June_19_409_002

We're gonna go ahead and get started since things, as we know, are very back -to -back -to -back here. I am Andrew Schroeder. I'm the Vice President for Research and Analysis at Direct Relief, which is a global humanitarian nonprofit, and I'm joined by my colleagues, Takahiro Yabe from New York University, Professor Sakimoto from University of Tokyo, and Alex Pompey, who's with the Data for Good program at META.

Alex will be joining us online, so hopefully that works when the time comes. We are gonna be talking about human mobility data and disasters in recent Asian emergencies. Every disaster, as far as we know, has some kind of mobility component, and in recent years, we've had significant advances in the ability to use mobile phone data, telemetry data, automobile data, a range of connected devices, social media applications,

that are able to allow us to understand and analyze mobility patterns in near real time, and so we're going to go through a few different dimensions of that, and then hopefully have some Q &A. We're gonna start with Takahiro Yabe from New York University, and then we'll cut to Alex online.

Takahiro, do you wanna come up? Sure.

Oh, perfect, okay. Thank you, Andrew. Hello, everybody. I'm sure you just had the coffee break, so thank you for coming in early after the coffee break. So I'm an assistant professor at New York University.

Full disclosure, I was a student of a first as a Sekimoto seven years ago. So we're kind of, some of the work that I'll be presenting are joint work with a first as a Sekimoto. So I think many of you have seen, oops, sorry.

Many of you have seen this kind of video, if it works. Thank you, of human mobility traces. And Andrew has been talking about these kinds of data, which has become more and more prevalent. Maybe, can you play it back again, please?

Thank you. So this is during the Great East Japan earthquake back in 2011. This is joint work with a Sekimoto. If you can go back, the earthquake hits at 2 .46. It's already 8 p .m. Thank you. Yeah, so now this is kind of the usual dynamics in cities.

And then at 2 .46, you can see the city kind of come to a halt and people are starting to walk back to their original locations or home locations without urban transportation. And this mobility data captures around 3 to 10% of the entire population.

And it's quite easy to detect a home, workplace, where they visit, things like that, especially during disaster scenarios. Okay, so one of the first things that we did back in 2016, when we had the Kumamoto earthquake in Japan, was to estimate where are people evacuating?

And there was a big issue at that time because you think about the current situation. If an earthquake comes right now, do you know where you should evacuate to? That's a hard question if you're just put on the spot of evacuation.

And people tended to go to familiar places like shopping malls, parking lots. And it was really hard for government officials to track down where are people going to. So we use mobile phone data together with Yahoo Japan, one of the biggest internet providers in Japan, and provide these kind of maps which were delivered to city governments.

From then on, we scaled up our efforts to the global kind of stage with Nick Jones at GFDRR, the World Bank. We scaled this up and we tried to do a variety of analysis using this type of data for, for example, cyclone impacts on transportation in India, commute analysis in Kathmandu, Nepal for city planning, and also land use analysis in San Juan, Costa Rica.

And one of the things that we've been pushing forward a bit more is to try to go more, not just the disaster response phase, but to go beyond the disaster response and go look at the more longer term disaster recovery phase.

So using this data at a macro scale, we can understand among the people who are, how many, first of all, how many people are displaced from their original homes and how, at what pace do they come back to their original locations?

And by looking at this curve, you can see the displacement rate go down, but even after five months from the hurricane in Puerto Rico, we see that 30% of the people are being displaced. And this is useful information for disaster response and disaster recovery.

Because we have, because of this mobile phone data, we can scale this up to different countries and different disasters and try to compare what are the generalizable messages that we can take home. So we compared different hurricanes and different disasters from the US and Japan, and found some universal characteristics of how people start returning to their home.

So if a new disaster happens, we can quite confidently say, probably after 50, 60 days, people would start coming back at this exponential decay manner. So these are some generalizable insights that in the academic field we're working on, using data from multiple disaster cases.

Another use case is to understand the impacts on businesses, so not just population displacement, but by looking at how many people visit this Walmart on this particular day and by tracking that and building models that predict what if the disaster didn't happen and taking the difference between the predicted and the actual data, we can try to estimate the causal impact of the disaster on different businesses.

So this is for one Walmart in Puerto Rico. You can see the big dip after the disaster. Also, you can also see the couple of days pre -disaster increase just before the disaster as well, and we can scale this up to many, many places or many, many stores or businesses within the island.

So this is what we did, and you can take a global approach and say, what are the industry sectors or regions that had a significant loss or gain in terms of foot traffic? Now this isn't completely one -to -one with the actual revenue, but this serves as a good proxy of how many businesses are starting to recover after a disaster.

And at the end of the project, together with the World Bank, we developed an open source toolkit to kind of redo all of the analysis that I just showed you, basically, if you have mobile phone data at your disposal.

So if you're familiar with Python, please go ahead and do pip install mobilekit. It's really easy to download and install, and if you're a data scientist, it's probably kind of very self - explanatory on what you can do and what you can't do.

It has a full tutorial, and we put a lot of effort in making this more accessible. So if you're interested, please check it out. And final slide. More recently, we've been looking at some of the issues that mobile phone data has.

So mobile phone data, of course, it's amazing. You can do all these analysis, but it's proprietary, and it sometimes has the risk of privacy evasion. So leveraging some of the AI -based and machine learning models, we're trying to generate models that can generate the synthetic mobility data, which protect the privacy of the individuals, but also are useful in terms of trying to predict what would happen in a disaster scenario that maybe has not happened before.

And some of the neat things that I won't go into detail, but that we can do using these AI models, is that you can try to predict, using one city's data, what would happen in another city if a disaster strikes.

So some of these counterfactuals that they were presenting in the morning session, the counterfactual simulations, what if an earthquake happened in Kyoto? Can we learn from Kumamoto? Those kind of questions we're tackling to try to answer at New York University.

So those are some of the things that we're doing. Some open questions for the panel. How can we scale this and further democratize or make it available the use of mobility data for disaster response?

Can LLMs or AI help us achieve that goal by narrowing the gap between data scientists and practitioners? Again, balancing the utility of data versus the unintended risks and consequences that this potentially harmful data could have.

And we're eyeing towards a more, and these kind of research has been supported by private companies providing large -scale data to academics in a top -down manner so far. But can we look towards more community -based, bottom -up approaches together with community partners, rather than that top -down approach?

So those are some of the things that I've been thinking about. And that's it. So Andrew, do you want to? Yes. We're going to switch now to balance. There we go.

Hello, this is Alex. Are you all able to hear me?

give you a thumbs up.

I don't have the ability to share my screen. It says only meeting organizers and presenters.

can share the screen. So if someone can grant me that access, my slides will be visible and you won't have to see my giant face at the front of the room.

sorry that I'm not joining you in person while we work on the screen share. My name is Alex. I am a research manager at Meta on a team called Data for Good. I'm joining you from my home near Detroit, Michigan.

It's some family items I had to attend to, so apologies that I'm not there in person. We'll get the slide share working here shortly and then

Alex, just one really quick thing. I'm trying to get you set back up with screen share, but it may take me a minute, so maybe I will go on to the next presentation and then come back to you. And while I'm presenting, I will simultaneously get your screen share hooked up.

Masterful multitasking. Thank you. Yes.

Oh wait, hold on. Okay, actually I'm going to let Alex go because Alex is the only one who's got it.

Thank you.

Actually, I think you're good to go.

Yeah, it has appeared.

As Taka mentioned, we have several dimensions by which we can consider a variety of mobility datasets. I'm going to speak to you about the dimension of privacy and more detail in reference to datasets that we release from Data for Good at Meta.

All of these are available for free. You can check them out at our website here or contact me directly. First, this is the full offering of the datasets that Data for Good at Meta currently offers. The orange stars are just to indicate those that are all derived from user mobility data.

This comes from the users who have opted in to sharing their location services data with the Facebook Blue app on their smartphone. You can see that it makes up over 50 percent of the various datasets are different aggregations of this same base source of data which is the user's GPS coordinates that they've opted in to share while they're using the Facebook products.

Let me explain to you the different privacy parameters that we can consider and you'll go through a thought exercise yourself on how we might consider the privacy versus utility trade -off in our newest mobility dataset.

Let's first start with this very simple version of the world that you see here. We're going to divide the Earth's surface into these grid cell tiles, which is very common in these types of GIS applications.

We don't have all six billion or so human beings on Earth that are using our app yet, we just have three users and we can make a distinction based on where they are using our app at a certain time of day to estimate their home tile.

Then maybe throughout the day, we get some more GPS coordinates that come from them that are shown by these stars. Now, it would be completely irresponsible from a privacy standpoint to just simply provide this level of data, this user -level GPS breadcrumb data of users.

We need to do better. How can we do that? Well, let's start in the not so recent past or in 2020. There is really this paradigm shift in the way mobility data with the scale at which mobility data was packaged up from meta and other tech companies in the private sector included.

Everyone had these different mobility datasets that were meant to aid with the pandemic response. Here, we're looking at an animation of a metric that we released as part of a dataset called movement range, that would estimate the percentage of Facebook users that we think are staying put just very near their home in a given day.

We can see that at the admin level two region for every single day and we can see Europe slowly freeze, come to a close starting in Italy, moving into France and Spain. Hopefully, these aren't so familiar.

We've moved out of this day and age. But the way that we package up these datasets, if you want the full details, you can read about them here. But really what we want to do is protect an individual's re -identification risk.

And what this is often referred to in a hand wavy way is differential privacy. But what that really means is that we're trying to quantify statistically what the risk of re -identifying an individual from an aggregate mobility dataset might be.

And this is a very, very important statistical parameter. You can choose the amount of noise that you would add to an aggregate sum such that it is proportional to the maximum amount of data that a single user might contribute to the aggregate in a single day.

And that is considered a very strong best practice because then if that user disappeared from the dataset, you wouldn't be sure if that was just the random perturbations in your noise or if that was that the user actually did disappear.

So you're not even certain if any single individual is present in the dataset. If you choose these parameters in such a way, and you also need to go to the actual lengths of publishing your methods so that people don't just have to trust you, they can actually audit it.

So that is where we were kind of in the pandemic sense. Let's take that view of all of the mobility data sets offered by Data for Good at Meta and place them in this very simplistic non -scientific thought model, where we are trying to balance at times competing elements, the utility of the data set, how useful it might be for a variety of purposes, including disaster response, and how much privacy protections are in the data set.

This was the view at the time of this movement range data set release. You can see that COVID -19 really kind of moved us off of this trend curve, this unscientific blue line, that many of our data sets sort of fell on.

When we pushed really strongly for extreme privacy protections, we really diminished the utility of the data set. This usually means that we have to remove a lot of regions because there are too many counts below a privacy threshold.

And if we do that, then there's many geographic areas that might not be reported in the data set. So that's kind of what we took the approach taken with the international travel patterns data set. And then kind of more utility driven data set applications, this displacement maps or population changes during crisis.

These natural disaster events kind of speak to the life and death nature of that particular crisis. We can make perhaps some more compromises on privacy put in place maybe more legal or programmatic protections.

We're not just releasing the data fully publicly, but maybe only the trusted humanitarian organizations. And this allows them to make use of that utility and still be confident in the privacy. So this blue curve is interesting that it emerges as a trend line.

COVID maybe moved us off of it a little bit where we have still strong privacy. We're taking advantage of the academic literature on differential privacy and mobility data, which is growing at a very strong rate.

And now we want to decide this new data set movement distribution, which I'll cover very briefly. In your head, you should think about where you think it falls in this curve, this trade off. Is it on the curve?

Is it off the curve? And we'll kind of do that exercise altogether in our heads and then maybe talk about it during the Q and A. So where will we put this movement distribution dot? So movement distribution is really a data set that will report on this research question.

How much do we estimate that people are leaving their homes? And it will do so by estimating the percentage of Facebook users who for a given day in a given say admin level two region, we're only seeing roughly in around a one kilometer circle around their home within 10 kilometers or maybe more than a hundred kilometers.

So we're gonna report for those four categories. So if you remember our simplistic view of the earth with our three users and their GPS coordinates, what we're going to do is take a random sample of those GPS coordinates.

We're not going to use them all. We're going to measure the distance between those random sampling of the GPS coordinates, the average distance between the home tile estimate and that random sampling.

And then we're going to sum up all the users that live in a given county, say San Francisco. So in our fake world here, the report in the movement distribution data set would appear as this. We have these four categories.

Those are the four radiuses of distance from the home location. And then we report the fraction of the users that we estimate to belong to each of those cohorts for a given day and a given county. And you'll notice because we're going to add noise to these sums when we're taking the average, we, this differential privacy means that the sum will not add up to exactly one.

And that is on purpose. We're specifically going to add noise such that the re - identification risk can be quantifiable and statistically zero for a given, roughly zero for a given individual. And this is what it looks like in the context of a natural disaster, like for five counties that were affected by wildfires in Northern California earlier this year.

For each one, we get these four trend lines that reports the percentage of the overall cohorts that has this general mobility. And this is what the data set looks, if you look like it in tabular form.

So, where might we place this movement distribution data set? And I think the three key components that come out of this when we consider the privacy angle in these mobility data sets is one that we want the utility to be high and so we want to be releasing it fully public without those legal or programmatic protections.

We want it to be available very widely, so maybe we post it on the UN's humanitarian data exchange. We also want to publish our differential privacy method, which describes in detail how we assess the sensitivity or the maximum contribution of a single user to a single segment of the data set.

And we want to publish that publicly so that it can be audited and considered against best practices at other companies or in academia. And also to kind of maximize the utility, I think this is a lesson we learned from COVID.

For many response, especially policymakers, interpretable at, say, a spreadsheet level, not having to go into a full GIS or maybe a Python pipeline. Those are very useful for a set number of data sets for a smaller audience.

But the maximum utility that can be derived for large scale crises mean that we need to do the simplification on our side. So I think those are the three main ingredients. You can be thinking in your head now how you might place the movement distribution data set on this graph.

You might ask me some questions in the Q &A if you want to discuss where you might place it. Here you can see a quick summary of the temporal and spatial resolutions and the sampling that we do across all of our main mobility data sets in the program.

So you can see the breadth at which we've chosen those parameters across a large tech company like Meta with many users. If you're interested in seeing the movement distribution data set, it's available on the humanitarian data exchange, as I mentioned, as a potential best practice if you scan the QR code here.

And I look forward to your questions and the rest of my panelists. And again, sorry that I'm unable to join you in person. Thanks so much.

All right, thanks, Alex. We'll bring you back on screen for the Q &A. All right, so I am going to talk about more of the implementation side of this from work that we've done through the crisis -ready partnership between Direct Relief and researchers at Harvard University through the Harvard Data Science Initiative and the School of Public Health to figure out some ways that we can better take data sets principally from Data for Good at Meta and mainstream them into Sudden Onset Disaster Response.

And this is a application called ReadyMapper. It's focused mainly on health resource allocation questions, but I'm not going to focus on that just on the mobility portion of it. This is Louisiana after a recent hurricane.

And so it's looking at the distribution of population and the change in population density for that population at time intervals determined by the data in eight -hour time blocks. And so we're going to then look at displacement patterns through there.

This work really emerged from the shock of the pandemic when we were doing sort of work with mobility data prior to the pandemic and when the demand spiked for mobility data during the early stages of the pandemic, a group of us got together.

And we're trying to figure out sort of what is the practical mechanism by which you would get this data into the hands of agencies. Public health departments throughout the world were being confronted with sort of proliferation of data.

They had too much to do normally and did not have the ability to kind of quickly absorb this. So we put a network together to pair up epidemiologists and researchers with people that were working in public health departments so that we could put that to work for them but not have that take up their cognitive load.

And this was some of the data that Alex was describing earlier. My personal role was to work with the governor of California on this. And so looking at the county level day over day at the change in the number of people that were staying home or going out.

And this was one way to look at sort of adherence to changes in policy guidelines that were being issued. And this is the director of Health and Human Services. And so they sort of got into a loop to use that.

And what we realized through there was that in order to make this work, we needed a framework to sort of apply to any particular situation where we needed to put mobility data to work relatively quickly.

I mean, the pandemic was sort of sudden onset, but it went on a long time. And so we had time to adapt a lot of these methodologies when hurricanes, wildfires, earthquakes, et cetera happen. We don't really have that same kind of time.

So focusing on the translational part of this, making sure that we have the ability to take established data pipelines and methods and then put that into a framework that is sort of easily adopted by decision makers became sort of the difference between using it and not using it often.

So, the readiness framework here implies the need for a set of data agreements, governance standards, and data frameworks, sort of the practical pipelines, a set of questions, so we're not trying to answer everything with this data.

There's a specific set of questions that this is geared around. The methodologies are, it's not necessarily an innovation, you know, kind of focus, it's on things that have worked over and over again.

And then making sure that these are then included in repeatable workflows that we can simulate how people actually do work with the data. Key challenges that come up, and Alex and Taka both alluded to some of these, but there's issues with bias, the distribution of users across different geographies, for instance, that's not entirely representative.

There's uncertainty in the data. It's a sample of a sample of a sample. There's access questions, not every organization can get access to the same data, so as Alex was describing that trade -off, that's a trade -off that's very real for every organization, and then capacity questions, so being able to integrate these data sources successfully into workflows sort of tests the capacity of a number of organizations around the world.

So we built ReadyMapper as a way to sort of take the sort of key data sets that would allow for sort of a basic real -time analysis of health, sort of where people are going and how they're affected by a particular hazard or crisis and how that affects their access to health resources in the area, and then make sure that we're not simply sort of turning dashboards over to people, but we're trying to get human -in -the -loop analytical reporting that would come out of this so that we're trying to get people to,

you know, interpret the data that is presented to them and not simply sort of fiddle around with a dashboard online. So just one really quick, then, example of this in practice recently, this cyclone Biparjoy, which hit Western India and Pakistan in June 2023, reasonably large -scale evacuation, and in this case we were working with the All India Disaster Mitigation Institute, which was also, which is based in Ahmedabad,

which works also with the Emergency Management Agency of Gujarat. You can see we've actually simplified the pattern by removing the middle part of the storm kind of approaches. The, you can see change quickly happen around the coastal areas in Koch, which is sort of, if you follow Gujarat out to the west, that's Koch, and then as the storm makes landfall, especially now around this, you know, kind of the outer band of the uncertainty cone,

that is where you'll start to pick up quite a lot of change in the population density. Koch itself declines quite substantially, and if we look at this in time series, Devbumi Dwarka, which is on the other side of that big body of water inlet, and Koch itself experienced these sort of sharp population density decreases and recover relatively rapidly, and then areas around there have some rate of increase, but it's actually fairly smoothly distributed around other areas. And just looking at that, there's these sort of key areas that have seen really drastic declines at the peak of the storm, but then this fairly smooth distribution elsewhere.

Just a really quick, you know, look at some, a sort of simple way to understand some of the representativeness questions. This is the relationship between the Facebook user population and the census population.

The, you know, that's not a great correlation, but it's actually dominated by a large outlier. For whatever reason, that's Hyderabad. So Hyderabad is, has much lower relative representativeness. And if you remove Hyderabad, you actually get a pretty high correlation between the Facebook population and the census population, which I think makes the, it's not perfect, but we have confidence in the reliability.

So this is an open tool. The older disasters have been cataloged, and we've been working with them to do simulation exercises with public agencies and other nonprofits to kind of work through scenarios where mobility patterns happen in various crises.

So happy to talk about that later. Now, Professor Sakimoto.

Okay, good morning, everyone. I'm very honored to have this kind of wonderful international conference in Japan. Thank you so much for joining. And did you already see the Himeji Castle? World heritage?

I also, unfortunately, I have never been into a Himeji Castle. So yeah, if I have a time, I'm going to try to go. So yeah, finally, now I'm going to talk about the special data distribution for disaster response.

Yes, I'm already focused onto the people mobility or something like that. So yeah, in my terms, I'm going to talk about the special data distribution in disaster. At first, let me introduce something about myself.

So yeah, my laboratory is the human centered urban informatics laboratory in the University of Tokyo. And yeah, of course, originally we focused on the people mobility and the whole urban area or something like that.

So yeah, we called, we renamed people for projects. Over 20 years or something like that. And yeah, of course, now I'm trying to kind of the digital twin of the urban, whole urban management, like as much the project or something like that now, expanding the people mobility analysis itself.

And also the, my main belongings is the CSIS, Center for Special Information Science. So now I'm the director from this April, but yes, CSIS is the center of the excellence all over Japan about the special information since the 1997 or something like that.

And then the, how to say, one of the most important things is the, yeah, kind of the joint research system called as JOLAS, purchases the commercial data from the national fund and provide officially among the researchers as a joint research, joint research is so that is why not only of course University of Tokyo researcher, but also all the research in Japan can use the commercial data for free through the joint research systems.

So, but now today I'd like to focus on to the actual geospatial information distribution among the industry and national and local government. So because I'm also the director of the kind of the NPO called as Association for Promotion of Infrastructure Geospatial Information Distribution.

So yeah, this association operates the Geospatial Information Center in Japan. So yeah, of course, we already open. the various kind of the geospatial information from the local government or national government, or some private companies.

So the number of the data set are now more than 10 ,000. And yeah, the number of files are 70 ,000 or registered. Institute is around the 500. So yeah, this is kind of the open data hub by the distribution of the business open geospatial data from the 2016.

So yeah, but this is kind of the NPO activity. So how to say, yeah, it's very difficult to keep the right business or something like that. So yeah, we sometimes forced to have a kind of side businesses or something like that.

But anyway, recently, many kinds of interesting data became open from the, this is one of the example from the building three dimensional data in Japan, provided by the MLIT, Ministry of the Land, Infrastructure and Transport.

And this is called as Project PLATU. So of course, Japan have the many local governments, maybe 1 ,800. But now 200 local government already developed this kind of three dimensional building data model based on the national standardization.

So yeah, of course, this kind of data set can broaden our possibility to have various kind of assimilation. Of course, not limited to disaster cases, but it's very important basic data in Japan now. But now our focus point in this introduction, I'd like to, now I'm tackling into a real time data distribution in disaster mode in 24 hours or 48 hours.

Yeah, by the collaboration of the private companies. So yeah, this is actually the case of the Noto earthquake on January 1st in this year. So yes, of course, we have some trials for two years, about four, five or six big disasters.

Yes, we have some promises with the two or three private companies. So they can upload some, yeah, emergent area photo by several area survey companies like this, right? And also, yeah, of course, road passage information by the current navigation company called the Pioneer.

Yes, it's also uploaded in that Noto area. Yeah, of course, that maps can enable us to know the possible route by the possible route for a logistic companies or something like that for rescue the severely damaged area.

But this trial is not enough, I think, so yeah, of course, I'd like to... how to say, enhance that kind of trials for 10 companies, 10 or several 10 private companies now. So in that sense, I have one national project called as SIP, Social Innovation Program for disasters.

So, yeah, of course, this is not pure research, pure research, pure academic research, but also, how to say, social implementation research activity, whether private companies can respond or not based on the official contracts, okay?

So, yeah, of course, voluntarily, they may provide some data or something like that, but how to say, we need more certainty for the emerging provision even in the emergency situations. So after the discussion with them, yeah, of course, they want to have a kind of, of course, budget from someone, right?

So yeah, of course, at this moment, we have a research project. So if we promised, if they provide data within 24 or 48 hours in a severely damaged area, yeah, we can pay something, some money, not so big money, but yeah, pay something.

So, yeah, if so, yeah, they, how to say, yeah, they received our proposal. So, finally, so now, seven or eight companies provided the kind of contract with us, but this is at the beginning of this year, so how to say, in case of Japan, the flat season, rainy season start from July, yeah, mainly July, so yeah, in that sense, how to say, yeah, we are preparing for the system and contract almost finished,

but if once the big disaster happened, yeah, we can ask them to provide within two days, yeah, but of course, delivery certainly is not guaranteed by them, yeah, of course, they don't have a strong confidence in the case of the disaster, right?

So, yeah, of course, data provision is not obligation, of course, but if they successfully provided to us, yeah, we pay in that case. So, yeah, this is all what I want to say today and yeah, so now I'm focusing on to this project.

So, yeah, about the open question is the, yeah, of course, how sustainable can be in this trial and, yeah, of course. This is, at this moment, this is the research project. So after this research project for five years, we'd like to move on to the official operation with governments, or something like that.

Because national government always provide some very big budget support in the Shibuya damaged area. So this kind of the data cost should be included in some part of the big support budget, I think. So yeah, this is the trial of my side.

So yeah, this is all, thank you so much.

Can we bring Alex back? There he is. I don't know if you guys want to come up front. We have time for about 11 minutes, 50 seconds worth of Q &A. So I'd like to open it up first to the audience if you have any questions that you would like to start with.

And, actually, we have a mic here.

Thank you very much for the amazing presentations. So, I have a couple of questions with regard to the use of data pre -disaster situation. For example, I can see we can use your data to train machine learning techniques, for example, to understand the behavior of people's movement, right, before that happens.

So, we can plan evacuation centers and so on. So, the first question is, are you already using machine learning to develop sort of predictive algorithms? Or, if not, can we use your data to do that? That is my first question.

Second question is, it is also important to know the people who didn't mow, and why didn't mow, right? Because it's about their risk perception, and also, we need to make sure that we generate some community programs to work with this community to make sure they understand the kind of level of risk that they're facing, and we can change the behavior of those people as well.

So, I was wondering, how can we use that data to really, during the pre -disaster situations?

I mean, just really quick on the people that didn't move. I think you're absolutely right on. And this is implicit in some of what Alex was talking about, for instance, in terms of distance bands from home and the way that needs to be understood in terms of the propensity of different groups to not be able to leave during a disaster.

So that's one source. You could look at it in a more granular level to see sort of at the sort of tile grid cell level in terms of areas where based on the hazard, you would expect movement, but you're still seeing a higher than normal population.

In the Pakistan floods, for instance, the greater problem was not people that were displacing longer distance. It was that the transportation networks were so disrupted that people were literally not able to get out of areas where major flooding had occurred.

And so the much more significant analysis was how many people ought to have evacuated, but were not able to have evacuated. And so that sort of flip side of the mobility analysis, I think, becomes quite central in those cases.

That's a little different than the scenarios we were talking about earlier, but it's right on. I'll hand it over to Taka on the machine learning question. Yes.

Sure, on Andrew's point, that's right on the spot. If you have people who should have evacuated on one axis and people who did evacuate on axis, the mobility data just shows one axis of that, whether they did evacuate or not.

But once we, for example, post -disaster, if we collect more information through surveys, we can try to triangulate who are the people who needed to evacuate but couldn't. And usually it's the most vulnerable populations who can't evacuate because of transportation issues.

So that's a very important point that just looking at mobility data, you might kind of miss. So there needs to be multiple angles to that. The machine learning stuff, yeah. There's a lot of research, I think, in kind of merging between computer science and urban science and social sciences, where they're developing more models to predict how would people move if this disaster happens.

And in our lab as well, we're trying to develop models using AI to try to predict hypothetical scenarios using data from past disasters. So one challenge is can we use data from India to predict, or this city in India, Chennai, to predict how would people move in Mumbai if there's a similar disaster that happens?

So trying to tease out the generalizability from these disaster cases, but also trying to impute the more local and context -specific things so that we can improve the predictability is something that we're working on, yeah.

So do we have access to data, sorry? Can we have the access to data to conduct further research?

Yeah, so we have one project actually with Professor Sekimoto trying to make this, because this predicted mobility is synthetic, right? It's not real mobility. If it's synthetic, we can, you know, try to open this up to the public and for research and for public agencies.

So that's something that we're trying to work on. Making like a global platform, we call it a global data commons. For disaster resilience, can we open up various scenarios of disasters in different regions for people to use for planning?

So that's something we're working on, yeah.

Yeah, not too much to add from my side. Just to double down on, I think, the question of the people who should have evacuated but did not speaks to the importance of having multiple companies or multiple data providers.

So for example, in the data that comes from the Facebook, the Data for Good at Meta program, there's no ability for you to know in the data and even us on our side, if the population density went down, does that mean people just stopped using Facebook or does it mean that they can't connect at all or does it mean that they left the area?

It's a gap that then you could triangulate if you had multiple data sources, right? So that speaks to the importance of these preparation exercises and this coalition approach so that we're not reliant on one.

And I would also, on the machine learning side, we have not done anything in our group on trying to predict the flows of persons. We have studied in depth the ability of different machine learning packages to create synthetic mobility data.

And we're quite happy with those promising results, convincing practitioners that they would rather have perhaps synthetic data rather than the actual data is a very difficult thing. Academics, of course, would want the actual data, right?

And our program already offers it, so why would we switch to synthetic? But that is where we pushed mostly on the machine learning side is increasing the privacy. We haven't focused on trying to explore the tip of that utility, which I think is likely better done by others than meta people with more experience in actually contextualizing the data and using it to respond locally.

Hi, thanks a lot for the presentation, and I think what was very interesting about what you shared was the fact that through the data you've kind of cracked the geospatial component of mobility and disasters, and I think what was interesting was the temporality of disasters, right?

All those graphs you showed on how much time it took for people to come back, and from a response planning point of view, this tends to be a very unanswered question that's not often answered, and so my quick question was to what extent could you use the data to start at least observing through various types of disasters, so cyclones, earthquakes, fires, floods, and start thinking of how much time do usually people stay displaced, because this would really give interesting temporal dimensions on the response planning side of things. Over.

I mean, I completely agree with you and it speaks to the kind of comparative taxonomy question that was discussed earlier around being able to look at multiple hazards and mobility patterns relative to urban structure that would then give you a comparative view on that temporality for the displacement and return dynamic.

Unfortunately, the data actually used to be better. I mean, the operating system changes that Alex pointed out earlier actually changed one of the best data sets that meta ever produced, which was the ability to look at a cohort.

So like there was a remarkable work done where you could sort of take a bounding box cohort and look at their distance from home and their then return to home over say a period post two weeks after a disaster and then look at even disaggregated changes in gender at the rate at which those return dynamics would happen.

But unfortunately, operating systems change. This is one of the challenges with private data for public good, honestly, is that there's lots of other things happening that have nothing to do with the problem we're trying to solve, but we are sort of subject to the constraints that occur with the companies.

And so, you know, sometimes progress is nonlinear.

Yeah, I would mention the trade -off was for the benefit of privacy. We don't build our own operating systems. We're beholden to the mobile phone operators that build it. But if we just looked at the privacy dimension of that change in the majority of location data handling, the privacy side went way up.

Now, in the disaster response, is that the proper trade -off on the utility side? Conversations such as this are very important in bringing that perspective back to the private sector so that they don't move solely in thinking of just privacy, just because that's perhaps the biggest legal constraint at the moment or

One thing I would add to that, the longitudinal aspect is back then when we had the data, we were able to analyze how different communities recovered their population, uh, across different, you know, different regions and different regions have different policies.

So you're able to kind of tease out if, or, you know, this community, this, that made this evacuation chapter or this, this community didn't invest in this. And then you can try to trace out or tease out the effects of that on the longer, longer scale.

That was actually my PhD thesis, but that's something that you could do. And that was, but probably you couldn't try to get more policy insights by looking at the longterm, uh, consequences of, um, after the disaster.

I think we had

time for two questions although I don't see anyone beating down the door so you know if we want to occupy the room we can do it.

Yeah, thanks a lot for these very interesting presentations. Going back to, I work more on the practitioner side and impact -based forecasts. And actually, for us, having mobility data will be really crucial to know the exact timing of a disaster because whether it hits during morning or afternoon commute makes a huge difference.

And given that we have very little to non -mobility data, I'm actually very interested in these synthetic data sets. Could you maybe give a short elaboration on how much data does it need for learning?

Because for us, having very little, even a synthetic data set would be a great way forward to including some time dimension.

Yes, thank you very much for your question. Yeah, of course. Our laboratory is developing nation -based, nationwide, synthetic should people data in Japan. So in that sense, we can simulate everywhere in Japan.

In that sense, that is very good. But at this moment, I'm not sure the prediction accuracy is enough or not for the actual operation of the rescue or something like that. So yeah, this should be measured with the national government or local government or something like that, I think.

Yes.

All right, that was about it, by the way. Alex, one final comment.

Yeah I was just going to second that the synthetic data study mobility data that we studied is very good at capturing regular mobility but as soon as you're trying to specifically triangulate outliers perhaps the perturbation caused by a crisis the noise is purposefully very destructive for those outliers it's trying to capture the overall trends over time so much more work would need to be done we feel we're confident that we could move into that synthetic data realm for just kind of general mobility but I think in the crisis response if I was placing a bet now we would be staying with the more traditional and some of the pipelines that our program has had around for over five years but they've been tried and trusted and true and there's whole workflows that permeate throughout the practitioner system that are kind of derived from that so we wouldn't disrupt that anyway unless we had good reasons.

All right, thank you so much. And thank you to Taka and Professor Sakimoto. I really appreciate it. And thank to all of you for attending. And enjoy the rest of the conference.