# Transcript Day 2 video 1

# Session 4: Outcomes/Applications

Jeremy Morley Ordnance Survey chief scientist

welcome to the second day into session four which turns more towards thinking about kind of applications and application areas for data science and geo data science

whatever that may comprise so on screen you can see the rough agenda I think what we'll be doing is going basically straight into the keynote for this session and Wendy's talk after that we'll have a comfort break while we kind of reorganize for the for the panels and so on so somewhere around 10-ish and then the we have a slightly different structure for the discussion groups today which is apparently in the Google tradition of when you want to think deep you get out and so we're going to in the to use an OS bus reboot and we're going to get outside and have some walking groups to for the structured discussion before lunch yes I think though we if there's groups that want to stay here and sits then that is fine not everyone necessarily wants to go for a walk but we're on the edge of some beautiful outdoor areas I understand so stretching over some not indoors outside spaces so we'll experiment with that before lunch as well so we have **an interesting cross section of thematic areas to explore today** I'll say a few words as well before the before the panel too but I think probably will go straight into the the first session the the keynote so could you please

## Presentation on Outcomes - Title: Integration of Geospatial Data: Examples and Implications

* Wendy Martinez Presentation: Integration of Geospatial Data: Examples and Implications

welcome Wendy Martinez from the Bureau of Labor Statistics

First of all I'm happy to be here today to get us started for session four on outcomes I have to say that I'm I'm pretty intimidated when well now and that especially when George asked me to talk because I unfortunately do not get the chance to work with geospatial data and so there's so many of you around here who do and and know a lot more about it than I do but fortunately I do get to hang around with people who work with geospatial data and this is through the **Federal Committee on Statistical Methodology** or **FCSM they have a geospatial interest group** that I've been involved with for many years and then also the FCS M had a working group that's been focused on how do we deal with data quality issues when we have statistical products that come from integrating multiple data sources together so for example data from surveys or say administrative data and I'm especially grateful to you this list of people here because they're the ones who who did all who do all the fun work and they're the ones who gave the applications examples that I'm going to talk about so I'll just be kind of scratching the surface on these and I'm sure it's probably okay with the people who know more than I do and and Where's George oh they you're right in front of me **if it makes sense you know maybe we could have seminars a web remote access you know events where these these folks can explain these applications in more detail** and then yeah I said we could even do it maybe so it could have international remote access - okay great all right so

I'm going to try to motivate the discussion that's going to follow my presentation and the examples I'm going to talk about I'm going to focus on a couple things first that they are products that came about by **integrating multiple data sources** together and then also that that these are tools that you could use to download the data yourself to use because I think this is **really important to be able to access these datasets** and then I'm going to conclude with some issues that you know hopefully will kind of inform our discussion

I'll be talking **about five applications from these federal agencies** so this is your tax dollars at work which is good and the first one is from transportation and I'm really grateful that Edie Strocko is here because you have any questions on this one you know Ed can answer them but I really like this one because it in my opinion it really is a true example of an application of machine learning data science with a geospatial data and you know integrating multiple data sources together and the other examples I don't think have so they don't really have this modeling aspect to it although based on the talk I think at the end yesterday about AI at the edge I guess he did man she did mention the different types of analysis and one of them was descriptive so of course we do have **descriptive statistics** so I think the rest of the applications are more in that so

**With Waze data**

this first one as I said **Bureau of the Transportation Department** and what they did was they integrated information from ways how many of you use ways I haven't yet so I guess I need to need to start doing that but they integrated ways data with sort of what I guess what I'm calling traditional transportation information and they had three sort of case studies or projects one was talk with more detail about it was **they built models to indicate or predict crashes in four different states** and then they also did a sort of a similar modeling task for the state of Tennessee and then also for Bellevue Washington so I'm not going to talk about Bellevue but I do have a slide at the end serve in the back up area but I won't talk about that so in the first example of integrating the ways data this was over for state so it was Virginia Maryland Connecticut and Utah and I'll just go into these so the details here but the overall goal was to see how they could integrate the waste information with with the traditional traffic information and then see if they could build better predictive models for crashes so they had data all for 2018 as I said it was Maryland Virginia Connecticut and Utah they built separate models for each state **the response variable was let me get this right the presence of a police accident report in a specific grid and in some specific hour** so you see here the it is so here's the list of predictor variables there they used the count of ways traffic incidents information about traffic volume information about the roadway socio-economic data historical fatal crashes and **they used a random forest which is a machine learning approach** they train the model on 70% of the data and then tested it on the remaining 30% so this is what I said this was in my view a **real application of data science where we're creating a model where they were creating a model to do some prediction** and here's the results they created dashboards in **tableau** and so they have these are sort of clocks I guess for each of the states and it gives you information about how the model performed in terms of the percent of crashes that were estimated and what's interesting is you can see that so it starts at midnight up here and then you know goes around just like a clock does but you see for these states Maryland Connecticut and Virginia sort of in the nighttime hours overnight there's only 70% was were predicted of the were estimated of what was actually observed and whereas in the times where there was commuting so the commuting hours you see that there's sort of more observing more crashes were estimated and I have to point out that the based on stakeholder feedback they did err on the side of overestimating so you sort of over are you know including false positives all right and then the so it's kind of interesting because the Waze signal is stronger here during commute hours and so then that's where they tend to overestimate all right let me I just want to go back to here to the map because you can see the different colors here dark blue and white so the dark blue is true zeros where there really was a Waze and no police accident or no ways accident our police report was given so those were true positives rather than false positives and the white are mean so these are true zeros or negatives and then the white were true positives where the correct number of crashes were estimated so I think this shows that you know the use of Waze data can give you better information and better models and that's I think evidence to buy this Tennessee case study where they did a similar approach to do the modeling and what I found was interesting about this one was that the State Highway Patrol of Tennessee was already mandated to use **models to help predict crash propensity and so then they could better target their patrols** so the department transportation included Waze data incorporated ways dated the models and they were able to improve the spatial resolution down to one square mile and then also the the time to one hour so as I said the incorporating Waze information **improved the ability for planners to make investment decisions and more effectively target patrol areas**

17:17

so the next one is about disasters we heard a little bit about this yesterday but I don't think we heard about this actual application how many of you heard of **OnTheMap at the Census Bureau** okay some of you that's good so this is a web-based mapping application that shows information about where workers are employed and where they live and it does provide a user interface to go in there select data import your own geography or use theirs download the maps and more importantly as I said earlier download the underlying data so **it's a way to you know select data and then download it** so this is based on the longitudinal employee household dynamics survey this is the documentation page I will give the slides to you George that people have the links and everything here's just a screenshot of the tool but according to the census folks who develop this some of the **emergency management agencies start asking for more I guess analytics in the tool** because they really weren't any built into the current tool that I'm showing here so **they added some I guess smarts** to it that's a great word but by combining and information from different on data sources so you see here that **they have data from NOAA that gives information about hurricanes and storms and weather they have information from I think that's FEMA about disaster events** that are going on so things like the wildfires here in California we also get that from **Department of Interior** and **USDA** and then of course socio-economic information from and demographics from Census itself so they created **OnTheMap for emergency management** and here's an example or just a screenshot of some of the information that you can get from the tool so again I don't know that **there's not a lot of modeling going on here** like we have with the example I just showed from the Department of Transportation but certainly this is **a way to describe** you know what's going on in an area region and accounting for disasters or other events

so the next one has to do with a public health this is from the **CDC they have an environmental Public Health tracking network** and this has data and information on environmental issues on hazards that are going on the health effects that come about because of those those environments and hazards and then information about the population they have a similar tool I guess to on the map that you could allows you to explore the data that they have and what I was really excited about this data this tool I guess is that they have data sources from **national state and local agencies** is the California person here she is that you know okay anyway she somebody yesterday was talking about she was from the state of California and you know have Public Utilities Commission yeah so this is you know I think this is a tool where such a commission could you know **access data and information that they could use for better planning and and health interventions** and so forth but the other thing that was really exciting to me about this tool was the fact that it brought together **data from different sources that I just said but it's standardized** and then once you you'll kind of create the you'll get the data that you want or need for your application or your study then you could download the data for your own use so it's another way to kind of get data from these different agencies so here is the website I was kind of surprised when I was preparing the slides that it has been there for 10 years so take advantage of it I don't know if they actually have historical data that would be interesting to find out but they do have this **Data Explorer** and here is a screenshot of it and when we had our GIG workshop last month they gave the somebody from the CDC gave an presentation on this and it was really excellent so that's something as I said I could set up for everybody here so just to say a couple here's a couple examples of how this has been used and I got these examples from their website they have an information sheet one was the **state of Minnesota they back in the early 2000s they banned smoking indoors** largely for the most part at you know public places and workplaces but I guess in the past couple years or so there was a movement to rescind or remove those regulations and some organizations that well that's you know not a good thing and so they used information from the tracking Network to gather evidence on you know how there's this decreased exposure to secondhand smoke and then the health benefits that resulted from that so you know now there's there they've kind of stopped pushing to remove that legislation and **in California there is a concern about nitrate contamination and drinking water** so they used some folks use the tracking tool that the network the tracking network data to help **map the entire water system in** **public water system in California so it allowed them to identify areas where there's concern for nitrate contamination** and then propose ways that they might fix that and what was also interesting was I don't know how they would do this but they said that there's this helped them also find out where there might be some unlabeled or unregulated wells so they might have these like secret wells that were using the water so I thought that was kind of interesting and also wondered you know maybe they do have this in the Network but it would be **interesting to create some models for the EPH tracking network to maybe do some surveillance to keep an eye out for bad things that could happen based on the environment that could impact our public health**

24:44

the next one is agriculture so that one of the speakers yesterday talked gave an example of you know hurricane flooding and you know how they were determining what were the areas that were affected by the floods so this is a kind of a similar application this is from the **National Agricultural Statistics Service** **or** **NASS** and they were doing **real near real-time flood mapping of cropland areas** so here's some pictures of recent hurricanes and what's one of the common things what's one thing that's common amongst all these pictures right swirling clouds or high clouds and this was mentioned yesterday so that's an issue because the you know nationally so NASS uses information from remote sensing which is from optical sensors and so those are affected by cloud cover and night so that makes it a little difficult to you know understand what's happening with floods and cropland so they turn to **SAR or** **synthetic aperture radar** because this sensor can gather information our data day and night and through clouds so these were the two sources of the data again this is **another application of datasets being you know integrated** so they use information our data from the **Copernicus on Sentine**l one synthetic aperture radar two things I guess I want to point out here first is that they are it's free I really love free data and so you could there's a link to it down at the bottom and secondly that it's **free within 24 hours** so there's somewhat rapid production of or availability of the SAR data and then **NASS has provides data on crops in** the US it's geo-referenced and they have land cover data sets as I said though this is from optical sensors so this is from satellites that have multiple satellites that are overhead and so they do have as we said earlier issues with a cloud cover and at night and here is at the bottom is a link to the **Cropscape** which is where you could get this cropland data layer and I'm pretty sure it's also free they have information about crops at different types of crops as well as nonagricultural elements or variables or characteristics and so what they do is they kind of **put these datasets together do some analysis and produce products like this so we have a map and the red areas are from floods** so we could see or I can let's see from there but you kind of look closely you'll see areas of red from flooding but then they also have the percentage of the you know the amount of area and percentage that that's inundated or flooded they do have a website for this information so they have maps but again what's exciting to me is they have the geospatial data so there's the data for you also I'm going to talk later a little bit or at least mentioned **the idea of transparency and metadata how important that is when we are trying to integrate datasets together like this** and you can get that information here at their website so we know about the methodology that they use in more

28:58

so keeping along with hurricanes I guess it's a recurring theme okay this example is a little closer to home this is from the **Bureau of Labor Statistics** we have something called the **quarterly census of employment and wages QCEW** this serves as our establishment population and it gives information about the employment count wages for different industries and occupations but it covers more than ninety five percent of the jobs in the US **and this information is available over different geographic areas for states and countie**s I believe two and the next thing is it's **it's easy to download the data okay my common theme** here so here's the website for you to actually download the data and you can construct maps so it's kind of a way to like these other tools on the map and the CDC tracking it's a way to kind of visually and interactively select the data that you're interested in and then download it so here is a screenshot of the website up here at the top you have drop-down menus where you could select the year and the quarter and I wouldn't meant to check to see how far back this goes but it does go back years so you could you know extract data from the past - and then you have different industries I think whether it's private the issue of ownership and once you do your selection you just scroll down to the bottom of the page and there's a little link so I'm telling you because it's kind of hard to find but you scroll down to the bottom of the page and says you could download a CSV file so the in the past years maybe ever since I guess Katrina the **QCEW folks have created these hurricane zone maps** so it's kind of similar to I think what NASS was doing with agriculture are you know looking at how crop lands affected by the flooding here we're looking at you know how is how are the employees and the employers the businesses how are they affected by in a hurricane so this these hurricane zone maps show total employment wages and establishment counts for every county in the hurricane flood zones and **they do this by linking information from the QCEW with data from the corps of engineers and other state emergency management agencies** so here is a screenshot of the hurricane page hurricane zone map page and here is I think this was from Dorian and so this is what is produced so we have the to see the establishments I think that's the first column and wages unemployment and then for different zones and those are dependent there's a related to the level of the hurricane

32:22 Ethics

so I want to just switch gears a little bit and start so these were some applications of you know I guess I'll say geospatial data science when we have integrated data products but **I want to talk about ethics** and somebody mentioned ethics yesterday and I was really happy to hear that and I've become more aware of this I guess in the past year because it seems like I keep your kind of running into it and but it's been a bit of an eye-opener to me and I really liked what was said yesterday it was you right yeah yesterday because that was another aspect of ethics that I never I didn't think about but so I'm just going to just highlight some things I'm hoping this will be some things that we'll talk about later in the discussion so as we all know hopefully **ethics is a study of right and wrong it's not laws** that we you know we have to follow because you know that's what our government issues to us but they're kind of abstract and they're guidelines and they really depend on the culture too so you know kind of the ethics that we follow maybe here in the United States would be different from you know some other country in the world but there's something now called **data ethics and so it's a branch of ethics that's it kind of looks problems related to data not only just the data but the algorithms** and then the corresponding practices and uses of the data so there is a special issue of I looked it up this morning I don't remember the journal now but it's something like well it's like it's **a philosophy journal but the link is down at the bottom and they had a special themed issue on data ethics** and this is kind of the intro article and they from the intro article they talked about the three axes of data ethics so one is on the ethics of the data that's the collection analysis of large data sets and I tried to highlight and read here the things that I thought were maybe a little bit relevant or more relevant not more **relevant but relevant to you know geospatial data so the idea that you know we could D and itemize our data and I think that's a little more I think it's more dangerous or more of a problem I guess I should say when we have information on location** right we talked we heard yesterday about the Internet of Things so another axes or aspect is the **ethics of the algorithms** and I'm going to talk more about that too so there's this increasing complexity of algorithms we don't really have a good idea maybe of how they interact and what could happen from that and then the **third one is the ethics of practice** **that's how are they being used** so there's the problem of privacy and surveillance secondary uses of the data set so you know **uses of the data that weren't intended** from the beginning and I think that's an **especially important when we're integrating data sets** together which I think also has an implication for ensuring privacy and then unintended use we was interesting because yesterday the one of the early speakers were he was talking about we had seismic sensors I think and was using those two which were used for to detect earthquakes but then he was using those to find to find locations of gunshots now that's a different use of data but it was a I think there was a good use of it right but certainly there could be you know bad uses of data so just to give some motivating stories there was a **recent paper or article in The Washington Post about this racial bias in medical algorithms** I guess I should back up a little bit because when I was going to school work in my PhD I got in computational statistics so everything a lot of the methods that I used were data-driven right and so I used to think well we should just always use data-driven methods because the human won't be involved you know these humans I and they're tweaking something this the data is driving at all but after kind of learning about **these the data can be inherently biased** right and so then the bias that's in the data just drives the the results and the models right and so this is kind of what happened in these cases so it turns out that they were some companies we're trying to develop models to predict people who would need more healthcare in the future so that they could potentially do some interventions and you know given them the resources so they wouldn't get so sick so the you know they wouldn't have to pay so much to treat them in the future and they did not have races a year but it turns out that **the algorithms that ended up being developed were racially biased** and they underestimated the health needs of some of the sickest black patients and I think that could be in part because the data that's used probably would had fewer black people you know patients in there because they didn't have the medical care so maybe more affluent people would have or non-blacks and but you might say well why is that relevant to a **geospatial group but it turns out that how this was in part discovered was somebody mapped out the most affluent scores you know where did the affluent scores land and they were in affluent suburbs of Boston** so another more obvious example of this was this type of thing is **in predictive policing software** so you know these models want to predict where our crimes likely to happen so what's the input you know where crimes already have happened which tend to be you know in certain hotspot areas so then they said oh we're gonna add police to these areas so that's where the police are they see the crimes because they're not somewhere else and so then there's more arrests and so then it's just this feedback cycle going on and **autonomous vehicles** I think somebody mentioned that yesterday how many of you heard of **the moral machine** just a few of you yeah later you might want to try it out because you can still I can take the quiz I guess or look at the case studies but this was done by MIT and they were trying to get up you know what would what types **how would people want autonomous vehicles to react in situations** like we have here where you're a passenger in the vehicle and you could either hit these pedestrians or you could swerve and hit the barrier in which case you might be killed or your car would be damaged or something and what I thought was really interesting about this is that they did the case study or the analysis in collected the data internationally so over 200 some countries I think and it was in different languages so you get you can understand you can kind of get at this this you know **how our ethics and how are these situations considered in different cultures** and one thing that that was I thought was kind of amusing was they said that most people would wanted a car to not hit the pedestrians but hit the barrier but they did not want to buy a car that would do that okay yeah so they wanted somebody else's car to do that but not the one they ran and you know and I've I keep thinking about this situation I personally don't know what I would as a driver I would probably just swerve and hit the barrier but um it's a hard question but so

40:58

I'm almost at the end and I just want to say a couple things here the working group that I mentioned last year we had a **geospatial interest group workshop** we have one every fall and **we focused on this idea of data quality with integrated data sets** but are there is there **anything in particular because it's geospatial data** that we need to consider and I'll just point out that transparency and metadata seems to be this the importance of that as like this recurrent theme which I think that's important regardless of the type of data but I also want to point out **error models** because we talked about that a little bit yesterday and I think that's really key or very important I think because especially **when you're integrating models and datasets together you know each one has its error you know how do you how make sure that you know we're kind of considering how the error propagates** through the whole process everybody knows about this here I'm sure in this room some won't say anything more I think these **are important issues (Geometry and spatial relationships) whether we're modeling and integrating** or not that I think it becomes even more critical when we when we are doing those tasks I want to focus a little bit more on this in my remaining minute or two in the article about the medical biomedical application the **bias in the medical application or modeling** one of the not the author but somebody a spokesperson said well it's really easy to remove the bias and I thought well you know maybe but what happens if you remove the bias for that particular data set that particular model does that mean then that you know you're just kind of adding some different kind of bias in there and what happens if you have a different situation or a different data set you know and you're doing some of this prediction do you still have you know the bias in there so that just is **always a little bit scary to me when somebody says well I'm just going to remove the bias** somebody mentioned this yesterday the idea of especially with **deep learning artificial neural networks and some where you they're very brittle** they could be brittle and then so they're not very robust so we need to **do more sensitivity analysis** and I think this is important and key when we are integrating datasets together and what about **data stewardship the validation the cleaning** especially when we this **is important when are integrating data** this idea of profit privacy and surveillance I have to ask Jeremy since he's from the UK but you know watch a lot of British crime shows and so there I was like okay CCTV and some thinking wow people are just like always under surveillance in the UK is that true okay yeah (Jeremy in the background) I mean imagine is in the States getting to be that way too so it's a little bit scary because you and especially if you have drones up there flying around you don't know they're there and they're you're under surveillance and so there goes you know your privacy and we just talked about the Internet of Things so I think that's pretty much it I just have some I just point this reference out here that we did have a **series of workshops for the data quality working group** you can get to the slides here at this link and we also have videos of them and so you could go back and listen to them you want and not for the panel so I'm done

45:00 Q&A

Question (Kumar): t I'm with Max our technologies we collect a lot of data one of the questions for you is in the past we would store all the data the government policy for us how do you see going forward when you have up as it's coming along **how long should we keep them** I don't know

Answer

that's a really good question that was something I think was discussed a little bit yesterday or at least at the table I was at because I think having that historical data is really really important and I don't have the AI don't have the answer I would love to keep it forever but maybe that's something **that we need to need to invest in** and one thing that that occurred to me yesterday when somebody was talking about data so some of the data sets are they **really stored just in one place** I don't know talking about that we said we had someone the Asia cloud and some anybody want to say anything about that because I worry a little bit about datastore okay good because I worry about data stored in one spot because what happens if

Ed: I think in the early days when you were storing your data in the cloud you didn't really have much choice but but **now you can choose regions actually quite specifically where you want your data to be hosted and that's largely for regulatory purposes** so in Europe for many states for many types of data it has to remain in Europe or in North America

Jeremy: so I was going to add similarly at least in **Europe** some of those issues of how long can you keep data for and now covered by the **GDPR** and so now that's to do with personally identifying information where you're only meant to keep it for an appropriate length of time for the task under which it was gathered you're not allowed to reuse data that's personally identifying without going back to subjects different when you've aggregated those that's a maybe a different question yeah and what about map might be a little bit different too but if you have something that's on the web and then it was that part of the GDPR that if you want something removed from the web that it has to be removed or if is person if it's about you then you have a at least a right to request it yeah

Question (Stephanie Shipp): I have a question about that because I think that **storage has gotten so cheap** that we should be storing all the data although I think it depends on the kinds of data the sensor data is really expensive I know Arlington County where we're doing a lot of work already destroying data because of the cost which you know is sad they don't know what to do with it yet and so they're just getting rid of it but the data and **the cloud what is what do you do about curation** do you still need to keep updating the format's for it or does by being in the cloud by default put it in a format that allows it to last forever so to speak I mean so we have data on floppy disk right I'm making a joke but no

Wendy: but that's actually true yeah that's a really good point about **data curation and the data stewardship and in the media** and the medium I guess I should say because I do have this will really date me you may remember Bernoulli drives and it's good when people know what you're talking about but yeah and then and then you think oh this is real this data it's here it's safe and then **you can no longer read it because you don't have the ability to do so anymore**

Jeanne?: so I was also gonna add that when we built **Data.Gov** one of the things we intentionally did was not to duplicate data because the problem is you run into data synchronicity issues and so what becomes the actual valve you know the actual true data set there's many instances in the government where people would later be able to correct a data set or add a season that they hadn't gathered before or give more explanation about something or they decided to change the format of the data in a later year and so they needed to go back and normalize it for earlier years so rather than **trying to sort of rehost things it's much better in some ways to either create that URI** on the Semantic Web or some other indicator back to the data source on the other hand on the other hand the reality of the situation is you can still create backups of your data like and that's a very good thing to do whether they're you know still on another cloud-based service or with that same cloud-based service or on-premise but the other thing is you have to understand that **sometimes there are political whims so for example when President Trump came into office there was a lot of concern about climate data** and so there were a lot of hackathons which I totally support to be able to say look we want to gather the data we have and may can make a separate instance of it like a point in time but it was very clear was a point in time instance yeah

Wendy: that's a good and I think isn't there it wasn't there recent stories in the news about EPA had to publish all the data that that was used in the study or to you know produce the results irregardless of whether you know I was going to violate people's privacy so I don't know how I think that's news I don't know how how far that's gone but hey when

just one question the Waze data I'm not expecting that it's publicly available is it

Ed: Waze does have a program for releasing datasets to local government to municipalities for the research as a partnership program Google it yeah is it true that Waze is part of Google okay

53:10

Quesiton: just a comment on if you everything I would say there's a separation between the data format and the data storage so the data format is usually associated with how it's going to get used and what's the best way of storing in order to meet the need so for example if you're collecting data you were streaming data and you want to collect as fast as you can receive them the format in which it'll get stored is like IOT kind of in a friendly format spatial-temporal and so on and then there's data that has to get I'll just say massage Tom manipulated to allow for better querying so for example if you're querying the data you want a fast answer to that question it's not so much how quickly you're receiving it how quickly can you give the answer so that's another sometimes both of these can be one and the same and then the third aspect is archiving the data which is if you want to go to the past and sort of like rewind and play you're probably willing to accept some time delays for the privilege of storing a large and massive amount of data so all of these three things are separate and then for each of these needs you find the best storage possible you know whether it's short-term but expensive super fast versus long-term archive cold storage they sort of like go from hot to warm to cold so those are the two things and with the cloud and so many other technologies coming in you have so many options to consider those as opposed to there's a one-stop place where you put everything and yes you decide and sometimes that decision is made by the cloud provider to be able to compact that data and to store it optimally for you so that you can make the request and then it decompresses the whole thing for you later on

Jeremy: I think something we found it Ordnance Survey recently though is how things that were archived data suddenly become data that one people are more interested in again that's after kind of 10 years of national mapping being kind of regularly produced in a in a digital form it's been produced for longer than that but the last 10 years is that **people now more and more interested in those past snapshots to do kind of modeling to present to prove the models into the future** as it were so as those archives suddenly become present - really again I think it's interesting yeah

Wendy: I would think that maybe with some of the remote sensing but also I guess it was yesterday somebody mentioned something in it but it seemed like that the data that was being collected would be thrown away but my thought was well what about see my crime shows I was thinking you know what about well if you if you wanted to you know you know find the person who you know was did the crime but also what if you wanted to go back and get evidence to say it wasn't you or so yeah so I mean

Amazon: I just wanted to add **a little clarity on maybe the cloud storage aspect** of it so my primary hat is the geospatially Don a specialist team but my other hats is storage and I've been supporting our open data team for like eight years now I think and so you know ed said it right so used to be very nebulous it's like you've stored in the cloud who knows where it was but a dubious was never built that way when you stored it in our cloud you were very specific about where you put it in what region you place your data right and then I just wanted to add that when you make those decisions all these decisions you know we don't make the customer very specifically so you want to put it in in Virginia you can put it in Virginia put it in Oregon you want to but we have a lot of **customers for example that have very valuable data they will you know make decisions about primary stores so where is the authoritative original copy of the data and it's a it's a trivial exercise to go well this is really very valuable and there's multiple reasons why you might do this but I'm gonna I'm gonna have a copy of this on the other side of the planet** just in case the whole point of the cloud is reduced radically have reduced the cost of that kind of storage so we're seeing customers that used to throw stuff away we have customers that you know thousands of tapes tens of thousands of tapes now that now they have moved that now they have **access to historical data at a scale they've never had before** plus you you can control placing it in multiple parts of the world with very simple directives right so I just wanted to add that plus and I'm glad some of the formats that came up but you know you again **there are formats that are that are you know driven by cloud requirements** cloud you know use cases that are not you know old-school on-prem formats that are that are new cloud formats that make basically in situ use of the data possible right and that's a very important fact I'm happy to talk more about that later hmm

Question: we'll move on to the last another IDI okay so I'm all about **data quality** and certainly metadata is important to that and when you're **doing data integration so your data from multiple sources that have different data qualities** it seems to me it's fundamentally important that we make sure we've got appropriate metadata to describe the quality of the result as well okay and somebody mentioned **provenance** the other day so part of that metadata it would seem should include the history of all the processing that's happened to this data and you know even detail to include the version number of the algorithm because sometimes those algorithm changed and that means your data is different that's like a point now I've had people argue to me that well you're gonna end up with this huge metadata file that metadata might be bigger than the native well that's fine because you guys just told me the storage is cheap so let's do the metadata and maybe because no human will ever look at this maybe this is an application for a

Wendy: yeah that's that's important I didn't think about metadata for the algorithm and pedigrees huge - somebody mentioned pedigree yesterday and and one thing I had a thought yesterday was somebody mentioned about standardization and ERDs did that and **I'm paraphrasing but it seemed like it was a standard stifle innovati**on did I get that right but anyway so I'm wondering if maybe met that having better metadata could that sort of alleviate that so we don't have to have that strict standards so could we have more metadata rather than and then fewer standards who was the standard person here now is somebody was like I do the standards all right that's my provocative question

I'm sorry we need to take a break now we can pick up questions in the panel Wendy's agreed to thank you we'll take a five or ten minute break now

## Panel on Outcomes :

1:13

### \* Regan Smyth, NatureServe - Perspective: The Age of Precision Conservation: Applying AI and Collaborative Science to Prevent Species Extinctions

I'd like to welcome Regan Smyth

Thank You hi as Jeremy said im Regan Smyth. I'm the director of spatial analysis at nature serve which is a biodiversity informatics Network we manage information on imperiled biodiversity across the Americas serving as kind of the hub of a network that has scientists out in the field collecting information to answer key questions about species and ecosystems some of those are up there on the side there but that's things like what is it so standards for taxonomy and classification **where is it which is what I'm going to focus on today given that location** and then how things are doing so we assess the health of you know we we help figure out what's imperiled what's doing fine and then that information is provided out to other users so they can make decisions about conservation and land management so while we are a conservation nonprofit we're really in the business of data but data relevant to species

so what I'm hoping to do with my five minutes today is tell you a bit about one of the projects we're working on that kind of brings us all together specifically in the realm of you know data science and how we can pull new information out to answer that where is it questioned and given that that this is the outcome session gonna be focused on **how we can answer this key question related to species extinction** some of you may have seen the headlines recently estimates about 1 million species going extinct worldwide we're kind of in this moment of biodiversity crisis and we're trying to figure out with the data we have at hand and kind of **the new spatial analysis methods a new proliferation of ways to measure the environment and make sense of that with kind of big computing solutions** can we can we address the extinction crisis and particularly with the project I'm gonna talk about can we do something about this mismatch between where you have areas that are important for biodiversity and areas that are protected **so for 40 years we've been collecting data** like this those are some of the species we track the numbers on the side are ranks of how well they're doing so something that's a g1 is globally imperiled there's not very many laughs they're in trouble the scientists in our network have been out there in their field boots you know with their instruments for 40 years collecting those dots on the map so those are delineations documented occurrences of where these things occur now the problem is you know people can't be everywhere there's not millions of dollars for field biology inventories and so this map is it provides really valuable information but it's totally inadequate you address that species extinction question we only know what's going on where people have been **we don't know anything about those blank spaces on the map** and there's a lot more blank spaces and filled in spaces because of that people like the Fish and Wildlife Service who are charged with managing these species often take a pretty conservative approach and they create big range maps like this to say well the species occurs in this region worry about it there so if you want to build the highway you know in the Northeast if you have an endangered mussel that's gonna be your concern you're gonna have to spend money assessing it conversely on the other side if you want to protect a mussel do it somewhere in there and maybe that'll work and that you know again **it's not going to address this critical question we have so that's where the data science piece comes in**

what we've been doing that nature serve in partnership with technology supporters is **building predictive models to more precisely map habitat for all these species** so taking just one species from that previous slide on the far side there you see those polygons where people have been out in the field and said this frosted flatwoods salamander which is listed as threatened under the Endangered Species Act has been **observed at these locations and what we can do with that is combine it with this wealth of geospatial information characterizing climate and soils and land cover run predictive models use machine learning approaches like random forests to bring that data together and then project out where you have similar conditions on the ground and produce a refined habitat map**

what this looks like in practice is something like that here's our **map for the salamander** on the side there you see our predictive our prediction of habitat suitability so the yellow areas are where conditions are suitable for the species and then on the other side we've just thresholded that multiple thresholds so what you can see is if we've gone from a map like that on on the Left **where we have we've seen it here it might be here** to this much more precise and that's kind of the power of what we're able to do now for individual species but the exciting thing is you know with with the the data science today we can start doing this at scale and for many species so this map is just stacking up models that we've built for all the critically imperiled species in fib Ian's in the United States and the areas that are pink and yellow or where you have concentrations you know places where we really can focus conservation resources in the past year we've done this for 2,000 species and we've done this by pulling together the data collected by hundreds of people in the field **building a kind of cloud-based environment where we could have a team of scientists collaborating on the modeling and then using tools to get that information out back out to our scientists** ok they could review it and tell us how well the models have done

taking that one step further we've done some further analyses looked at where not only you have many species occurring together but also where those species aren't protected or where there's only a few opportunities left to protect them this map is showing the outcome of that the green areas are areas that have been protected and the yellow areas or where you have concentrations of biodiversity that that need protection the most so these are our **conservation opportunities if we're going to address that extinction crisis** just want to point out you are here if you go in a really long walk you could get into some pretty cool places this is it's one of the places in the country where there really are our concentrations of cool stuff but

what's you know what's been kind of **keys to the success** of this you know where we're producing information that people have never had before we're producing it at scales that are actually actionable for the first time but for this to work you know there's there's several things that are necessary that I think are probably **commonalities across the various work that a lot of us do** so for us because **salamanders don't use Waze** and there's not you know this proliferation of data from the NAT a world that there isn't some other realms having **standardized ground truth data** about the environment and what's going on the ground so there's 40 years of field collections from our network are really the basis of what makes this possible and having that information come in in ways that are standardized you know people mentioned yesterday the 70% of the project time that goes into dealing with the data **when communities get together ahead of time and set some data standards it allows us to do work like this you know at scale** in ways we couldn't otherwise the other thing that's made this work possible is really that **partnership between you know big tech and tiny nonprofits** like nature serve as the woman from the California utilities group said yesterday there's so much that's possible now and so many pressing problems that can be addressed but often the people who are at the leading edge of the the analysis to do so aren't connected to the people on the ground who either have the data or who know the questions to ask so having that partnership with ESRI and with Microsoft has allowed us to do this not just for one or two species like we had been doing in the past but really **massively you know at 1/2 continental scale** and then the last point I wanted to make of kind of key lessons learned is with this kind of modeling and this has come up before to the **trust and transparency is a really big issue if your results are going to get used** what we found with this is well it all sounds great test data scientists is there's a lot of **distrust about you know what can these models actually tell us** is this information good you don't know what's going on with this plan I've studied for 30 years and so finding ways to make the process open and to **pull the people with expertise in in a review capacity** and we created models for 2,000 species and then we put them all out in front of people and they won learn that you know good real useful information could come out of that but on the other side the data scientists we learned that you know these are the situations this work four and then a lot of situations it doesn't work or you know the field people can go out and get additional data and feed it back in so that's really one of the biggest takeaways for me is just that **for us to have maximum impact in our work we need to be open and we need to be listening to the community of end-users** and able to communicate how what we are doing and can and what the limits of it are in in answering whatever question we're trying to address

### \* Megan Furman, Defense Digital Service, OSD

Megan Furman I'm the deputy director for the defense digital service (DDS) and the office of the Secretary of Defense

what the defense digital service or **DDS does is we are a SWAT team of nerds** inside the office of the secretary of defense and the comms here what was that SWAT team of nerds we are primarily folks who came from industry to do two-year terms with the federal government and so what that means is focus we have a bunch of folks from Google I spent six years at Palantir technologies we have folks from Akamai **really excellent technologists coming to the government and saying here's what good looks like and helping people make better decisions** because at the end of the day at the end of the day the people who show up get to make those decisions and trying to bring a technical voice and technical expertise whether it is data science or engineering or user research really really matters so what DDS focuses on is a combination of technology acquisition reform because the way we acquire battleships and aircraft carriers is not actually a good way to acquire software tools or data sets and so what you can end up with are two or three-year requirements generation processes where at the end you've actually defined a way part of your solution set and part of what we do is try to change that and make it better we also take on our own projects where that means **we bring those modern technology practices into the government** sometimes we are taking projects that are floundering a little bit other times we take projects that are new so when the government thing is about transitioning to the cloud and cloud migration how do you go about doing that when what you've always known is a data center and your procurement officer has always you know bought bare-metal servers and then you go and you install those things that's actually a big transition to ask someone to make and so we do sometimes a pathfinder an illustration

we're working with the Marine Corps right now to say you have a particularly challenging logistics system let's do some user research let's look at what changing your UI and your data management system would look like on the back end and in the process of doing that we're going to migrate you to and we're gonna show you how to do it this is going to be a well-documented approach we're gonna help get through all of the policy and security hurdles which often exist for a reason but sometimes don't sometimes the biggest thing you can do is **just ask the question why is this like this** can you show me the rule that says I can't actually do this thing and often people can't and so one of the values that our team brings is well we don't know all of how to work in the government we bring kind of a fresh perspective where we can ask those questions and then say well here is actually here's how Google does it or here's how Akamai doesn't done it or here's how you name another company has gone about doing these things and in the process of bringing modern tools modern technical stacks and starting to be kind of a pathfinder in illustration there so we have been part of from thinking about outcomes perspective

the Jedi Cloud acquisition started from this office we have done a variety of capacity-building programs so when you think about starting to **hire a data scientist how do you go about doing that** how do you make sure that it's not just someone who put data scientist on their resume but doesn't actually have the training how do we set up those processes and empower folks to to do those things better so it's an office of I say it's a SWAT team because we also work on you know will be called out to do work on things like the PCS process so what a when a military service member is moving usually every two to four years they and their family pack up their whole house and they move but that is not done by the service member that is actually coordinated almost entirely by the US military and in 2017 coming into the summer PCS season the website that folks used to manage their moves only had an uptime about 20 percent which is crazy and unreasonable and so someone want we had started hearing about this and finally someone walked in our door and said **hey my family is supposed to be in Hawaii in two weeks and I am starting a new job and I can't schedule my move because this website is down** I'm not really sure and so that gave us the space to work with us Transcom which was responsible to do an initial fact-finding mission to go and get those servers back online for the near-term and then in the long term work what would it look like to actually have a system that works for a service member and their family when they need to move to give them the information they need on the time that they need it to allow people to log in and do things that are very very simple and basic and then if you're storing the data correctly project out what should this look like in the future how can we pre stage things how can we know we're in the actual bubbles of time are when people tend to move everyone thinks it's the summer but there's also a couple other times in the year maybe we can plan ahead for that so those are a couple examples of the types of things that we work on

**data is absolutely foundational to all of these** things there is a huge push to use AI and machine learning in the government right now **but if your data is not ordered if it's not available if it's strictly across silos that you can't access it's really hard to do effective machine learning** we're also discovering that when people say a I am ml often they just mean you know a linear regression some like very simple statistical methods and you're kind of like well you know I could do that for you you don't have to you don't actually need machine learning for that just do like when your aggression let's see if we can find some statistical significance in here like don't make it more complicated than it is sometimes it's a very simple solution it's asking why can I earn can I not do this what does the data actually say and then helping people get access to the right data so that's the kind of work we do I'm excited to talk about it

I should also say from a **capacity-building perspective** oftentimes we've seen people don't always know what they're asking for so we were talking about security and privacy earlier you combine two data sets and you can unexpected outcome one of the things that we try to inculcate into everyone we work with is thinking carefully before you do those things and saying what is the security implication of this what is the personal and security implication of this so that we are actually you **know bringing those crack ethical moral practices along the way** which is something Wendy talked about someone to mention it thank you

### \* Steven Ward, The Climate Corporation

1:30

climate corporation is a subsidiary of Bayer think of us as the digital arm of Bayer that is sort of at the forefront of this agricultural revolution that's occurring right now many of you may be aware but **the AG industry over the past hundred years has gone through a couple of pretty extreme inflection points** one of those being mechanization another being the implementation of new chemistry's and new modalities and modes of action and chemistry's the third in the 90s very hot topic right now being bio tech and biotech traits and pairing those with chemistry's **it's our belief that the climate corporation and within Bayer that the next inflection point the next real shift in the agricultural industry is going to be data driven** data is the new currency in agriculture without a doubt farmers seed salesman chemistry retailers they are swimming in information

to give you an example we offer a handful of products on the market we have over 1 million downloads of our products we collect over 5 million hours of machine data every month we processed over 500 million image images last year from drone satellites we have over 1.8 excuse me 180 trillion data points across 100 million acres of the globe and guess what **there are no standards in the AG industry** everyone's doing it differently we are dealing with a hot mess and it's taking every bit of brain power and technology that we can combine to make sense of this all and that's really what we're trying to **do we want to offer the ability to the growers for them to see their activities and their operations through a different lens their lens of data** that's always been there but it's been buried

it's our belief that every pixel of information every square meter on the farm holds a story right now only about three or four chapters of that story have been written or interpreted it's our job we feel to help interpret the rest of that information and the rest of that story and bring it to a sustainable end **our goal is to allow farmers through data to more sustainably improve and increase their practices** we've probably all heard that by 2050 we're gonna have so many more billion people I've heard every different number in the world we can argue about those numbers if you want **I think we can all agree that there gonna be a lot more people in 2050 than there are now we're gonna have to feed those people** what that means in terms of the amount of production or the amount of reduction of waste is up for debate but the reality is we're gonna have to do one of the other probably a combination of the two to make things work so

I've been very fortunate in my career despite don't believe this baby face I've been doing this for 21 years okay in the geospatial tech industry I've had the pleasure of working with some of the folks that are sitting in this room and I've learned a lot along the way started my career in habitat suitability analysis love seeing what you were talking about loved seeing that you weren't showing County and Parrish maps I'm from Louisiana we have parishes down there and that you've graduated to actual real concrete delineated boundaries move from that into the intelligence community spent quite a bit of time running **fusion and pattern analysis teams and predictive analytics teams at Special Operations Command and Joint Special Operations Command working for a dia CIA and some of those other three-letter agencies** grew really frustrated still having to work with Internet Explorer six in classified space so I got lured out of that concrete box in Tampa by the Climate Corporation to get back to my roots where I came from an education standpoint that was really looking at **spatial statistics in the environment particularly with agronomy and soils** and have been **running a data insights team for climate now for five years** that team has grown from nine remote sensing scientists now to over sixty weather scientists remote sensing scientists geospatial statisticians AI data scientists throw in every INT you can think of and we've got a good mix of it but the beauty is and what I really enjoy right now is and the approach that we take in this world of no data standards our mission is no different

I'm able to combine that traditional intelligence mission that I worked so hard on for so many years in the government and bringing into industry to solve growers problems we're taking the traditional approach **everyone can tell you kind of what's happening some people can tell you where it's happening we're really striving to tell you why it's happening** and so getting down to farming by the inch is is where we're trying to move now that means that we have to work with multiple sets of data multiple pieces of information qualitative quantitative different resolutions it was mentioned earlier in the talk this morning **there are no good ways there are a lack of standards for combining data that suffer from MAU issues they suffer from spatial non-stationarity issues that suffer from all of these problems that we face every day** and so we're lucky that we're not only doing the research to help farmers right now but we're also focusing a significant amount of our efforts in **establishing these standards we're very active in the STAC community we're very active in the open source community we are rewriting pySAL for distributed systems right now and I think it's incumbent upon all of us in the room to contribute back to the community** that we're all that we're all working in all of these are gonna help drive data standards and we'll get into it during discussion I think on the panel I have mixed emotions about data standards someone mentioned the **constraints on innovation and I that they potentially cause I am a proponent of that I do believe that they constrained innovation but at the same time they're necessary evil** so with that said I'm going to wrap up happy to talk to anyone else about climate of what we do I'm excited to be here and thanks for having me Jeremy

### \* Edward Strocko, USDOT Bureau of Transportation Statistics

I'm Ed Strocko with Transportation Statistics in the United States Department transportation I run the office of spatial analysis and visualization so we're all about telling the story transportation through data statistics Maps my group we do a bunch of cartographic products we put the National Geographic Atlas database together about 80 different players and we also do a bunch of their spatial analysis we're one of the **13 principal Statistical agencies in the federal government** and we're all about being objective and policy neutral and the currency we trade in is trust we ask you very sensitive questions if your business or your person or household and we have a lot of protections that we put on that we we have civil and criminal penalties **we pledge a confidentiality** we will not share it with department justice with the Department of Homeland Security with IRS we take the really good care of the data and we really that if we don't have the trust we can't get the data that we need on the other side in this era of fake news **we have to make sure that you trust the data that we're putting out there** that this is authoritative this is real you can believe in this stuff Wendy puts out when you put out the unemployment statistics or you think about that I mean that moves market so we need to make sure that we have you trust that this is real

1:40

so as we enter the space of data science and geospatial we're very timid and we want to make sure that we get this right and we've been thinking about statistics and statistical methods for 100 years so we've refined how we sample things and how we deal with error and as Lauren was talking about bringing science to the data science so how do we we learn this new science and how do we do it safely and when can we accept bias and really it's about fitness abuse when we get in and start using data science in in this space I think yesterday Kathleen and Todd integrated examples of how we're using data science and transportation Wendy this morning showed another example so I'm not going to go into a whole bunch of examples I'm going to throw out a couple slides some of the bigger picture things that we're thinking about I think that's right for us to be thinking about right now as we look at **geospatial and integration of data science**

so when you think about **the transportation system** we cut it a couple different ways you got freight and passenger you got public and private you got real real time like I'm going 75 on the 101 in a vehicle and I could die or I got real-time I can matching a system or I got I got archived and doing a model so we can kind of think about those we also have a system level and then a project level and some of what's driving us in this area is the **scalability we've focused a lot on GPS in the freight** area so we we like that a lot we're starting to get into the imagery but it's **tough to scale at the temporal pace or the for the quality at a national level it's good for cities** a little tougher video cameras we don't control video cameras for the whole country we're decentralized if we're Singapore they travel you awesome but you know we're not so you know it's easier FAA is part of us we get all their GPS data Coast Guard were able to get all the AIS data from all the ships we have a partnership of the Trucking associations we're able to get a million trucks 60 billion points and their GPS every year

so that's where we've been working in general **the freight system** that's why I think we should continue to look at back in the 70s and 80s they read irregular do you regulate totally change industry very low barrier to entry for trucking you just have to have a license and a truck and so people are competing on technology and you saw these farms as leaders and how you're going to separate themselves and so when you think about package and parcel delivery you can track your package anywhere they've been forced to really invest in location and data science how **am I going to do optimization of a million packages I got to get them there by 9:00 a.m**. I see that every day and whether it's the trucks on the road or the planes gotta get all these planes in to manifest into Louisville at nighttime unload it back out and to another airport in the country at 7 a.m. it's a lot of optimization there it's not just that we see it in the shipping lines the ships we see the **container ships** gone to 18,000 TEU that's a twenty-foot equivalent unit that's one of those short containers big ones are 40-foot that's a lot how you packed them on a ship in Asia how you get them over here how you offload them how you deal with them in a yard how you stack them how you move it around we've seen a whole lot of that and then of course just management of a fleet whether it's a truck you know what something's wrong phone it in you know ahead hey I need a part get it on so we can get the truck back on the road or **railroads** railroads a very capital-intensive having precision railroading all those tracks where do I need to be spending my limited resources to actually make those corrections so I can run trains at speed I don't have slow orders or that I I'm not derailing a train and blowing up a city with a train for the oil. So as we talked about **edge commute computing yesterday I see that's where we're going to see it first**

the next thing I wanted to touch on a **shared mobility and connected and automated vehicles** shared mobility wouldn't exist if we didn't have this conversation right we know where our doctors bike sorry scooters were we wouldn't wouldn't have ride-sharing ride-hailing an area very interested in is the data sharing we talked about data trust yesterday and how do we bring that data together everybody's asking for this data I mean when you think about the transportation system it's teeny it's very hot very sexy right now how do you how do you bring that together anonymize it so we can look at it together and also some of the standards we've seen coming out of there a group called **Shared Streets it was created by the World Bank and a National Association of city transportation officials** **kind of looking at a global standard for linear referencing** how can we do this so we understand what's going on in the street how can we go between different geospatial layers connected in autonomous vehicles to things they're not the same they're related edge computing works well for autonomous vehicles but you need something more you need that mesh for connected vehicles and it's also you start to think about who's in control and really want to have us think about **HD maps** and you think of the layers of **HD base layer** of where the roads are you got your geometrics then you got some additional layers up there and it's a semantic that **semantic layer** of what are the with a sign for their signals well you know we got to create that but what's how would it create that standard and there **91 thousand units of government in the United States** and a lot of them own roads so everybody has to get in with these standards and the person at a little municipality who is the Department of Public Works Director is the chief data officer also drives a snow plow so how we're going to do this that everybody can actually accommodate us not just 50 states then then **who owns the Maps is it the OEM is it the map provider is the federal government** I don't know any you think I own one mile road out in Virginia where we test cars so is the states again lots of people **where's that liability** we need to be thinking about that now as we are looking at the semantic layer one of the efforts we're doing is **work zone specification** we're putting out some grants so anybody's interested in money or trying this out come and look at our website now we want to figure out how to get this work zone standard out there and adopted widely autonomous vehicles and connected vehicles they don't like uncertainty and works in the very uncertain you have a utility truck parked for 10 minutes fixing a line you could have something closed for six hours six days six months how do you deal with that in their

**transportation and planning engineering** it's kind of inside infrastructure owners and operators the two things here the convergence of GIS and SIM and BIM the civil integrated management and building **information modeling and to think about smart cities for transportation** how do these two come together different purposes different scales different Precision's as we're looking forward what it's kind of the Wild West and BIM and SIM right now what do **we need to be thinking about on standards and how'd it relate it to the GIS world** I also threw up a transit map this was a really good example of why standards and specifications work back in the 90s one of the things you know what's the extent of the system well where all the transit stops and and Ed's mentioned hopefully we won't have trans stops in the future because we have inefficiency but for now we do so we hired a university did the whole country every bus stop and every bus stop in the world or the country took a lot of time a lot of money and it was wrong before it came off the plotter so we couldn't do it again so we just maintained the transit stops so then **GTFS** came along great great opportunity it allowed us to very seamlessly easily pull them all together to **create a national map and actually start doing data science on something we couldn't have done it without that standard everybody had adopted and was using** the question I have is **why was that so successful** how can we make that successful in other areas of transportation or other areas of geospatial and data science that everybody was adopt everybody uses it and it's not some standard that sits on a shelf

I want to wrap up with **transportation safety** this is really at the end of the day this is it **we're still killing over 36,000 people on highways every yea**r got better who killed a thousand left last year but still way too high you think of Boeing two planes crash to ground the whole fleet right we're still killing a lot of people I put these five examples up we did a **data visualization challenge** we had some different players in here and using data differently University of Central Florida using AI TS data all that data we've invested in over 20 years they've used **machine learning to do short-term predictive analytics** on where our crash is going to occur where do we think we need to be focusing our attention changing the variable speed limit signs giving the infrastructure operators some information Ford and arity they're looking at the behavioral information coming off the car hard stops hard starts fast starts hard stops people using cell phones and using some data science in there to get that get down that number of fatalities until we can **really achieve that that vision of a much safer road with autonomous and connected vehicles** so with that I thank you and let's let's have some conversations

### Panel Q&A

1:50

Question: My perspective is as an investor. IM Todd Simon with geospatial alpha and also a venture partner with London VC called the n capital so my perspective is investing and also nurturing startups that derive or create geospatial intelligence spatial data a lot it's a lot of it's satellite oriented so I actually have **four different questions** then you can maybe decide if they should be done offline again so Steven I think I wanted to we're considering an investment in **hyperspectral** satellite and I'm thinking about a lot of different things one is they're really strong correlation between you know the narrow bands and signatures for biology I that's part A Part A Part B is is the cost prohibitive and you would know if that is the case so that's my question should I invest then your Megan we do some work with the DIU and I'm curious if you're more focused on sort of **enabling and onboarding existing technologies** or exploring and discovering new ones which is what DIU does a lot of and then Regan we got a partnership with a group out of Stanford that's developing **panda sat** which is a dual technology one micro microchips for small species and two micro satellites called chipsets to track them and it's starting with pandas and I'm curious if you actually well actually contribute capital to projects like that that can you ingested into your data system

Jeremy: take the hyperspectral one offline and seeing how the specific sure so

Megan: so Defense Innovation Unit (DIU)Diu is a defense innovation unit folks aren't familiar they are focused on like you said a lot of **investment and finding small companies** that are not traditional like traditional systems integrators large scale contractors and bringing them in to start working with government the way I think of like DIU versus DDS was just like the office of the CIO and DARPA and our IR buttons and things like that is Diu is really focused on how do you how do you nurture the companies and the technology from outside government to bring that in which is incredibly important there is so much specific information about how do you do government contracting that just is a learned skill that that needs to be spread more broadly so that more people can compete and we can get better technology coming into government not just for the defense but across the board otherwise you know you leave giant holes of things that you should just be able to get your hands on and work with **DDS is more focused on the people side will take projects** we do work with DIU but unless on that investment side they're really more focused on investment the CIO shop for the Department of Defense very policy oriented and then DARPA very long-term projects so I think there's **there's space for all of these different types of innovation organizations** I think you know we we need to all make sure that we coordinated together and we play together well we know where projects and people are moving across the ecosystem and that that is sort of the way I think about that that whole space which is incredibly incredibly important I married

Jeremy: Wendy and Ed you want to come in from different places in government here in the in the u.s. bringing in innovation and so on well

Wendy: I just I guess I want to echo what you said that um it's it's it's hard to navigate both sides it's hard to navigate the whole contracting you know process and I used to work for the Department of Defense so I know how important yeah well I know how important it is and they do strive to you know bring in new entities and you know to work with small businesses so I think that's really important for innovation so

Ed: **we like to do both we like to take calculated risk and we think it's okay to fail it's okay to fail for $100,000 not 100 million dollars** but we want to take the risk that the market I won't subsidize that to get that new technology in there but then when we go into more production **and we scale up we're looking at how we do we adapt the existing technologies** and frankly existing data so I love trash and so I can collect there's little data I use other people's trash for almost everything and so you have to make sure you use it correctly but definitely **reuse of other data is key for u**s and

Jeremy: there's one one other question

Regan: about specific tracking yeah so we we are not heavily involved in kind of the **micro sensor world** and though there's a lot of interest there I think the challenge too often is you know we're relatively scrappy not particularly resource rich nonprofit and we're we're also primarily function as the hub of this network so within our network we have programs within academia and within other even state government who are definitely utilizing and like dipping their toes into those areas nature serve itself is kind of **thinking more about how to pull that data in and use it in unified ways and to systematically answer questions across jurisdictions**

Skymatics Question: I actually have two questions sorry I said that we are allowed now to we'll see how that scale so I was super interested in what you mentioned we can about **human involvement in the generation of machine learning models** as a success factor I think I think and I think we all agree here that when you create a machine learning model you do **need humans to analyze the outcomes and then to the model to make sure if the outcomes met with the observations and then correct or tweak** in the in the model but once that a machine learning model is considered valid and then you keep it running and then I am linking here with a question that Jeremy asked yesterday night over beer said okay you have a model that works with this data but then imagine that the data changes okay and machine learning models by nature they are adaptable they can learn from data but **what happens if the data changes so extremely** and with weather with climate we're seeing more and more observations of extreme occurrences that have never been heard before so your model that worked with the old data is not valid anymore with this new set of data or maybe the geographical scope change is completely so the model starts to diverge now about the **maintainability of the model** the what do you think about human involving in the maintainability of the model versus putting more effort initially into making a model more adaptable automatically

Regan: I have a lot of on that I mean one of kind of **our dream infrastructure would to have be to build this dynamic modeling system where we create our maps of probabilities of where people you know where you're likely to find something people go out collect more data that feeds back into the process and the models get better and better over time** that sounds fantastic in theory then you put those even the first iteration of the models in front of people and they say well you know for the salamander this really worked but for those other species us didn't and if you tweak this and you get into that human mediated component that makes the outcomes better on like a per object basis but is hard to build into a you know a dynamic updatable consistent system so it's I **mean there's a balance right between tweaking the computing to optimize your outputs and tapping into that human knowledge that's a little bit more variable** and so that's something we're thinking about a lot is **can you use the input you receive an initial iterations to figure out systematically what's going wrong and update your methods to address it or do you somehow need to structure your data** so you're pulling in both measurements and human input and somehow develop algorithms I can kind of balance the two and adapt over time so there there's not a solution but I think of anything from the big project that I've been engaged in this past year is we've really seen that that human **the human knowledge is invaluable and that without it we could produce valuable apps for the country but they would essentially be meaningless** and not addressing the question so it's something we need to figure out how to do

Comment: one of the things to keep in mind is this is something that the newer machine learning frameworks mostly based on deep learning have built in they talk about machine learning pipelines there are built-in checks to say it keep running the test thing on the if when you new data comes in to run and check a prediction again and if it's beyond a certain threshold trigger the retraining based on the new data but once again I'm not saying it's going to work for every kind of models right and I mean what we decided in our thing is to just **use a combination of use machine learning for like specific parts of the model not for the whole pipeline so you'll be able to automate retraining specific parts of the model but not the entire model** so if your variance comes from the individual pieces you will get the retraining automatic but once again it only works if you are in the using something like AWS sagemaker one of the in-house once I am drawing a blank on names but there are frameworks that do that you know

Climate: for us it's very simple no **no model should go without being paired with a calibrated eyeball** right and this is where people can get confused I think at times **too often people read in the machine learning as human replacement when in reality it's a force multiplier** I can do now with one what I could do with a hundred before and if we don't take that calibrated eyeball test that does this make sense test and and evaluate the results evaluate the output look no one knows in my case no one knows their field better than the guy that's been farming it his dad's been farming it his grandfather was farming it for the last hundred years so when I put something in front of him he's going to be able to tell me if it makes sense we might not know why yet but he's gonna tell me that there's a level of fidelity there and and I think that's the that's the truth **this is the way that we conducted human terrain analysis** this is the way that we conducted human intelligence analysis and and there's at this **point I don't believe it's responsible to remove that human element**

Andre: from a financial investment perspective I would ask the same question from an intellectual investment perspective about hyperspectral because we're all many of us are struggling with we see the potential and we have certain visions what one could do with it and then you just don't find the data or they're just not the quality or they don't come at the right temporal frequency so I'm really interested in that question

Climate: you know it's particularly industry and I can I'm happy to talk to you about it afterwards we've done significant testing **I've spent a lot of money testing hyperspectral data** on our research fields across our research partner programs and the reality is I think at the moment the data is fantastic and we know it is but we have to ask ourselves is the industry ready for it is the equipment ready for it are the chemistry's to pair along with the decisions from it ready for it right now it's not and so it's it's a decision about do I get ahead of the curve or do I not and so I think in the spirit of keeping the rest of conversation alive happy to talk to you guys afterwards and and and fill you in on what we found and what our thoughts are there

Question: one last question position that conversation with umbrella OTC what are your thoughts in terms of **standardizing HD maps** itself that is contrasted with you know today is mapping system which has this call it has **SD map standard** definition being these are just layers of information today this piece of information is used for autonomous vehicle is going from level one to level four and so on right but I think in the larger context of data science collecting the data doing analysis and all of that it's super valuable as you know from a DoT perspective kind of solve transportation problems safety problems and so on but there's more to it than that so any thoughts in terms of you know maybe are in fact push this as a standard that data to make there is you know everyone was trying to build these autonomous vehicles and keep that data in a format that can be understood by not just the cities but software in general and the other scientists at large

Ed: good question and I definitely think we're behind the eight ball on this one where we spent a lot of time figuring out how our times vehicles can work and how connected vehicles can work done a lot of research a lot of projects on that but not really looked at the bigger policy and how this and the non-technical pieces come together and that that is we need to do that right now and how that that comes about I'm not I'm not sure what form it is but **it's definitely the collaboration across the spaces** there to get to something and it's it's you know the ability to **go up and down between levels** and also **what type of standards are we talking** about we spend so much time on the communication standards on autonomous vehicles and connected vehicles as we're looking at these **are these communication standards are these geospatial standards and I think some of those basic conversations still need to happen** and can they both coexist in this environment

Jeremy: I think is **from a national mapping agency perspective** as well for the UK at least I think there's interesting questions also as to whether when we talk about **HD roads** or just interested in that CAV market and so on and there there's still a way to go to find out **what is a good spec between the different technical for just doing the localization and piloting countries itself yes not kind of one map that each of these fleets would want at the moment** but then for us as well as how does it serve other purposes as well whether it's the IOT market or simply local authorities better maintaining their assets and so on as well so I think there's a way to go to accumulate enough evidence as to what is a good product or standard in that space

Ed: I would also say **without a standard I think end up with a lot of sub optimization there are optimization for the individual user but suboptimal from a systems level**

### Break out session planning.

unfortunately I think we're going to have to stop the panel session per se but that doesn't stop the discussion we're going to move on to the more of a breakout session in a discussion session so we have until lunch the opportunity to continue the discussion in breakout groups and so we've identified them by application areas if you will and you see them in front of you now so insurance we have you ting both leaders and insurance guys gather at that table or go outside we'll get to the outside option in just a second transportation Edie is at a table maybe we'll put that right here or outside natural resources and AG so Regan and Steve where were you [Music] okay let's put it on that corner table down there and oh my Antonio if you're involved or not but that'll be the natural resources in Ag and I think they want to go outside anyway defensive Intel so Megan you're on there along with David and Mark so over here to kind of begin that discussion on defense and Intel and then Jeremiah was going to lead up data sharing and I added ethics to that just to make it interesting and

so you can either stay so the two options one is to go to those tables well start at those tables and then if you collectively want to go outside we have a Google map to show you where that's with a 3d model that's just playing glorious okay so if you want to go outside you can see this is the building we're in now hopefully quite obviously this is the building you're in now if you walk across this car park which is the car park in that direction you get to a path here which is the Stevens Creek or I zoom out you'll see it and it basically runs out to the bay in this direction how are you could walk over in the other direction it goes out towards the Ames Research Center and you can see some of the that's the wind tunnel and all the NASA stuff I would go out towards the bay if I receive and Vice President Pence is visiting Ames today so there may well be Secret Service guys buzzing around in that direction so if you want to go for walk and off in that direction out towards the bay it is actually a nature reserve so you might find some some interesting things there amongst all the old bits of NASA hardware at the same time if you are going for a walk remember we do have lunch at 12 o'clock back in the cafe we used yesterday if you're a little bit late that's not a great a great issue it's just gonna be a bit busy as you know from from yesterday but if you go for walk it's a bit cool out but it's nice bit of fresh air and you'll get to see the neighborhood

I like somebody to report out that's what we do when we get back one o'clock [Music] and preferably quite frankly people that haven't already been on the stage if you will so nominate somebody will recruit somebody else to be the voice rea

# Session 5: Actions to Take

## Presentation on Actions Satoshi Sekiguchi, AIST

## Presentation on Actions: Andrew Brooks, NGA - Title: Designing the Future of Data Science

## Panel on Action :

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

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

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

## Discussion Groups on Actions

# Summary Session