj that I would have used-- personalization is going to be extraordinary.

[01:57:51.04] But the qualifier I'll add is that the extent to which this translates into improvement will depend on the extent to which the capacity and the power of the government to do something about it, challenges and overtakes the capacity of the private sector and private companies who have different interests to do something about it. I'm thinking of this in the context of the work I'm doing on financial decision-making.

[01:58:14.79] I mean, a lot of what we end up finding, if it goes where we go, will give a lot of power to a lot of state and federal governments to regulate AFIs, and actually banks, more effectively. But the banks and AFIs will have those data too. And they'll have a very, very strong incentive to avoid all of that. And so there's a bit of an arms race component in my perspective to this.

[01:58:36.18] And so my optimism is about the capacity. My caution is about whose capacity. So I'm hoping that for every one of our students who goes into the private sector to make a ton of money through big data, there's at least another one who goes out there and tries to do some good.

[01:58:51.63] ELIZABETH PHELPS: OK, well, thank you so much. Dr. Bobo, do you want to say anything to conclude?

[01:58:55.98] LARRY BOBO: Fabulous. Thank you, Liz. This has just been fabulous. And thank you for curating the questions and for your own very provocative question. If you hadn't asked it, I was going to ask it. So I'm glad you put it on the table for each of them. And I share that optimism about the future as well because the level of data we have about repeated patterns of behavior of particular individuals that we can now link over time and connect to lots of other sources is truly extraordinary compared to where we were a hundred years ago, let's say. And that just opens the door to remarkable developments.

[01:59:36.11] I feel greatly indebted to you, Liz, to Raj, to Gary, to Mario, and especially to Peter for the kind of perspective he brought to thinking through these issues. He had to leave early because of a conflicting meeting. But this has been a great start to these Dean's Symposiums in Social Science Innovations. I think the next one, I'm going to work very hard to make on the subject of fighting truth decay. And we will have a number of our colleagues engaging that question, I hope.

[02:00:06.95] But thank you all very much. Thanks to our presenters. Thanks to the audience and their questions. I really appreciate all of your work. And Jennifer Shephard, in particular, who worked very, very hard to pull this together, along with others in the office who were here-- Bev Beatty, Cherisse Haakonsen, Addie Connelly, Kirsten Colton, and Richa Mishra, thank you all. Have a great night, everybody. We will--

[02:00:34.85] RAJ CHETTY: Thank you.

[02:00:35.75] ELIZABETH PHELPS: All right, bye-bye.

[02:00:37.86] GARY KING: Thanks, take care.