Transcript

Title: Webinar - Social Data in Action - Road Testing Community Data CO-OPs
Creator: Prof Jane Farmer, A/Prof Anthony McCosker, A/Prof Amir Aryani
Year: 2021

Audio/video for this transcript available from: http://commons.swinburne.edu.au

JANE FARMER: Thanks, everyone, for being here. Obviously some more people will trickle in as we move along here. Welcome to our fifth and final webinar in the Social Data in Action series. Hopefully you've been to all of the other ones. And Amir, this is the culmination, so we're looking for something fabulous-- no pressure. Yeah. So it would be great to see your faces if you feel brave enough at any point, and definitely at the end of the Q&A session. So let's move on Paul.

And so, I just want to acknowledge that we are hosting-- or I am located and hosting this webinar from the lands of the Wurundjeri People of the Kulin Nation, and I also acknowledge the Traditional Custodians of the various lands in which you're all working today and the Aboriginal and Torres Strait Islander people participating in this webinar. And I pay my respects to Elders past, present, and emerging and celebrate the diversity of Aboriginal peoples and their ongoing cultures and connections to the lands and waters.

So Amir's going to speak for 25 to 30 minutes. We do encourage you to stack your questions into the chat as we go along. And obviously, we'll call for questions at the end, which we'd love you to put into the chat. And Anthony, if he ever makes his way through Zoom and into this webinar, will be managing the Q&A at the end. And we will be recording the session. So if you have any challenges or problems with that, please get in touch with Paul-- paullavey@swinburne.edu.au.

I might start to introduce Amir. So Amir Aryani is the head of the Social Data Analytics Lab in the Social Innovation Research Institute at Swinburne. The lab applies contemporary and emerging co-operative data analytics techniques to provide insight into health and social problems. Amir is a computer scientist by background, and he has worked with illustrious international institutions, including the British Library, Orchid, the Institution for Social Sciences in Germany, and on projects funded by ARC, NHMRC, and the National Institutes of Health.

I just wanted to note before Amir gets started that Amir is a great example, I think, of how community and social data projects and innovation benefit from a mixed team with a mix of different inputs on the team. So if anyone was present at Sarah Williams's talk recently in this webinar series, she noted how project teams should have data scientists, social science specialists, and lived experience. And a lot of her projects have community organisations involved in them.

So in this way, high quality social data projects become a space where the vital facets of knowledge are melded for innovation and insight. And so today, Amir is going to focus on our innovative work with community data co-ops and collaboratives. I'm now officially handing over to you, Amir.

AMIR ARYANI: Thanks Jane. Thanks for that very nice introduction. Let me get the technology. Can everyone see my screen? That's good. Thank you. Thanks a lot. Sure. Thanks. So following what Jane said, I'm planning to do in the next 30 minutes give you a bit of background about Swinburne's venture into the data cooperative projects.

But also, I will tell you kind of an overview of the tool box of data science tools that we have available at Swinburne. And also, this is a partnership with four other universities. A lot of these tools and platforms that we're going to talk about, they are available-- some of them-- in your institutes directly, or indirectly through your collaboration and kind of collaborative links.

The concept of data co-ops has been something that's been around for a while. But the concept of data co-operatives, basically. But in the context of this presentation, we are talking about the concept of data co-op more abstract as an element of bringing different data collaborative and data co-operations from different groups together. And from that point of view, it's not just about data cooperatives as what we as a form of organisation, it's more about the concept of actually enabling collaboration between groups, teams, and communities.

Now, to start, let's talk about the couple of background things of how we got here. As Jane mentioned, my background is computer science. I have a lot in a different [INAUDIBLE] with a scientists from the [INAUDIBLE] physics, to chemistry, to biology, to social science. And now more than ever, we have access to the advanced data analytics capabilities in different fields.

Social science and humanities are the one that are kind of in a very interesting position. When we look at the commercial sector, there is a lot of data capabilities out there. When we look at the research sector in these domains, there are a lot to be desired. And I'll get to this in a moment.

But in a commercial space, data analytics and AI now is-- well, in the past, it used to be a game changer. Now it's essential component. In 2018, I used to have this quote about, in five to 10 years, AI will be integrated part of a lot of different systems. Given the pandemic, a lot of those things has been substantially accelerated, and a lot of those things is already in place.

The biggest changes have happened is, the unstructured data previously working with that was a big challenge and information was in silo. Now, they are all interconnected. If you scan a coffee cup image on your phone, the AI system underneath-- it knows where you are standing in the coffee shop in a shopping centre. It knows you have previously purchased coffee from that coffee shop.

And all of these elements come together to tell, OK, I know exactly this a Starbucks coffee cup. So provides a very, very high accuracy given other connected information. So there are two transition of unstructured data to the structured data and disconnected data to knowledge graph is now embedded into a lot of commercial platforms you're using.

The other concept that more and more get traction is a concept of augmented intelligence-- that is, bringing AI to the point of actually being an active assistance in a lot of day to day activities. The best kind of example of how this operates is, not only AI drives your car, AI would tell you that you are going too fast. You need to turn right. You actually-- have you paid attention to the roadblock ahead?

It knows about the different climates, and if the road is raining, and basically provides substantial assistance to the operation of the car. Same thing can happen in organisation-- already happening a lot of commercial sector, in supply chains, in transport, in a lot of fields that enables effective decision making.

Now, these are all in a big corporate space. But also, a lot of activities happening in the social and urban organisations. Across the globe, we have initiatives like Nesta in Europe that works a lot with the platforms like collective intelligence and open data platforms. There's a lot of effort around the data collaboratives in the United States. And we have the urban institutes, the DataKind, and also in New Zealand we have the Centre for Social Data Analytics.

In Australia, this Social Data Analytics Lab, or SoDA lab, we start doing work with the not-for-profit sector in creating similar capabilities. The main driver was a lifting up the data literacy and data capabilities in the sector. And that also signified that well, that is apart from the data skills, there's also lots of infrastructure components and the governance elements are missing. And that has been the motivation behind the whole data co-op platform.

Now, what is data co-op platform? It's basically the methodology based on the idea that to create effective data projects that make a change, we need data-- yes, of course. But that's not the only thing that we need. We need people. We need domain experts, data scientists, we need researchers. We need people who can actually transform data to actionable insights.

And they cannot do it by themselves. They actually need analytics capability. So that's where usually the university come to play, and sometimes commercial providers. By the Centre point of a data co-op are actually people who make a difference, who make an impact, using all the insights that they derive from the data sets.

And during this course of projects that we have done in the last couple of years, we kind of built a model. If you need to do a trusted data partnership project that leads potentially to a data collaboration, you need infrastructure that supports this. And that infrastructure talks about things like a data storage and data access, talks about artificial intelligence, it knows capabilities of how to dispose of sensitive data after finishing the project.

We have the infrastructure but we are [INAUDIBLE] data governance model. That is what we're dealing with the ethics problem, is how do we actually manage the data lifecycle? Answering to the questions like the risk management, five safes model, how do we actually make our data findable, accessible, interoperable, reusable data set, or FAIR data. And this are the operation pillars, if you like. It you like up from the concept of trusted data partnership.

When you have them in place, then you can actually focus on the creating the data collaborative projects. And in that layer, in the data collaboration with a number of different initiatives and concepts, to name a few of them here, like the knowledge transformation collective intelligence-- this is the space that you can actually look at the data communities, and also, data cooperatives also goes into that layer.

Now, I thought this is actually going to be useful to look at different perspectives of communities regarding the concern around data sharing. In 2019, we ran a workshop in Canberra with a number

of different government departments, not-for-profit sectors, researchers-- the main question on the table and the roundtable discussion was about the responsible data sharing. And we had in the room researchers who needed access to the data. But we also had a lot of data custodians, data providers.

One of the things came out of that conversation, as we recorded, transcribed, and then analysed the text later on. And we also in-- it was a two day's workshops. So on the second day, we went back to the result of the analytics, was there is a lot of discussion around the data access.

And this is not complaining about we want data, we don't actually have-- we can't get access to the data. It's just kind of other way around. A lot of data providers, they wanted to get their data reusable, usable, drive value from it. And they wanted to provide data to the resource sector and the commercial sector.

The problem is, there is always this kind of shroud of doubt about, how do we actually make data reusable in an ethical way, and how we can actually drive value from data without compromising the privacy and security. So that has been one of the topics of conversation.

There was a lot of interest of making data as a first class citizen of the research and science. No roadblocks to make that happening. And then, when you are going to go past the ethical conversation around this, there's a lot of discussion around the governance, and data linkage, and data value by both sides of this conversation.

Now, if you think about the government data, at the moment you have the five safes model in place. This is the model-- it's a recommendation by data commissioners in Australia, which basically says, any data projects that wants to access the government data should answer the risk assessment question around five different pillars or components. The first group is that, is it a safe project?

So you have to justify that the project you are doing is a safe project. The second thing is that are they the safe group of people who are running this project? Then, there's the safe settings. That comes back to concerns some data sets cannot be stored in servers overseas. Then, there are the data security. This is kind of like, if everything goes fail, what do I need in a data to protect itself? Do I need the identification, encryptions, and so forth? And then the output. Who's going to manage the output of this?

Now, this is a very good model to think about what can happen with a lot of our research projects in the university sector. This is not currently applied outside the government. But there's a lot of appetite for kind of expanding this model to the not-for-profit, to the commercial sector, and to the education sector. This is also a very good segue to look at the risk assessment model. The risk assessment-- a classical way of managing a risk of the data project usually comes from the likelihood of something bad happening to the severity of the impact of that problem.

And this is kind of a classic model. When we look at all data projects, we look at the data and output, both of them as the content that we need to manage the risk for that. And then, there's the use cases and the settings of their project, and the operation of this, and also the content of the project operation. And also the people who are involved. So we can look at the five safe concept in this way. And a lot of research projects enable that conversation about, is it a safe project to go ahead?

Now, with all of that background, that is where we got to the idea of the data co-op platform. So we knew that-- one of our operation at Swinburne initiative was based on the core design model. And that was kind of a recipe to success for a lot of our projects. That was also the beginning of the concept of the data co-op. So we said, we want to create the trusted data partnership, and we want to create value out of it.

We establish iterative model, that in that iteration we were actually reading information from different sources, from the government data, to the not-for-profit sector, community data sets, and the social media. We had a model for running a number of different co-design workshops.

When the data was actually answering specific questions, we are workshopping these with the community and project partners. We will getting their feedback, we send the feedback to the data engineers, and we basically, from that conversation, we produce data products that later on lead to insights. Now, this process we will repeat in these workshops again and again, in order to basically provide the richer insights, and more impactful actionable insights.

Now, the idea of the infrastructure was create a platform that enabled this, even all of the things that I mentioned, as a requirement of a trusted data partnership. And this transformed to a funded project by [INAUDIBLE], so seeing them as leading the brand. AMU, Griffith, University of Melbourne, and UTas are the partners. And we built a platform-- I will take you through that-- that basically provides the capability for enabling these data co-op projects.

Now, in the next couple of minutes, and I want mainly focus on the data infrastructure activities. Because these are kind of new components, and a lot of previous webinars, we've talked about that in and out of the data governance and the challenges around collaboration. But the data infrastructure is kind of new from our kind of operation point of view.

We almost finished developing a lot of components, and now they are at the point of providing service to our projects. This part is the one that kind of-- if you like-- are the shiny objects in our toolbox. We are going to talk about artificial intelligence, data visualisations, and some of the other elements that enable these components that come together, such as data access and data linkage model.

Now, we have the hybrid data co-op infrastructure today, based on the needs and requirements for a system. I know this is looked like quite a techie perspective of the whole different thing goes together. This is actually one of our interesting-- that's now one of our internal documents. And there's a tool we are using that transform 2D to 3D. So we have a 2D version of this we already use for a kind of status checking of our servers.

And then, this produced this nice 3D visualisation. But it also has some important essence of how the operation divided to two different layers. So we have the social media layer of data that we are managing, and running, and collects continuous information from the media and social media. We have the public data layer, as gateways to main government and education sector data sets. Or is [INAUDIBLE] in platforms, the ABS data sets, the data.gov.

And then, we have a secure data layer, that predominately is running on Azure cloud. And that's been quite a good instrument to basically enable a lot of data co-op projects to get running. Which I

will mention this was quite essential to actually create, curate, and hold data in a secure environment that can be useful producing insights.

Without going through the detail of this, the main function of a lot of these boxes is that make the unstructured data to the structure data, and collect that information together, and provide a tools for us to create data insights out of the mishmash of ideas and a lot of different disconnected information.

For example, one of the things that we have in our secure data space, is that we are plugged into the cognitive search from Azure. So when we get the audio files, or we get the PDF files, they basically transform those information to text, and then provide knowledge graph on the content of the text. So that provides analysable material out of the unstructured information.

And we also have a secure space where we actually get information from our clients or partners, we can actually store them in information that is disposable at the end of the project or securely archivable. Depends on the process that we will confirm in the ethics.

Now, a lot of these infrastructure that the [INAUDIBLE] produced two main front end. Either we produce analytics dashboards-- I will show you all of those. Or we produce Jupyter notebooks, that basically shows exactly what information has drive to what kind of insights. And how we actually can walk back into those process really in the interests of reproducibility of the science, but also for any fact checking if you need to know exactly how they got to a given number.

Now, a data insight that we can get out of this system usually connects to multiple different elements. So it is linked to original data source. It links to a software that actually drives the data inside. This exactly tell us how we got from the data source to data set. It produce a transformed data set, that often is a result of the work.

And I'll show you some of these kind of transformation in a minute. It connects to the organisations that are linked to that data in the site. It tells you what other publications are linked to this, and who are the researchers. And in our ecosystem, everything is linked to Orchid and [INAUDIBLE]. So this is quite a good transparency of this connected graph.

Now, these are some of the examples of public insights. One of the things I mentioned, if you remember, everything comes out of Azure. It's kind of a private data insights. So it's come from the private data sources. So I'm not going to present those. I'm just going to go for the public insights that from the course of our data projects has been useful to our partners.

So these are some of the examples that you can see. They are actually on open source code. You can go to the GitHub, and you can see how the code it runs. For example, this is the insight we derive from the AIHW public data set. It relates to the mental health services. As we know, more than 38% of Australians in some way actually contacted mental health services during 2018. Now when you look at the insight like this, you know exactly the way the data came from. So you know the source of data.

You would see the transformation of this. You see the visualisation. And basically, it's not just a text. It is all the steps that takes you to that fact. And that is quite important for a lot of projects. As the

times goes on, we often would wander around, and well, how did we got to this statistic? And this process actually answers that question.

Now, as exciting as Jupyter notebooks are for data scientists, is not always useful for everyone. So to make it more useful, we have the visualisations and data dashboards like this. It's one of our other projects with Bendigo. And same data sets, just transformed to a data visualisation. This is run on top of the PowerBI, and again, connects to our Azure infrastructure. That's where we see the access of people to internet in Bendigo.

And basically, this tells us the story that about 20% people that are in the greater city of Bendigo, they have no access to the internet in 2016. That's based on a survey done in that time. But you can also see this in different suburbs and different areas in Bendigo on the right. So the graph that you see on the right, the red bar is the number of people who don't have access to the internet, compared to the kind of red and yellow to gather the total population of that area.

So this is much more tangible for people in Bendigo. And when we run workshop, they actually work with this. So it's kind of like there are two phases in this one. The Jupyter Notebooks as a quick way of producing insights. And this is the one that we take full of to a lot of our workshops.

Now, thinking about the-- I initially promised to talk about AI. So it's just-- this stuff getting more technical. We use AI to actually tell us what data sets are actually linked together. For there's so many ways to link information together. One of them is that using their place based concept. There are different social variables or incidents that happens in different areas that AI can actually find correlation. This is a lot of different social variables.

And when you see blue here, it tells us people or characteristics of two different variables live together. For example, the chance of having a high income and having three different cars. Yeah, maybe. How about home ownership and having two or three cars. What about your jobs and a chance of renting? And this information, using the power of AI, can be calculated in so many different ways in all different kind of spatial lenses, from the suburbs, to states, to country, to different regional areas.

So you can basically look for all the different permutations of potential connections of different communities together. And we have done this in a kind of bigger scale. I'll show you the result of that. But potentially, it provide us some insight, like when we were doing a projects in the city of Glen Eira.

At the time we find very interesting patterns between the different industries that people work in, and different kinds of dependencies that they had to different social services, and the number of cars that they own, or their ownership of their properties. And some of those may or may not produce valuable research outputs. But for the policy makers and actual not-for-profit organisations in that area, they're quite insightful in providing the right services to the right people.

Now, when we put this into action, this is kind of a picture of Victoria. And we're using the same model. We kind of ask AI to tell us what are the communities living here. And AI kind of gave us this colourful pictures of different groups. We didn't know in the beginning what they are and who they

are. We give them some names, like the green ones are-- we call them CVD, since most people living here are in the high rise buildings and are renting.

A lot of them are students. But they are not just in Melbourne and Sydney. And in the boxing areas, on the right side of the map, you can see there are other CVD type of structures with those green boxes in that area. I just want to take your attention to two of the big major communities. We have identified the orange ones are the sort of area that have high incomes.

Families living there are-- mostly they are in the age range of 40 to 54, and then they have teenagers between the five to 19. Unlikely that you find people in this area that are 25 to 34. They move to other suburbs. And they are unlikely to have moved recently, so they're established. They have high education. They're professionals, and they're in managerial jobs.

When you look at the other arm, the purple one, all population almost double the orange area. This is a lot of area full of high level of immigrations. People came from overseas. And compared to the orange area, they have a lower median income. Also, they are less likely to be aged 55 to 69. And they're more established, [INAUDIBLE] to move their house, but they're more likely to rent, and interesting enough, they're less likely to engage in volunteer work.

So there are lots of different variables that you can look for people living in different areas. Now, this is just one slice and dicing. There would be many other ways that we can do this with different focuses. And also, this is just one data set. So there are different layers of data set you can add to basically achieve this kind of computation.

Now, we transform similar information to 3D visualisation. This is one of the capabilities we have in the data co-op platforms. It get the data and basically produce 3D maps. We have a pipeline for it. The data cleaning is still quite mechanical and done manually. That's always the most expensive part of a lot of our projects. But then, the remaining of these things are automated.

Also, it's a good time to mention that this multiple data layers are very, very important. Both in projects, like the infrastructure layer. This is, for example, a screenshot from some of our vulnerability layers that we have for one of our projects. You can see the different data sets. And they can be enabled different ways to provide information to the participants in a data co-op project. But also, these layers and a combination of them can feed today, or I put a cluster in the dimension. So just kind of like providing both for the human users to understand what is happening, and also for the AI to provide that kind of documented intelligence.

Now, speaking about the social media that I mentioned earlier, that is another source of information that we have available for a lot of our projects. We have-- I think at the moment we have more than $2 billion tweets in our data lake. And this is just one of our projects that uses the Bushfire data. That has 700,000 tweets we collected in a specific period of time.

This is all related to the Bushfire happened in kind of East Coast. And I think it started from Victoria. At all of this information analysable in so many different ways and are accessible to machine driven API. But also, they have dashboards like this that you can query the data, and read it, and pull information from the system.

We also have access to the commercial Twitter API, which enable us to actually for a given project run a specific queries, and basically get an archive of the entire Twitter data in the last 10 years, and find out exactly, for example, how many people tweeted about the government policies and on COVID-19 in different areas? And that information can be mapped effectively and provide that kind of specific lens for a place-based research.

Now, I mentioned briefly the secure virtual machines and secure data access. That's one of the boxes in our ecosystem. Well, one thing we have done is that a secure data is essential for working with sensitive data. So where we have sensitive information, one of the problems we historically had, was people putting this in a notebook, and you walk around. And if you lose a notebook, you actually potentially losing the sensitive information on your hard drive.

Now that we have this pipeline, we also created virtual machines on the cloud, that has all the data analytics tools readily available. On the Python, to the graphic, to this, to machine learning tools. It provides one of the key functionality for the people or staff who are using this, is that they basically have the work resume at any time that they want. It's just sitting on a cloud. When they're turn it on, it's immediately available.

When they kind of finish their work, they can just turn off their notebook. But this doesn't have to turn off the machine. But it also enable us that basic have that bubble of data compute in one place. And when the project is over, we can either archive it or dispose it. So it's very much kind of provide the assurance the data is not going to leak or get lost. Also provides another function, and that's for the projects that have a longevity of that, we have a reproducibility of the data science. We can always go back. It's the same environment.

The code is there. So we can actually run it, and we basically get the data in place. Now, there is a lot of sensitive data platforms in Victoria funded by Australian Research Data Commons and other groups around the medical field. This is not up to data standard. It provides our needs, and it's not really designed for working with interconnectivity and interoperability with other systems like hospitals. This is just a space, secure space, and very effective model for working with sensitive data that lands into our ecosystem.

Now, a quick note about the data co-op projects that we are running, and experience that we got from them. So this is one of the examples. We recently had a work with basically three not-for-profit organisation as part of the funding by Lord Mayor Charitable Foundation. The data project that we have run here is started from the concept of working with the local data to produce insights.

And this was also a closed ecosystem work with all the domain experts from those organisations and our data science team, to look at the public and private data sets to say, well, OK. We have all of these data sets. What does it actually tell us? It was quite a good data exploration exercise that led to a very actionable insights for the organisation participating in this exercise. It also provided a good way for us to understand, what are the requirements for the small to medium enterprises to actually engage in data projects.

So previously, we had a lot of experience with the local governments and bigger organisations, like Red Cross. This was exercised at a more contained level. And one of the things that we learned from

this exercise was, there is a lot of value in the public insights. So a lot of information that we actually search and kind of curated for them, they are coming from the public sources. So yes, their private sources are very useful. And we got a lot of insights from those.

Unfortunately, I cannot share those insights in this presentation, since they are private and coming from the secure data sets. But these are examples from the private information that we kind of during the lifetime of those projects, we try. So we find out, for example, people who are earning between 2,000 and 3,000, they are basically driving around or commuting around a 20 kilometres in Australia. Now, interestingly enough, if you make more than 3,000 [INAUDIBLE], then you actually drive or commute less, based on the ABS data.

We looked at information related to the mental health and anxiety. For example, we found that about 32% of females reported experiencing some kind of anxiety as some part of their life. It's all data back to 2007. But given the nature of the insight, that has been valuable for the partners and the project. The same thing about a disability, and so forth. And they are basically during the lifetime of those workshops, they produce a number of these insights for the group.

This is one of the examples of the private insights that is not very sensitive. So that talks about a good cycle, and we mapped their information using engines that we have. And we found that they are basically-- in the way that they measure the travel distance of their staff, they have basically contributed more than $4,000 to the community based on this saving transport time on their staff.

Basically, the way the good cycle works, is that they send the services to different areas. And they are very much focusing and hiring younger people, and basically make them kind of job ready-- if you like-- for society. So in that context, one of the things they were looking at was all the travels of these people do, and the way that services that they provide.

Now, lessons learned from this particular type of projects. And also the infrastructure that we are running. Well, the first thing was with on data acquisition and data cleaning is the most expensive component. And that's not a surprise for industry to a great degree. But I think for the education sector, that's to some level it was surprising. Another thing that we found is that-- and we knew this from the beginning, so it's going to get firm or kind of initial understanding, that data collaboration is an iterative process.

So you don't get the data, you analyse it, and write a paper, and give it to the client, and walk away. There needs to be done an iterative process, that you're continuously working with them. And you refine the results. You get their insight. You go back to data sets. You basically build a pipeline of data human interaction in a way that actually produce value from data. The other thing is that data visualisation is-- it's not the goal, but that is what make a difference. To transform data to the actionable insights. Without visualisation, you just have data that no one understand and no one uses.

We also find that there is a great value in public data sets. I cannot emphasise that enough. And there has been a lot of investment in Australia on the reusability of research data. And there's a lot of kind of efforts and activities by different groups to kind of tap into the existing data sets rather than collecting the information again and again. And our own journey is kind of highlighting the

same thing. That looking at existing data sets, reusing them can provide a lot of value for the research and for the not-for-profit sector.

And finally, data language is what you need to do for a lot of projects. But a lot of times, the only way to connect data set together is based on the sense of place. So if you're aggregating information for a given area, then understanding of correlation, of different phenomena, and social variables in a given a space, that is the best way of looking at the connection between information. And that is what we use for a lot of our kind of collaborative exercises to bring information from different partners, from different organisations together.

I think on that note, I think I can finish this presentation. And I can just say that this has been so for an amazing journey. There's a lot of capabilities here at Swinburne, and also across all of our partners. If you have projects that you think can benefit from some of these, I will definitely want to hear from you and work with you.

Jane has been quite effective and amazing going around a lot of different projects that we are doing, and keep us always busy, and we don't complain. So if you have something, and you want to collaborate with us, just raise your hand. On that note, I can just pass the microphone back to Jane and Paul. Thank you.

JANE FARMER: Anthony, I think, is in charge of this next bit?

ANTHONY MCCOSKER: So I will help to coordinate questions. And please feel free to add a question either in the chat or raise your hand through the Participants tab. Do you want to exit the screen sharing, Amir, just so that we can see each other-- each of us a little bit better. And yeah, very, very welcome. Very much welcome anyone to jump in with a question. If you want to a little bit more, if you have thoughts of your own in working with data in this way, I have questions. I always have questions.

So I'm just going to kick off, because I get to decide who asks the first questions though. And I'm just jumping in. And so, one of the things that we always find frustrating, Amir, and I'm wondering if you can just tell us a little bit more about your experience in this space. You talking about the difficulties in the sort of I, guess operational layer, with partner organisations coming together and working together in the process of sharing data.

But also looking for what kind of insights might be-- shared insights. Not just insights that will help their organisation specifically, or their mission, et cetera. I have a bunch of questions around that particularly in terms of the difficulties in building trust, and building data agreements.

And also, but the question that I'm kind of interested in at the moment is, what you think about the increasing role of data stewards or data custodians in these organisations? People who seem to take responsibility, or are most interested in pushing forward with data projects, and what your experience has been around that.

AMIR ARYANI: I think that was a very, very long set of questions. I have to try to remember. I start from the back, and going forward. So in the stewardship positions, and data custodians, I think one of their problems that we had with a lot of projects is, when we start engagement with the

organisation, doesn't matter it's a government department, or within a small SME, there's always a shroud of mystery of what data do they have, and who owns the data, and how we can actually access that information. That is one of those areas, that when we start a project, often even after signing the contract, we don't know what it is.

Now, you're right. As we actually start tapping through those data sets, then different interests, or sometimes competing interests, start to bubble as people start to share given data sets. So that, in a way, you would always have influencers in those workshops and in those conversations, that try to actually take the directions of, if you like, the whole workshop, take the direction of the conversation. And this is not new to the concept of research. It's actually new to the-- even in the commercial sector, that is given that anyone who share a resource would have kind of agenda attached to it.

Now, in the context of universities partnering with industry, it is almost gets a bit more complicated, because you often operate as a kind of like you're providing research services. And then, that puts us in a very strange position, because in one way, they expect us to kind of be fair and do the ethical research. At the other side, there are components underneath moving around that make things difficult, because there are different rules and different expectations goes with it.

Now, this is something that we deal with. And I'm sure you have experience in doing this a lot. We deal with this often during those workshops. So often, during the workshop, that is where the main work happens. That we basically try to showcase different features, and take attention of people in different facts. But end of the day, a lot of people management involved in actually coordinating those activities.

Now, I want to step back before in this, is that we need to sign those data sharing agreement and lease documents and contracts. And that is the most complicated part of it. Because we often get into lots of challenges and difficulties around accessing to given data sets when it gets to the legal requirements. And those requirements often come with expectations attached to it. And that's usually is the most complicated process for projects like this. So I don't know how much I managed to go into the depth of the questions that you asked. If I forgot something, let me know.

ANTHONY MCCOSKER: No, absolutely. There's a question from Paul. I'm not sure if you want to jump on Paul, and ask the question. But it's a pretty quick question. Yeah, Paul?

PAUL: I'm just happy to do it. I was very keen since the Australian Research Data Commons and the federal government's initiatives around a humanities research infrastructure, [INAUDIBLE] to manage research infrastructure. This is obviously being going for some while. Have you had any connections with that and some of the initiatives that are starting to be put forward?

AMIR ARYANI: Yes, that's right. Thanks Paul. That's actually a very good question. So interesting enough, actually, the first component of this project was funded by ARDC. So we are closely working with them. Some of the components around the data governance of this project has been done in a direct consultation with ARDC, so they are quite involved in that data governance layer process that we're establishing. And in the infrastructure layer, we are quite connected to the increased facilities.

We are working with the Australia Archive, we are working with [INAUDIBLE] in space. We are getting information from [INAUDIBLE] into our system. And also, we are working with this kind of increase facilities, and on the concept of the [INAUDIBLE] common infrastructure. So in that way, we are quite aware of the sector, and we're working with the players in that domain. Also, [INAUDIBLE] part of our system, all coming from the ARDC services. So thanks Paul.

ANTHONY MCCOSKER: We have a couple of questions. Leigh, it's a long question there, I think. Did you want to jump on and ask that one?

LEIGH: Yes. We have the frustration in a cross-border community of Albury Wodonga, which really operates as one community. But it's just-- it's so incredibly difficult to get simple data that tells you basic things. So our presentation the other day on Victorian breast screen participation data for the catchments that we serve in Victoria. And there was an assumption made from that data that the rates in Wodonga or in Indigo might be lower than the state averages in Victoria, because people were going to New South Wales.

But you shouldn't really draw that conclusion unless you can test that assumption with the New South Wales equivalent data to know where people were coming from. And that's just a small example of a zillion things every day that are very frustrating in this community. So my question is probably really about just the model that you're using. That overlay, would that sort of be applicable in an environment like this, or have you come across those sort of issues before?

AMIR ARYANI: So Lee, you mentioned two different things. Let me just rephrase this just to be sure that if I understand your question correctly. So the first thing is that you kind of had a problem of accessing the data in a complete form. So almost kind of a small slice of data that doesn't tell you the whole story. And the other thing is you are looking at all the different data sets from other sources, that they can basically provide the big picture. Is that what you are asking?

LEIGH: Yes. Mostly. That's a very good summary. Thank you.

AMIR ARYANI: Thanks. So on the first one, and this is one of the classic risk of data science. Is almost like if it was a biology concept, thinking about you are taking a very, very small sample test, and then drive a conclusion. And say, well, this drug is very safe, because we just had five people testing that. All of them had no problem. And then applied it to 1 billion people. And that's a classic example.

In the data science, you want a data set that it is complete, and it is based on normal distribution, has been collected correctly, and basically provides the coverage around the majority of cohort or population that are subject to [INAUDIBLE] study. I remember without mentioning the name, I was in one of the workshop presentations, one of the commercial providers.

And they had a data set about Australia. So I asked to zoom into a given area. At that time, we had a project in [INAUDIBLE]. And I found that, well, actual number of people in that area who have answer to that survey are only two. You don't drive information or any conclusion about the whole community with just two people.

So this is a very huge risk on the social science, because it's not recognised as much as it's been recognised in the biology, and health, and other sectors. That a lot of papers in social science get

reviewed, and you look at the sample size, it's like, in 1,000 people that answered a survey, and then drive the conclusion if you have the similar drug test example, is not going to get approved in any shape or form.

Or the paper is not going to get published. So this is one of the problem. And part of the reason is that data collection in social science is complicated. Now, your problem that you mentioned is slightly different. It sounds like you have a problem in the area to actually access the right data set.

Now, the other thing is that this overlaying data sets from other sources is definitely the way to go. In a lot of ways, we use a concept I call proxy data. I may not have access to the information about people commuting in a given area, but I might be able to access the petrol purchases, and kind of the energy consumption in that domain. So that can be something I can proxy to find out the usage of car.

This was an example of information can be used in different ways to actually draw conclusions about things that we don't have data about. This part-- it is a very risky activity, because using the wrong proxy you might drive a wrong conclusion. But it is a way to actually kind of cover the gap.

JANE FARMER: Can I just say as well, that it's lovely to see you again Leigh Rhode. And that this whole kind of issue of rural areas data is something we're really, really interested in. And we have kind of dabbled with this kind of data bricolage kind of concept, which is like chucking in all the data sets that you can get to see if you can get some findings, to put it kind of crudely.

Because of what Amir said, that there's often small numbers across massive areas. But I also get what you're saying about borders, because you've got different data sets and different ways of collecting the data and different accessibility of the data set. So I think that what Amir is saying about looking at other data sets that we might use together is probably a way that we could go-- or that you could go ahead.

AMIR ARYANI: I also-- one thing else to mention that might be useful, sometimes the data from the other sources provide a very, very important complementary part of the picture. Like example of this in an urban area is a homelessness data. If you look at just one local government, you might have a picture that relates to their services.

But if you look at all the other neighbouring LGAs, then you actually see potentially a different story, given that people who are dealing with that problem moving from area to area. So that's sometimes in some of these data co-ops, it's actually a necessity of getting the data from multiple different sources, especially when it gets to a geographical area to get a better picture.

ANTHONY MCCOSKER: Jane, did you want to add your additional question that you popped in the chat as well? I have more too. So go ahead.

JANE FARMER: Oh well, I just took conscious that the talk maybe sounded like, oh my God, how would we even start doing this? So but by the same token, I know that we have started at 0.0 or scratch with a number of organisations. So I wanted to make it seem not super scary. So my question Amir is, what advice would you give to a small organisation that maybe didn't have specialist

workforce, but was really interested in trying to look at what extra value they might get from the data that they collect?

AMIR ARYANI: So there are two different things that might help. But one is that the only infrastructure that I mentioned, there are-- basically we did the engineering work. So it actually makes it extremely simple for people who go to data co-op workshops to make those data products and insight happening. So I think all the things that I mentioned is kind of running behind either [INAUDIBLE] or if you are actually running a not-for-profit organisation, tries to do a data project, not even data co-op project, just a data project. this is-- you won't see all of these infrastructures in detail.

You just see that look at all of these insights and services are just working, and that's the intention of this. The other thing is that, as we have done my example around [INAUDIBLE] and Good Cycles, and other projects I mentioned, they started in that context of being small data co-ops with a small number of data sets. And it growed very gradually. Now, the recipe for this is that those workshops are very good vessel in a way to get to a bigger plan. So you start with a small project.

You basically go into a number of different iterations, but confining at about three or five months. And then, you basically from those you would have a much better understanding about what is possible. And I think that is a very good opening for any data project. And that is just, start in a small pilot space when it's manageable, and produce some useful but limited number of insights. And that is kind of first give a taste about what can happen. But also it provides insight about what is possible. And that's where you are going to plan and go in.

ANTHONY MCCOSKER: There's a question here from Erin. And it's a really good question, because we've just started a project where it's about those connections between data sets within government and access across government departments, where there is essentially a goodwill to data sharing, but still a lot of concern and a lot of issues around trust in that sense. Erin, did you want to ask your actual question?

ERIN: Thanks Anthony. That's a great intro. It's really about whether or not we can piggyback off these previous successes in passing the five safes type risk assessment check, when the next data opportunity comes along to access state or Commonwealth government data sets. So does it stand us in good stead? Is there a way we can leave those past successes in, or do we just have to go the rounds and complete those processes every time?

AMIR ARYANI: So Erin, I'm not actually aware of a formal process for this. So government right now, the federal government looked at the process of kind of creating different organisations to access to the sensitive data from government. But that's not based on five safe model. That's much more kind of detailed verification of organisations' capabilities.

Now, when it gets to the five safe, at the moment, it's still it is sitting at a level of recommendation. It's not implemented as-- it's in a framework, but it's not detailed framework enough that you say, I'm going to pass these things by basically going through these steps one at a time. And as a result of this, then as I said, you're passing this again and again for every data project, for every single data sets.

There is no record of it somewhere to say, look, well, I've done a five safe project for this project. It's safe. So I can have a go-- I can-- it's not like ethics, that you've done your ethics, and then you go and access many different data sets. Is a conversation to have for every data set at every government department. And I know this is frustrating and expensive to do.

But at least this provide a framework. Because previously, we even didn't have that. We were talking to different data custodian, and they were not even-- they didn't know what questions to ask. So at least there are now a way to communicate, and the way to actually ask the right questions. But unfortunately, there's no way to record or reuse the answer to those questions.

ANTHONY MCCOSKER: So I'm just building on that Amir, there's a lot of uncertainty, I guess, but also interest in how we do move from principles to clear processes around ethical questions. And I guess some ethical issues in dealing with data at the site of community sector level, health sector level, et cetera, outside of government, for example.

And I'm just kind of wondering about your thoughts on whether or how you see building those ethical questions into the design around the data engineering side of things, the kind of work that you would want to do in the background in order to set up those processes smoothly. But how do we build ethical practise in at that level.

AMIR ARYANI: So there is established practise of the research data management that taps into the very superficial level of this conversation. Say what you do with the data in a sense that follow those rules. But doesn't actually answer the question that what rules apply. So those would be dynamic from project to project. But that was the whole idea of trusted data partnership model. This is the idea of that model that we follow. We try to actually kind of build-- well, so far, we have kind of built it into our infrastructure.

But we are also trying to produce that governance model into this set of practised question. So following what just Erin asked, we don't want to actually get the same problem internally in our own ecosystem. So if we know what questions to ask, and if we know what rules to follow, and we know what procedures needs to be in place to cater for different type of projects, the preferred model would be at least-- probably automating is a wrong term-- but put them in a rail in a way that you know exactly what needs to go where for a different type of questions. Kind of a decision tree in that way.

So that is the intention with what we are dealing with. I would still, as probably mentioned in a part of the presentation, we are building the building blocks. So what you have seen today are the kind of building blocks of the much bigger plan. And we are almost like people sitting in a house. But at the same time, we are building the house. So it is kind of-- we don't have a luxury of just going out in a tent, and then build a house, and then move in. We're just sitting here, and Jane comes a different question, Anthony comes with the questions, and we have questions.

So the projects, and we have projects from all different partners, soon then partnership networks. And as we are actually going through this project, we are putting these things together. So the house is getting built together in at the live time. And there is always the drawbacks, otherwise a lot of redone work needs to happen.

But also, the advantage is that everything we are building is 100% applied. Because you are just drived by the usage of those. We're not building something, and then cut the ribbon, and then they start using that to figure out, oh, that was a mistake. We find mistakes much earlier. That's in some way saves money.

ANTHONY MCCOSKER: I'm just conscious of the time. We do have one very big question right at the end there from Fiona. But I think we don't have quite enough time to answer that question, which is really about where this can lead us with really big issues around data leaks and data security. And I think that's partly addressed by your approach to private and secure platforms as well as open and public platforms for data sets as well.

But I just want to thank you Amir, for your time today. And thanks, everyone, for coming, and for the insightful questions. I hope that this has been a fruitful series for everyone. We've had some really great seminars, I think, in the social data and action seminar series. And this was a really great way to end it, I think, because it was very practical. And these are projects that Amir, and the team, and we are all implementing working with non-profits, and health sector, and government, public sector.

So thanks, everyone, for your involvement. Have a look for the videos via the Social Innovation Research Institute website, and Swinburne Commons, as well as the Centre of Excellence for Automated Decision Making and Societies YouTube channel. Please like and subscribe. And we hope to see you in further webinars. Thanks, everyone.

AMIR ARYANI: Thank you very much. Thanks Anthony. Thanks Jane. And thank you, everyone, for joining the presentation.

[END OF TRANSCRIPT]