# Transcript Day 1 video 2

# Session 2: Foundations/Motivations

## Presentation on Foundations - Title: Fundamental Issues in Geospatial Data Science: Emerging Trends in Data and Analytics

Yolanda Gil

### Introduction

Thank you very much for inviting me my research is in artificial intelligence and I'm always interested in connecting with new communities and I've already met several connections that are very exciting so I'm very happy to be here I thank you for the invitation to speak George asked me not to talk about AI so I won't talk about AI but what else am I going to talk about what else am I going to talk about that's what I know about so I will tell you about exciting things that we are doing that have AI at the core and that are significantly improving our ability to do data science that 8020 proportion of time that you spend focusing on the data and so on so I call this knowledge power data science and this is something that we teach at USC this is the slide that summarizes my message my message is knowledge power data science is extremely important I see a shift that that moves from focus on data to the focus on models I was talking with someone from the White House OSTP they have Kelvin Drogo Meyer and Lynn Parker eminent scientists directing the office and they are very worried about how do we describe models we derive models from data all the time how do we describe the limitations of the models the assumptions of the models when do the models need to be updated the entire government is being flooded with models how do we think about models

 **and so this is a very big shift and so I'm interested in these problems of characterizing reusing and integrating these models and I'll talk a big deal about that today and I think if you look at models and how you harness data to develop useful models for decision-making you have to incorporate a lot of science knowledge about these models and the data that they use and its knowledge about physical geological chemical biological ecological social economic factors and also knowledge about what a user might want to do with that model so I'll talk about these topics more but infusing into our systems more information and more knowledge about these models and this will enable new forms of reasoning integration visualization management learning and discovery which geosciences data at large** I have a very big focus on Geoscience modeling but I think that these ideas and these trends apply generally so as my message models are extremely important and in order to harness the power of models we need to bring in these knowledge power to data science I think what George saw is a paper that we had on intelligence systems for Geoscience is really calling to infusing more knowledge coming from AI knowledge representation planning workflows etc into geosciences so that paper is available if you're curious about that side so I'll motivate this need for knowledge and data science by talking about integrated knowledge modeling that's something that I think about quite a bit and then I'll talk about why the models are so different across disciplines and why all of this knowledge will help us bridge along the way

### The need for integrated modeling in geosciences

I'll talk about work that we've been doing in mint and I'll show you examples of the kinds of knowledge that we're capturing and using to improve the time to result from the time from a problem formulation and then I'll finish talking about this broader picture of intelligent systems and AI for geosciences so if you if you try to understand how natural systems climate hydrology interact with human systems the people that do the planting and the harvesting human migration industrialization the growth of urban systems you try to understand things that these in this sphere you need to integrate models from different disciplines that model different types of processes and so I have examples here we have a strong collaboration with UT Austin they have a project called Planet Texas 2050 that looks at how can the state sustain the growth in population so how do they manage groundwater resources how do they manage the industrial need for water agriculture needs for water and they have a lot of connections also through the Rio Grande with Mexico and so they need to be able to understand how to make a sustainable plan for the state for for 2050 that's their focus so we work a lot with them in terms of creating integrated models at the bottom I have a map of California there's a mandate to have a plan to manage groundwater resources across the state and USGS just cannot generate models of those groundwater resources fast enough and there's very few people that can actually do this and harness data and run the models so modeling is always a bottleneck in terms of data preparation and data harnessing and so on we work a lot also with cell Saharan Africa I'll talk more about this project but the effects of climate dynamics droughts and flooding and and the general weather forecasting and how that affects the ability to have food accessibility and availability there's a lot of human migration in the region and so we look at the intersection and the connections between all of these different natural and human aspects of that and I think at the bottom right I have examples from the Mekong River where there's a lot of implications across the entire River for different regions different countries there's fisheries involved there's many different aspects and a lot more disciplines than I've been mentioning so far and a lot more administrative regions it's it's virtually impossible to create a comprehensive model that that looks at all of these perspectives so so so a grand challenge

**I think for geospatial data Sciences integrated modeling and you think about agriculture models economic models social models natural models and also models of the infrastructure that deliver some resources there's increased demand for these I think there always has been great demand but there's always been this human bottleneck of manually setting up and integrating these models it takes months or years to put together models for these kinds of problems and it's really a craft very few people understand how to do this and so it's very hard to integrate a pair of two natural models my first attempt at this was between limnology and hydrology so lakes and river models and realizing that those communities never talk to each other they have entirely different views on the characteristics and the behavior of the water flow and so really really hard to integrate models across different disciplines even those are both physical natural models imagine if you try to integrate social economic models with agriculture models so this is a very very huge problem a grand challenge because there's so much demand for it I think we'll have a better world if we manage to do this integrated modeling and if we can do integrated modeling well then we can do many other things really well we can do data preparation really well we can do single model reuse really well etcetera so why is this so hard models are complex but when you talk about models across disciplines they are very very different so I'll give you a flavor of their challenges**

### Diversity of models across disciplines

 so you look at **hydrology models** they're looking at building maybe an irregular grid where they can look at physical variables in each one of those grid cells there's a lot of physics involved the complexity can be enormous but the system can also be simplified and modeled on a coarser grain level it needs a lot of data historical data to adopt the general physics and fluid equations to that particular area but it generates very useful information about the times and days where particular grid cells are flooded so they're very important but they have this flavor of very rich physics based modeling

in contrast **agriculture models** tend to be more focused on biophysical processes so the growth of the plants weeding practices different crops behave differently a lot of these models look at different versions of the crops or not all mazes there's many different genetic variants and so so they look at all of these bio geophysical processes so this is not so much the physics but their processes that are dynamic and they look at the the race of this variable if there's less weeds for example

so how do the different variables relate very different processes **social models** tend to look at societal behaviors through agent-based modeling where you have different groups of agents doing certain behaviors so you can define groups of agents that have children and the children will go to school and so they're able to do the farming or something else and so you define all of these behaviors and you can see the dynamics of how the system evolves and behaves over time \

so if you're trying to understand and integrate two of these models they work as such different scales they have such different methodology some of them are very based on theory some of them are very empirical some of them are modeling variables that are very very different in the physical world some of them the more data and the more types of data that you have they do a better job but there's not so much data availability and the the ways in which you integrate two models from that both look at physics is very different from the way that you would integrate with a social model so the challenges are many and what

I want to emphasize and talk about today is the ability to bring knowledge that we have as humans when we do this by hand taking years to do and how can we incorporate this knowledge into our data science systems to improve the way that we do it so the diversity we actually organized in several levels and I will describe each of those levels in term from problem framing to selecting the right models to mapping variables across the models and to bring in the data that the models need there's a lot of heterogeneity in terms of runtime coordination across models and model coupling etc that's a whole another bowl of beans now have a very colorful graphic on the right if you're interested but I won't cover that I will focus more on the this knowledge power data science ideas why is this so hard so you frame the modeling problem well what variables should I consider what models are available so there's a whole back and forth you may go down a path and then come back to reframing your problem differently the scope of the modeling problem can be very very different but this is all done manually so when a modeler comes in they don't really know what's available to them and and how to think of the problem and what can be simulated and run the model selection there's model catalogs they describe what the models do sometimes you can run them sometimes you can couple models if they're very similar very close but typically you'll read through documentation and you'll do a manual selection of your models there's lots of tools out there they're not necessarily well integrated and so there's a lot of passing data from tool to tool etc it's a very manual process as well then models can talk about different variables so for example a hydrology model is very different from an agriculture model in many ways but one way is that a hydrology model looks at a grid of data the agriculture model is a point model there's one point for that entire area that has corn plants and so just being able to transfer things about soil moisture and other kinds of information is very hard so the mapping of the variables happens mostly by hand and through manually crafted data transformations and then finally the models are very thirsty for data so the creation and the ingestion and the integration of data there's lots of tools as well but they tend to be disconnected and you try to find the ones that help you to whether it is regreting or something else so so we can do a lot better in all of these aspects so

### MINT: knowledge-powered data science for integrated modeling

* Intelligent systems for geosciences

MINT

I will talk about work that we're doing towards this because if I tell you just words and I don't give you any concrete examples then it's a lot less helpful so I will show you examples from the work that we've been doing so how do we use knowledge in all of these pieces of the problem from framing to to model selection to variable mapping and data ingestion the project that we have is called mint it's a DARPA funded projects a pretty large project we're focused on South Saharan Africa we also work with the Gates Foundation and it's very exciting because we have modeling experts we have data experts we have workflow experts we have high computing experts etc all looking at this problem of facilitating the model integration so going through this pyramid I will start first with the bottom aspect of this so imagine we've picked our model we know what kind of variables and data it needs let's see how we deal with data ingestion so so knowledge guided information extraction and integration so one of the things that we discover very quickly is that for example in a region like Ethiopia or South Sudan if they have gauges in the rivers historical data they don't have a lot of years and they don't have a lot of gauges so we're looking at creating digital gauges that come from satellite data or remote observation and what's very interesting is that a lot of the things that you can learn through machine learning they do a combination of supervised and unsupervised models this is work of vipin **Kumar and his group at Minnesota and you know they can they can extract information about land use and what kind of crops grow in different areas they've created this concept of virtual gauges for the river so they can tell the depth and the width of the rivers and what's most interesting is that they have figured out a way to incorporate knowledge about physics that constrains what the machine learning models learned so I'll give you very quickly an example where part of training these neural networks is to incorporate knowledge about physical constraints so this is an example from looking at Lake I think Lake temperature and so they start to look at the data that they have at different layers in the lake of different depths in the lake and the the neural network does learn a very good model but there are some places where you'll go from very warm temperature to cold to warm again that don't make a lot of sense and so they incorporated knowledge about what is the I think you can see it on the right so for example the depth versus temperature and then other energy conservation laws and so what you see at the bottom is what the machine learning method learns**

versus what a physics-based lake small lake model learns and they're very very close together but the machine learning model is really learning from the data and some of the laws that the the lake model does so the machine learning model is much cheaper to develop as long as you see that what it's learning is consistent physically with what's going on in the lake so this is an example of where using knowledge about physics and knowledge about the world guides the machine learning method to do better very exciting work also in terms of data preparation and data management let me go to the

next slide first we're doing a lot of work on using probabilistic models to learn what tabular data contains and so tabular data tends to have a lot of structure you will say this is data from a country a this is data from country B and then of the rows and columns are organized in certain ways and so there's a lot of structure within them how can we discover that structure and tease out the content of those data tables so we've done some great work where extracting metadata the variables used on those tables creating semantic representations of those tables can be done mostly automatically there's human intervention of course but it cannot be done mostly automatically once we have descriptions of these tables we can say okay this model ingests data in this format the data is available in this format so we've been working on automatically transforming data from one format to another so we use AI planning techniques but if we know the initial state and the end state we know that we'll have to go through some set of actions to do those transformations so we're including a lot of the transformations that are typical of modeling into our architecture to be able to do this so that's this layer of data ingestion ingesting knowledge of physics into machine learning models from data and then also ingesting semantic models of data that are automatically or mostly automatically created by the system the next layer

**that I'll talk about is mapping variables and understanding modeling variables so the work that we're building on is by Scott Peckham and others at university of colorado looking at what can we possibly measure in a physical system and organizing and describing the patterns that we typically use to describe variables so there's objects that you're measuring things about there's processes that involve those objects it's very important to understand what process you're trying to get to and you're trying to measure the processes you can observe certain quantities about them the state also have quantities about it and there's properties about that object that you may also want to observe or measure in some way we've also discovered that in order to connect modeling to decision-making you want to be able to have indices and define indices that tell you something interesting about the system so for example a drought index index tells you a lot more about what's happening in a region than the raw object properties in that region so we're building on this way of defining variables physical variables using patterns and so in our system if you include a model or if you include a data set we're always trying to map when it says you know channel depth into some particular combination of patterns that describing the object which is a channel and its properties etc etc so this is fantastic work that is very very helpful**

**Model Selection**

Slide 29 and 33

1. Model invocation

2. Data formats

3. Model variables

4. Constraints

5. Adjustable parameters

6. Interventions

 let me move to how we select models and how we describe models this is very very new work on using semantic descriptions of models we've done a lot of work on describing machine learning models dating back to the year 2000 and before your interest is in that particular aspect I'd be happy to talk to you about this also describing software and describing data transformations and other kinds of software I'm gonna focus here on on describing models that are these dynamic simulation based models and how I'll focus on six particular aspects of these models that are very important for us to capture knowledge about so the first one is model invocation so we say here's a big software package for example modflow from USGS it has many versions many different processes that it could model for groundwater so here's a particular configuration with seven select processes there's no snow melt because in this particular configuration we're not going to worry about that so we won't need that kind of data etc and then we'll do set ups and calibrations for different flavors of problems or different regions or areas and so how to invoke the model how to map this this majestic piece of software into specific ways to invoke it and use it in a consistent way is one of the first steps that we follow the next thing that we look at is the data formats so this you probably all know very much about and we try to characterize not just the major data sources or weather data sources etc but every model will use specific kinds of data that we try to describe and standardize properly the model variables are also very important so just saying that you use a netcdf format is fantastic but we want to understand exactly what variables are used and so I showed you before the example of how we use these ontology x' and patterns to describe variables but we also describe you know that you the particular boundaries and types of values that they take etc we also describe constraints and preconditions so a lot of the models will say you know I need daily rainfall data versus monthly rainfall data these things turn out to be important but if you're combining a hydrology and agriculture model they need to be using the same data so we want to capture these constraints that are not specific to one model but that bridge across different models number 5 is very very important and you don't normally see in models which is what are the drivers that I can use to do forecasting of different scenarios and what are the parameters that an end-user or decision maker or modeler may want to adjust in this model and so we put a lot of care into describing how to adjust parameters in a model and make it very use very simple to do so and also what driving data we can use related to these parameters we also tease out potential intervention so here's an example where in an agriculture model if we change the planting windows if we do an intervention that says we're going to plant on these dates at the earliest then that might be a way to avoid the drought with more likelihood or if we plant earlier we might avoid the flooding well the you know that may happen before the harvest or weed control and management practices and so on so any potential interventions are mapped to the model specifically so this is a flavor of the kinds of knowledge and the kinds of information that we find are very very necessary in order for the tools to understand how a model fits and how to use it and save ourselves all these manual configuration and setup of the model

**Problem Framing**

the final thing that I will mention is on problem framing so depending on the the framing of the problem you will select different models you will have need for different kinds of data so we want to and the problem flaming framing it's kind of a negotiation so I'll frame a problem if I have this data if I don't have the data then I won't bother to frame the problem this way so it's very important for users to be able to explore what's there in the system so we're using kind of a general decision-making process framework and the modeling sits in the middle and we're simply trying to see how much of that bigger context we need to incorporate so for the moment we're allowing users to state their problem to create tasks that tie response variables with driving variables what is what are the variables that you're interested in varying and seeing the response to what kinds of interventions you're interested in so will weed out and we'll show what models and what data are required for each one of these so the users can explore different paths throughout the system so they can do forecasting they can look at interventions but all of this information that I've showed you is extremely useful to make the actual going through the modeling exercise is pretty fast so just last week we had a DARPA evaluation with several users that are third-party users running through a lot of different problems through mint they identify what results they're interested in they look at what relevant models are shown to them and very quickly they're able to find the system transforms the data for them and they're able to set up and run models pretty easily so with two hours of training they were able to actually use this and and come to their solve some challenging modeling problems so we're very interested in tying now that we understand all these variables that the models look at in tying the different variables into causal models in understanding how to do different kinds of forecasting interventions that will also require the forecasting and quantifying uncertainty this is a big topic in modeling but certainly very important here so my summary of all these things that I've told you about meant is that we are using a lot of knowledge to enable this more assisted data science this more automated and less manual kinds of data science so three highlights to use knowledge guided machine learning how can you use machine learning systems that will take in knowledge that you know to be the case for a domain in our case physics for example knowledge rich catalogs of data and models

so I showed you how we use probabilistic models to automatically extract metadata from tabular data I've shown you what kinds of information we keep about models and we reason with and then the third piece is this knowledge driven problem framing so understanding what are the indicators and the variables of interest to a user and how those map to the variables in different models

**29:50 intelligent systems for geosciences**

I want to kind of come back and conclude by talking more generally about how all of this knowledge can create more intelligent systems for modeling and for geo Sciences in general. This is the article that I mentioned at the beginning so if now you are more interested please visit. It as it came out of an NSF workshop on intelligent systems for geosciences but it is a very active network so if you go to is - geo org you will find a lot of people that are similarly minded about how do we build these knowledge power data science tools for modeling.

Slide 45

In this article you will see a lot of really exciting ideas about opportunities across the scales that all of these knowledge can bring to bear so workspaces where you have more interactive model building more support for model building the ability to specify high-level questions theory guided learning so machine learning that incorporates knowledge of different kinds the idea of trusted threads where someone doesn't just hand you a graphic or an or a visualization but actually you have access to the entire thread that generated it and then on model driven sensing I haven't talked very much about this but it's really the idea that the more that you understand what data your models need and not what granularity and where your priorities are the better you can guide expensive sensing for example

Slide 46

I've talked about the project that the bottom center called mint but we have a lot of ongoing projects about infusing knowledge and better representations throughout different aspects of science of whether it is the software the collaborations the workflows the provenance and of course the data we've done a lot of work on crowdsourcing vocabularies we have a project on doing automating discovery with AI we have a meeting of the Turing Institute a worldwide meeting on **how could AI make discoveries worth of a Nobel Prize.** I think there's a lot of automation that we can move towards. If you're interested in these topics I heard this morning questions about link data and and Semantic Web **everything that we create our URLs and web objects** so if you want to see what model were using what software are using everything as you are eyes and they're interconnected so you see a visualization and you follow all these URIs to see where it came from the provenance being very important. I'm a very big fan of infusing a lot of knowledge throughout all the objects that we have and have them really better interconnected

Slide 47

I will conclude putting back the first slide that I showed you my message, my conclusion is the same that I started with so **I see a shift from data to models trying to understand models** characterize them better we use them integrate them. I think we need to incorporate more science knowledge into how we think about models and data so this is knowledge about all sorts of aspects of science and also knowledge about the users goals and context what do they want to see what do they want to do how do we link all these models to things that users care about.

Through all of these knowledge I think data science will see a new generation of tools and infrastructure to really support better reasoning better integration better visualization learning and discovery. I mentioned the is is-geo.org site

Slide 48

My last slide is a pointer to a 20-year roadmap for AI that I just co-chaired it's public and it looks like twenty years out what do we need to accelerate science what do we need to accelerate research what do we need in terms of infrastructure to support AI technologies and AI research so if you're curious about this I know there's a section following up on AI but I thought I would talk about AI anyway if there's interest. I'll conclude there and take questions

[Applause]

we have time for maybe two three questions

work slide you listed **uncertainty** can you give us a motivating use case and why that's important for your future work yes so in a lot of ways you know we can generate a lot of flattened maps we can generate a lot of predictions about flooding that look at you know what if there's 10% more rainfall than in previous years etc etc what a decision-maker wants to really understand is how likely it is that this flooding will occur in fact and so to do that we need to do a lot of different model runs consider different models as well so the uncertainty comes from the data that we have to start with the forecast that we start with being able to see if many models tell us that the same kind of flooding or flooding is predicted so I'm not a scientist but I understand the uncertainty comes from many different sources and certainly the decision makers want to hear about this and see about this and get a sense of that right okay so I'm certainly familiar with these sources of uncertainty and all that kind of stuff I'm really interested in the decision makers who are banging on the table to say I need to understand the uncertainty so a lot of reasons is because we're considering the **interventions** as I mentioned right so if you're going to spend quite a bit of money and effort into a particular intervention you really want to make sure that that's a likely scenario and so that's the main source of the demand for uncertainty

any other questions I do have a question for you in terms of **data and knowledge** who are using those two how do you separate those two how do you represent I think I know how our data HR turned red **how do you represent knowledge** so I don't know if I know knowledge when I see it I certainly know data when I see it I think what I mean by knowledge is symbolic representations probabilistic representations that speak to a conceptual understanding of a system and that you can link together to create new knowledge and so I'm thinking about concepts I'm thinking about inference I'm thinking about laws I'm thinking about you know a different layer the data is more linked to observations or predictions but it won't tell me that knowledge based understanding or knowledge level description of how the system behaves

any other questions I'm sorry okay so **some models are proprietary and black boxes** and we you know we don't know the inner workings of them could you address that issue with your MINT framework? so yeah so we looked very extensively at this the the area that has us more worried are the the social and economic models there's a lot more models in the physical natural realm that are public that are open they have their own problems they're buggy there may be three people that know how to fix your bugs they have a whole host of an other problem but the **proprietary models particularly in infrastructure and socio-economic modeling** it's a very grave situation so I'm not an economist but I've talked to a lot of economists and there's very little of this concept of model reuse so you take a hydrology model and you use it for Ethiopia you use it for Pennsylvania and it's the same set of physical laws and principles and the same software that you're bringing to bear in economics this is what I'm told you know if you've published a model why would you use it for a different problem right then you don't have a paper it's not modeling right so so we're we're trying but it's not very easy to develop economic models that do capture basic market principles that we can reuse across different areas but that's a challenging area in itself I will tell you that a lot of models are open and in principle accessible so for example in the do-e labs there's a lot of models about infrastructure that in principle are accessible and open to us but if there's two people that understand how to use them and it takes them two months to set them up it's not a very you know they might as well as well be closed so so yeah this is a huge challenge

40:10 - please join me in thanking professor gil.

## Panel on Analytics :

### Lauren Bennett, Esri - Perspective: Spatial data science moves from the avant garde to the mainstream

hi I'm Lauren Bennett I work at a Esri. I lead up our spatial analysis and geo processing software development team so we build the framework for doing spatial analysis within ArcGIS platform.

I think what I want to talk about follows up what dr. Gill showed really beautifully because I am similarly passionate about the idea that a model sitting somewhere where three people know how to use it is has minimal use and is disappointingly I mean it's and it's very useful but not actually in practice so kind of I posed this question and I'm gonna stay very high-level because I want to make sure we actually have time to kind of have a discussion but

you know to me the question is will data science and specifically spatial data science live up to its promise I don't think it certainly don't think it has yet and I think that it's really important it's not just about the algorithms right it's the algorithms are very very important and I'm really glad there's lots of people working on them and certainly we've got some folks that are passionately focused on building algorithms but for me especially when I think about building software that hundreds of thousands of people use the minute we we put it out there I think about how important it is for it to be able to be applied to a wide range of problems that we find you know we focus a lot on finding those things that are out there that can be broadly applied and aren't may be quite as niche that's certainly where where we spend our time also really important that they can be explained to a decision maker so that they can really make an educated decision about what that knowledge really means and that's increasingly important because decision makers are being bombarded with knowledge and being able to to really create things that can be explained is critical and then also that they can be operationalized and I think that gets it the idea that we need it to be something that many people are able to use and that within an organization you know you don't have to hire a team of PhD data scientists in order to make use of those models at that at the kind of organizational level so we spend a large portion of our time trying to make things simple and approachable trying to take widely use methods and think about how do we boil this method down into something that pretty much any analyst can understand you know

every local government state government every public health department and public safety organization has a GIS analyst or a public health analyst who has a lot of subject matter expertise but not necessarily a degree in statistics or degree in you know operations research and we're trying to take these things and make them approachable to to all those people because a lot of the most important change is gonna happen at the local level and so we have to put the powerful modeling capabilities into the hands of those people that are at that local level and then finally a lot of our time and and we think a lot about the idea of not just you know we build tools that we think will be very broadly applicable because that's kind of our our audience but of course those niche models the things that are very specific to a small group of people are still incredibly important and so it's it's we put a lot of thought into the idea that the things that we build have to be really easily integrated and extended because you know there's no way that the complex problems that any one of our users are facing can all exclusively be solved inside of ArcGIS and so we spend a lot of time thinking about how can we make sure that our users can you know make use of the things that we've done to make things simple approachable but then also reach out to this broad ecosystem and use our use the stuff that's out there in Python to extend and integrate what we've done so that they can solve those problems essentially find the best possible tool to solve their problem and just use it to solve that problem and we want to make that really easy for them to so I mean I think that that's at the heart of how the work that we're all doing is going to actually live up to its potential is that we go beyond publishing papers and you know it's building the really powerful algorithms and **we get into getting these things into the hands of people that are going to use them to drive decisions** at that very local level that's my position

Thank You Lauren I like your slides

simple keep your next thank you

### 47:20 Keith Hare, JTC 1 SQL/GQL convenor

The real reason I'm up here is because of what I do for fun and a few decades ago I got involved in the SQL Standards Committee and hung around long enough until I'm actually the convener of the International Committee that creates standards for database languages for years we've thought of that as SQL and recently we've had a new project I'll talk a little bit about called gql property graphic query language so I'm going to talk about some things not quite as succinct as other speakers.

When George emailed me about doing this one of my questions was how much time do I have and basically I need to represent 30-plus years of standards in five minutes so I wanted to focus on some highlights of things that we've done in the SQL standard for years in the SQL standards world we thought of anytime new technology comes along we've integrated it into SQL so we've added online analytical processing capabilities we've added temporal capabilities most recently we published a new part that supports **multi dimensional arrays** and as as column and the table and operations on those arrays and a lot of this work came out of a couple people in Germany who have been active in OGC and TC211.

We've also started looking at property graphs and about three years ago I talked at a location power seminar in Orlando on big data and one of the questions I got was **when are you going to integrate graph queries into SQL** and that was sort of the first of about three or four other discussions that I had in the next couple months about graphs and property graphs and we've ended up creating two projects in the standards world. One is a project to integrate property graph queries in SQL we're calling **SQL PGQ** the reason for that is that a significant portion of the world's data exists in SQL tables there's some really interesting analysis capabilities that have been developed in conjunction with property graphs by giving users the ability to represent the relational data as property graph data we give them access to those analytical capabilities we think it's going to be important. What we're really doing there is we're integrating a property graph query language with SQL but we didn't have a property graph query language standard. So we've now created a project to create a declare a standard for a **declarative property graph query language that we're calling GQL** and then it's going to include the ability to represent data in property graphs to add modify query delete in a transactional context a path language to do all of the queries wherever possible we're going to use the SQL specification. The other issue is SQL **traditionally you have to define the schema upfront before you start storing data** it turns out that there are some valid use cases for stick the data in **figure out what the schema really is later** I have to be careful with that because you could end up with complete garbage but we're going to support in the standard the ability to have either an upfront to find schema which has some distinct advantages or the ability to be schema less. I don't know if it'll make it into version 1 but eventually we'll have the ability to mix those in so you could have a partial schema.

That's the high-level view of this property graph query language

The use cases this picture you can think of the SQL pgq project is really the intersection between GQL and SQL and so we have these two things going on and we're working on the intersection as we build the gql standard what are the use cases a company called **neo4j** published a press release back in September after the project was approved and some guy by the name of George Percival said something about **OGC building standards on geospatial standards on GQL** just a tweet. We have a couple a couple of other interesting use cases that I don't have time to go into but this is one of the places where where this kind of technology we think is going to make a lot of sense. We've had a conversation since then and one of the wg3 people got the name of tobias linda kerr from sweden will be at the the meeting next week in in toulouse to particularly talk about the geo sparkle work.

What else are we doing?

We've been talking about **streaming data** for a number of years and we're beginning to make progress on it although it's not nearly as visible what we've done isn't this visible is ready to be visible as the GQL work the idea is that you want to be able to query the data apply all of the SQL analytical capabilities to the data before you actually store it or maybe instead of storing it you want to do input from zero or more streams output to zero more streams or tables and then provide some additional analytical capabilities across that streaming data. Looks like some interesting work and I'm looking forward to somebody actually doing more work on it.

We've also talked about adding **statistical functions** to the SQL specification and the basic idea there is that there's a lot of statistical functions that would be useful in data but they're already defined in textbooks or standard references with the algorithms and so we don't want it to find the reference we really just wanted to define the signatures how you would call them in SQL and leave them up to the individual implementation to call the appropriate library or to define the appropriate library.

Other than that we're not really doing anything in the standards world.

### 54:40 Todd Mostak, OmniSci

I'm Todd Mostak co-founder and CEO of OmniSci, some of you may have formerly heard of us that's MapD, so we've been around for a while we kind of focus on the intersection of interactive big data analytics and geospatial and location intelligence.

I just want to quickly talk about you know what's changing what's changing in the world of geospatial analytics there's a lot of things you've heard a lot today and I just think some of it can be summarized and **three points**.

One is that unlike the old world of GIS we're moving into a world where we have all these new sensors devices phones IOT today their data coming off the cars planes etc power lines satellites. This **proliferation of data** is creating located the proliferation of these type of datasets from these sensors and devices it's creating location data of unprecedented scale and velocity right so instead of assets that you might measure in the thousands whether they're power lines or you know real estate properties these things can be billions or even trillions of data points and they all have a time and geo stamp to them and so I think that's the first thing that people are struggling to deal with

**second** is I think one of the themes here has been that we're seeing a convergence of traditional geospatial analytics right with traditional just analytics what people might call the data warehouse and bi space as well as you know **data science** workflows. Everything you know people talk a lot about data science and how cool it is but it's really just people trying to solve their problems augmenting the human with the machine. I think in particular location data is is very ripe for leveraging machines and the power of modern computing and modern hardware to actually give answers that a human could uncover run models etc and then

**Third**, finally **new use cases** these new use cases demand real-time and interactive interrogation of this data right so no longer is a day are running a spark workload often enough where things might take hours days or even weeks people need real-time answers because they're trying to figure out where there's fire risk if your utility if your tell code you're trying to figure out why calls are being dropped is it related to can you tweak the cell tower the location or the direction of the antenna is it a handset failure a firmware update what's causing the issue. so I think there's a need for **real-time scalable analytics across geospatial non geospatial data alike**

know our whole reason for existence is basically to allow people to do scalable analytics using different modes of understanding sequel queries visualization and data science workflows to allow them to get insight in time spans that they wouldn't be able to otherwise

so I thought it might be interesting just to walk through a quick **workflow** that we see sometimes in our customers we have open source core the large number of users using our platform different use cases so we've been working with a major federal agency that works with domestic **aircraft traffic** and so one of the things that people are trying to do for example is figure out you know hey what's the **best route** how can we actually reduce delays and optimally route flights both not only for on-time performance but also for safety. How can we make sure that we can dynamically space out these planes we route them around weather and make sure that they get there on time and in one piece. This is an example of our platform it's **four point five billion records** we're running on a workstation and here we're leveraging not only this fast sequel running on GPUs allows us to get millisecond response time and you know here I can instantly drill down click Southwest I'm rendering all this all the state is being **rendered on the GPU** because it's not scalable to send billions of records over the wire and I'm gonna see the max eight here and we can immediately basically see you know when **the max 8 was grounded** can zoom in here and say you know I'm in actually brush over and see you're up basically it's grounded on the 12th of March I believe and then us grounds on the 13th of March.

One of the things that we've been looking at which is at the **intersection of analytics and data science** is basically this ability to do things that it's great to be able to see this data but imagine we wanted to actually visualize you know the routes between two given airports so say we wanted to look at Chicago to Boston. This Jupiter lab I can't one of the things that I can do here is what I can do is I can actually now convert to a **pandas** like **workflow** and actually take this data and take all the flights that are between ORD and Boston Logan and actually bin them and what I'm doing here is now I'm getting all the spatial bins between these things this happens in seconds and now what I can do is I'm **pivoting this table** into a generating this huge SQL we're behind the scenes this panel is like syntax that basically is creating 200 columns of the top features the tops facial bins between these two airports and then effectively what I'm doing is on the GPU running a **GPU accelerated k-means algorithm** which takes seconds. I'm not gonna step through the whole thing here and then finally load it back into the database. now I actually can look at these flights here and yeah **showing them clustered** here so I can actually see the directories and the next step would C be to say hey what are the trajectories of whether the flight paths that actually lead to the best on-time performance so you can see I can just go in at it and now I can visualize some of these different trajectories this is just a quick demo of what you can do

I think this shows the example of an **analytics data science and location intelligence converging happening at scale in real time** and the things that you can solve for with these types of capabilities

1:02:45

### Hamed Alemohammad, Radiant Earth

Hamid Muhammad I'm the chief data scientist with radiance foundation. I'm not gonna talk about are sorry I'm not an expert in are so but

I'm gonna talk about training data and a little bit maybe Python later on but before going to that I wanna share an experience from my PhD time how I'm thinking about the user dealing with data and particularly cloud versus an FTP server so I started my PhD about 10 years ago at MIT my research was fusing precipitation data from six different satellites over CONUS over eight years of data that was massive amount of data and I had to do everything on a local server where was the data and an FTP server behind the web. There was no API you might be familiar with in your class is the comprehensive Large Array something system I could request data through that but there was a limit on the number of files. So I emailed Emma saying I'm a researcher I'm doing this and that can you increase my limit of files I said yeah a factor of two not more so I had to submit like 60 different requests and for each of those requests I would receive an **FTP link** that I heard that download the data but I had to search each time myself so limit the time filter limit the spatial filter we get the real the correct number of files so you can submit the request because if it is even one more than one you can submit the request. **It took me almost six months to get that data** put it in a server process it from an orbital degree to a regular grid match all the six different satellites after six months I could start doing the actual work of the modeling.

**So with that mindset I'm thinking how can we do better** I can we come to a point that that that process is much reduced it is more benchmark it is reproducible I think that is also the factor that we don't talk about it much in science it's more talked about which is a paper comes out there's a lot of data being used there's a algorithm being built how can we make sure there's reproducibility there and I think with with data science and ML that's becoming more challenging because we can build models more easily probably compared to the physical kind of paradigm we had in the past so we need to be more careful with that that being said

I'm gonna say what is ready **in Radiant Earth** before jumping into what I'm gonna talk. Radiant is a nonprofit organization it's a foundation established about three years ago our mission is really focused on empowering users both as individuals and organizations who are working to use Earth Observation data for global development challenges we were funded originally by the Gates Foundation and Omidyar Network and through the years we have attracted more funding to do that but that's really the key mission for us to empower them through tools that are we did a new focus we have recently focused on **machine learning applications** so you want to use their observational learning and IRA culture on land cover surface water socio-economical impacts how can we empower them to do better and we do this through a ecosystem of work

what is the motivation here so this is a different panels of **agricultural** practices at the global scale from u.s. to Europe to Africa to East Asia to Australia and Latin America especially you see a lot of patterns they're diverse patterns there's also farming practices that are different climate is different weather is different the seed that is used is different probably fertilizer application is different if you want to build a machine learning model to basically do a monitoring system and these even a simple crop type mapping nothing more you need a training data that matches that diversity right we need to make sure there is diversity in the data that we are dealing with and that is that is a lack in our community in terms of how can we do this better

a similar example of that goes to **computer vision** work so where the data science comes from right some of you might have heard about the **imagenet** which is the famous label data sets of humans objects animals about 1 million objects with bounding boxes across about 14 million images and there's animal competitions that dye data and that is deriving the innovation in that field because there is a benchmark everybody builds a model they know the best of the basically the year and then next year people come and do more innovation than that

how can we do similar stuff in the geospatial were thinking about the type of challenges that we have the type of data that we have we have a temporal dimension that the computer vision world typically don't have the video have but the classical imagenet of the world don't have the temporal dimension so some efforts are going on

you might have heard about **big earth net** which is a training data so data sent you know - it contains about 600 thousand images patches of 120 120 pixels so that is a very good effort but it's all over **Europe** there's no other parts of the world in that data set let's let's do better SpaceNet is another concern you might have heard about that there's a lot of good data particularly from where commercial imagery that each the globe max or data in there with a lot of buildings and roads it has more diversity is not six cities in different continents but still we need to do better of that so that's why we basically to address these problems lack of diversity accessibility of data interoperability of these data and the ML readiness

established what we call **Radiant ML Hub** which is basically a hub to share these data make it easily discoverable and accessible and benchmark and document it so users can easily trust the data the hub is really the key part of our work but there's also a community element to that it is not just Radian doing this we are doing this as a as a community effort we are not going to generate all the data we are going to serve the data through an API but the key part is really having the community engaged in that process and we are doing that actively and partly also doing an education work particularly with the user base that we have I don't want to get to the details of this but the whole architecture that relies on STAC so Mark mentioned the stack which was an effort started exactly two years ago in October in state of the map in Boulder and that's what sitting behind Radian mo ha because we don't need to have all the data on our repository you can sit anywhere what we need is the stack catalog so users can easily discover the data and be able to serve it so you started to say

I want to close with an example why again this is important think about I'm Group one I'm doing a modeling on a problem so it start generating my own training data I do my own training I do my own prediction and then I have an inference I provide to the user group to comes they don't have access to my training data so they do the same thing again from scratch what happens at the end if a decision maker comes how can I compare the prediction one and two if they are based on different training data the model performances are different the ideal situation would be we have a benchmark in the beginning so we can then compare and benchmark the models at the end and that's what we are trying to get to that point so users can use similar training data sets and show how different models are performing and

this is basically my dream for line of code you import a library in Python which is the mo how you search on a bounding box with a keyboard like maize. I need crop data and the next line you load the data you don't need to spend six months to get your data from an FTP and then I greet them and then get to the line five of that we will be there I mean we will be here in a month you can get the data out the Python won't be finalized by then basically a Giri is our target will release of their crop data we have from Africa and then we have a land cover training data set coming up in February and the Python package will be more in in march/april timeframe but it's all open source and we would love to have your feedback and engage more and Rhonda thank you Thank You Hamid

1:10:33

### Q&A

I will kick off with one question and then I will turn it over to the audience my first question is let me announce that I'm actually on the board of OGC so that's what it's coming from where our standards helping you and where do you see a need for new **standards**?

Todd: I mean I think it's in terms of OGC standards it's been very helpful that yeah I think sometime in the past people were doing obviously each one had their own custom implementation of geo functions it's so it's really great to have one standard that you can effectively adhere to you know. One of the things I think we see is on the **raster** side I'm sure you guys are working on this but it's a little less specified and there's people with different implementations and so we would love to see more effort there because I think it's a critically important data type particularly as all this satellite data comes online

Hamed: I think there are two aspects of the standard very critical for our work one is the data catalog is standard so stack efforts so one thing I want to mention before getting into that was there was a spring last week that George and the OGC community was there to integrate **STAC** as part of the OGC definition particularly the API so that's openly moving forward and we look forward to that integration when the stack comes to 1.0 it will be more official but what stack enabled was there was a problem people were the simplest task is you want to search for raster data if you will go to five different data providers in the US the big ones even both government and commercials you have to write five different API calls with different parameters to get those imagery the simplest definition of cloud cover someone was with means Iran someone was between zero and hundred so you need to make different changes for that but the stack enables that you have a single API definition to find the data and then serve it back so and that has helped us with their mob now because there is an extension for training later so we can have a catalog a standard for training data easily discoverable

Hamed: the second is under actual data I mean the cog that Mark mention is very innovative in that says it's very helpful to serve the data easily just get a chunk of data the traditional raster data another satellite data it's usually stored in hdfn netcdf they are very good they're the perfect for kind of geospatial dimension but they are not cloud friendly hdf is getting their mark mentioned the the s3 object thing netcdf has improved as well there is or coming up and a new format there's also X array implementation of netcdf in Python we are getting there but I think what would be helpful is more data being converted to those formats more **cloud native formats** that would enable and streaming the data through the cloud

Keith: I write standards

Lauren: I guess from from my perspective it's about making a **really easy for people to go between these different platforms** to kind of bring together these diverse sets of data and a diverse set of methods models algorithms that are coming from all over the place the standards kind of allow that integration and that's crucial as people we need people not to be stuck in one place cool

now any question from the audience

Question: so Lauren you mentioned the kind of the importance of making sure that models can be explained and him and you talked about the difficulty decision-makers face when they have you know multiple models and they don't know how to make sense of them I'd be interested in hearing from the panel kind of what you think we as a community can do to communicate what we do in ways that kind of maximizes the **uptake of you know like types of models** that we're all building

Lauren: so I guess from I think that on one hand we have to put a lot of emphasis on our communication I think it's really easy to as people who are passionate and I mean large portion of us are passionate about the analytical methods and so it can be easy to kind of assume that that's where things live and die by how right it is and how powerful or sophisticated the algorithm is but if we can't effectively communicate it to decision-makers and it doesn't matter if we do the best analysis that's ever been done in the history of the problem that we're trying to solve if nobody can use it then it really wasn't that useful and so that that work I mean one of the things in a lot of those the **Venn diagrams about what a data scientist** is the one that George showed earlier I think is a is useful but one of the ones I've seen that I really like about what makes **the best data scientists is their hat there is an ability to tell stories** that's inherent in what it means to be a data scientist and it's why data science is sexy for kind of the first time ever right now is because there's this group of data scientists who are good at telling stories and I think we have to embrace that it can be easy to say well no but if if it's if it's that's kind of playing into what's sexy but actually if we want the best analysis to be the one that's taken the most seriously then we have to buy in to the importance of storytelling so I think that's a key piece

Hamed: I can use any example to answer that so there are a lot of applications now being built for agricultural applications using multispectral data so they provide for example recommendation app to the farmer based on the performance of the crop what to do next the farmer gradually adopts that solution when they see the benefit over the course of one season or two season but then that company I've talked to many of them both in US and outside us they go and say ok we want to aggregate this information provide a recommendation to the policymaker to the Extension Office but they don't accept that and I say ok what do you present it and what is this I say yeah we say we have **R square of 0.6 that doesn't help the policymaker** right because the policymaker the typical way they do it is they do sensors they have all of people on the ground they see all of the papers they know how the process has been evolved to come to the conclusion they they don't see that **transparency in the model building in many cases and the benchmarking** I think these are the two key words we have heard from the policy so we also work with the World Bank team and the group that does the LSMs survey it is really the benchmarking and the transparency from them to understand what is really going on behind this ml thing so we can trust it and then put it into our decision-making process

Todd: Thanks yeah I think there's always gonna be a **tension between wanting the most powerful model and having a model that's more explainable or interpretive** all right and I think it's one of these things where people just have to know when it makes sense to get that additional degree of accuracy versus doing something simple like a regression or geoweighted regression that has maximum explained ability. I'm a big believer and also in **visualization** in allowing people to tell these stories in the sense of a lot of people use our platform both to actually just see they're all data and see ok is this actually yeah garbage in garbage out is this bad data is it the right day to put in but also to actually interpret the model so in one example one of our automakers who work with actually takes their model I think they're using extra boost and then basically feeds in every possible input so that they can see the shape of the output and then visualizes that and cross filters on that so they can see potential biases in the market which could either be in the model which could be inaccurate or could actually lead to legal issues for them. I also think that seeing the data is really important the sense of another example CPG does all this deep modeling of what census block groups otário with demographics a bunch of you know pretty heavy geospatial analytics but the end of the day they won't make any decision about a campaign to run unless they see it because the model might still spit out like you should this deodorant you should go after this demographic poor demographic whatever it is and it could just be that it's people in a certain region of the US right so it's really important to have a **human in the loop** still I can actually see the data and its full breadth and depth.

Keith: to add to the the machine learning model discussion one of the things that I've been told is that the **property graph technology is actually really good for allowing you to record the internal steps and the machine learning** so why the machine learning worked and that gives you a reviewable some low-level machine learning model

Lauren: I also think it's really interesting I think at the end of dr. gills presentation about **what's knowledge that question** you asked which was really interesting to me and I think it's particularly important when we get into AI because in a lot of cases there is this question which is **are we actually gaining knowledge because a prediction is not the same thing as knowledge** so I think framing what our questions are it plays a really important role if we go in we're not even really sure what our question is we get something out but we haven't gained knowledge I think there's there's some value and stepping back and really thinking about and having our decision-makers critical have a critical eye towards **what what's the difference between prediction and knowledge** and how can we be comfortable input times where we're just gonna have to be ok that it's a prediction and we haven't that there isn't knowledge that comes from it and when we actually need there to be knowledge that comes from it

Question: I have a question about the **ML Hub.** So the data you have are for training and Finch and benchmarking right yes okay so the data sets you have there they're stored on your servers so they're not like stored somewhere else and then you go after them?

Hamed: the ones that we are generating ourselves are on our s3 bucket AWS I should appreciate their support but there is also data that comes through partners for example we are working with Microsoft ai they have the data coming from their own grantees those are sitting on Azure they are not sitting on our own bucket but they are served through the same API for the user so as a user when you come to mo how you don't see what is going on behind the scene the data is served through one API

Question: okay are there any sort of similar hubs or **repositories for just geospatial data in genera**l you know regardless of where it comes from and it's so not necessarily that it's all stored in one spot or even just a few spots but you know so there's a way to kind of say well I want data that like. I love your little pipe says very simple saying this is what I want go forth and find it somehow from whatever resource. I don't know if I'm making sense they are big

Hamed: I see your point but I think that the platforms that have been so far has been they make a copy of the data on the repository to serve it I mean Google Earth engine is a good example right you have a replicator all the raster data from different satellites climate model data on Google server but when you go to Google Earth engine you get all the data but because there's a replicate behind the scene you don't get for example a copy from a AWS there's the open they done a double guess the same thing they have all the data there and there's a stack API and those that illuminate you for put it up there air search I think it's the name that search for all the data juice on AWS but that in practice can do search on other dashboards as well if they are stack I think that the key part that we are able to do that is the STAC definition that enables us to search any catalog if it is a stack no matter where they are sitting but I top of my head I don't remember anything similar to that other than the data being replicated really behind API

Lauren: we also work we've got a team of folks that work on that project called the **living atlas** which is very similar and that it brings together a lot of datasets a mixture of raster satellite imagery kind of stuff but also lots of socio-economic demographic infrastructure kind of data and we kind of just act as the **steward** of that data and largely it's datasets that the actual agencies or organizations that own that that authoritative data provide. We make sure it kind of works together and is easily accessible to anyone within the platform so that's the living atlas which is kind of our answer to that because it is. We want to make it less than 70% of the time people spend on these analytical projects as much as we can

Kumar: I can comment on that from a Maxar perspective space net I think I'm not somebody mention here we one of the things we are doing is we're making the data open data so the licensing is so you can actually do whatever we want to do but the way I look at this is this is similar to face recognition right somebody created database and took after 10 years when it actually the algorithm started working I think that mentality is slowly coming around but I don't think there's one common repository today so

Wendy: so I work for the federal government and Ed here also sorry it works for the federal government so one of the things I'm you know realize is well first of all the data that we collect is a is a public good so I mean it's out there because you know we're collecting it for the public so should be available as much as we can as long as we protect privacy and all that other stuff but I've just it's really hard to know what data sets are out there and so and I know there's been efforts in the past like **data.gov** so forth to kind of be sort of a hub I guess or a clearing place I don't know for for data sets like that but just because you're right if you don't have the data then you're really gonna have trouble building your models unless it's physics pace models maybe but so I just think having something like that would be would be very useful so I just wondered if something like that exists and I just don't know about it but ya know

Todd: I think it's a big I think it's a big thing in a problem right now I mean even if the data does exist and it's public assuming people know where to find it just a case in point we've been working heavily for some time with census data right and yes you go to census I go you can get the tiger shape files but then when you want to try to join that to the metadata it's a laborious process that's very convoluted with the way the data split out between three or four different files and so one of the things we've been working on it's like hey we're working we have this data **we've got the whole pipeline let's open source that let's put it behind a there's something called an Intake which is like a Python open standard API for data cataloging** and let's actually put that out there on the web for anybody to use in common formats and so they can pull it in there Jupyter notebooks or whatever that's obviously one small piece of things but I'm just even key data sets that people know what they are sometimes they're very inaccessible you know much less together in one place. I would recommend **Harvard world map** I don't know if the project is still that alive however Center for Geographic analysis

Hamed: there's one more thing we are at Google so let's give a shout out to them there's **Google dataset search** so beyond the Google search they started a project I think maybe it knows more than me maybe two years ago that you can now search datasets specifically search for datasets but behind the scene you need to have that data with the similar schema so everything should be registered so it's getting there it's a it's a gradual process but that's another portal you can use okay

Question: so recently I joined the IEC JTC 1 and SC 41 is artificial intelligent the standardization organizations and then recently we have some issues **what is AI data so do you think there is a difference between the a idea and Big Data** if we yes how can we represent it of the a I'd address for example dictator we can explain the characteristic you three V right volume and velocity so maybe I think if we can define up to certain you know characteristic the AI data so that is my questions

Keith: I've really been looking forward to requirements from the SC 42 AI people I've I knew something about what the big data working group has been doing and to some extent the property graph work is coming out of some of the big data stuff but I really would like to see requirements

Kumar: I will comment as the moderator I think the line the definition of AI ML are blurred when we are talking about this. I think most of the time it's machine learning is what we are talking about **artificial intelligence from my definition** is when the machines start making decisions for you or making some intelligence inferences we are not even close to that for geospatial I think we're all talking machine language right now in that case the data that's the reason why I was asking the question about what is data versus knowledge there are schemes for representing knowledge as well so for machine learning that's how I look at it

Hamed: I think UC Berkeley professor wants use the terminology called **intelligent augmentation** IA and he was like what we have used ml for so far has been really augmenting human intelligence we haven't really built artificial intelligence in that sense so yes **we are really using machine learning not artificial intelligence** as a bigger thing in many cases I think Radian does that as well we use AI as a business term but really in engineering data science then we are using ML there's no there's no idea what I would see a idea versus or ml data versus big data is the **big data might be the more kind of unstructured unlabeled data** that you might get I mean the term was also used a lot in the social media kind of platforms. There are a lot of data coming in they are not necessarily ready for ml but there's data so that is big data **but ml you need really to have labels and inputs a more structured way you know getting into technical training tests** bullets so it isn't more kind of I would say higher level prepared data compared to just raw data coming in

Lauren: another thing that that I would add is that really machine learning is just this it's another way of solving problems they're just a it's really just a series of algorithms and tools and so I don't know that there's a special kind of data maybe since I've had three minutes to think about it but off the cuff what I think is I don't know that there's a special kind of data I mean unless there's a special kind of data for all analysis because I have a hard time with the idea that we're not just really just talking about analysis or now it's machine learning now it's data science and who knows what we'll call it five years from now **but we've been doing this which is trying to solve complex problems using tools** and so I think that all of the same kinds of rules about what kind of data we needed are there now there are some new kinds of data that are allowed in maybe that weren't allowed in before but I think really **it's about trying to solve problems** that though framing the questions and that sort of thing are some of the bigger challenges

1:30

Questions: I have a question for Ahmed on our spaceship actually my question is related to this so beta we have an ambition to build imagery data set for spatial analysis like I said okay what we really want is not just a data label data so I mean what label data tends to be very application specific right and what a label data looks like for I didn't I think water bodies. So yeah so one of the difficulties in **building a spatial image repository is it's not as simple as ImageNet.**  Imagenet it was like a very ambitious project I'm not trying to minimize it but it just rivet us identifying objects here is the chair here is a dog it is said of a car right here you got to be able to say what you want to do with the application as well it may still be a very good goal for this group but we are nowhere close to that no

Hamed: I agree definitely the type of data here is application or vertical is specific so I can give you two examples how we targeted those so we started with the land cover 1 and it was like okay we don't want to be the organization to define a taxonomy for land cover like I want to generate a global land cover training data but I'm not in position at land cover field is much older than my age right so I don't want to be in a position of defining that so we hosted a workshop of experts on **land cover and machine learning people** I said ok we are going to generate such a data we are investing in that we need your guidance what should be the definition of that data what should be the classes what should be the granularity of the data so we get community feedback to do data standardization the other one was an **agricultural** data so agriculture you can do image annotation you need to be on the ground collect data field boundaries potentially crop type crop yield some temporal information when was the for example planting when was the harvesting that data are been augmented with imagery to become the training data everybody who does that ground referencing does a different way they have different standards so one of the things we learn after one year project we had was we need to develop a community standard about ground referencing and we drafted that actually two months ago we took it with the CGIAR team who is the big agriculture group with 15 different offices across the globe we went to their big data convention we had a session getting community feedback is the security standard for doing this thing on the ground then we took it to Grand Challenges even in Ethiopia we had a discussion there next is a NASA workshop we have next is the ESA workshop you're going to have in April to get a community feedback and how can we do better ground referencing specifically for AG data so we can have a better training data we're hoping we can do more of that in domain-specific but it is it is an extensive process and we need to do

1:34

Question: I think this question is it's about **synthetic training data** and I'm curious to what extent you are currently encouraging ingestion of synthetic training data number one number two why don't all the standards begin with synthetic training data so that you can then move on to real and say this doesn't even serve the purpose of what we really need to ultimately do the analysis or achieve the you know the actual predictive behavior I mean

Hamed: personally I like synthetic data so the challenge I think as a community we haven't invested much in the generation of **synthetic data for training data applications** particularly we have a lot of imagery can we generate synthetic labels out of them so we can ingest it we are investigating them you're just getting into that process we haven't invested in it in the past but I haven't seen that much of effort in the community either but if there is data we are happy to ingest it because at the end of the day it's a good catalog data people will build on it and there will be a lot of lesson learned out of it but yeah that's that's the way of future we should try it not that we should not do the ground referencing and labeling but we should augment that so we can have more diverse data

Lauren: oh it's interesting because we work on building tools and **we often start validating whatever we've built using synthetic data** and I can't tell you how often we're so excited about something that we've built based on that synthetic data and then we throw it throw real data at it and we you know have to take the rest of the day off and come back with clear heads because we're very disappointed so I often wonder it's like why do we start here why we keep doing this to ourselves because at the end of the day it needs to work with real data and and it's messier and it's it's it doesn't meet these kind of it doesn't fit into these neat buckets that but of course it has to also work under the perfect conditions. So I guess it really is it depends on the problem like we've been working on a time series forecasting method and you know we create this synthetic data that makes sure that given this nice set of this **beautiful time series data sets where we've introduced noise** and there's a cycle that's obvious it works perfectly and then you throw real time series data at it which is nothing like that synthetic data and it's like well this doesn't work at all and actually it's these really complex methods work beautifully on that synthetic data but it's the **simplest methods that do the best job on the really complex data** and so we end up we got went down these very expensive roads and we end up back at methods that have been around for 25 years

Question: so kind of going back to this question of data sets and this is something that's really interesting I think to us agree because you know even like people doing real-time work like mission-critical work like weather are also looking at you know where are these their availability of curated data sets for historical data but a question I have to the people in the room and people in the panel is how much **has this community explored what was done in the** **astrophysics and astronomy** community you full disclosure that's where I come from but you know there's a lot of you know work that there's certain things that are easier certainly to do with Astrophysical data but you know being able to use location being able to use metadata that tells you how it was collected when it was collected having standardized time stamps and so I'm just wondering like kind of you know how has that worked in the location data community they

Hamed: I can comment on one difference between astrophysics and our board and that is the **human aspect** and that makes everything harder because I mean AGGA go back because that's my daily job now think about AG if it was in a non human word the patterns would be more defined right you would be able to explain them with your models or synthesize them more easily but when you have the human factor in there it makes it much more complex we are learning through. I think astrophysics has been very good in adopting a lot of these data science statistical approaches I think this is my perspective you know another organizational one **geospatial world is the way behind astrophysics in terms of adopting those but we are getting there the human aspect as just making it more challenging and also interesting at the same time**

Lauren: I guess I just say we hire a lot of astrophyics we actively seek those resumes I'm very excited when I get one comes across my desk with that note

Kumar: one of the things we are noticing is especially geospatial conferences 10 years ago it used to be few of us talking to each other now we see a lot of practitioners here which is very encouraging I was talking to somebody from health industry we never used to have them five years ago so bringing that expertise together basically makes it very valuable which goes back to this sessions title is about how do you prepare the data to actually do the analysis based on what I hear is we still got work to do but it's encouraging that with all the compute and other aspects we are looking at we can actually make this happen to the next session is how do you analyze this data

with that please join me in thanking the panelists and dr. Gill.

So this last portion of Session two what we're just finishing up and the like is to **encourage an active discussion at your tables** and so for the next half hour or so focused on your table take a break if you need it but come back and so at 310 come back and if representative from your tables each you know you want to like somebody or somebody is of you know a real specific thing when I will invite each table if they want to give you know just a comment as to what you discussed from the group if you haven't been heard this is now the time to make that happen so please talk at your table take a break we'll start up three 305

English (auto-generated)

* need for new **standards**?
	+ Need more for **raster**
	+ Use **STAC** as part of the OGC definition particularly the API
	+ **cloud native formats**
		- hdfn netcdf are not cloud friendly
		- ZARR new format there's also X array implementation of netcdf in Python
	+ it's about making a **really easy for people to go between these different platforms**
* **uptake of you know like types of models** that we're all building
	+ **the best data scientists is their hat there is an ability to tell**
	+ we have **R square of 0.6 that doesn't help the policymaker** right because the policymaker the typical way they do it is
	+ **transparency in the model building in many cases and the benchmarking**
	+ **tension between wanting the most powerful model and having a model that's more explainable or interpretive**
	+ **visualization** in allowing people to tell these stories
	+ a **human in the loop** still I can actually see the data and its full breadth and depth.
	+ **property graph technology is actually really good for allowing you to record the internal steps and the machine learning**
	+ **are we actually gaining knowledge because a prediction is not the same thing as knowledge**
	+ **what what's the difference between prediction and knowledge**
* **ML Hub and data sharing**
	+ **repositories for just geospatial data in genera**l
	+ Cloud hosting
	+ **living atlas**
	+ data open data so the licensing
	+ **data.gov**
	+ **we've got the whole pipeline let's open source that let's put it behind a there's something called an Intake which is like a Python open standard API for data cataloging**
	+ **Harvard world map**
	+ **Google dataset search**
* **what is AI data so do you think there is a difference between the a idea and Big Data**
	+ **artificial intelligence from my definition** is when the machines start making decisions for you
	+ **intelligent augmentation**
	+ **we are really using machine learning not artificial intelligence**
	+ **big data might be the more kind of unstructured unlabeled**
	+ **but ml you need really to have labels and inputs a more structured way you know getting into technical training tests**
	+ **but we've been doing this which is trying to solve complex problems using tools**
	+ **it's about trying to solve problems**
* **building a spatial image repository is it's not as simple as ImageNet**
	+ on **land cover and machine learning people**
	+ **agricultural** data
* **synthetic training data**
	+ **we often start validating whatever we've built using synthetic data** and I can't tell you how often we're so excited about something that we've built based on that synthetic data and then we throw it throw real data at it and we you know have to take the rest of the day off and come back with clear heads because we're very disappointed
	+ **beautiful time series data sets where we've introduced noise**
	+ **simplest methods that do the best job on the really complex data**
* **has this community explored what was done in the** **astrophysics and astronomy**
	+ **geospatial world is the way behind astrophysics in terms of adopting those but we are getting there the human aspect as just making it more challenging and also interesting at the same time**