# Transcript Day 2 video 3

# Session 5: Actions to Take

Chair Nadine Alameh

## Panel on Action :

### \* Jeanne Holm, City of Los Angeles - Perspective: Building a Generation of Government Data Scientists

I want to focus in specifically on the idea of how do we generate a new generation of data scientists not just what we're doing in LA although clearly I'll talk a little bit about that but I really want to just sort of bring in a lot of the background I've had before I came to the city and and try to understand how we are doing something different in Los Angeles that may or may not be useful for others but I think it's particularly pertinent for the fact that we live in a completely dynamic time or the way in which a lot of feels are coming together

so these are some of the **technologies that we're using in LA related to data** and I'm sure that you can see some things on here that maybe relate to other things we've discussed or things that you do at your organization some of the things we haven't talked about at this workshop are gamification gamers in the group no seriously no like world of art like okay so I'm a big gamer world of warcraft Assassin's Creed like ecology okay so but **gamification** is super addictive and it's a great way that I'll talk about on how we get our youngest citizen scientists engaged with the city and the world around them and then of course virtual and augmented reality okay great all right so another way in which we start to change the experiences that we start to understand how people want to not just have the world around them change or like a fantasy world but really how the future the city can be imagined together and so we actually learn a lot about how people design things in Sim City for example and how they want to have you know sort of new generations and new ideas about how a city should be designed not all of which are actually physics based but it's really interesting to think about the concept of taking a rollercoaster to work so we you know we talked about all these amazing things

and we all in addition then we've got this team of heroes right you know who has a data hero in their organization okay who has data sidekicks in their organization okay because every hero is only supported by all those sidekicks right so there's Batman is not really Batman without Robin and Alfred so let's think about like the fact that there's a whole infrastructure to these people who we often think about oh my gosh they've made this amazing breakthrough and so we call open data a team sport in Los Angeles and so we have and it took me forever to **get the city council to approve the mayor and I put together an idea to have a city data and predictive analytics team that really looks across all the city departments for big issues much more than just our chief data officer kind of looking at policies and aspects around dat**a but really how do we look at harnessing all of that data synchronizing it putting into data like and being able to really change the face of how we use data at the city and underneath this are some very basic values and **number one for us is equity** because we all have smartphones we're all data savvy we're all very data literate **50% of the households in Los Angeles and some of our neighborhoods don't have internet acces**s so the idea of talking about data and data literacy and digital services needs to really take into account the fact that we have this **huge data divide and we have to significantly work on data inclusion** we are also working on two documents that are draft stage and we're about to go to public review one is a code of ethics so we've talked about this a little bit today but I thought it was super timely because I was in the back of the room madly updating it so our **code of ethics is for our ecosystem of developers** that is as a rule of series of guidelines by which we have to behave in the values in which we uphold in order to be able to create systems because of inherent bias unintentional bias unattended consequences and so just because you've been directed to create a system just because you've been directed to grab data sources you also have a responsibility because with great tech comes great responsibility and we have to make sure that everybody in that ecosystem speaks up whenever they think there may be an issue with the consequences of what they do the third aspect and the other piece is a **digital Bill of Rights for citizens** how they can't expect that they are monitored he was talking about monitoring earlier today right you're not just monitoring the UK you're monitored everywhere right **if you're carrying a cell phone you are monitored everywhere** and so we are working on this digital bill of rights for what citizens can expect as they move work play throughout Los Angeles and what kind of sensors are in our network we also work a lot around inclusion which is making sure not only that we talk across the digital divide but that we are diverse teams so our teams are staffed with gender equity across the spectrum this is Los Angeles so we think about you know a lot of diversity there we think of age diversity ethnicity 220 languages are spoken in our city we need to start thinking about representation and co-learning is an important aspect as well where we want to take young people with new ideas and partner them with let's say maybe older or more more entrenched cities it's more serious as out you said experienced city staff who maybe haven't been exposed to some of these new technologies but do it in a way where everybody learns together and it's not judgmental we are grateful for all of our city staff and for all of the work they do every day and if they haven't been back to school recently it's not anything that we hold against them we want to make sure everybody learns together and then that becomes a powering experience

6:40

so **we start young with data science** we have programs with girls who code just an amazing organization if you aren't part of it you should sponsor a local group we have over on the far side you see three girls looking at their phones they're actually playing discovery agents which is one of our gamification x' for helping kids explore the parklands la is a super biodiverse place we have mountain lions and bears and cougars we just don't want the kids to actually interact with them but we want them to catch augmented reality versions in our parks and understand more about the flora and fauna we see a young man over there working in a robotics lab at our public libraries we have 37 makerspaces and you can see sort of our multi-generational learning activities going on in there and then we've just opened up our **Tavia lab named after our Octavia Spencer who is one of the most amazing black women American science fiction writers** and she did a lot of her research at her Central Library for her books Octavia lab has 3d printers and scanners and podcasting stations and all kinds of amazing new technology totally free to anybody in Los Angeles young or old to come in and learn these and we have all these sorts of teachers in space **so that people can start to cross that digital divide**

then we also think of data science as part of our ecosystem and this is hack for LA have you heard of **Code for America okay so our local brigade is called hack for LA** it's the most active Brigade in the country we meet two nights every single week city and county are there as well as a lot of local technologists who work together to hack on using open data different kinds of solutions for the neighborhoods and then really kind of coming bringing starting to bring a lot of this together

**I founded a data science Federation** so I've been a professor at UCLA for 25 years and my students like for those of you who teach you know you want your students to have meaningful assignments that inspire them and generate really interest in the field and then we use these data problems where the data is you know just the same data that every other set of students has you before them and the same outcome is just this classroom assignment and so instead I said what if we can solve this problem by having students make real change in their environment like in their neighborhood what if we could have them do homework on their hometown and what if we can combine that with the fact that our city really needs to have different changes and this co-learning experience and so what we've done is created a **partnership with 18 different universities and colleges in the local area between California and Arizona and then partnered those with 88 other cities** that we've expanded to and students and professors each semester sort of bid on or suggest different programs that they would like to do and that's based on what our city departments have needs of so right now we're doing a ton of programs around the zoo **we're doing programs with our LAX Airport** if you've flown into LAX it's not about the traffic at the airport and so students have a chance to make a real difference and then we have this pipeline for job placement in the city because really **what I want to do is hire these young data scientists into city government or at least get them exposed to government data and how they can be part of their future government experience**

they work on all kinds of different projects each year we do about 20-25 different projects everything from homelessness to earthquake early warning to taxation issues to economics and we provide all this data on our **open data portal** so shout out like I think I partner with almost everybody in this room I'm just like so surprised so shout outs to ESRI AWS Uber even if we're having a little disagreement right now normally we're best friends let's see University Department partnerships USC was here the California Public Utility Commission USGS and NGA you're like on my family so we bring our open data through our open data portal but we use lots of data or services or support from lots of you and actually making this happen

so one of our programs for example the students have worked on we just won **a 1.3 million dollar grant from NASA it's called Predicting What We Breathe** and so the way that we pull all of these pieces together I know you thought didn't really make sense when I talked about kids and citizen science but we're doing what we breathe generates connect collects data from satellites airborne instruments and ground data sensors from our Internet of Things across the city and across the region and then federates that data in a way that helps us to start to use machine learning to create predictive analytics about the quality of error and we are **partnering with other mega cities around the planet** to look at characteristics that we share to see if we can evoke patterns from satellite data that may help us to understand things that are happening on the ground in Mumbai for example where they may not have as many ground data sensors so it's a pretty interesting experiment many of you know the the program sponsor for this **Mike Little from the NASA AIST program** and and we have lots of different partners on this one and

**ShakeAlert LA was one we rolled out last January** and this is an earthquake early warning system it gives you up to minutes warning that you will about to feel shaking of an earthquake in Los Angeles and so we have over a million users which is a lot for a little city app and we partner with the USGS shakealert system which has sensors up and down the west coast and the trick here was really the technology and the latency and the network but it's a lot about behavioral nudges and behavioral science to get people to be safer during an earthquake

of course we have autonomous vehicles on the streets of LA and thanks to Uber air and Bell helicopter's I think I'm not over a lot today and fellow helicopters we will have autonomous airborne vehicles carrying passengers next year super exciting but the idea of how we pull all that data together and we talked a little bit about this in transportation group how we think about the ethical and the implications of that how we started to think about working across the digital divide getting people familiar in areas where they wouldn't necessarily think about getting in one of these vehicles how do we get them comfortable with it it's all part of what we're doing as an **ecosystem and gathering all these people to work on data science together**

one of the ways in which we're doing that in a very super geeky way is with a datum of the mobility data specification that we have launched on github we've just launched in addition the **Open Mobility Foundation** so we put this specification out on data sharing with micro mobility which is scooters bikes and then Uber and Lyft as well and as part of that we just have these data sharing agreements we need to know like anonymized data like where you're dropping these things off and where you're picking them up so we can build public spaces to make it easier for people and safer and then **43 other cities have adopted this we've now opened it up as a nonprofit** and this is a really a precursor to how will be doing **data sharing around autonomous vehicles** and then just to kind of like spin a little bit from la very quickly I know I'm out of time if

it's not just about like what we can do with hi-tech but so I also have a nonprofit in Africa and we recently got a micro grant and completed a **project called clean streets** based on the same idea we have in LA around my le 3-1-1 app which is letting people go out and crowdsource we're streets need to get picked up more trash needs to get picked up but we funded activities in Kampala Uganda Accra Ghana and Bo Sierra Leone and mapping was the key as it is on many many of these projects underneath all of it essentially almost all of our data is geospatial data at the heart and it's all got a location because we're a city like it's a place and so **the clean streets project was really looking at how to get trash picked up faster** and so we have super creative solutions in **Ghana** they created a mobile app that help people be able to identify a report in BO they worked with municipal government to get trash pickup faster and the best I my personal favorite in **Kampala** Jeffrey Qatar Agha enlisted the boda boda drivers so these are the little motorcycles you jump in the back on as like taxis super dangerous super fun the boda boda drivers they're traveling all over the city and so we just asked them to **turn their GPS on and snap pictures as they went through the city of all of the trash heaps** and then we aggregated that and then reported Anisa palliative which i thought was super creative

15:30

so really as you pull this all together we think about the **future training of data scientists** in three communities first our bigger community which is around things like citizen science and if you haven't seen the **federal citizen science toolkit it's available online has been for a few years Lea Shanley led a great group of us working on that** we work with our local Code for America Brigade but and we do a lot of these data literacy programs through our LA Public Libraries and we also have a volunteer corps called data angels because we're the City of Angels it's a great name you know so our data Angeles Program in college we say once you're on the track of something in stem or public policy or health like we don't care where you're coming from but we want to make you a data savvy person and so we have the data science Federation we for every graduating senior of LA Unified School District and we have 650,000 kids we give to four years of Community College in Los Angeles and so we can launch people into a free community college for two years to get them on to a UC or other system and then our cycle le is a program or **give away computers and Wi-Fi to low-income families and kids so this year we give away 6000 computers to College promise students** and then we have the tech talent pipeline which is a partnership with our local Chamber of Commerce that takes kids from low-income neighborhoods and non-traditional backgrounds and we fund them each summer to come to the city or Google or snapchat or blizzard games his games are great and we help them get into data science and then offer them jobs and then we have all these great programs with for K through 12 **including cyberpatriots which is a national program available in all of your cities in the US that helps kids as young as third grade start learning cybersecurity skills** all of this helps us kind of organize ourselves

**we are a city that has adopted the SDGs** which I probably were gender equity be my favorite STG and and we have adopted the SDGs as a city driving towards successful decrease in poverty hunger and increase in equity and climate action and that's what we're doing in LA

Nadine: thnk the plan is to save the questions for the panel I missed the introduction but I don't know **how amazing you know you know this woman is she's I was just like she's got the NASA exceptional award twice** and my favorite **she's a top-15 woman in tech** and she's with us today so thank you and I'm what I'm learning is the secret to impact is this energy even at the end of the second day

### \* Patrick Griffiths, ESA - Perspective: Earth Observation data and analytics supporting policy and geospatial industries

Nadine:so let me challenge Patrick to see if he can top that this is a very weird event you guys usually you know the energy goes down I'm seeing the total opposite it's like the afternoon has more energy so a while Patrick is setting up so

Patrick comes to us from ISA the European Space Agency and he's in the Earth Observation Directorate and he does development and support of digital platform technologies which as you'll hear involve analytics innovation and partnership with industry on things like data analytics data science and AI just in time

Patrick

19:30

So I come from ESA from the Earth Observation Directorate and I want **to talk about three things** here one is kind of the ESA perspective on the data deluge or the **big data of Earth observation** and the requirements that result out of that for **data storage accessibility and services** and then bring an **example of a major policy piece** that is strongly affected and taking advantage of this new data rich environment

so you know at ESA there's three big programmatic lines there's the science mission the earth explorers and there's the Copernicus mission that you all are familiar with and then we have the meteorological missions that are operated by you met sir right and that's currently 25 satellites under development 15 operation 12 in preparation and if you look at the you know this exponential curve of the ESA data archives you know it's we're kind of getting to several ped 10 15 petabytes per you know and if you look at the distribution here across these missions you see that you know the Sentinels are obviously the big game-changer here in terms of data volume whereas you know if explorers are great missions but the data is a little bit less accessible and requires a little bit more expert knowledge maybe in addition to that

so the **Copernicus sentinels** as you know this is a co led by the European Commission and ESA so ESA builds and operates the satellites and the European Commission owns this so you know the central one is a you know all-weather capability of a synthetic aperture radar with two platforms Sentinel one a and B in orbit and operational Sentinel to multi spectral building on the Landsat heritage with you know 10 and 20 meter bands and 99 main spectral channels also again here sentinel-2a be in orbit and operational since then all three more like the modus instruments or global observation capability with daily acquisitions global acquisitions Sentinel 5p is the precursor to central 5 but with the Trapani instrument that provides some really interesting data on atmospheric composition trace gases and so forth and you see here already that you know the C and D components of these missions are planned financed and almost ready you built and will be launched you know this covers the time period until 2030 roughly but the Copernicus program is also going into the next generation and Copernicus 4 is being discussed these are the six Copernicus high-priority candidate missions there's a hyperspectral mission there there's a high temporal high spatial resolution land surface temperature mission atmospheric missions l-band radar mission and so forth so this is really the game-changer program as you as you all know and but

at the same time you know ESA is realizing that there's more **like other space agencies** also are realizing that there's **more players new players** in terms of new space a new space companies and all of these constellations that were to be expected during the next year's you know I mean planet or Terra Bella or is this new you know these great new synthetic aperture micro SATs constellations really amazing things so you know the data situation will be further increase you know it's a you know the great thing about sentinels of course there's freely available open access freely available global acquisition capability so that's you know let's see to what extent these datasets will be available but of course you know a sentinels also precisely calibrated geometrically and radiometrically and it's kind of you know always considered a gold standard against the slightly less well calibrated measurements from these micro sets and small sets and

then the last days we heard a lot about the IOT and that we kind of expect a big impact but we're not really seeing that yet and also for Earth Observation you know this is kind of the the vision for the next year's you know there will be a big impact when we can actually **take advantage of IOT and then we're gonna have like several levels of data acquisition here with the satellites and then perhaps the high-altitude co2 satellites drones** **and the IOT sensors on the ground** and all of this is supposed to be connected and interconnected and you know if we really have this this will be a big change

24:55

so the questions a little bit you know with all this data you know how do we how do we handle store process these things and I think this **layer diagram that we work with here that depicts the earth observation platform ecosystem** **is similar to what we saw earlier from our colleague from Japan** here a lot of common elements I would say so you know I mean obviously the old model of downloading files and running processing on the on your local workstation is kind of obsolete so we have you know distributed storage environments co-located with distributed compute environments and then on top of that we have basically this layer of platform and exploitation tier layer with platform services and you know **data cubes and analytic environments** to take advantage of all of this data

really **essential with all these with this high data volume is basically data access and one of the important developments here oh of course data cubes** and I think you know it depends on the definition of what a what constitutes the data cubed varies by a person to person and application to application but for me coming from the kind of optical land domain for me the main point here was always that we can you know align the data in a way so that we can directly exploit the temporal dimension this is a figure here from the Australian colleagues right and once you have it in this shape then you you know then you have this simple data access that's basically **one line of code in Google Earth engine right where you just define a collection the time frame and a spatial you know rectangle to access the data and you don't have to worry about anything like projection reprojection or file overlaps** or stuff like that so you know this is a big boost in productivity if you have if you have the data at this level and you can just ingest it and **access the whole archive without having to worry about pre-processing data management which often took up to 70 or 80 percent of the earth observation scientists time** so but

the main improvement with the sentinels is and I think everyone in the LAN community would agree to that is that the more data we have temporarily the better the information retrievals and the application results in the end you know and with the operational sense the won and sentinel two we have a **global revisit of five days for both missions and we have you know due to the orbital alignment of the tracks you know we have many areas we have revisits every two or three days** and that is a big big game-changer and you know with more temporal data one of the application fields that has benefited probably the most is agriculture because that's of course a temporary dynamic phenomenon for which you need a lot of you know observations through the growing season and this was in part also part of

the **motivation for the Copernicus program to support European policies and one of the major European policy element that has really benefited from the sentinel data is the Common Agricultural Policy** I don't know how much you know about this but it's you know the single biggest fiscal budget item in Europe it's **43 percent of the total budget of the European Union and it's mainly you know it mainly pays subsidies to the farmers to make agriculture still economically viable** for the farmers but it also has elements for you know it's enforce environmental policies and you know to to lead to more sustainable agriculture so there's a lot of problems about the CAP but it has come a long way and so traditionally the **Common Agricultural Policy compliance checks so to see if the farmers are actually doing for what they are receiving subsidies** this is traditionally done through interpretation of a very high resolution imagery for a 5% sample per country and then for a smaller percentage of those parcels there's on the sports check in in addition to the VHR imagery interpretation to determine the compliance with the subsidies that were received and **this is really complicated and inefficient** and you know very problematic kind of way of doing things and yes since 2018 we have a regulation in place that basically pushes this system now into permanent so and continuous so throughout the growing season and for all parcels in Europe that receive subsidies a **fully automated monitoring based on Sentinel time series and you know and data analytic tools that come basically from the data science community** there's a this is a big step and you know there's a couple of interesting aspects to this because of course this is kind of **moving into the question of legal liability of you know remote sensing based evidence for compliance** and

30:15

here's you know just a simple example of three cereal parcels or parcels that grow cereal here in 2080 and it's kind of the density that you get from Sentinel 2 we see the harvest and all three parcels but then there's another thing so these farmers can grow catch crops right but catch crops have to be destroyed mechanically according to the Common Agricultural Policy requirements after a period of two or three weeks and so this lower right that parcel is obviously not compliant and that's just a simple example to show you that with the temporal detail we have now we can see these things in in very good detail and then

I think there was in an earlier talk there was something on optical lens our fusion and of course you know here already you saw that there's you know a lot of clouds so you know optical data is limited to a certain extent and with central one we have this **all-weather imaging capability** throughout the year day and night descending ascending orbits so this is amazing and here I think this is really one of the areas where I think machine learning come you know new machine learning methods can make a big difference so I showed this example here **from a company in Belgium called Vito** and what they did here is they basically have a model that derives FA power so the **fraction of absorbed so photosynthetically active radiation based on observations from Sentinel 2** which you see in the topography and at the bottom you see the coinciding central one observations you know look at that density of observations they look at the sparseness of observations in the Sentinel 2 data but what they do is they have a **convolutional neural network that takes the Sentinel 1 data and makes predictions for the FA power value for Sentinel 2** and you basically get this nice you know uncertainty estimate across around your measurement you know and the more you know it's then corrected through time with incoming Sentinel 2 observations but you know this is I think really **we haven't seen a lot of good fusion data fusion approaches but you know these are really exciting new applications for bringing together optical and SAR data and then basically further increasing the usefulness of the Sentinel data** so that's all I had

so you know in summary you know **more Earth observation data the better and more temporal data** in principle you know we get better applications better information retrievals and algorithms are increasingly moving to the data that's nothing new but you know which in 2019 we're almost there and **pixel based data access to analysis ready data is what the users want to need and that increases productivity greatly** and finally you know this example with the Common Agricultural Policy shows how you know Earth Observation data and methods from the broader science data science domain can be used to monitor enforce major EU policy I forgot to say that there's also now a new regulation in place that will look at the do an environmental impact assessment of the cap so look at whole Europe and see to what extent like the greening aspects of the **Common Agricultural Policy actually leading to improvements in the in the environment in Europe** so that's really quite exciting I think that's all I have [Applause]

Nadine: I think he did well what did you think matching Jeanne yes plus I loved how many times he repeated how do we how it helps us improve productivity so going back to the basics

### \* Stephanie Shipp, U. of Virginia - Perspective: Harnessing the Power of Data to Support Community Health and Well-Being

Nadine: so my pleasure as you know Stephanie is hooking up her laptop to introduce Stephanie our last Speaker of the day definitely not least so Stephanie is the deputy director of the social and decision analytics division at University of Virginia the bio complexity Institute and she I love this she creates and implements new program to build data science capacity and cities and counties of all sizes and what she highlights in her bio like the one word she highlights is you know she focuses on innovation and tools for using all and that's what she has all data to address social science policy questions so all yours

34:50

Stephanie

thank you so this is really exciting to be here these last few days I thank George and Nadine for inviting us to be here

**we work with local communities to build their data science capacity** and while we do that we're building our data science capacity as well **we start by working with them and asking them what their issues and questions are and really we start these local conversations with saying what keeps you up at night** what questions can't you answer that you would like to right now

so as Nadine introduced us we are part of the bio complexity Institute which is part of the University of Virginia scientist in our Institute model from they scale from bench biology at the theoretical level and we scale to societies and populations striving towards evidence-based policy my **social and decision analytics divisio**n is obviously on the social on the society some populations side of this mission **our work brings together a variety of local state and federal data administrative data opportunity data and procedural data** and we come use as I think Jeanne had a great slide that was looked at all the different methods that we use but **we primarily focus on statistical analysis combining that with geospatial and machine learning** so I'm gonna present a few examples from our early work so they might look a bit naive but I want to set the stage for how we got to where we are today and how we got hooked on working with local governments

so when we first started our lab in 2013 **we thought we would be working in the megacity space** and involved in a IoT of working with the spence sensor data and the internet of things but when we walked out our front door in **Arlington County Virginia** and we began talking to the fire chief and the chief information officer and the people that collect our tax data we began to get very excited and so **the fire chief came to us and he said I have all this 911 data and it all sets in silos and could you link these data and to end in order to recreate incidents and if you could do that I would better understand my day-to-day operations** so he didn't really want predictive or dashboards who thinks he just really wanted to understand his data so that was actually the easy part and this was one of the first things that we did we had a group of summer students and they **put together this map it's three years of data and it's looking at 900 incidents by time of day in day of week and we were so fascinated by this the first summer and you sit there and you can see that Monday is different than Tuesday and Tuesday is different than the weekends** and the evenings and the during the night are different than the early mornings and we got very excited by this because also the fire chief had said I understand who my resident population is in Arlington but I don't understand the hundreds of thousands of people that come to work in Arlington and flow through Arlington and when we were looking in this we were struck that not only were you mapping 9-1-1 incidents you were mapping flows and densities of people that come through our lling ttan by hour of day and the fire chief golf so got very excited by this

so moving on we began to work with more and more agencies within or Arlington County and in fact at one point we had a **data discovery workshop in which we brought together 90 of the leaders across Arlington County's government** just to begin to teach them about data science and the possibilities of what it could bring to them and one of the big things that we were doing was trying to break down the silos across these agencies so let me just give one example of how that worked the parks well uniformly I'll have to say that what they were most interested in is not who they were delivering the services to but who they were missing **who qualified for their services but either didn't know about them or didn't engage in them** they care passionately about that and so the parks and recreation were the first and the cue to come and say we offer these subsidies and we want to understand why certain families families that would be eligible for these subsidies are not participating in our program and so if you look at the top map on the right hand side that's all the families who are participating in just the subsidies to participate in parks and recreation programs and Arlington sorta has a divide the north part is considered quote rich in the south part less so but you can see that these subsidies are sort of spread it doesn't really tell you much right so we decided that **what we needed to do was to compute two different kinds of composite indexes both focused on vulnerability and actually that's portrayed in both the maps** and we combined some socio-economic variables for example percent of families whose rent burden is greater than 50% of their income or their participating in social benefit programs and we map that across neighborhoods in Arlington and so the darker the color the more vulnerable the neighborhoods so Arlington sort of knew that but you know this helped confirm firm that and then we created a second for mobility indexes by the schools and we use the school level data such as percent of children participating in English proficiency classes are those who were eligible for free or reduced lunch programs and so the size of the circle represents how vulnerable the students are there and you know there's somewhat overlaid and they aligned with our economic vulnerability indexes no surprises there but then **what happened was that not only did the parks and recreation understand where they could target outreach or improve their programs the Human Services Department and the schools began to get involved too and not just target outreach but began to reshape their policies** so that they could reach out to these children in need in and ensure that they also were participating in the programs across Arlington so this was the second story where we really got hooked and **we got convinced that these data these administrative data at the local level when you combine them with other sources are very very powerful**

41:05

so then we move on to Fairfax County and they go well this is all really interesting and we like the work that you're doing but **we don't really find it actionable to look at data at the census tract level** we are really much more interested in looking at data at by high school pyramid boundaries you know your high school your two middle schools in your elementary schools looking at those districts or looking at our political supervisor districts those are the areas of geography that are of interest to us **those are the areas of geography that are actionable to us we can take action through our school system and through our political districts** so we thought about that and we said well you know right now the American Community Survey and some of the other federal sources of data they're not aligned to provide data you **you're sort of stuck with either zip code or census tracker census block and so that's when we took our synthetic technology that we had been developing for other reasons and we realized that it applied here** and so what we did was you take the public summaries of these data and you can match them with the Micro data and you impute synthetic people the 1.2 million people that live in fairfax county and then you can place them in their households and now you can read you rewrite it and you control for a handful of variables this is very similar to sampling in our case we were using age income race and poverty so that your totals some back up to your published totals so we didn't have the same problem which i would have been more interested to delve into with the time series but this is a pretty straightforward **method that allows us to now draw new geographic boundaries that allow you to look at data so in these two maps for example it's just one variable the percent poverty in the top map you can see the percent poverty by the high school pyramid boundaries the twenty four of them and in the bottom by the nine supervisor districts** so that's just using one variable

we then went on to create what we call **CommunityScapes** and so this is again reflective of the fact that communities are composed of very diverse neighborhoods and not just census tracts they help us conceptualize where we live learn work and play and they **lay a foundation for a quantitative understanding of the kinds of the social determinants of health that policymakers are interested in** so in this case fairfax one of the issues that keeps them up at night is the **increasing** **obesity rate especially among their youth** and so they asked us to look at and create a composite index that would help them identify where risk for obesity is high in Fairfax county and again that way they could target policies to these specific areas and so again we created a composite index that looked at not just socio-economic factors but also their access to healthy food their access to physical activity and so this looks somewhat different than the vulnerability indexes although some of the areas do overlap which is not surprising as well and then in part of the overall analysis we were also able to capture the analysis of use from a Fairfax County Youth Survey that's administered every year as well

so bringing this all together we have worked with many different kinds of sponsors from industry to Federal Statistical agencies to DoD in the US Army as well as the local government examples and **out of working these really applied problems we call this a research pull we have created a data science framework and what's a little bit different than what was in the conceptual paper that you provided to us ahead of time or what you see in many data science frameworks is that most of them start at the third step that data acquisition or data ingestion and governance** and so we what we think is novel about it we actually had great discussions yesterday about this is what's key is working with your sponsors to identify what that problem is and so Fairfax came to us and they said we're concerned about obesity but then what is it that we're trying to answer about obesity we're trying to understand their access to healthy food their access to physical activity and so you begin to sharpen what those questions are you do that by both of course looking at the literature but also talking to your sponsors and talking to experts as well and then **what we think is really novel is our data discovery you don't just start with the data that you know you don't just start with the data that's readily accessible** and here's we're having our young people work with us because they are masters of data discovery and identifying unique sources of data that come to bear on these problems **and then the process becomes much more like what other data science frameworks are where the data wrangling you're profiling the data to assess your data quality and you're preparing and linking and exploring and all of this is iterative** and so we think it's rigorous it's flexible but it's also iterative in the work that we're doing that data wrangling takes about 80% of your time and we're hoping with what we're doing to create the science of all data is that we'll move that needle down it would leave more time for statistical modeling and analyses as well

so finally the last piece is the **ethics review** and we've heard a lot about this we heard Andy Brooks talk about how that's sort of the last stage what we think is that **the ethics review has to be built into every stage of your research** and so we have set up a very **generic checklist that is provides a guide for this discussion** and I will have to say I'm going to talk about our **data science for the public good program when we introduce the ethical dimensions** this summer a lot of the students where it really didn't get it they get data discovery and they get learning how to code and do this but the ethics they're like if all the data are public what I don't understand the ethical dimensions so we had to have some long deep conversations about this throughout the project lifecycle

47:05

so getting to **training our next generation of data scientists** here's our data science for the public good group from last summer we do a national search for undergraduate and graduate students here's some examples of the applied problems that they work all have a sponsor they always start with the sponsor can give us data but that's not a requirement unlike other programs that they do we **emphasize the first thing they learn is our data science framework** many of them come to us and they don't know how to code and at the end of ten weeks they have all learned to code in our proficient in that as well as the broader data science world so they worked through these their problems using multiple domains and we're very excited about this program that we established in 2014

so this is an example of some of our **sponsors** you can see a lot of them or local governments or federal at the Federal Statistical system one of the things that we're most proud of is that we at the beginning of this program thought that if just 10% of them went into civic activities or civic jobs that would be a success but in fact 28 from 2014 through 2018 **over 25 percent are now working in either for not NGOs or for primarily local governments** but some for the federal government as well about 20 percent are in graduate school and the rest are in the private sector but we consider that a **huge success in terms of creating a next generation of data scientists for the public good**

so we are beginning to ignite a national movement we received a USDA grant to bring the data science for the public good program to Oregon Iowa and more broadly across Virginia and we're doing this through the Cooperative Extension Service and through a process that we **call community learning through data driven discovery**

so we were asked to make **a list of actions** I through our discussions over the last two days and through things that I care passionately about have suggesting a list of actions that **data literacy** is as important as becoming a full-fledged data scientist I think that some of the actions that we need to address our **data sharing agreements** they often get in the way and hold up our work there's ways to overcome that that are not terribly hard **training** that we've been talking about through both of our through our talks and then of course the **ethical dimensions** of data science and geospatial research and finally something that we haven't talked about too much is **privacy** we are an advocate that you don't mask this data but instead you **punish misuse** and you find ways through either data on clothes or research agreements that allow researchers legitimate researchers to have access to these data though we'll end there

## Discussion Groups on Actions

50:30

Ajay: you mentioned a number of projects where the data was sourced by just people just citizens on the ground do you have a sense for how accurate that dataset is how high-quality those datasets turn out to be because I think they turn out to be really accurate

Jeanne: I totally agree with you **citizen science can be super accurate** so when I was at the White House before on the data.gov team we created with a bunch of other folks at the federal government the **citizen science toolkit** and so I just referenced that as a way of trying to bridge this gap between National Statistics offices and science organizations who said citizen science is interesting but not really accurate and so we said look if you follow these methods for collecting the data using metadata taking it in certain ways then actually a lot of the scientific organizations would be you could use it in California for example we have a huge issue right now with decimation of our butterfly populations which is really important actually and so citizen scientists are really important for that even in our **air quality work with NASA so I have asthma I have a GPS device on my inhaler** we're giving those away to school kids and our most asthma problematic populations and so they use their inhaler when they have trouble breathing we anonymize that map and share that data so that as a city we can go and send investigators to see where those issues are that are causing population problems around breathing and then individually that information is useful to the kids because just knowing where you get irritants in the atmosphere or where there's particulate matter for you our allergy issues **you can decrease your daily asthma attacks by 85%** so so **we strongly believe in citizen science** sometimes it's more for the getting kids engaged and understanding about data and data collection and the power that they have to contribute and sometimes it's cuz we actually need them to do the work

Stephaine: I'll just add one comment to that is that I'm on a **National Academy committee that's looking at prizes and crowdsourcing and one of the biggest challenges is just getting government agencies to embrace doing the crowdsourcing** so I think that's a reverse challenge to the sense

Patrick: from for the Earth Observation community is also great to hold on citizen science the problem is the Quality Assurance part you know you don't you know and **there are some things to work on that you know like the gamification aspect or crowdsourcing or you know you can do a revaluation of citizen observations** you can look at you know how a certain citizen voted for a certain thing also and the other problem that we have with it is that the labels that citizen science provides are often kind of qualitative rather than quantitative but then you know things like you describe or a small hand you know small sensors or so that you actually give you a quantitative reading are really valuable right for air quality there's something even for you know a for attribution also so these things are really

Nadine: open data it appears to be a crucial element in the data science workflow so what **can we do action again to further increase and ensure availability of open data and make recommendations to them**

Stephanie: okay is this not on just close okay thank you so I think that one is the question of **availability of open data** which I think increasingly has soared in terms of quality and quantity but the other is making it **more standardized if possible it comes in all different formats and I mean it's not these are challenges that can be overcome but it also needs to come with more metadata** as well it seems to be you know compared to five six years ago it's much better than it was it still has a way to go in terms of being standardized another place for standards

Patrick: I can maybe add to that only the Earth Observation perspective and again you know we have the **big open archives of observation data free calibrated readily available but what's lacking is the you know availability of in-situ data a reference data and training data** and here you know I was mentioning that earlier to a few colleagues here already especially **the land applications community is very unorganized** you know because it's so heterogeneous so you know research groups have their field data on some weird storage media old you know **not according to you know common protocols for the collection of the data and also you know in terms of metadata and actual data formats** too so you know it would be great if these if we could **motivate these rich research groups to reprocess their field data their reference data and bring it into the standardized format** and we actually have a pilot activity on this now to do a community review to look at you know best practices and data set specifications for a reference data and training data and then basically to create incentives for people to reprocess their field data and make it available in a curated and long-lasting way so basically you know in the era of open data archives for Earth Observation at least you **know it's the reference data and training data the calibrated data that's really lacking and that's also kind of hindering the major breakthrough of new machine learning technologies**

Jeanne: I would say and I'll just um those aren't great so I'll just add to it that there's this whole aspect of getting people to publish the data in the first place and there's **issues around both you know sort of expertise and liability** so like we should just be really clear that governments in in the US it's not true but it is true in Europe that if you publish as a government person data that's flawed and we know that all data is flawed or incomplete to some extent let's just be honest that that if something happens if an autonomous vehicle follows your road data but it has an accident and you're somebody you are potentially liable individually the US laws are slightly different and so **when we started building Data.Gov I had to convince the folks at 175 federal agencies that it was okay to publish messy data it was okay if your data was incomplete it was okay you know if you were publishing the data the best way you could today you just also had to be open to getting feedback on how to improve your data** and you had to be really honest about where your data was incomplete in nobody was gonna well we weren't going to hold you liable people might judge you and and I think that that issue of **change management behavioral science is really important and how we open up data and keep it open** and then we can't also ignore the **political issues around keeping data that may be inconvenient for one administration or another keeping that data open from you know over time** because people are running businesses and they are running in analyses and data journalists are linking to this data all the time and they need to have real-time accurate data that is you know standardized to the extent possible but accessible over time

Hammad: I have more of a comment than a question and the open data and sharing data out of research particularly in many cases I mean researchers received particularly federal grant money and they are supposed to publish data openly right but the challenge that you briefly mention is the incentive and resources that they don't have so they do the research everything ready results are published in a paper at the end you see in a statement email me to get the data right the reason is that the PI and the student or postdoc don't have the incentives to put it in a public repo or documented or don't have the technical resources usually they're more of a scientist than an engineer **I think one thing that would be helpful is policies and resources in those grants to help grantees to share their data openly beyond just a policy but on the implementation side**

Patrick: I think that's a good point you know and I think we're seeing more and more that **final results and products from scientific papers actually made available already you know there's data sharing environments where people host scientific result dataset** and link to it and you know I think what we're now also beginning to see is that things on github also are linked into scientific journal papers and also you know in **with Jupyter Labs or so you can actually you know share a methodology and data at the same time so that's like I know some people refer to that as the gold standard of scientific sharing** **and reproducibility** you know that's them but I like the point about the you know creating an incentive already in the under project formulation that you say like this has to

Stephanie: so it's taken us a while to do this but for the last two years we're finally at a size and structure that allows us to make our data available and our code through github and so we're beginning to do that as well more and more but it actually took more work than we expected it to and as you say that knowledge and sort of having that extra etra resources to be able to do it so it's not it's not trivial especially if you're just you know a research team of two or three

Question: the statement you made Patrick about the **need for ground truth data** I'm curious a little bit more on your thoughts on that and opportunities to address it and efforts like **Geobon** and kind of standards that help the community writ large generate data in a way that's usable in those applications

Patrick: yeah you know earlier in the in the session where you spoke and the colleague from the climate corporation I was gonna ask that question I wanted him to specify you know he said they have these something trillion data measurements right and he said and we're working towards defining the standards and everything I wanted to hear his opinion but he left and I was also gonna ask you because you said that you have worked on how to you know protocols you know best practices also for how the data should be curated and generated so you know I think you know as I said **it's very heterogeneous depending on the community how that's done** so I think like you know we're doing this community review to see how different initiatives that doing that **Radiant Earth is of course already going along that line** and I think you know overall everywhere we're hearing the same thing that people are realizing that this is kind of lacking and I think there's a couple of efforts underway to do that I mean you know it really depends on the application example if you look at **land cover there's a very elaborate protocols** and you know in in Europe we have the Lucas rank of a survey it's very detailed established methodology and you know metadata agreements and everything but things like you know **yield and biomass and these kind of things they're much more interesting** and you know if you look at I always like to bring up the example of you know mapping **biomass and the South American Amazon there's been ground campaigns then since the late 60s or so right measuring you know doing forestry inventories and doing field forest plots and stuff right imagine you can bring all of that data together right that would be such a rich datasets** really depends on the community in the application but you look at I think

Regan: I think there is I've noticed in my own work in the past couple years that there is a push for that there's in the biodiversity world there's Darwin core and other you know **we've maintained our own standards within our network for a long time but now we're thinking more globally about how can you apply these in these new kind of international frameworks** to maximize the reach so it's an exciting time yeah

Patrick: and also you know thinking ahead a little bit in the future like you know you're gonna have maybe you know the aspect of **training data interoperability** would also be interesting you know because you might have one environment with GPUs one environment with distributed storage and compute and then a very environments that hosts reference data **and you might want to query different reference data sets for the same type of measurement you know you know federated access to different reference datasets would be super cool** so but you know so that **aspects of interoperability would be interesting to see if OGC would want to pick that up at some point**

1:04:40

Antoino: so I've seen that **two of the main problems about access to two data is versus availability to today that is actionable and the other is making sure that the data is good quality** do you see that there is a **role for entities that behave as data authorities in the future** that say that centralized sources of data validate them they ensure that the data and the metadata is good quality follows the standards and they simplifies also the data sharing agreements or do you think this is not a viable model and that the data will still remain in a big lake of data where you will just have to look for it by yourself and figuring out yourself if the data is good or not well

Jeanne: I think we've in some of the groups we've talked about this **issue around data sharing brokerage's or Tinder for data sharing agreements** or something is that we need to start making the data that is not fully open or is open under certain conditions accessible so I know we all have like Creative Commons a different licensing that we can apply but if we don't sort of **proactively figure out someone has a need and someone has some data how to put that together so we have you know sort of aggregation sites** like Data.Gov of and the UK sites and then we have all these national data portals and others around specific areas but I think there really is this idea this **potential need for that you know sort of provenance checking but it's it's sometimes difficult to figure out which international entity could do that** at a national level like I did a lot of work for the World Bank in central Africa at a national level there's things that you can do like in a country level and certainly I've seen that some of that happen in the US but you know internationally it's probably a little more difficult

Stephanie: the data **the quality of the data are useful depending on your purpose** and sometimes the data may not be the highest quality but it may still provide important insights into what you're trying to answer or address and so we talked **about it's more understanding the strengths and weaknesses of your data** so you look at Twitter data it's young man or no I mean that's now expanded unfortunately to older man as well but anyway but you know we know we know sort of the demographics of Twitter data to some extent and that's also evolved over time so I think that sometimes I want to be careful that we don't just throw data out because we think they're poor quality we want to maintain those with respect to you're sort of been organizing an organisation and we have some of those in the United States some are very bureaucratic and hard to access and others are less so **Pittsburgh has a model having data warehouses seems to be pretty dynamic and researchers have access to this data without having to go through the gathering and data cleaning** and you know those data are there but you could see that quickly our **research data centers that the Federal Statistical agencies run are valuable but it takes you over a year or longer to gain access** to those data and that just can't be when I was at NIST we had a valuable source of innovation data that we made available through the **North data Enclave** that was actually created for that and since then many other federal agencies are putting their data in there because the turnaround is less than a week or a month it's very short compared to the research data centers so we need to figure out a model that's dynamic and flexible it allows researchers quick access to these data so they can get in and not wait the whole year or two years to access it so I don't know if there's a perfect model yet but I think it is something to strive for so that we're not all reinventing the wheel of finding these data sets and describing them and creating the metadata for them so so no easy answers

Nadine: actually it's two questions for each panelist so we'll go maybe Patrick gene and Stephanie so the first one is what would be your **top recommended action** and the second question would **be what do you think or OGC can help with as a community** Adam get ready

Patrick: I think honestly you've probably already heard that between the lines you know that **I think we should really think about you know having this reference data ground truth data Cal Val data training data in the end to think about that direction to provide you know data set specifications protocols** and then also you know environments that are dedicated for working with training data and curating training data on the long term and you know I don't I'm not so involved in all the standardization activities but I think it could be good to think about that too and as I said before also **the interoperability of these reference data repositories or training data repositories is something that could be really really valuable** so I would that would be my two points to put on the list

Jeanne: I one of the things I think that would be helpful you know **schema.org** and others have helped to create searchable like file type data is now something that you can search for in Google and so **expanding that standard to include the data sharing agreement** aspect of it like like right you can now you can **look for images that have certain copyright or open new standards** having that applied to the data that you find on Google would be really useful and a relatively straightforward thing to do and it would take a long time to eventually get it to be adopted but I think that would be important and then I think the other thing is we've talked about this but I just I want to lift it up again is the **digital code of ethics** is you know if we could figure out sort of those things about inherent bias and transparency and you know unintended consequences and figure out a set of values that maybe the community writ large could you know step up to then organizations could potentially sign up and say I also you know like the Paris climate Accord um as an American I can say that that we we agree to adhere to these because it's a value to the community

Stephanie: so something that's been swirling in my head the last two days because I've been really **impressed by how many people are here from the private sector** is what about private sector data and we do know some companies provide that some provided if you're a researcher and you know maybe then the data are on quality are unknown but that's part of what you as a researcher are doing and hopefully contributing back to the community but there's also **data sources like CoreLogic or burning glass** that we're that charge a lot of money to get these data the data quality are really poor and yet they're really valuable data and so the core logic data provide you property tax data and characteristics of housing units and in some localities you can even get the rental units but the data are why every researcher would have to reinvent the wheel on that is unknown and not only do you have to reinvent the wheel you're spending lots of money to access that data again I don't know what the solution to this is but **I do think that the private sector needs to step up and make data accessibility and data quality an important issue** **and be able to be part of this research community in terms of providing good high quality data for very important issue**s so so that's one of the areas and I was so passionate about that I now can't remember the second area I think it's related to the **data ethics** though and I do think that the data ethics is just absolutely well

Nadine: you're happy Adam - happy George you're happy all right thank you

so much thank you the panelists for the

energy at the end of the day okay

# Summary Session

1:13:30

so we've learned a whole lot that means there'll be a test

actually the lucky thing is we already had people preparing the summary throughout the day so let's

spend a couple of minutes reviewing what we heard for you all to take away with you

## Session 1 Summary

Marc Armstrong gave us a really great fundamentals presentation Nils gave us an inspiring kind of look at the history of video we had an excellent panel Kathleen Anand, Marc Jayant and so each of these summaries will be I think if they all use the template in three parts one is what works second is open questions and third is next steps

so what works you know lots of things are going on that are valuable the data science massive quantities of data parallel analysis video analytics the whole ability now to do space and time look at changes in data time series trajectories data cubes on imagery we got a little bit of that at the end as well

open questions so you know it was very interesting to hear the throughout the discussions about edge computing and cloud a very good use of object stores to support data access but also then what do we do with computing that happens at the edge like the ISO committee that addresses us as cloud computing cloud is computing it may happen at the edge and so you may see containers being really deployed at the edge video one of the pithy quotes there was MSIB works great but it has fundamental items that need to change in order for some of the future video analytics that we need a lot of discussion about models brittle bias and benchmarking digital twins patterns of life lots of open questions about models that continue throughout the day

so next steps entirely new classes of sampling strategies the heterogeneous computing architectures with higher-level program interfaces that hide that complexity we already do that which he GPU use right but you have the whole set of standards for GPUs that make that possible mark gave us some other computing types that we need to have additional capabilities with the relationship between digital twins and models for data science we can explore that quite a bit the deeper understanding of mobility and harnessing that mobility and lbs data and then lastly both in this opening session and throughout I think we heard we should be looking at Jupyter notebooks with respect to some way to develop to promote development and reuse of workflows we know this happening across the community what should a standards organization be doing with respect to Jupyter notebooks to promote that I think is an open question that we will take on we actually will do that next week so

that's what I heard session 1 is there anything that is glaringly missing you know partly a refresh because you guys will walk away with a richer understanding of what happened it's part for Trevor because he sat out front the whole time it didn't hear any of it so this is all for Trevor you all met Trevor earlier that's what I have on session 1

## session 2 Summary

Annie Burgess

1:18:35

so very very briefly before I go into session 2 for those who I didn't meet I'm Annie Burgess from ESIP for those of you that don't know it ESIP is you all should since you came to this event it's earth science information partners we work very closely with OGC we're community of data and informatics professionals if you were ever wanting to say hey I work at a federal agency wouldn't it be great to chat with people across federal agencies private industry nonprofit academics and just talk Tech shop around data and metadata we are your community so **ESIPfed.org** if you don't remember that you could look it up or science information partners

all right so this session was analytics and representations Yolanda kicked us off and really focused on knowledge and models and then we had this cross panel from Todd Lauren Rob and Keith Rob wasn't there I was like I don't think he was there all right

so what works Yolanda really focused on adding knowledge and how it enables more automated and less manual modeling another thing that worked was just the idea of provenance and capturing that provenance and model and she noted that everything within her modeling has a URI to point to and so that being able to explain models and have that is really critical and something that is working Radiant Earth pointed out that they kind of realized that they needed a standard and then it was that shopping it around and really by bringing it to different communities you ended up with a community vetted standard probably much quicker than building the standard and then just getting comments back throughout the years so I thought that was another good something that worked and then we heard from the SQL right here and I think it was just another thing that worked was that the standards community hears what people are saying and that they evolved and build new standards based on what the community is bringing back to them so I thought that was another good example of something that worked

open questions the idea **of differentiating between prediction and knowledge** and when it is okay for it to be a prediction versus knowledge there's a question of what is **ML data versus Big Data** and I think the very quick consensus or maybe not consensus but suggestion was Big Data focuses on the bees but it can be unstructured and unlabeled whereas ML data must be labeled and structured so no one can quote me on that as the difference but I think that was kind of maybe an open question always questions of **ensuring quality of training data** we've touched a lot on this of what does it mean to build up a next generation of data science how to maintain training day bro data that has a temporal component in other words data that are changing over time and then someone brought up sharing data with local entities and just that at that kind of fingertips of the arm of government that those people really need high quality data and so are there different ways to share data with that crew

next steps **compile training data that matched the diversity of landscapes** for example agricultural fields around the world and other kind of changing geophysical entities there was a I can't be more specific than saying **we need to work more in the raster side of standards** so I'll take I'll put that out to OGC also I mean STAC has kind of been peppered throughout this I mean it's fairly new it's super fresh but I certainly see I mean we don't need you know that STAC will kind of continue to be in the conversation and then let's see **EO data in cloud friendly formats** again the kind of cloud optimized getoTIFF HDF netcdf discussion that will continue and we ended with the idea of synthetic training data sets and kind of pros and cons but more generally that there's probably a need for more domain-specific synthetic training data set so I would put it out again like George did anything that I missed that you think would be critical to adhere right 1 2 3

1:23:15

## Session 3 Summary

Kyoung-Sook Kim

that's it and I heard a call for synthetic people session three i'm i'm sure i'm i missed a lot of things for that do my best for that okay my name is council came I'm also the from AIST and I responsible for the OGC GeoAI I to week and also moving-features WG co-chair now lots of we work in OGC

so is the session 3 is live trends and especially is really interesting talk and it's palatable as the AI at edge so I always think about the AI in the cloud but he gave a different aspect and then the panel we have a full great interesting talk and more detail the next so it's pretty he introduced of the mapping of the information and customer necessary from to prove it for emerging technologies he introduced last Java Martina a hottest market sabe and and especially he covered he said it's AI will be the game changer but in IOT area steer is already steps I really surprised that because IOT is more than you know has some history nowadays but still the market is not really increased increase that but he expected that maybe 2021 or 2020 maybe IOT in industries become increased and one of the key use cases he introduced up the transportations I think the mobile smart mobility is one of the big target to combine of **the IOT and the AI at the edge** I think and what is the new rearm when we think about the eye at the edge and data security is one of the side may be is a relative cybersecurity and the bandwidth and latency is we have to think about that and also yeah better interesting workflow a data floor he showed that and between the cloud and add we may be that is very interesting yeah the flow so maybe we have to think about how can we support those kind of different workflow M hundred discussion is one is a skill set diversity and so and also special temporary intelligence and situation awareness and crowd a I especially cry the AI his they they introduced large over use cases for analyzed of the image Nuria P do to detect of the change of the infrastructure after disaster so I'm also in Japan we have lots of disaster I think sometimes it's a department store of disaster but anyway so we also very many people working on to this area and IOT data and in a purpose problem and the last part AI can be asked to manage of the Latino data problems and determine value of the data at edge so that is kind of this panel skip asked this where are these contents and this

is open questions I don't want to read all of sentence but what is that official robot AII pen that times what but we notice about dag of the terror so I think it is special especially I think the training there is that so for you know we have lots of data but still it's not enough for training the motion I think and and especially the hardware player so an Nvidia team he mentioned about the heterogeneous computing architecture you know now is **GPU CPU efforts FPGA and Google is TPU** and so let's up now is community or computing Union it comes up so we I think the hardware peoples also think about more heterogeneous you know the architectures and so the meaningful safety project is responsibilities disaster response and this one is everybody cannot yeah answered and everybody said all of the things we maybe need to to combine and data science and AI and another exhibit open question because this is I just put a date because we shared of the this question on the mail and all but

I want to share with you the next steps is first of all how can we create value chain so how can we think about the ecosystem is very important part and how can we support optimal development and another things issue is geospatial data science education problem so how can you know what kind of skill we have to consider so cross-disciplinary over convergence of eg domains and in the communication skills part teamwork so and then I also want looking at the New York University they have a good you know education system for this area and then what when we need to retrain models so maybe today is just important some people mention about the quality and but we usually discuss about the quality of the data but now is a people's mention about think about the quality of the models so not only the performance you know the accuracy or some fairness over lowest knees and then a a community peoples is sort of the disgust of the what is quality of the models and the how can support of that may be yeah that is kind of next step and on other things because I'm not you a I too that's why I wanna put because whenever eyes mention about the Gio AI and many people ask me what is it zo a I so we need of definition of the toy a oh yes okay so now it's doing have developed some definition what is the GeoAI

anyway I wanna share the one of the things of definition of the AI so this is not you AI it's AI definitions and I also you know the 1995 they I spoke every and they defined of the true definition of the AI and recently the JTC 1/sc 42 is artificial intelligent groups community star they developed a vocabulary I'm very surprised because to definition is a little bit different and then to define the GeoAI I then kind of things I have Trevor but anyway I want to yea discuss with you in August a meeting so please join the OGC TC meeting thank you

Session 4 Summary

1:31:55

name is Ajay Gupta and as a previous report or mentioned want to put a plug in for the TC next week in Toulouse France I chair the health WG and so we are meeting Wednesday at 4:15 a.m. Eastern time 10:15 a.m. France time but I do work at health care as our domain we are a population health firm we work with providers and departments of health or ministries of health to find solutions to reduce the cost of care we do that through a geospatial analytics on health outcomes and health care costs in America we use Medicare data we have access to the latest Medicare data from CMS as well as all of the social environmental data that we've been talking about today to find out why people get sick and stay sick and do something about it outside in the communities so that's what we do this

session 4 on outcomes and applications that's not a really good session on why we can do some of these things Wendy got us started with a really good overview of things that are taking place in the field and Regan Megan and Steven and Ed talked about a lot of these kinds of things that are working in the field and holistically there was I thought a lot of overlap in what they were presenting in that to get something that works a geospatial application that works for a positive outcome essentially starts with **merging standardized data from disparate data sources** that may come from government or one of many government agencies could be crowd-sourced it could be field collected or could come from private industry and then can **serve a broad spectrum of applications** road rail safety emergency management I owed you see just did a pilot on emergency management areas tracking of containments public health issues natural resource management and logistics

there were of course open questions and while this slide looks a lot less dense than other repertory slides on open questions that's not to suggest that these aren't deep and nested questions but **questions of error modeling and bias** I think we've had a lot of discussions on how exactly we're going to address bias right and in one thing that we can say on bias is what does that mean bias as the term is itself not a negative term we've given it a negative connotation right because we associate it with discrimination and those kinds of things but bias just means what's the inclination so sometimes the bias may be positive towards whatever it is we are looking to identify we're **human knowledge and involvement can be in place** of course there was I think agreement amongst the panel that human involvement in the use and development of models and the application of outcomes for models is important but the degree to which human involvement is in is should be involved as a question there's a **calibrated eyeball test** or the sniff test that the panelists mentioned and quickly **an example from Healthcare's University Oklahoma did an analysis of ophthalmology Co** readings from an AI system versus actual physicians who would read would read these test results and found that the physicians had an 87 percent correct accuracy rate the a I had an 82 percent correct accuracy rate right so we're still better than computers but **when AI and computers are working together to identify and diagnose diseases of the eye or the retina it was a 93 percent rate** so **if we work together man and machine we can be we can certainly be better** so human knowledge is not necessary replaced but **augmented data retention and curation** the medium and stewardship all of that is an open question as is **ethics and privacy** that came up in all of the sessions what are we doing with data how are we gonna give it and make it openly shared and how much privacy do we really need and we have so much privacy on data is it going to be available for the work that we want to do and the research that we want to identify and

next steps obviously the next steps is to address some of the open questions right so that's clearly there but there's **certainly need for more standards on data and data collection and for those standards to be balanced against potential risks to innovation** but the panelists did mention a number of places where innovation has led to positive I'm sorry **standards has led to positive innovation in a number of different fields so the right standards at the right time is certainly a model that we can we can follow**

**model validity** itself as an area for next steps and communication on the model and machine learning right that it is again this is a quote from the panelists not a human replacement but a force multiplier and as it was mentioned earlier there's probably stuff and content that I've missed if there was anything that I missed please let me know or let George know so we can put it into the final report thank you [Applause]

## Session 5 Summary

Adam Martin I work at ESRI my colleagues had to catch a flight so I apologize they're not being here and then I press what to get the screen presenter local present info ah great okay so you're still here thank you yeah I don't spend too much time since it's all somewhat fresh but

what I heard from your speakers is that government investments in infrastructure are working at some level or are starting to work to accelerate the use of AI you know both in Japan and European Space Agency to help with experimentation and providing better Earth Observation products but and it continued into drive to provide more analysis ready data you know continuing to move up the value chain from bulk downloads in you know random formats to usable services available through APIs etc so that's what that was a big theme you know **starting to actually use AI to impact real-world policy compliance efforts** and in the agricultural sector and the you that was a great example which I think you know as ultimately the aims right and were these problems that we're trying to solve you know in this case the value chain is super large lots of capital investments in terms of satellites you know for specific purposes this is one of the outcomes that somewhat drove when I heard drove those investments as well as **are helping to reinforce the value and the ROI of some of this data** and so the more we can implement that the better those kinds of practices a lot of discussion with our university and local government panelists around and in NGA **how to train data scientists within your organization** and doing that partly you know through this hub-and-spoke model understanding where you are in the solution design landscape process but ensuring that you know it continues to be a team sport that you're getting a diversity of perspectives and partly doing that through partnerships both with University and in the community so we heard about the **Data Science Federation** and the growth of that partnership with the universities to tackle actual problems that local governments had as well as the **data science for public good program at UVA** that's also scaling out so it seems to be you know a good model you know that points to I guess some of the other next actions which is essentially keep convening across these sector boundaries yeah

open questions you know you just heard some of them right just as few minutes ago **what are the standards we need for training data a reference data** how do we make sure that whether it's the STAC or others the specs allow for easy exchange and exploration across addition at different national government investments are there existing **codes of ethics** or value statements Bill of Rights etc from from la that that we can we can adopt as a community or shape I'm sure they're out there there we don't need to reinvent the wheel what are those what are those that we should adopt how can we lower the barriers for for **private sector data sharing** especially for the research community not to call out core logic but sorry these are fresh notes I just copied and pasted it in there and and then how do we accelerate you know the population the metadata that came up in some of the conversation and that stuck out with me because I'm actually working on a **project Esri to to enhance metadata management** and search functions within our ecosystem so I'm very interested if people have good ideas about ways that the software can populate metadata and in a formalized way because that's what we're looking at

so the next steps you know similar things right how do we ensure that the Jeanne brought up specifically like data licensing and making sure that those are standardized and adopted throughout the web so that we can generally find those easier because we've heard that a lot as we as well as like **hey great you have open data but I can't easily find which ones I can actually use which ones have the right license types that allow me the user commercially or otherwise** and so we've actually been working on up for the data that flows through our system making that more available but there's still need for that to propagate throughout the web and yeah bringing keep convening how do you OGC doing what you do which is bringing industry and academia and local government federal government together to tackle these problems Thanks thanks for doing that

# Thanks

all right well we're almost done last steps if you presented something please send it to me including the ones that you just presented and the like please send those right away because I got a report out next week so the video we've been recording video and the like and Eds got that it going to go up on YouTube the video and the presentations will be linked back into the webpage where you saw the agenda and everything for this lastly we're going to generate a report or a white paper builds off of what you just heard from the summaries and other stuff and and gets us to the recommendations that we ought to carry forward so that's kind of what's going to happen coming out of here

so in terms of thanks and thank you to folks involved the panel that I mentioned a few minutes ago the organizing group that I presented the names at the very beginning of yesterday and they're on the website as well Ed thank you buddy for thank you to Google for sure but thank you to you personally so he's British I'm gonna embarrass him just by thanking him right and thank you for coming and participating and being involved here because that's really what you know made it work is your involvement and the like so I will of course leave it for the last word to our CEO

I think there's one big thank you that we owe to George because he's behind all of us actually being here so thank you and you know I you know we can't do better than this to show you what OGC is or actually who OGC is so if you're not familiar we're here talk to anybody get engaged i you know ii everybody who said you know we need to convene some more so thank you again for all of you and we thank you George for putting this together as part of our like future-looking tech trends activity at OCC and safe lights everybody thank you