RAW FILE

ITU

AI FOR GOOD GLOBAL SUMMIT

GENEVA, SWITZERLAND

DAY 1, 15 MAY 2018

15:45 CET

ROOM K

IMPLEMENTATIONS OF AI TO ADVANCE THE SDGS

PANEL 3: DATA FOR GOOD

Services provided by:
Caption First, Inc.
P.O. Box 3066
Monument, CO 80132
1-877-825-5234
+001-719-481-9835
www.captionfirst.com

***

This text, document, or file is based on live transcription. Communication Access Realtime Translation (CART), captioning, and/or live transcription are provided in order to facilitate communication accessibility and may not be a totally verbatim record

***

>> Good afternoon, everyone, and thank you for all those who came for Data for Good.

So as you probably heard me explain earlier in the big room, we are five different tracks today that are more practical. It's an opportunity for all of you to hear examples from different speakers about their domain and what they are doing in applications of AI for Good, and also we are going to be focusing on data. Obviously, the conversation on data is an important one, and we try to go bottom up during this conference, where we identify data issues for the examples that are given to us by different speakers and different experiences so we can understand and hopefully formulate some strategies for academia to work with industry, for industry to work with government, and what it means in terms of data.

For that, we have this first session today, which is read by Urs Gasser from the Harvard University. A number of speakers will be introduced as well.

We also have in the room data Rapporteurs. Those will be tomorrow distributed in all four tracks so they can listen throughout the day, where there is a topic on satellite, whether it's a topic on health or smart cities or trust in AI, listen and capture feedback on issues related to data. And you can help them also by suggestions or questions or ideas so that on Day 3, when we report back, we have two reporting back on Day 3. The first reporting is on each track, what projects were meaningful and impactful that we want to push forward as a group, as a multistakeholder issue, but also what data issues have we identified, and which ones are structuring enough that we would like to propose and push to have more multistakeholder support and what it means for us. Your input and suggestions are very helpful in helping Rapporteurs to bring back the right feedback and to share with the whole attendees.

So with that, Urs is your master of ceremony, host, and moderator for the rest of this session.

>> URS GASSER: Thank you so much. Thank you.

(Applause)

Hi, everyone. I am super delighted to be here. Some of my favorite people are in the room, so let's make this fun. It will be a good session, I promise.

First of all, of course, we already solved the big problem by calling it Data for Good session, which avoids the problem of definition. What is AI? We make it easy. We talk about data, and you have a good sense what data is. Now, of course, if you look a bit back in that history, what governments, what companies, what civil society have been doing for many years, and academia as well, is using data to create a better world, to improve transportation systems, to build more efficient healthcare systems, to provide education, whatever your area or example is. But at the same time, I think everyone recognizes something has changed over the past few decades. The role of data is a very different one in today's society as we enter the age of AI and other emerging technologies. Of course, we are all very aware of how much data is collected and aggregated and analyzed and it can be used and product advised. We are also increasing the aware of some of the risks, of course, of course, of these data collection practices.

What we hope to do in this session today is really to look at the sea change. As Big Data is making big headlines, as Big Data meets AI, and the way we'd like to approach it is essentially, as Amir introduced it, by featuring a series of very concrete use cases from very different perspectives, from different organizations, from different continents, to have a collection of snapshots, how data can be used for the good. We will do so through a series of lightening talks, and we will see in a second on the slide the list of lightening talkers who will give us very short presentations, about seven, eight minutes long each, and highlight different aspects of the data ecosystem. So we will have spotlights, but they will add up towards a more holistic view, that's the hope. And over the next two days, we also hope, with the help of the Rapporteurs, to add other use cases.

And as we move through these lightening talks and go through the days here in Geneva, we will hopefully be able to distill a few common themes across these different stories, narratives, examples, and use cases. So we will see benefits, of course, of the use of data for the social good. What are powerful narratives that demonstrate the value of data and how it can be leveraged using AI-based technology.

But we will also talk about barriers, what are current impediments to making full use of data for the social good. We will look at and hear about stories related to more technical challenges, to interoperability problems, but we will also address questions further up the cake-layer model, where it's about policy challenges, what's the role of privacy as a barrier sometimes to data sharing? We'll talk about the human factor, when we may talk about data, but who are the people who are able to do data analysis and put the data at work for the social good? What are some of the educational challenges we face in different regions? So we'll have a whole set of issues that hopefully add up, even if not perfectly, in a perfect systematic way -- which I would like, as a Swiss, being very organized -- but nonetheless, it will be an approximation triangulation, I think, for some of the key themes.

And lastly, and perhaps most importantly, as it was already mentioned, it's also very much about how do we address and overcome some of these barriers and challenges? What are kind of new models that enable unlocking data that may be currently in silos? What are practical experiences to build new models for data sharing and for collaboration across private and public sectors, across geographies, for different industries? So what are practical examples? How can we build a repository of good models, practices, and experiences that help us to use data for the good?

Again, this today is a starting point. We start with concrete stories, with concrete examples. We'll add some over time, but hopefully also have a broader conversation about some of these connecting points and commonalities and differences across use cases.

So as you see, we'll have five lightening talks, followed by an open discussion. I would very much also to invite the colleagues from the various UN agencies who have made very helpful contributions to the preparation of this session to weigh in and add their examples that they see on the ground, how they engage around the world using data for the good, for building a more inclusive society. So hopefully we can have a very open conversation in the second half.

Now, without further ado, I think Sam would go first. Sam Molyneux. He is a scientist by training and entrepreneur, has found really fascinating work on studying the knowledge ecology in the modern world, looking at scientific information, how it is shared, and sometimes how it's siloed as well. And he will give us an example of application, how AI-based technology can be used to open up knowledge and put things together and make visible things that otherwise may not be visible to the human eye.

So Sam, if you want to take it from here, please. And just before you start, maybe you can tell us, up here as you walk to the podium you can think about it, what's your favorite AI-based application that you use in your daily life.

>> SAM MOLYNEUX: That's a great question and a good opener to the talk. So one of my favorite AI or applications that sort of involves AI is actually Spotify. I think it's a fantastic media example, and it would be fascinating for everyone to understand sort of like all the machinery behind the scenes that basically organize and make available and discoverable music. And some of the principles that we've used in our project is effectively applying that to scientific research. So that's my answer.

So it's an honor to be here to talk about our project as a case study. I am Sam Molyneux. I help lead the medi group at the Chan Zuckerberg Initiative, and we build tools for scientists, scientific researchers, on the belief, basically, that most progress in society and medicine and towards Sustainable Development Goals comes from scienceers routed in science ultimately.

In scientific information, there's two major problems -- there's a lot of minor problems, but two major problems. One of knowledge complexity. So today it's tremendously difficult to be able to read enough articles and contain the complex knowledge, especially in biomedicine, in the human brain. It's impossible over very large numbers of articles. So this is one problem that we work on, knowledge complexity, which I will mention at the end.

The other problem is just a basic problem of research awareness, and this is what my talk is about.

So working on the research awareness problem, a couple stats about this. 4,000 new papers published every day in biomedicine. There is effectively a global conversation happening in real-time amongst researchers major document by major document, which are peer-reviewed articles, and it's basically a truism in science that nobody knows what's actually sort of being published from each others' labs, but everyone thinks that they do, and it's basically because they can't tap into these global streams of papers that are coming out. These are built on the back of 200 years of modern scientific research, over 70 million peer-reviewed articles across the scientists who have seen an exponential increase in the citable entities as well and thinking about the speed of science and the speed of progress, if we want to accelerate this, we need to give researchers just sort of powerful basic tools to be able to tap into the global conversation and handle knowledge complexity.

Researchers recognize this. 858% of scientists believe they are missing new research on a daily basis. 75% of scientists wish they were more up-to-date. At the Chan Zuckerberg Initiative, we believe that tools like artificial intelligence can be applied at scale today in the context of a philanthropy to build a product that can solve a problem like this and then give it away to everyone for free for global good. So we are doing this also as part of the Chan Zuckerberg Science Initiative's goal or to support the goal of making all diseases curable, preventable, or manageable by the end of the century. Of course, that's an audacious goal. The only way that you can start to work towards that goal is actually enabling all scientists and helping the entier ecosystem go fast -- entire ecosystem go faster.

So we imagined a Spotify-like experience for research, where you could stream research, tap into the global conversation, discover arbitrary intersections or subfields of research instantly and be able to browse the entire landscape of discovery in the sciences over those 200 years. And we've been using AI and working with scientific publishers to enable this. I am going to talk about a tool that we have been building for about eight months.

So MED is a tool that consumes scientific research at scale and builds an evolving map of scientific knowledge out of it. And we use techniques from the knowledge representation, NLP, and reasoning field, and we -- effectively consumes papers from publishers, from PubMed, knowledge that we buy or crawl for on the Web. It involves predictive models. For example, we have a neural net-based predictive model for anticipating the impact of research the moment it's published or even before it's published to be able to rank it usefully in news feed-like settings. Each one of these capabilities, whether it's predictions, trends, detecting emergence on new concepts that are just being mentioned for the first time in some very, very small subfield but rapidly escalating in science, we are effectively bringing all of those data sets to bear as part of the knowledge graph in a consumer quality interface that enables scientists to stream and explore their personal world of research. And the knowledge graph we have been able to build over this project has gotten quite large, so it has billions of connections, for example, amongst papers, concepts. We cover the full genomes of organisms, text files, and we are trying to span the panoply to enable them to be discoverable, followable, and explorable in the product experience.

So we worked with some of the best talent in Silicon Valley and elsewhere to build a beautiful consumer experience. Not to sort of understate how difficult this is, but with the integrated efforts of product managers, designers, people who understand the best qualities of the data, et cetera, we've been able to work with biomedical labs at Berkeley, USF, and Stanford in the Bay Area, which we are located in Palo Alto, so they are right at our footsteps, to build and evolve this experience and make it really, really good. So META is a product that allows you to follow research at effectively any level. You can follow departments, pathways that contain collections of dozens or hundreds of genes, for example, and this ability to arbitrarily discover historical research and follow research at any level really is the capability that happens at the intersection of all of the various data sets that we've been able to put together.

Here is an example of a feed built on the Stanford Department of Genetics. You would think that this is something that's really easy to do today. You can set up an RSS feed from a journal or you can create safe searches or all of these sort of like duct taped together tools that people use, but it's very, very difficult to just follow every researcher in the Department of genetics, and this should be easy in this day and age.

Also, personalized feed, so personalized based on analyzing a researcher's entire historical research set or the last couple of papers. These are the sort of features in a Spotify where you can create a radio station off of a song, for example, we think that should be possible for researchers in this setting as well. Then we've built a beautiful recognition system to enable you to discover subfields of research, discover leaders in research areas, each one of these followable, streamable, et cetera.

So again, this is a tool that in the context of the philanthropy, we've built for free for everyone. It is designed for biomedical researchers at heart, but it's available to everyone else as well.

So maybe I will just switch gears for a second, talk about a couple lessons from our project. The first lesson I would have you go away with is to go out and in a concept like this, go out and find a data set. The data set is the articles, articles are owned by publishers, and it's been true for most organizations that want to work with publishers that it's extremely difficult or very, very expensive to get access to the articles at scale for text mining. We found that in the context of a project like this in a philanthropy, given we don't have business constraints or strategic goals with respect to the industry, we have been able to align publishers and align the industry around our goal and develop basically mutual exchange of value. So we want to drive readership to the articles, that works with business concepts and strategy, and that is ultimately aligned with helping scientists move faster.

So the second one is AI and data doesn't necessarily equal value. This is a hard lesson on the ground for projects like this. As you go out and analyze data sets at scale, you discover the data sets you produce out of that, the recommendations, whatever the outputs are have bad qualities, false negatives and false positives, that in different settings can have a really negative impact. So being able to overcome this and discover the value at the heart of the data set is an integrated design and user experience research problem, so I would encourage everyone to think about that. These two things together are not necessarily magic.

The last piece, ironically, it's all about the people. It's all about the quality of the talent that you put together towards a project like this on the team. Research, engineering, design, user research, which is such sort of an underappreciated competency, project management, data cureation at scale, et cetera. You need the very best mission-aligned talent. Of course, the type of talent that you want to work with is being highly competed for in the big tech companies, and so the approach we've taken with this, because we are not going to pay at the same level as those tech companies, is we are looking for people who care about the project, care about the problem, aligned with the mission, and we find that if you can help candidates self-select into those types of -- into projects like this, you can actually get tremendous quality talent, but you have to have a strategy going to get the talent involved in a project like this.

So the last piece is the research may not be -- the edges of the research that you are bringing into your project from an AI perspective may not be at the level that is sufficient for the quality of product or quality of system that you want to build, so you may need to fund additional research, which is the kind of thing that we are doing in a project that we call computable knowledge. We put forth a $5 million grant to Andrew McCallum's lab at UMass. We are trying to push the field forward to be able to enable the future product.

I guess I am out of time. Thanks.

(Applause)

>> URS GASSER: Thank you, Sam. It's so interesting that even the first lightening talk highlights the role of the human. Also, you can talk about data where you think it's a technical story, but one of the top lessons learned is humans mat tern.

We have one question that came up if you may give a quick response to that. That is there has been a lot of criticism with respect to Keer review as a mechanism in science, so the question was whether the tool you have built here is somehow amplifying some of the problems related to peer review or is actually trying to mitigate that criticism or some of the problems that are involved?

>> SAM MOLYNEUX: So there's a lot of criticism of peer review. Most of the criticisms we hear about are the speed of peer review, which I think is sort of upstream of our product. The predictive

models that we built that drive news feed, for example, or the feed in Meta, they take into account the sort of whole system of what's happening in the biomedical sciences, and they do encode biases in the system. What we've discovered is that the ability to sort of, like, usefully rank new research in real-time that uses all of those signals, it raises the efficiency of the system far above what you experience in other products today, but we are sort of very, very cognizant of those biases, and what we'll do over the next couple years is learn more about them, learn how to train better potentially to counteract them, or even use additional signals from the ecosystem to be able to compensate for them. So we are sort of thinking a lot about this.

>> URS GASSER: Super interesting, and again, I think a general theme that hopefully emerges, thank you, throughout the presentation is really this contextualities and how do you deal with some preexisting problems that may be amplified but at the same time there is an opportunity to fix it or at least make progress.

We have more questions in the question online tool. Please vote up or down. I think you can only vote up. We'll get back to it during the conversation.

Next I'd like to invite Ed Su from the World Bank, who actually is a senior advisor to the President of the World Bank group. You have done amazing work. As you walk up, I have a question for you as well. You have worked in finance, in health, you are dealing with emerging technologies at the World Bank and trying to make sense of this changing world. If you have to narrow it down to one area where you feel that data will really make a difference and help us to build a better world, what area would that be? Is it finance, health, environment, where would be your bet? Thank you.

>> That's a hard Question. I mean, I would say health. I would say health. I think you look at the health field, and it's rife with so much inefficiency right now in how we do things and the potential for data to really revolutionize that I think is -- and I hope so too because I want to live longer. That's my personal goal.

All right. Thank you so much. My name is Ed, a Senior Vice President at the World Bank Group, and my role there is to help the World Bank Group think through or approach disruptive technology, and how can we harness the benefits of disruptive technology for all of our clients, particularly the poor and developing countries? And so that's the approach we are looking at. So I am pleased to be here to talk about a couple projects that we have been working on that use Big Data, machine learning, and analytics. So just a quick intro to World Bank Group. We do about $40 billion to $50 billion a year in projects on developing countries, and we run the gamut of health, infrastructure, social protection, governance, jobs, education projects. So we see incredible scope for Big Data, AI, disruptive technology in general to really revolutionize the way we do things. So the reason we are here is really to learn and be able to try to incorporate all this technology into our projects.

One project I want to talk about first is a project we did in Haiti.

This is a bus in Haiti. After the earthquake, we looked at how can we rebuild Haiti's transportation system? What we found in Haiti, it's what we call a data-poor environment. It's really hard to figure out where do people live, it's really hard to figure out where do people work because much of the jobs are informal. Yet if we are in there, we are investing millions of dollars in a new transportation system, if we get it wrong, this has real world implications. Many of the poor who commute to work face very long transit times, one to two hours per day or more, and some have to spend 30, 40, 50, even 60 or 70 percent of their income just on transit cost alone. I think the stakes are high for us as we think about how do we build a transit system, how can we be much more effective in doing it?

So what we did was worked with flow minder and also the mobile company to get CDR data, and the first thing we looked at is where are the commuters and where are they coming from? This graph here in the yellow in the daytime you can see the population is concentrated in a few areas, then at nighttime you see a wide disparity in where people actually live. The way we conceived of this is that we need to look at -- as we design a trant system to look at accessibility, and can we measure as we build a transit system how much are we improving the accessibility of jobs for the poor, and how can we improve their commuting time to get them to be able to get to their jobs much quicker and get to where jobs actually are? And that's the challenge because many times when you design a transit system, you are not really sure where the jobs are. You can go out and do surveys, and maybe even in developed countries you can do that, but in a place like Haiti, that's impossible.

What we looked at is we started to analyze the data, not just where populations are, but also we could track where people actually live and where they actually work, and we could track it on the individual level. So I think we used the CDR data, partnering again with GSMA and using their standards to anonymize the data to track how people get to work in a much more efficient way. This helped us then to design a new public transport improvement plan using buses, and we were able to measure that for the city center, for example, we were able to improve the average commuter's accessibility by 52%, and this is measured both in time and cost. So this has real-world implications that I was mentioning that the poor then have much more time to be able to get to work, more time to spend with their family, and also job mobility, able to switch jobs easier based on different opportunities.

What did we learn from this? We learned that even in data-poor environments like Haiti, you can find data that can lead to useful and insightful policy decisions, and we also learned that data can be used to reframe the discussion. The mobility question -- and it's not just looking at concentration, but it's looking at access to opportunities, and with data, you can find those opportunities.

And in terms of next steps, what are we looking at? We are looking at how can we use advanced machine learning to derive -- to look at other questions like can we actually determine the income, literacy rates, or look at other behavioral questions of the commuters, particularly of the poor, to even continue to improve the transport systems? We are going to scale this up. We are going to uing similar Big Data analytics in all of our transport projects, and we have over a $38 billion portfolio in over 72 countries. Now we are actively looking towards partnerships. The ones we developed with GSMA, with Mobike data, to look at -- also design better urban cities -- I mean better urban centers, and also Grab, which is a ride sharing company, and we looked at that to help design better transport systems in Manila.

So I am going to go to a project we are working on now, and this is very much a project in process. And the goal here is can we use AI and Big Data to predict FA mins?s In a collaboration between us, the UN, and the private sector. Can we enable a more impactful famine response by acting much earlier? You can see here when you look at the yellow line, that's the excess deaths per month, deaths that are a result of famine, and you can see the donor funds on the bars didn't come in until mainly lives were lost. What if we were able to intervene in the green box? How many lives could we have saved? Can AI and Big Data help? We think so. This is modeled on another project we had, an insurance project which enabled us to work with insurance companies, and we used a parametric trigger to trigger funds coming from insurance companies so we could intervene at at pandemic at an earlier time.

What are the questions we are trying to answer with this project?

What is likelihood of a crisis actually to happen? Here you have to look at many factors, poverty, infrastructure, the political ones. Are there data that informs conflict or violence? And also environmental climate factors, of course, vegetation or agriculture or crop prices. A third question, how can we actually measure how good we are at mitigating impact of humanitarian assistance? Can we measure that better? Then using all of this, throwing in sources of Big Data and new machine learning tools, can we actually build a forecasting tool that will enable us, the World Bank and the UN, to come in earlier, donors to come in earlier, then even working with financial companies, like insurance companies, can we build insurance product where they can come in much earlier too?

Just to wrap up in terms of lessons learned for this, given the scope of this problem, a wide range of data sources can improve predictability. The sky is the limit in terms of what data could be relevant. You know, we wanted to look at border crossings. There's no real way to get good information on what is the flow of traffic across borders, but now with satellite imagery, maybe you can. Can we use social media? Can we use -- actually, we learned today, can you use radio and use text analytics, voice recognition on radio to actually also get early signals?

I think we also looked at there's a lot of confidential and public data sources, so we need to build a new data platform in order to deal with that. And the combination of AI and econometric models can lead to more powerful insights.

Next steps. Right now I think Sam mentioned the difficulty of finding good tale ent, so I think that's why we are -- the good talent. Look at partnership with technology companies that are willing to come in and take a look at this. Their teams are motivated to help with a problem like this, where we go to them and say do you have a team that wants to help us end FA mins and save lives? It's a very powerful proposition to find partners. If anybody in the room wants to partner with us, please let us know.

Next steps, we have a model. We are trying to validate that. Now with six countries, we are trying to scale it up to 20 or 30 countries. We need to expand the number of predictive variables, like I was mentioning, that there are many variables that can predict famines that we haven't looked at yet, but with AI and Big Data and new sources of real-time data, we think that can open it up for us.

Thank you very much for your time. I am very pleased to be here again.

(Applause)

>> URS GASSER: Thank you very much for this fascinating case study.

I have one question. There were several questions, but one I think you could answer maybe quickly at this stage, and that is about the involvement and engagement with local communities and local politics and local stakeholders, referring specifically to the Haiti example. What you have proposed, has that sparked discussions? Was there an engagement strategy, or how do you deal with these local contexts?

>> At the World Bank Group, we have a commitment that in every project we have, we have local citizen foyb. We started to explore how do we use technology to help that. With Haiti, there was a huge push towards looking at local communities. The problem we have is it's so difficult and exspendsive to do that at scale. Now we know where they live, how much it costs for them to get to work, and we can help them directly. But citizen feedback is always a huge core philosophy of the World Bank as well.

>> URS GASSER: Fabulous. Thank you so much again.

So with these two presentations, we have essentially concluded the first mini segment within the lightening talks, and we will now actually zoom out one level and look at data collaboration and also some of the educational topics that have come up in the previous two talks. And Nagla Rizk is a good friend and colleague of mine. She is Professor of Economics, is also the Director of the Secretary cess to Knowledge for Development Center at the University in Caro. She will share details on what enables data collaborations on project she has been involved in and that she's been leading.

Nagla, the last time we met was actually in Rio de Janeiro at the symposium on AI and inclusion organized by the network of Internet and Society Centers.

My question to you was this symposium took place in November of last year. Now it's May, we are here in Geneva. Are you more or less optimistic between November and now whether we can use technology and data to build a more inclusive society? Are you more or less optimistic?

>> NAGLA RIZK: I am certainly more optimistic. Just being here is actually proof that between Rio and now not only the discussions that took place there, but the work that we have been doing on the ground following on what everyone else is doing. It's amazing the speed and the vast range of activities that are taking place are certainly making me more optimistic.

Well, thank you very much for having me, and thank you for joining the panel. Can I have the slides? Yeah.

>> Little clicker.

>> NAGLA RIZK: Oh, sorry.

So actually what I would like to bring in the few minutes I have is a perspective from Egypt based on work we have done at the Access to Knowledge for development Center, research we have actually done with partners and also research we have coordinated with partners as we are the open data for development node for the Middle East and North Africa.

So basically, I will start, I will talk around three axes, very quickly, the challenges given the overview of the topic. I will talk about lessons learned with examples from work that we have been engaged in. And I will have a few questions and points on the way forward.

So let me just very quickly start with sort of a simple framework I thought of. The earlier sessions today made my job easier because some of these points were already alluded to. We are all aware that data as a prerequisite of technology for the global good, but when we come to think of and I think of challenges in my part of the world and the voice from my part of the world, clearly there is the clear overarching challenge of the risk of the digital divide inequality that we heard of earlier, and I remember clearly the speaker in the previous session spoke about outward versus inward AI for global good. I would argue that not only should we be aware of the challenge and try to mitigate it, but also we should be proactive in bringing change that actually makes the gap smaller and makes better livelihood for people in different parts of the world, not only globally, but also within the national internal divides.

The second set of challenges come from the data itself. Quite often in parts of the world, including my own, the data that will be either not available, it will be scattered in different offices, it will be either with large companies or with government state, data is usually not available as open data. It is sometimes politicized and sometimes filtered. But most importantly, it's not sufficient to provide material for technologies like blockchain or for AI applications. So just the presence of data can lend itself to some challenges, but also once we have the technologies, step further, you can face challenges that have to do with enabling environment. Data is quite often political. And you can face challenge starting from the very first part of the spectrum of lack of political will all the way to data surveillance, and the use of that not for benefit of the citizens, obviously.

Endemic challenge that is available in parts of the world is actually the asymmetry between encouraging economic freedoms and promoting data-driven innovation yet, on the other hand, curbing civil liberties. And there it is problematic that is faced in some parts of the world.

A recent example in Egypt has been the ride-sharing law that was passed just a few days ago, and the bottleneck there was the government's request to have the users' data housed physically inside the country. And this was a point of debate. Eventually it was resolved, the data is in the cloud, and the article was not as pressing and not as limiting as it was originally requested. The point being that it is the users' data, and the users were not really part of the conversation. That's an issue.

So when we think of the ecosystem, when we think of curbs to civil society, for example, we are looking at challenges facing data for the global good.

And then I can't speak about the ecosystem without clearly talking about the infrastructure, physical infrastructure, but also the human element which was clearly mentioned in this session, the importance of the human component of the technology.

So within that, the lessons learned from the work that we have done is that, A, there is a need for data collection on the ground, data that is organic, that is collected in innovative ways, in different ways of collecting data, because quite often parts of reality is not captured in state statistics. Novel technologies, a mix of technologies, layering the data to come up with the full picture. But also the second lesson is that -- oops, sorry -- the second lesson is that we need to have the proper enabling environment and encourage the ecosystem in a way that enables data to be -- the applications to be achieved for the common good, and overarching challenge is clearly the need for capacity building, but very important -- sorry -- very important lesson learned in this process is actually the importance of the interdisciplinary subject. We work with people from civil society, technologies, academia, and other areas, and in the private sector.

So very quickly, I am going to show you examples of projects that we have had. We had a study of the ring road, accessibility on the ring road in Cairo, and one of the findings that we have seen is that these are the bus stops that are built by government, but these are -- quite often you will find the informal stops and informal waiting areas that are actually constructed by citizens, and they are quite often next to informal housing. So this information will not be part of the state statistics. And if you, we, want to study transport and roads, we need to look at what is happening on the ground.

This is a picture of these stops and these stairs that are actually built by people in order to access the roads. And it impacts safety, it impacts jaywalking, it impacts stops of public transport that are not at all on the radar of the state statistics.

Another example, it's a simple data set, but it looks at informal innovation, informal enterprises. This is part of the research we are doing in the open Africa research, where we devise alternative methods to complete the picture of innovation in Africa. This is a pilot we did in Egypt and clearly shows that the informal practices of skill development are actually prevalent over the formal training and vocational training for the people. So at the end of the day, what you will see in government statistics are the top-down statistics that do not provide the full picture. So I suppose my last example, very quickly, is an example done by partners at the American University of Beirut, collecting data on health statistics on refugees and compiling based on a needs assessment and compiling, cleaning up the data, and presenting it in open data format to be used by policymakers.

In all of these examples, and I apologize I am going a bit fast in the interest of time, these are examples of projects remember we are doing. We have others on putting together an open data set for the solar energy market. We have studies by partners in Tunisia on gender-related discussions that dates in the social media and news outlets. And we have also a study by partners on the air quality, the carbon emissions air quality, layering the data over land use in Egypt.

I present these examples really in the hope of emphasizing the need for grounds-up collection of data that is open, that needs to be open, that needs to be organic, and that is devised for local applications, for suitable, simple applications that address issues and problems on the ground. So how can we encourage that? How can we address app enabling environment, and can we do it through an apolitical lens? That's an open question. How can we actually develop the ecosystem, and instead of having a brain drain, we have some AI companies actually end up in Silicon Valley, how can we develop ways of training our human capital and encouraging the businesses to stay within the country to have a brain gain rather than brain drain? And finally, how can we harness AI for the global good? And the point I raised earlier, mitigating the digital divide. So the key words is really open data for development, you know, address issues, have conversations, engage, localize, you know, coordination and collaboration. Thank you very much.

(Applause)

>> URS GASSER: Thank you. So I would have questions for you, but the question got so many votes, and I think it's a great discussion question for the panel afterwards, with your permission, I will save it for the entire group. It's about the role of traditional knowledge and some of the issues that may create, but we will discuss it later.

So it's also actually your points about ecosystem and capacity building are a perfect segue into the next lightening talk that highlights the role of education and is really reporting from the African continent about an effort to democratize machine learning skills. So it's my great pleasure to ask Kathleen -- with the clicker, can one move forward -- to join us, who actually is reading amazing work as the head of the -- what is the title exactly? Machine learning and data science group in the Nairobi chapter of Women, and you are also the head of data science at Africa's Talking. Is that correct? Excellent. Welcome. Thank you.

>> KATHLEEN SIMINYU: Now that I know how that works. My name is Kathleen. I am enjoy (?) and this is how I ended up working in the field of data science. I am the head of data science at Africa's Talking, as Urs has mentioned. I am passionate about the democratization of machine learning with a particular interest in ensuring African women have the knowledge and skills to meaningfully contribute to the fourth industrial revolution by building AI power solutions to the problems that they face. It is this passion that has led to my organizing the my yo bee women in machine learning. I will hereon refer to it as Nairobi WIMLDS.

It's been one and a half years since we started. I remember the first meetup, we had 11 people, 9 of whom were women. At the second meetup, we had four individuals present, three of whom are women, and three of us also happen to be the organizers of this community, and then one guy. I mention this just to be able to proudly say that despite the slow start, we kept at it, and now the picture you see on the screen is what our meetups look like.

We have an average of 60 to 70 individuals who attend our monthly meetups and consistently have 50% of them being women. We also have a considerable following with 1200 people who follow our page on meetup.com. Not long after starting the community, we realized if we wanted to have a tangible effect, we needed to do more. The content of our monthly meetup was and still is the speakers available and their areas of expertise, so it is difficult to curate sessions that would speak to the different levels of attendees. Beginners would get intimidated by advanced topics, and the more advanced topics would get bored by beginner content, so finding a middle ground was difficult.

So we decided to build the community online as well. We started a Slack and now have various channels on there that speak to the different levels of our membership. We have channels that speak to the needs of beginners, guiding them through gaining programming skills in R and Python, recommending online courses and material that can aid their journeys, challenging them to commit to learning goals, like a hundred consecutive days of coding, and have cohorts supporting each other through all these activities.

For the intermediate members, we have them find groups -- we have them form groups and take on data science challenges, mainly on cargo, as well as encourage them to facilitate the data science sessions for beginners in the community as well as in external events in the wider tech community in Nairobi.

Over time, our community has gained credibility as a center for knowledge and expertise in data science in Kenya, and as a result, my co-organizers and I receive numerous requests to facilitate sessions and workshops, so it was quite a light bulb moment when one of my co-organizers suggested we pipe this request to our members.

So far the results are stellar because our focus is on women, we prioritize them when making recommendations for speaking opportunities and have found that this not only reinforces what has been learned, but also builds their confidence.

For the more advanced members, we have groups that meet to discuss research papers and collaborate on projects. We also receive numerous job listings from companies looking to hire data science skills, and also somewhat serendipitously have found ourselves connecting our more advanced talent to job opportunities. My whole data science team at Africa's Talking, which is now three people, is 100% composed of talent that we found in the community.

Another initiative which I'd like to highlight is a deep learning endeavor, which is a grass-roots volunteer-driven organization with a mission to strengthen African machine learning. As a deep learning endeavor, our dual principles are to ensure that we as Africans are owners and shapers of the coming advances in artificial intelligence and to work towards more diverse representation in these fields. The three main pillars of the organization are learning and teaching, which is driven by the main annual deep-long master class that is a gathering of people across the country. The first one was held in Johannesburg, South Africa, last year. This year will be in Stellen bush, South Africa, and I am happy to squeeze in the fact that in 2019, it will be held in Nairobi, Kenya.

The second main pillar is leadership and community building. This is driven by small, independently organized events created to help build local leadership, spread knowledge further, and make those communities more visible.

The third pillar is policy and guidance. We hope to support and contribute to the policy framework around AI that will affect our continent and that will need to be developed. This is the slowest process, but through international engagement and engaging the bilateral relationships between our countries, we hope to make the underlying issues and challenges more visible.

The third initiative I will speak of is data science Africa. This aims to create a hub in the network of data science researchers across Africa by among other things providing an index of researchers and practitioners in the field of data science in Africa as well as being a leading resource of lectures and notes. They also have an annual week-long conference that involves a summer school to promote technical skill as well as platforms for researchers to present work demonstrating the application of data science and machine learning techniques to problems relevant in the African content.

I chose to highlight these three initiatives only because they have directly contributed to me, Kathleen, the data scientist who stands before you today. If we zoom out of my personal experiences, initiatives like this are present all over Africa. Data science Nigeria's vision is to accelerate Nigeria's development through a solution-oriented application of machine learning in solving social and business problems. The Lagos women in machine learning and data science community does much the same as we do with the Nairobi community.

What is my point you may be asking yourselves. The first industrial revolution is upon us, and Africa does not want to be left behind. Not many people have specialized training in the field, but these communities bring together individuals who are mostly self-teaching, creating room for collaboration and further expiration of solutions to the problems we face.

Some of the lessons learned on this journey, the first is that intentional, targeted, and sustained efforts work. In April of this year, the Nairobi WIMLDS community hosted an event that brought together stakeholders from government, academia, as well as industry. We had a national task force on blockchain and artificial intelligence, a professor who is also a former ICT permanent Secretary, was in attendance and commented in an article he wrote about the event that unlike most technical events he has attended in the the past where men were the majority, our attendance was evenly split between men and women. This did not happen overnight. It took 1.5 years of work the community has been doing to achieve that.

I silently give myself a tap on the back every time I mention this, so allow me to do that now.

(Laughter)

Stories like these are what I am most proud of.

Another tangible outcome is African attendance at the neural information processing systems conference, popularly known as NIPS, in 2016 versus in 2017. Of the approximately 300 people that attended the first deep learning endeavor, 20 went on to attend NIPS in 2017; whereas, in 2016 only four African countries from presented at NIPS, in 2017, there were 31 African countries present. Combining efforts makes for wider, more impactful reach.

The Kenya event which we hosted allowed us to connect with government stakeholders, something we had previously not been able to do and has opened the door for us to contribute towards a policy framework in the country.

Now on to some of the challenges that we face during this work. The first is technical capacity. Like I mentioned earlier, not many people in Africa have specialized training in the fields of data science and intelligence.

Most of niece are volunteer driven and most volunteers have full-time jobs, a lot of this work is squeezed into our spare time.

And finally, funding. These three factors are the main hindrance to the scaling of our activities.

On opportunities, I cannot help but wonder what a coalesce sense of all these efforts would look like. The Nairobi women's collaboration with the deep learning endeavor to host the event saw us gain access to government stakeholders and expand the effects of our work. What if instead of several grass-roots efforts in different parts of Africa we had an African body that brought together and helped intentionally drive these efforts? What if?

And finally, because AI is nascent more so in Africa, we have a chance to put emphasis on several best-case practices. Having teams -- put a very strong emphasis on ethics and inclusion. Thank you.

(Applause)

>> URS GASSER: I am sure you will get many more taps on the back for the great work you are doing. One question for you, though, from the audience. Is there one data set that you think people should be aware of outside of your community that the world should know about that you or your group is working with?

>> KATHLEEN SIMINYU: I cannot pinpoint one, but there is a gentleman from our community to find out what people buy. They built a database from receipts thrown away to see what people buy, when they buy them, where they buy them, stuff like that. It is very interesting to see the insights that come from that and got us thinking maybe we should create a repository for some of these data sets that are created locally.

>> URS GASSER: Super interesting and fascinating. Thank you so much.

(Applause)

All right. So we are coming to our last speaker of the lightening talk series, and we take yet another step back and arrive definitely at the ecosystem level, looking at the ways Big Data can actually inform better policymaking at the economic level, in other words, how can we use Big Data to better understand the digital economy and also kind of to shape it.

So Silja Baller kindly agreed to share her experience and perspectives with us. Silja is a practice lead at the World Economic Forum focusing on the digital economy and innovation, is doing amazing work, bringing different stakeholders together, and actually also has a lot of experience, as we will hear, considering some of the issues that came up earlier around metrics, and what are even ways to measure this changing world in which technology and data play such an eminent role. So thank you for sharing your perspectives.

>> SILJA BALLER: Thank you, Urs. Good afternoon, everyone, and thank you so much to the organizers for the invitation.

Like Urs was saying, I am an economist with the World Economic Forum, and let me maybe start out by giving you a little bit of context on our work. So some of the recent work that we have been doing at the World Economic Forum looks at the challenges and the exciting opportunities at the intersection of emerging technologies and economic policymaking. And we are convening stakeholders from technology and economics and also other relevant and related disciplines to broadly work towards four outcomes.

So on the one hand, we are trying to create a common language around issues. Sorry, this is actually not -- yeah. So on the one hand, we are try to go create a common language around emerging issues. We are also working on converging conversations, where the debate is very polarized, and try to create consensus amongst the different stakeholders. Then in areas where there is consensus already, we are working on building partnerships. And finally, of course, there are still many knowledge gaps in this area, especially when it comes to economic policymaking, so we are also try to go catalyze new research partnerships that address some of those knowledge gaps.

Now, one very important area where emerging technologies are intersecting with economic policymaking is, of course, the measurement of economic activity, which includes tracking economic aggregates and trends, like market transactions, for example, and prices, but also things like consumer welfare and distributional outcomes.

So to link to the conversation today, what I am going to try and do is to frame some of the questions related to the ecosystems and related to collecting national statistics. So to provide a very high-level view. And I guess we are still at the very beginning of this conversation. If you look at how national statistics are actually being collected today, a lot of it is still being done simply by administration of surveys and a lot of in-person interviews as well with people actually going to stores to record prices, et cetera, and of course, there's huge opportunities that are arising from all the emerging technologies to complement this data and these kind of statistics and eventually potentially replace them.

So in what follows, I am going to highlight some of the exciting approaches that are out there that are being explored from across the economic policy spectrum, and what I want to emphasize here is really the great heterogeneity of approaches and of data sources that are being considered for policymaking and the real shift in thinking that this will eventually require. If you think that we are starting from a system where we have one means of data collection through census and survey to then eventually being able to navigate an entire data ecosystem. And there's also important issues in terms of coordination to shift towards this new data system.

So maybe quickly, what does the current data landscape look like for economic policymakers? I think there's two developments that are important when you look at the intersection of Big Data and economic policymaking and measurement. On the one hand, Big Data, when it comes in the forms of services, is actually making measurement a lot more difficult for policymakers. So if you think of digital services that are being consumed freely or that don't really have a market price, this raises big questions in terms of how we capture this in measures such as GDP, and it is actually throwing a banner in the works of policymakers who are using traditional metrics.

On the other hand, Big Data is also helping policymakers in many ways by complementing traditional metrics, and I think there are several dimensions in which this is helping. On the one hand, digital data sources offer much higher frequency, so we can move a lot closer to real-time policymaking, and this is also important as things are shifting very quickly underneath and the value of historical data is diminishing.

Big Data is helping us get much greater granularity on certain issues, so both at the geographical level and at the individual level.

And finally, you have much greater reach from these kind of data sources. So you can reach further into the informal sector, for example, and this is especially important in environments where there's big gaps in national statistics.

So just to illustrate, some of the ways in which traditional statistics can be complemented with these exciting new data sources. So for example, today, if you look at prices, as I was saying, a lot of times this kind of data is still being collected by people actually going into shops and doing surveys. There is an intermediate step where people are using scanner data, but there are much more exciting opportunities today to actually use AI to scrape company websites to get real-time movement and data. There's some really exciting work going on at MIT, for example, with a billion prices project, and it's been shown that inflation measurements based on this kind of data are much quicker to detect trends. And it also provides a lot of times alternatives to government-reported data.

And in terms of the accessibility of this data, and since we are talking about the data ecosystem, this data is publicly available from company websites, so here no real barriers in terms of the access at least.

Then a second area that has concerned economists for a very long time are growth statistics, which so far we've been basing on GDP growth. But recent technologies are opening up very exciting opportunities here, namely if you look at satellite data that can tell you -- it gives you just a very different view on what's the economic activity on the ground by using nightlights, for example. Of course, there's huge infrastructure involved in collecting this kind of data, but a lot of it is being made available publicly through process data sets.

Another area that's of great interest to economists is our distributional issues, and so far with distributional issues, we have been mainly basing our analysis on income data. Now, if you look at financial transactions data that's being digitally collected, you can get a very different view on things by looking at the expenditure side, and it helps you answer important questions in terms of access to goods and services, for example, so there is a study that was done lately -- recently -- that looks, for example, at where different income groups spend their money on essential services, and it was shown that through this kind of data one could show that low-income groups have to travel much further to access basic goods and services.

And then finally, one very exciting area is welfare measures. So far, I guess we've been using GDP as a fairly rough proxy of welfare, and again, new opportunities opening up. One of them being the possibility to ask consumers directly how much they value digital goods and services. An example is Eric Brune Hoffen's work on massive online surveys to get a better grasp of this dimension. But then, of course, with everybody, every consumer or a lot of consumers having mobile devices, there's a lot also huge potential in terms of getting data directly from consumers about their well-being and their welfare rather than having to go through a very imperfect proxy of GDP.

Now, in terms of the challenges, I think if you are focusing mainly on online data, of course you don't necessarily get a random sample, so it's still not a perfect replacement for offline surveys, especially given the huge digital divide that we still have. And so one has to be very careful not to bias against the offline population, and the example would be for the price tracking, I mean, this is only going to work if on average online prices are actually the same as offline prices. As soon as that's not the case anymore, you are potentially introducing bias against people who WHO are consuming only offline.

Then as I said, there's big advantages in terms of having highly disaggregated data at the individual level and also at the geographical level, but of course -- so this is great for policymaking, but of course, every time you have high-level Big Data, there is issues in terms of privacy and ethical questions that need to be considered. The data is potentially very resource intensive to collect and also to analyze, and it's not necessarily resources that national statistical offices currently have. And if you look at trends in funding, it doesn't look like there's a huge boost in funding to national statistical offices, so this is also something that needs to be kept in mind as a potential challenge.

Eventually, if we are really thinking about this as a replacement for national statistics, there is going to be a need for huge international collaboration on new protocols and standards to really shift to the new system.

Just very quickly, on the way forward, I think there is great opportunity in these kind of data sources to get a very real-time picture of the economy and to get a much closer picture of welfare. There is a great opportunity to draw on the entire data ecosystem and to really tailor data collection to the exact policy question at hand. This is not a luxury that economists used to have with very imperfect data sources. Then finally, like I said, in terms of the national accounting, I guess the -- in terms it of the way forward, the biggest challenge is going to be to make the shift to the new system. Thank you.

>> URS GASSER: Thank you very much, Silja.

(Applause)

I have a question for you, but I would like to ask the other panelists to join the question again while I ask the question. So the -- please, if you just join in the meantime.

The question is the real-time data that's becoming available to policymakers may also have a downside in the sense that it may drive towards short-term thinking or looking at the trees rather than the forest. Do you think there is some possible downside of having this highly dynamic real-time updates with respect to that sort of thing?

>> SILJA BALLER: I guess it could potentially shift the focus, but the other data Stowerss are still there --

>> URS GASSER: So it's a complementary approach?

>> SILJA BALLER: I would see it as a complementary thing, yes.

>> URS GASSER: Thank you very much. If you could have a second microphone for the panel, that would be great.

So there's been lots of questions, and one question really stance out -- actually, one set of questions really stand out -- and that is the problem or referring to the assumption to say use data for the good, but that suggests somehow that the data is there. Ed was optimistic in the Haiti example under lessons learned saying well, even if the data is scattered but there is something to work with, but there was a little bit of pushback in the question tool saying well, on the one hand, quite often we have data sets where the most vulnerable individuals are not represented in the data set itself. So when we are talking about using data for the good, aren't we somehow just amplifying the existing inequalities as we look for new policies and solutions?

Similarly, there was a question related to some of the examples about -- and stories we've heard about traditional knowledge, where we also have maybe a problem that traditional knowledge data of sort is not captured like other types of data that may be put towards kind of these new technologies. So how do we deal with these existing inequalities and avoid the pitfall that we are all, you know, excited about these data sources that we increase actually inequalities as opposed to reduce them? Who would like to start with that problem? It also, of course, applies to the question of scientific knowledge and what's published and what doesn't get published, so it's really a common theme, I think.

Ed, do you want to start?

>> Sure, I can just jump in quickly there. I think at great point that when we use data, we want to be very careful of any biases, as well as are we ever at a point when we are increasing inequality?

It's interesting when we work with some of the data providers, they have actually come to us and said we think we know where the poor are, and we have poverty maps. Can you help us to ground truth it? Actually, so at the World Bank, we are committed to still doing our traditional work, which we go out, we do poverty surveys, we do household surveys, and we think that's a very important component of our work because we then use this to ground truth other real-time sources of data to check does it capture all of the poor? So I think that's an important part of anything that we do at the World Bank is how confident are we in the data, and you know, at the World Bank, we have a lot of economists who are constantly getting questions. I think it's a great point for us to continue to be aware of. Yeah, so I think that's our position at least.

>> URS GASSER: Great. Thank you. Nagla, you want to comment on the traditional knowledge question particularly?

>> NAGLA RIZK: Sure, and if I may also very quickly follow up on Ed's point, I think I agree with what he said, and this also emphasizes the importance of finding alternative means of gathering data and layering the data and comparing the different sources because quite often you will find if you use novel technologies, there are things you are able to capture that may not necessarily come from the

traditional mainstream collection methods. So this, actually, I was thinking about the traditional knowledge, and I am sure you all know traditional knowledge is a subject of other debates, like intellectual property, for example, the protection of intellectual property with respect to traditional knowledge. One thing is India does have a traditional knowledge Digital Library, so I was thinking about this after my talk, and I guess there are two sides of this. On the one hand, the importance of finding a way to document this knowledge, you know, how can it be documented? How by means of collecting pictures, stories in native languages? On the other hand, perhaps we are looking at it from the perspective of today's AI technologies. Maybe in tandem the technologies themselves will be developed in ways that I don't know, we don't even know. At the same time there may be novative ways of capturing that wealth of knowledge. That being said, it is definitely a very important subject, and it is a wealth of knowledge production, especially in the Global South, that actually goes into metrics of innovation. That's also part of our work had a highlights the wealth of knowledge and innovation in Africa and other parts of the developing world that is not captured in mainstream measurements.

>> URS GASSER: Sam, this goes right to your passion, the knowledge representation issue. Do you have any comments or thoughts on this one?

>> SAM MOLYNEUX: Absolutely. So one of the things I would say

is we have to be prepared to look for funding and then go and create projects to create sometimes incredibly tedious-to-assemble data sets. In our case, we will create projects around every sort of aspect of annotation of data, we'll judge the scale of data that we need to produce, we'll look for a grad student population or other populations who have enough specialized knowledge to be able to participate and label data creation. I think a lot of the value that you unlock when you create these systems is really, really mediated through sort of like these initial very high-quality data production projects. And in some cases, it may be quite difficult to obtain the data. And so designing for that, designing -- sorry -- doing good system and experimental design up front, acknowledging sort of where bias would be if you use the easy-to-get data and then going and defining a project around creating label data is essential, but it really is -- you know, that's the root of the value production in the project, I think. So I expect people should be willing to fund that as a defined project associated with any one of these types of projects.

>> URS GASSER: Thank you.

So I would suggest unless you want to weigh in -- directly or --

>> I was just going to add that from an industry perspective, I think this is where multidisciplinary teams are important. Because as a techie, I may just go on the data, but someone who has more domain knowledge would point out that there is, in fact, a bias. Then we can begin the process of making up the gap.

>> URS GASSER: Excellent. Thank you. Let me open up right away. Do you have a question you would like to ask or a comment to make or a story to support some of the themes we've been hearing? Anyone? Yes, please. I think you have a mic there. I assume they work.

>> Hello.

>> URS GASSER: Yes, it's on.

>> Hi. So I had a question for the -- for Ed on the World Bank, the project in Haiti. So in terms of the mobile penetration in Haiti, can you give us some numbers, and if that number is not high, then isn't the redesign project excluding the poorest who don't actually have access to the device and admittedly would be the ones who are most affected by it?

>> Yeah, I don't have that number right now, but let me get that for you.

>> URS GASSER: Someone can Google it or Bing it maybe in the meantime.

>> The other point I was going to make is that mobile penetration is often very high, and this is actually one of the big advantages of where we are in the world today, where we see just the amazing increase in the number of mobile phones that people have. That enables us to get much more accuracy. But I know that when we looked at it, we did try to correct for that. I think it's very important that we use standard sort of econometric models to try to adjust for how much of the poor are we actually capturing and how do we, you know, adjust for when -- when we know the poor especially don't have

those phones and then how do we find other methods, then we have to go through more traditional methods. But it can be much more cost-effective when you start with, you know, if you can get 70% of the population, you know you are still missing 30%, but at least you've gotten that part, then you can find other methods to get the other 30%.

>> URS GASSER: Excellent. We have a question right here in the front row or a comment.

>> Apparently penetration rate was 62% as of 2013, a Latin American inventory, which might be broadly applicable. It's certainly changed since 2013, and it might be closer to the 70% that you just mentioned. But while I've got the mic, I would also like to ask a question either appealing to the panel or anyone in the collective audience. It's got me thinking about data literacy among people who are collecting and processing data, and whether or not there are tools, inventories, or even an AI solution through which you can input data sets and their corresponding metadata, and adapting based on their usability, robustness of that. If not, is anyone working on that?

>> URS GASSER: Thank you. You have a response to this question? Okay. Are there responses to the tools question? Anyone? Anyone in the audience has the answer to the question? Or knows about the tool or approach? If not, okay, it's part of the list of open questions we should work on. Great.

Yes, please. We start with you, yes.

>> (Off microphone)

>> URS GASSER: You have to press again. I can give you my mic.

>> Yeah, my question is about data integrity. We are building a lot of models based on data, and you know, a lot of these models, once they are built, we use them to make conclusions; correct? And as machine learning, person who has done machine learning, these are black boxes. You put the data in, and you get conclusions. So what are we doing to ensure that there's integrity in the data that we are using to make these decisions? Not at just the collection level, the curation level. And there's a lot of data, so it's a challenge, I think. I don't know if there are any tools or methods that are currently being used to do that.

>> URS GASSER: Who would like to take that? Sam?

>> SAM MOLYNEUX: I will comment quickly. We are starting to design quality control in the production process, both in sort of -- so label data set production, have multiple people annotate the same data so you end up with internal standards. You either have internal standards or you end up with replicates. And then quality control sort of like downstream after the data's been processed, and then quality control in the application. We are starting to build these in more and more, and at each stage, we want to have an expert, so maybe somebody to do with a PhD or a masters in, like, the requisite background, whether it's immunology or some other type of information we are providing to be able to even at a small scale give us some data on that, and we would like to scale it up over time. That's

how we are thinking about it.

>> URS GASSER: Thank you.

Unfortunately, we have time for only one question, and this will be yours.

>> Thank you very much, and I'd like to hear from the panel what your personal experiences have been with governments, especially knowing that, you know, data is information, information is power, and governments will want to come in wanting to say that they are protecting their people? What has been your experience with governments and even legislators and legislators trying to catch up and legislate on data management? And should governments even be legislating on data in the first place?

>> URS GASSER: That's such a great question that we will turn it into kind of a last round question, and your challenge will be to give a tweet-length response to that question.

(Laughter)

What's the role of government when it comes to building infrastructures, projects that allow us to use data for the good?

>> I complicated, but I will tell you very quickly, I think the best way is to get them to buy in, so make sure to present the story as it is a citizen participation data and present it in an apolitical context as much as possible. That's my experience.

>> URS GASSER: Thank you.

>> I have more, but --

>> URS GASSER: That's good. That's tweet length. Perfect.

>> I mean, I would say from our perspective, this is really where the multistakeholder approach comes in; right? It shouldn't be in the hands of one entity, but you should listen and consult with all different groups in society, and I think that's going to get you the best outcome.

>> Thank you.

>> Government is not one thing, so depends which part of the government.

>> URS GASSER: That was a retweet, anyway.

(Laughter)

>> We've had little experience, but they have been very encouraging of our project. That's all I can say.

>> URS GASSER: Very good.

>> I'd say governments need to put emphasis in data protection education as well as implementation of policies.

>> URS GASSER: Excellent. Ed?

>> At World Bank, we think data -- I mean, personal data is critical, actually, for expanding government services. We know that 1.1 billion people don't have ID today, or are unable to be identified by ID. We see this as critical. Governments are catching up. But we need best practices on how to protect the poor, enable services but protect the poor at the same time.

>> URS GASSER: Excellent. Thank you so much. We have, of course, a lot to continue to discuss over the next two days, so these questions and initial inputs, as I said at the beginning, are just the starting

point. We have opportunities, again, to go much more in depth tomorrow looking at specific use cases. We will come all back together on Day 3 to add to the stories we've heard, to build upon some of the insights that you all shared today, so there's much more to come, but hopefully that set the stage and stimulated our brains that we can do this work.

I don't want to miss to thank Elena Goldstein on our team for doing much of the heavy lifting for this session. So thanks for getting us here. Thanks to the panelists, thank you, everyone, see you soon.

(Applause)

I think now it's back to the main room for the plenary session, I guess.

(End of session, 17:25 CET.)

***

***