# Transcript Day 1 video 1

# Introduction

## George Percivall

Welcome to location powers data science thanks to Ed Parsons for his introduction from Google and hosting us here and Nadine CEO of OGC for welcoming us. I'm George Percival I serve as the CTO of OGC we run this series of activities called location powers as summit a series of summits aimed at topics that are very relevant ripe trends that we think need some attention and perhaps some development based upon community consensus and the like so we've been doing this several years is probably about the tenth one or so I have to say it is the largest one I believe that we've done organizing committee Nadine mentioned that as well I'll put the names up as well as OGC staff who has contributed greatly to the development of the organization quite a few of these people are in the room and thank you all for the six eight months of development activities that have gotten us here today

So what are we here to talk about location powers data science **the explosive availability of data about every aspect of human activity and the ability to gather that information and analyze it and be able to put it to use in decision-making that makes us a better place to live** right so a **shift from data scarce to data rich lots of computing capability that we never had before cloud computing that exists in ways that we never you know imagined the decade ago even right and nearly all of that has location location and time as the key aspect** associated with it

so I make this regular contention that we are in **the golden age of geospatial data** science from a mythological perspective Golden Age was this garden in which all of the resources were readily available people were in a you know a happy garden if you will where resources were widely available data is widely available we're able to make use of it and live a splendid life I'll remind you that the Golden Age was Prometheus so where do we go what are we learn and what are we do with it this Golden Age that's what we're here to find out right figure out how we're gonna go from here

a diagram that make my screen there we go so we could argue about this **diagram from NIST but generally as in many innovations it's a convergence of multiple trends** statistician zwu argue they've been doing data science for a while president of American Statistical Association Wendy's here and will speak tomorrow the folks are to the cloud computing large scale computing alike will argue they were the the innovators and the like but we also know we're here for applications right that's what we're here for is to put this to you

so this kind of gets after that in some way and so it was with that kind of concepts in mind that the data that the committee put together a **outline of sessions** and session one will provide us the Foundation's looking at what are the fundamentals that are behind existing kind of geospatial data science and what are the driving trends for that session to is going to get into a little bit more the technology in order to do analytics you need good data representation the first step to really good analytics is is data representation so session two is titled analytics and representation but I'm learning how to use this on the learning machine right I may have some intelligence so session two analytics and representation we'll get into some really good math models some good discussion there about some really innovative trends with respect to data science and data analytics and then everybody's favorite topic at the end of the day AI and machine learning our chair of the geo AI working group is here along with quite a few folks that are leaders in the development of application of machine learning to geospatial we'll have a reception no host tomorrow night at steins down in on Castro Street in in Mountain View so we welcome you all to come down there tonight did I say tomorrow it is tonight time and space I got the space right I'm as important so tomorrow morning we turn towards applications and outcomes what does is make a difference with so how can we apply we've got a good panel of folks that are addressing multiple different applications of data science and how that can make you know better decision-making then actions to take at the end of the day tomorrow we want to be able to leave here with recommendations from you as to what should be done going forward with n o GC for sure but much broader than that as well the end of the day we'll have the Rapporteur reports the tell us what you tell them tell us what you told them and tell us what we heard kind of thing that also sets the stage towards the report that will come out of here so if those Rapporteur are not reporting back what you think you heard let them know because we want to capture that and get it into the report so that's what we have as an overview for the for the workshop that's the workshop for the plan for the next two days the five sessions and the reports so that ends kind of the opening and

**so now we turn to session 1 on Foundations** and so we need to shift towards first presentation and the like ED is our moderator for session 1 oh I'm the Rapporteur gonna be the first presenters any questions about the sessions no I mean the overall workshop generally obviously there's a lot of hi I'm on the line can you hear me we can great so I'm I'm having some real technical problems both to it's a computer being able to get in and I'm also having some problems with the phone is the presentation actually being presented through the meeting app it should be like let me have a look at it in a moment once we get the next presentation going on I'll check what's happening right okay thank you so much okay Adam Martin **reminded me that you can nominate a question and you can vote on a question** there so we're looking for you know what do you think are the important questions that should be asked and discussed during the workshop the link to so if you go to location powers now there will be the link to this meet so you can get that too so login if you want there and see the chat that's going on there it will be when we plug things in

# Session 1: Foundations/Motivations

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

Ed: Mark Armstrong for the University of Iowa I think he's going to give us just what we need which is the the foundation presentation here so we all I guess on the on the same starting point from the

perspective which I guess you know simplistically it's all about the data

I'm very happy to be here this morning I think George mentioned something about one some Greek person I'd like to refer to Sisyphus the dude with the rock right and I think that's kind of where I am rather than the Golden Age so I'm going to present a somewhat more pessimistic view of things I guess I want to do I also want to understand understand I'm gonna go pretty quickly here because I have a lot of slides and we're gonna get through them somehow that was expecting 45 minutes and I know that I'm behind 15 so we have to go faster

10:28

so what's the problem I think we've been describing some of this in the distant past we had software we had processors we had static data for the most part and then software improvements took place over time performance in computing steadily improved over decades and we our datasets grew in size and complexity **but now what's going on is we've got a real problem because the data is outrunning our capabilities our software** has moved forward but one of the main factors in addition to the onslaught of data is that Moore's law has been repealed and so we're seeing a lack of capability in terms of keeping up with computing performance so that's the motivation here is to try to reconcile these things

what I'm going to do is **describe some trends in big data and particularly on fast data which is data in motion illustrates some computational complexity of geospatial analytics this is even more important because of the repeal of Moore's law** and not that it actually is a law but it certainly talked about a lot and then suggest some strategies that we might pursue to overcome these problems

there are some clear roots in this area going back to temporal GIS and **dynamic GIS and I'm pleased to see that Kathleen Stewart** is one of our panelists and giving a shout out to **Kathleen's book** here it's old but it's relevant but we have a **fast streaming data** that is orders of magnitude and change and so this is where I think we are now we have fast data streams we have computational complexity we have the end Moore's Law and now we are going to spend weeks processing things and so what I would like to do is suggest a different path one in which we are going to use experimental architectures develop new methods in order to try to get things back into control so that we can do analysis of large data and fast data and time frames that would enable us to use the techniques in the frame of a meeting for example

I think we referred earlier to the **National Institute of Standards and Technology they have defined Big Data and the they also have defined the the v's of big data and I'm going to focus clearly on the velocity aspect** of it but there are lots of other fees that have been developed I actually went on the web and found one definition that had I think 13 different V's someone went to the thesaurus and did a very good job of making up these so every V word is Big Data the Big Data is also having an effect on research and so we have t**he very famous book that came out of Microsoft that is describing the fourth paradigm of research which is data intensive discovery** and I also want you to and that what I'll be talking about is **Internet of Things because that's a big enabler but there's a lot of hype** here surrounding the Internet of Things depending on where you look you can find estimates that are perfectly mind-boggling and they're I don't know exaggeration I want to say just the things that I've looked I've highlighted here the Bain and company says revenue is set to exceed four hundred and fifty billion by 2020 and then Gartner which is famous for hype is only saying twenty billion so there's a big discrepancy here disparity but the Gartner is self-aware and they talk about the hype cycle and that's what's the top of that illustration and a lot of the technologies it's very small relating to the Internet of Things and big data are in that graph so

when we start talking about the Internet of Things and particularly **the geospatial portion of the Internet of Things** satellites play a key role we have terrestrial devices of course fixed and mobile we have aerial UAV kinds of devices that are flying around all over the place preventing forest fires from being put out and so on so forth but satellites have really changed the game in a certain sense because they are enabling technology and they're gathering data so they have two roles to play here that are very important I'll talk just for a few minutes about the satellites I read space news which is kind of a strange publication dealing with satellites and and the like and if you look at there's a lot of hype there too but if you look at the constellations that aren't being planned by a lot of different companies including Amazon and the SpaceX version SpaceX is going to put up they're planning to put up according to the sources that I read 12,000 satellites so the sum of those is **16,000 new satellites that are going to be put up and next few years** that is mind-boggling and so what is proposed is that there's going to be some failures and these will be both economic and kinetic and if you've looked at the satellite literature you know about the Kessler syndrome and that's the debris cascade so one satellite runs into another it generates a hundred thousand particles fragments and that hits other satellites and that hits other satellites and pretty soon you have a bunch of junk circulating the earth and that's sort of **the nightmare syndrome the Kessler syndrome** and so that is a problem there's already been one major catastrophic collision with let's see I'm trying to remember if it's iridium and I think it was either a Russian or Soviet depending on when it was put up satellite and that sorry - oh yeah okay but there have been some big ones and they've generated a lot of space drunk

16:50

**imagery** is the other aspect here again there are plans to put up a bunch of satellites and one of the major things that it's truly remarkable when I first started looking at remote sensing was back in Earth's that's Earth's was before Landsat they turn earth into Landsat they didn't actually they just changed the name but the **revisit was around eighteen days** we thought that was fantastic you could get imagery every eighteen days now you can see that **black sky is proposing 40 to 70 times each day** so this is just truly amazing in terms of what's going on here **in addition to the static imagery there's a lot of streaming video** that's going to be provided as well so there are really really important things that are underway and in terms of what is underway there's a lot of consolidation that's taking place in the space as well I have to have a sort of a running scorecard here but things are being acquired and merged and so on and so forth so there's a lot of activity here in in space now we'll go back to earth

18:00

so I talked earlier about the **Internet of Things** and some new data ecosystems that are being created and I suggested a few minutes ago that some of the sensors are fixed there's talk about smart refrigerators refrigerators are mostly fixed electric meters are certainly in play here in terms of trying to monitor power usage at businesses and homes other devices are mobile and they have radios and that's where the satellites come in as well as Wi-Fi and other communication technology and so we're finding is that there are **new data ecosystems that are being created and OGC is certainly contributing to support of that** if we think about data again when I was coming up as a graduate student and maybe junior assistant professor I was over on the right side of this graph though it's **data at rest** and you can think about census data as being a early precursor of big data census data used to be stored on nine track tapes and it was just big slow data and you'd have to wait an hour to get a tape mounted and you know so on so forth using IBM JCL on all kinds of scary stuff **now we have fast data where the value of an individual data item is very high in in the when it's captured and it degrades over time** and then usually these things are aggregated and that's where the aggregation comes in and data at rest is comes into play so we're getting these streaming data where we have high value with fast data but

**this poses a problem to analytics in terms of standard kinds of GIS analytics we when we have fast data** we don't know what the sample size is it's effectively unlimited we have non-stationarity that is changes in the moments of distributions that are at odds with what we would like to expect and we also have algorithmic space and time complexity and so we have in the bottom here an arrow going from left to right that is illustrating computational complexity of different kinds of analyses I implemented a screaming data application last year that was published in cages I was published in January I believe Volume one this year in which we were just testing for exceedance so if a value in a data stream was exceeding a minimum threshold that's easy right it's just greater than are greater than or equal to but when you start getting into other kinds of analytics that's where we run into real problems

so just one a few examples I should say P median problem is a combinatorial optimization problem where you select P facility locations from n demand sites if you do a brute force analysis it just explodes combinatorially but we do have heuristics that can be used to address the problem but even those heuristics are space and time complex GID the GI star statistic is order N3 + cube sorry and then creaking depending on the kind of **Kriging that you're doing can be up to order into the fourth power so computational and complex**

so what are **some new strategies** I'm not going to spend a lot of time here I'm going to introduce a couple of ideas that were new to me at least **the first is reservoir sampling second is approximate computing**

**reservoir sampling** is a rather straightforward approach I think one way to deal with fast data would be to just throw out every other observation or something like that but that can lead to bias and that's called load shedding so reservoir sampling is an attempt to reduce load shed bias and basically what we're looking at are elements in a stream that are coming in and the location their temporal index is evaluated and in terms of their location in the stream and there's a probability distribution that we can drive in order to determine whether something will be sampled or not and then there's another modified approach that I've run into that increases the likelihood that more recently added observations or will be included in the reservoir that is recent data is more salient than older observations the downside is that reservoir sampling actually requires computation or to be achieved and that's always a problem

**approximate computing is heresy** so I'm here to be a heretic the premise here is that some analyses do not require absolute correctness and that is true we're trying to here to exploit the gap between the accuracy required by users which is sometimes not that much and that provided by the computing system which often is great and so this can reduce Starring requirement Sarge requirements and improve performance and and also it's being used to save energy and in terms of efficiency energy is sort of one of the new things that is being very seriously considered in terms of in addition to time and space complexity this is not a new idea we've used lossy algorithms for a long time to encode pictures and music to some people that's in a thermo particularly the music part of it but we often have been aggregated data and that's an approximation so if you're interested in learning more about approximate computing and the heresy that is involved there's a very good article in computing surveys that's referred at the bottom in 2016 written by matal and it's very thorough and very good

**(Parallel Perspective: Copy multiple streams)** there's a different perspective that I'd like to think about that is if we have a stream we can actually copy the streams and have different processes operate on these copies I assume here that storage is infinite or free or both so we can make multiple copies and we can process these different streams in different ways in order to achieve different objectives so this is one thing to be thinking about we can have one stream that would be just dealing with exceedences and things like that would that would be processed in real-time or near real-time and then a heavy duty processing that would be done sort of offline or in background mode so it's a different way of thinking about it

24:35

so we have a problem what are we going to do about it well going back again to our Greek friends we have the need to have a longer lever and a place to stand now this will provide some leverage this is the **Aurora machine** this is a stylized representation of Aurora which is a project that is being funded by the **US Department of Energy it will be located at Argonne National Lab** to the tune of 500 million dollars so this is an enormous investment it's **one of three exascale systems that are going to be developed** these are **called coral systems** a collaboration of Oakridge Argonne and Livermore the idea of exaflop or exascale computing is a very very large number as you can see I think I copied my zeros correctly I sort of get zero fatigue after a while but these things are crazy fast right I can't wrap my brain around that

(Race to Exascale CS&E) and it's not just the USA this is an **international competition for supremacy** China actually has three parallel there's a pond there parallel programs that are underway and they're going to pick probably the best but they're certainly throwing a lot of money at this and then Japan and Europe the European consortium are also in this game this is the cover from as you can see computing and science and engineering from earlier this year but there's a real problem here with these these systems that are being developed there's a bazillion cores and they've got you know all kinds of things but **having software that will actually exploit this parallelism is a difficult challenge** and then there's actually talk it's talk now but you know it seems to be inevitable that **Zetascale it's 10 to the 21st computing is in the is in the offing** so there's crazy crazy stuff but it's very very interesting to watch and it's going to happen

now there are problems so we have large expensive systems with limited access so at the bottom here there's a letter I get the National Science Foundation news feed every morning in my inbox and this is an announcement that you can apply for time on the Frontera machine at the University of Texas I believe that's where it is but **you have to have a long time frame here in order to get by time on these machines** and so that precludes any ad hoc kinds of research and in use in in ways that are unanticipated in the future so we've got a plan to use these machines **the energy requirements are enormous** or enough to run a small city just one of these large machines and what we're finding is that although they're throwing more and more nodes at the problem each node is affected by Moore's law which I'll talk about in a few minutes I mentioned earlier but more in more detail in a second and then I mentioned also the practical scalability of these things

27:50

so **there's a real tension underway (between centralized and distributed computing)** there is the what I call big iron that goes back to the IBM 360 that I showed on the last slide that is centralized computing resources and then what is not known as horizontal scaling which is distributed and there are also experimental architectures and something called heterogeneous computing that I'll talk about in a minute so that's a hint had things to come

**cyber infrastructure and cyber GIS are very important NSF is supporting this they're all in they have an office of advanced cyber infrastructure** and the basic ideas there's the use and development of middleware to connect computing storage instruments and people together and support the interoperability of these different components and in the geospatial realm the preeminent as far as I'm concerned **exemplar here is the cyber GIS project at the University of Illinois at urbana-champaign** does that is spearheaded by Sean long and I think Anand is here as well shout out to him the

**cloud computing** we heard from is a different distributed model it's it's very effective for some applications it's **on demand elastic heterogeneous measured service** so you can get you can buy cloud services from Amazon Google and others and there are certainly private clouds and so on and people have demonstrated the effectiveness effectiveness of cloud computing in the geospatial realm there's two citations there **but there's a real problem with cloud computing and that is latency** computing if it was a communication if it were perfect would be at the speed of light but even the speed of light is too slow so but we have switches we have the problems with media and we have distance so distance is a problem and then this is a **real issue for applications that require real-time response and I'm referring here to two examples of augmented reality and autonomous vehicle control** if you have ever used a VR system where there's latency it is very very very disorienting I've done it and I make yourself sick pretty easily so I make myself sick pretty easily anyway but and any in any event it is really it affects your inner ear big-time

30:30

**so there's a compromise that it's been suggested edge and fog computing are two ways to put processing in between Internet of Things devices and the cloud** in order to cut down on the amount of data that's transformed or transferred I should say and improve the performance of the networks so it **turns out that geography really does matter here** and the little picture on the bottom is a sort of a large credit card processor that Nvidia is developing that is execute at 21 trillion operations per second so that's an amazing device that they are using to implement the edge devices or fog and fog devices

now I've mentioned Moore's law so **Moore's law has been a big deal for a long time** it was sort of how BestBuy made their fortune every week best buy ads would advertise more gigahertz right for the same price and so that was a thing back you know a decade or so ago but that's been repealed it's not just me saying that and there's also the idea of **Amdahl's law which is limiting the number of cores that can be used effectively in any application** so people have begun to look at **alternative architectures that can be used to boost performance** and this is where things get kind of icky is **we have to get closer to the machine**

so what does Moore's law well transistor density doubled approximately every three years in accordance with Dennard scaling as this density increased power consumption per unit area remained near constant which enabled clock speeds to increase steadily over the over time until around 2007 we began to fail and then the solution was well we need more cores so more cores happened but problems with **more cores are that we had speculative exploitation** that didn't work out was wasted processing and excessive power consumption which has led to dark silicon where there's selective power reduction that is enabled to shut down some of the cores so this is becoming a real problem we are the gigahertz has slowed and the number of cores that are on each chip are sort of limited and

33:00

this is a **famous diagram from Hennessy and Patterson** this is actually out of their article not from their book but these are the two I don't know what to call them they're the gods of computer architecture they're both in the Silicon Valley and you can see how **the curve is flattening as we move past 2015** and so this is not me making this up this is a real thing now Hennessey and Patterson also were awarded the touring lecture a word redundant but they gave the Turing lecture and they have a few suggestions that I think are worth paying attention to because they're really smart people so one is to rewrite programs for efficiency we can probably do some of that but the other is to wring performance out of existing and experimental technologies

so rewriting software they cite an example where there's a **matrix multiply code where it was written in Python and they rewrote it in C to improve cache hits and exploit multiple cores and they got good performance increase** being as old as I am I've said wait a minute I've seen this before so I went back and looked at an article that Bruce mcdougal wrote in 1984 where he cited some interpolation software that was written and interpreted basic that took seven hundred and forty seven minutes to execute this was a really small program or example something like 80 by 60 or something computing with like 15 nodes or something data points but he recompile didn't see and it got down to 5.9 minutes this was running on a Z log z80 processor so these are scary things to think about so programming intervention can result in performance improvements but production code is really normally pretty optimized so I don't know where we're going to get it in terms of our software systems in terms of our research systems we can probably all do better jobs at bringing out some performance but they also point to domain-specific languages and this is where I think we can make some hay and there could be **geospatial domain-specific languages** we have kind of languages and map algebra and things like that and map calculus I think that has been promote proposed but certainly we can do better I think in this area

**in terms of domain-specific computing** they point to some historical improvements that we were using so if you think back to the 8086 8086 processors for those chips and there were GPUs back then again I'm old enough to recall that the graphics processors the original graphics processors were graphics boxes and then they got reduced down to cards and now I understand that they're back to boxes again but you know things go around in circles I guess but the key here is to map an application or a problem class to a particular architecture in the example here is a that that **Hennessy and Patterson site is the tensor processing unit that is used for Google for their neural net applications that's optimized outperform GPUs and CPUs** what do they do they strip it down they remove things that aren't needed they added a large two-dimensional multiply unit because that's what you're doing when you're doing these neural net applications and then they don't need a lot of precision here so they're using 8-bit integers so again this is limiting the amount of precision that you might expect to use

36:45

there are other architectures that are somewhat similar there's the **Nvidia has a a device graphics card that uses tensor cores** this is an article I'm citing from last year and computing and Sciences and engineering but a citing that that you can get an improvement of 16 times faster than normal computation in part because they're specialized processor that's using limited precision arithmetic so the idea there is that we can improve performance substantially but that's for a limited class of problems basically matrix matrix multiplications so a single architecture may not be the best choice

I mentioned **heterogeneity** this has its origins in the 1980s there was a person at Purdue who pioneered a lot of this same was Howard J Siegel and he was concerned predominantly with using SIMD and MIMD kinds of architectures this has come to the fore again there's a brand **new book that's public by the Association for Computing Machinery written by Mohamed Zahran** that I've just recently obtained I haven't really had a time to digest it in any detail but he goes into a particular kind of hardware called a field programmable gate array which is really a soft computer thing where you program the in Hardware sort of a low-level Hardware after the manufacturer these devices compute consume less power than CPUs or GPUs and there's a an example here is the Intel Stratos 10 DX as opposed to the Intel Stratos 10 which is an old product

this is a a an example diagram describing a **heterogeneous system** also provided by same author in an article that was written in ACM queue describing how different architectures can be used and there's you can see multi-core GPU and so on with FF FPGA there's a unified memory and storage and there's high speed interconnect

39:05

so is there a **heterogeneous example well yes I think so some parts of geospatial problems** are well suited to SIMD and others are suited to MIMD and we can develop spatial middleware that would align the characteristics of geospatial algorithms to particular types of Hardware environments

and I will assert that I've done this before in 1994 I wrote an article with Paul Densham which we sketched out steps to implement but did not implement I want to be very clear here this was a think paper not an implementation paper to deal with the heterogeneous vertex substitution P median model I looked at that paper and I said wow there's some interesting things here I think I'll update it and so I matched the different kinds of analyses that are required in order to do one implementation of the P median and that requires the computation of a lot of shortest paths through the network then we create the strings we do what is called a an edit which truncates the strings and then the actual teats and Bart algorithm that is used as a vertex substitutions step and then we would visualize a result and so we could have GPUs and MIND architectures involved in this kind of application so I think there's scope here for other kinds of **processing to take place that would use multiple kinds of architectures**

is this completely crazy? I submit no there is something called the **HSA foundation this is the heterogeneous systems architecture** I believe that's what HSA stands for is founded by some pretty major hardware vendors AMD ARM Samsung TI and others and then they've developed something called live or **LibHSA** which is compliant with their parallel data model and the goal here is to make it easier to code heterogeneous programs and **they're using high-level pipe like Python and abstract the low-level details to simplify the user view so you don't have to get into all the icky details**

41:30

**a word about machine learning** and I realized that I've reached my limit here in terms of time but I'll try to fumble forward other people are going to talk about this so I'm just going to brush lightly on it but I want to acknowledge that it is a big deal I'm not trying to downplay its importance but there are problems with machine learning and one solution to the problem there's a neuromorphic computing so on the left-hand side of the the figure there is the idea of Gihon NCN and that's convoluted neural networks and deep neural networks and so we have Nvidia which has the volta card this is not a Volta card but it has tensor cores the TPU has tensor cars they're very efficient at supporting neural networks but the **problem with with neural networks is that they're very brittle** if you throw something novel at it they break and so they don't know I don't know what to do with that so **neuromorphic computing is unproven but it's less brittle** and there's a lot of money being spent on neuromorphic computing intel is spending a lot of money on it and they're developing a very large system that is one with a hundred million neurons by next year and IBM has what is called the TrueNorth project that is also very well underway so there's a lot of action in the space by major players

42:54

lastly the quant what I call the **quantum quandary** so this has the potential to completely disrupt everything it's sometimes called a game changer and there are certainly applications that could be used here such as optimization and machine learning when I took physics I didn't study any of this stuff so I was a Newton guy so and he wasn't Greek either so don't ask me to explain this I know it very superficially don't know if I get into the weeds I'll turn myself inside out don't make me go there but the basic idea is that **if we have a fully functional quantum computer this is going to break the internet because they can factor integers and decrypt public key encoded messages** and that would just break everything and that's a big concern for security and the **United States is spending 1.25 billion on the National quantum initiative Act** that's intended to spur research and development in this area and these national research centers are at some of the National Labs which indicates the strategic importance of this activity

44:10

there has been a recent **google gambit on quantum** and there's been **a Nature paper** that was published I think three weeks ago now in which they are **claiming quantum supremacy** and so this is a very big deal their system was 53 qubits which is not that big but it can assume the two to the 53 to the 50 third state switches nine quadrillion states and the example that they use which was very compute-intensive but not data-intensive took three minutes and they estimated it would take ten thousand years on the Oakridge the highest performance of bridge computer at currently available IBM said hey wait a second we're working on this stuff and we don't think that it's right we think it only would take 2.5 days so even so 2.5 days is a lot and **what's being thrown around here is that this is Kitty Hawk all right so the Wright brothers didn't fly us here right we got on jets that were very well realized fully realized not some rinky-dink thing so this is a rinky-dink thing but it is important milestone** but there are large barriers still to come when I start talking about these barriers at Google headquarters I feel as though I'm about to receive a very powerful electric shock please please don't taze me so there are these are smart people that are writing about this the will go backwards here the op-ed in the New York Times on October 30th was written by one of the reviewers of this article the reviewer of the nature article he outed himself in the op-ed but he was also one of the authors of this National Academy of Science report that was issued last summer and so they are saying there are **lots of real problems to getting a fully functional quantum system** one is noise because quantum machines cannot reject noise unlike normal computers so the noise is confused with an actual state of the of the system the there's a lag in terms of converting real data into data that can be used in quantum systems and that's really a problem with geospatial big data we have to have an entire new software stack I don't know anything about writing software for quantum machines but I am sure that it's nothing like what we're used to debugging is also difficult because if you want to get a intermediate stage the superposition collapses and thing halts and they're also it's talked about the lack of an integrated system or a circuit virtual cycle not being present and that is the sort of feedback between market and research and development that drove developments and Intel and others for years and years the quantum initiative Act is intended to jumpstart this but IBM and Google they don't need this money alright so there is moving forward on their own without the need for government money

47:20

so I've reached the end of my rope so to speak so here are some **concluding thoughts (Two slides with bullets)**  big data is being created at weight in ways that we I certainly never expected to see methods are computationally complex and they're not really able to be applied to these new data sources we need to develop fast data analytics we currently don't have good methods in this space Moore's Law is no longer a thing and that's a problem the x-axis scale systems are coming but they have issues that I outlined now in terms of moving forward there are new programming models that match heterogeneous architectures to model parts and that's really the idea of decomposing software into into the chunks that can be processed on doing and using different architectures but we need a lot of work to make this happen quantum computing is a known unknown at the present but the National Academy is saying maybe not so much in the near future but certainly it's being accepted as the future thing and that the Google claim is but a harbinger and so there are I think lots of opportunities available to us in the future and

I want to assure you that no children were harmed during the preparation of this material and that's the end thank you very much

[Applause]

Q&A

Ed: I think we've got time for a couple of questions I want the microphone here so people online hopefully should be out of here us so maybe time for two questions for for mark

Question: so the heterogeneous architecture portion them different modular like interesting you also mentioned that there was already some development of writing Python that would allow you to make use of those headers that mean that to hide that complexity from people that are writing say geospatial algorithms in Python do you see some development of that

Armstrong: I think there certainly Han okay yeah I think there's opportunities there I mean **python is a de facto language for developers to be using these sort of at a high level and so if they can make use of this this LibHSA or whatever it's called that certainly is an opportunity** I don't know how well developed it is I was just grazing at the surface I'd never written anything in and don't know much about it Thank You

Question from Jeremy Morley from Ordnance Survey: I think an interesting kind of pair of things that could be put together in in the things you've talked about here is in the space of heterogeneity is both the heterogeneity of the computing but also where it's happening as well you know that kind of **edge fog core** kind of computing and where do you put different types of processing as well as that the computer itself I think would be interesting for us to model because a lot of our real world applications will be distributed in that way and that you want some of it concentrated as well it's a data centers

Armstrong: I agree **the problem of course is latency** when you're dealing with the network and so you know that has to be adjusted and has to be you have to be well aware of what's going on with it but yes

51:00

Question: Thanks for the presentation was really very enlightening you mentioned a number of challenges with quantum computing weren't all of those challenges with classical computing once upon a time as well

Armstrong: yes that's why people are saying this is kitty hawk and that there is a future I'm not suggesting that it's not real but there's a big difference between the Wright brothers and a 747 so I think there's a lot of work that needs to be done I didn't fly a 747 but the first time I got on a 747 I I was sure that this plane will never take off it was so big yeah there's a lot of room for improvement that's why I ended with there's opportunities here and there I think there are really interesting opportunities as well

Question: I'm **Aarthi Garg from Cray** so first **I'm actually happy to talk to anyone about some of the heterogeneous computing** and I'm coding paradigms but a question for you sort of at the beginning you talked about all of the new satellites and the new data streams are coming from there and some of the commercial ground-based data I'm also curious how you think about some of the other government data like stream gauges all these other earth observation data that I think are relevant but you know I don't know how well-connected that stuff that's right yeah

Armstrong: a lot of it is Wireless and there there's I know a stream gauge that's very close to my house that has a satellite uplink so I mean there's definitely stuff going on and it's all aimed in real time so I think **everything in the future everything will have a sensor and a radio right and they'll be locatable** I mean there's no question about it it's that's where it's headed and I just see it very clearly and maybe I'm crazy but so I think yes

Question: so there's the technology side and there's the algorithm side and I think decided that a lot of us are on **then there's also the side of kind of the analysts** at you know every city and every county in every state and how I guess how what are the challenges that are gonna manifest for them and how do we because ultimately we can build this stuff but how is it going to end up in the hands of someone who isn't a quantum physicist that we've built and yeah

Armstrong: so what I'm talking about is really where the research is going yeah not where the user community will be I think that certainly there is very **great scope for commercial vendors to come in and implement a lot of this stuff** and it would just become then routine to them they wouldn't have to worry about it they just do their **work so the user interface why not even change** to them if stuff could be going on on the hood maybe maybe not my crystal ball is a little cloudy on them yeah

Question: Todd Simon with Geospatial Alpha or an investor and the mentor program the question about **satellite based edge computing** and I guess the summation of my question is going to be is it a fool's errand to pursue given that you know your chart about fast data slow data for calling slow data **it's really about fast data everything that's captured from Earth Observation remote sensing altitude has a latency that can't be overcome** is efforts put into now and in the future for edge computing a fool's errand

Armstrong: I think that there is **streaming video that's coming down I don't know what its latency** is I was under the impression that it wasn't there wasn't horrible latency but I could be wrong I'm willing to stand corrected on that but streaming video you know it's it's it's happening it's a thing and the satellites are in low Earth orbit predominantly because of that because if you think about the time that it takes for a signal to hit a geostationary it's far lower in terms of latency to hit a low Earth orbit and that's why they're only up like 200 kilometers or less in some instances and so latency is not as big a deal there I think it's something on your forty microseconds versus making this up six hundred microseconds which you know I don't know it's right it's relative but I think it's it's **much lower latency in the low Earth orbit** and then of course there are these tethered and hovering things and then there's UAVs that are streaming so i think sensing imagery is certainly an important component of all of this

Ed: great well thank you so much for that he was excellent start for the day

## Presentation on Foundations - Title: Scaling machine learning to handle visual data will result in more powerful AI

Nils Lahr

56:24

Nils will present over thick a very different perspective but one that's probably actually really a useful segue on from that last presentation where this is a real application of data science to a problem space that I'm supposed to slightly familiar with but I'm from the wrong side of the Atlantic you know we don't really do basketball I'm more of a rugby guy but I'm sure the same technology can be used

yes hello I was thinking about what to say whether it's gonna be technology platform otherwise and I figured you know what I do which kind of makes sense to me is go through the arc of where I've come from I'm a mathematician by background I worked for the Air Force on supercomputing for targeting which you know had to do with location kind of **when I was invited I was thinking what it you know what exactly have I done that to do a location and and I kind of went back through and I realized like everything and so it was a little bit of an epiphany for me** to go back and so I wanted to do a little bit of that journey for about five minutes and talk talk about that and then and then look forward and in terms of what am i doing now but in a nutshell **I'm on my eighth startup company that is surrounding mathematics analytics video and otherwise** and all of them are trying to deal with these issues in different ways and and the newest company is OrionSystems which is the culmination of a lot of the things that you'll see here and and we'll talk a little bit about that

so agenda essentially what

58:10

I just talked about a little bit of a flashback and this is kind of cool for especially where we're at right now a couple exits up on 101 is and I'm better at walking around is there I'm a walker [Music] so here's East Palo Alto maybe some of you know it some of you don't it turns out to be one of the most dangerous places in the world in the 80s when I was in high school every day we would hear gunshots and my father you've worked for the US Geological Survey also in terms of earthquakes in geolocation and things like that and so he **took his knowledge from locating earthquakes and applied it to locating gunshots so that's actually me and my dad right there figuring out where gunshots were using seismic equipment that we duct-taped to people's chimneys** and essentially we designed something that would actually locate all these gunshots and we became for about three years we became Grand Central for the Palo Alto police we had police stationed in our you know radios and everything you know and in my bedroom and we'd call it all nights all hours of the day and we really tried to figure this out now if you go here interestingly this used to be called Whiskey Gulch and it was it was and credit it's like literally there was like four inches thick on the on the drive thrus four gunshots you know for protecting people now it's like a Four Seasons and IKEA so times have changed but this is kind of where I started in in thinking about these things and interestingly there's now a company our neighbor quit his job and said what a great idea and started a company called **ShotSpotter** awful lot of this so they're trying to continue that work

5:45

so then we kind of move into the into the 90s and was **working on analog to digital video so this is me at** **CNN** saying you know analog isn't good enough in satellite how do we get a digital you know system out of this again using my mathematics and how does that have to do with location interesting Lee I was thinking about that one of the things that was the most important is although we got video online we had to figure out where people were coming from and so there's **a huge effort in the 90s to map every IP address you know back to the geolocation and figure out then and then we could take that geolocation say if there any spell out though than their demographics or XY and Z and then you can figure out what your audience is** and by your audience you can tweak what you're broadcasting and now that's it that's a huge endeavor obviously now video wasn't good enough back then so I left and I went to Stanford University used my mathematics joined a company called V extreme we wrote what is now demonstrable a a decent portion of what's **h.264** we ended up at Microsoft as founding members of the **Windows Media Group** so I ended up spending a good portion of the early 2000s trying to scale things like the Olympics the various broadcast groups we converted most of the satellite components and so I was going a lot with locations in that but it was all about his **moving video and then figuring out where the video went** was that part of this kind of journey and it was it was really hard as a really difficult thing but it it was fun

but then kind of looking ahead I was getting all these calls being a video expert saying hey I know you're moving video around really well but **what's in the video matters to us** and so kind of in the in the 2010 you know eight to ten timeframe a little bit before I was it kind of it didn't occur to me I was literally getting calls saying I have all this video and I need I need to know what's in it and so kind of looking ahead I'll talk a little bit about that attacked looking ahead I think the previous intro was perfect right I mean **we're in a sensor overload I think there are now stadiums that have somewhere near 400 to 500 cameras** that are actually circling the entire stadium every you know **every athlete might have an RFID on them** it's pretty crazy so and I'll go a little bit forward and back I guess here so one of the

one of the solutions offered I think that was just discussed was ml so ML is a great way to say look we've got all this data there's certainly enough information in this data that if **we can create machine learning algorithms to parse that data it kind of gives us superpowers** the problem is it's very brittle and those superpowers when they break down they break down really fantastically they don't just break down a little bit and and **when you're talking about data science in terms of guesswork or actionable data** there's a huge difference so in Google's kind of search they're guessing you type in I want a red car it gives you some red cars in the government you type in red car you want all red cars you know in sports you type in how many times there was a shot you know based on the certain defense you want all the shots you don't want some you don't want maybe **you want to know very specifically the analytics around that so that you can make future decisions and be informed** so there's a lot of challenges with this not only are those **all the challenges but on top of that the languages we have the architectures we have are generating this huge barrier** which is you know it's pretty bleak I think it's actually worse than this in the video world they're saying ten percent of funded machine learning projects right now never you know art I see the day of light that means **ninety percent of all the research all the information that's going into this never even makes it into production because there's no production networks that can really handle it at scale** so we have a lot of challenges ahead of us

63:45

another challenge this is a military one but interestingly this is the same infrastructure of almost any corporate you know 50,000 so it's complex **it's not it's not only in the cloud it's not only on the edge not only in the middle** we talked about you know last five years we've been talking about cloud and all of a sudden we're talking about edge and we're kind of vibrating between these two well it's it's **it's a continuum between all of this** and interestingly I've you know I've been visiting Disney for example and this is **a Disney park** and and the edge actually is you know Splash Mountain and interestingly I you know it was amazing in the tour underneath the avatar ride is a data center and it's not a small data center it's a 10,000 square foot rack you know data center that's powering this sucker so it you know the **every one of the ride elements is the edge** and the data center which is five floors down is a whole nother in a segment of this challenge and then t**hat data center interesting leads to another bigger data center that runs the park and then that data center is actually been connected to all the other data centers around the world** and all of those are then connected by fibre to AWS so so we've arrived at this and the data is going through it but the intelligence and knowledge is is near zero it's it's it's operating the park but it is certainly not telling us what to do how to make decisions and there in the amount of analytics is very little if not kind of essentially non-existent

65:20

so my latest company's **Orion systems** it's the culmination of ok I had one company which was trying to get video places I had another one which is trying to figure out how to compress the company I had another one was actually a satellite company trying to get it over satellite and I realized going back to the Olympics the Olympic Committee came to us while we're at Microsoft and said hey the CODEC and this is a little bit about optimization another good segue into that the codec itself **you couldn't run the codec four times and tell it to create an iframe at exactly the same moment in time** and so we had to go back and fix the codec so that you can create essentially what is now multi bitrate that generated Netflix and Apple and all these other **smooth streaming technologies that see today would not exist if we didn't tweak the CODEC you know it given a couple months in order to create that iframe barrier so now you can you can fit every five to ten seconds** you can decide which fragment to get next and so you can be watching a low bitrate video feed and you can jump to a high bitrate but you're guaranteed that that iframe is is the border between those different bit rates the lights went on for me and it was like well **if you can tweak one thing in the codec and all of a sudden you create five billion dollar companies or you know multi-billion dollar companies what else could you tweak** in the infrastructure because all we were worried about was one-to-many but what if the problem wasn't at one of many especially when you're trying to figure out what's in the video the problem is more about how do you in how do you add video over the course of the over the course of the stream and then how do you scale that there's not a lot of places in the world that have content where you can play with and so we'll talk about that a little bit

there's an image here the first place I thought of is let's learn you k**now let's pick a content type that's not secret that has an abundance of it that I can play with** you know all these other things that was the NBA and I'm not a sports guy I think I fell asleep in my first baseball game with my dad so but I was I in talking to everybody about content you're like look basketball has a thousand three hundred games a year they have back-to-back it's the only sport where they get on an airplane they land and then they're basically getting into the next game and they need that t**hey need the intelligence from that game the night before in the next game** **and so that's real time enough and it's just the velocity and veracity** of that I was thinking well thousand three hundred games isn't too bad and so the I launched a company called Synergy sports and this is now you know ten plus years ago **there was no ML and other things and so what we decided to do was plug in computer plug in humans but plugged the humans into an algorithm so that the humans were actually doing work that essentially has become what's now called crowdsourcing** but but we were crowdsourcing a little bit of a different way we actually built algorithms that plugged expertise in at different levels so that one person can watch a low bitrate feed you know another perfect slide that you had so you're creating all these different feeds one person's we're watching low bitrate feed but that is fused with the courtside data another person is only looking at defensive another one only offensive all the way down **to we now have experts we have six thousand people who are trained and that are online every week** and all the way down to we have people who are experts at Paul Pierce in the game the individual person because they know his intent they know they know his body language before he's gonna take a shot so they're figuring out **their they're adding that that's something even ml can't do** so **all of those layers come in now actually believe it or not in near real-time so the lower layers tend to come in real time and then the upper layers the cognitive layers come in and within about five minutes of the game**

60:00

**our architecture** is covering the different levels so in order to do that it's a whole another discussion a whole another five hour talk but essentially you have to **you have to figure out the storage layer the transport layer the how to add intelligence over time that impacts everything** and so one of the one of the challenges I had was people would come to me and say hey can you solve this problem with the open source tools or ffmpeg or the streaming tools and the Netflix tools and and and **they're treating video as a done architecture** and then saying now that now that we can't video everywhere how do we use that video and how do we transport it using all the tools that were made well **those tools were not made for intelligence they are extremely under optimized** and so we've been trying we've been in that space trying to figure out how to optimize them we have done for example in our sports company every year we look at AWS Google Microsoft and we do cost comparisons what did it cost to run on our network we have our own data centers our own satellite systems and everything about five years ago I think we were at 10,000 times less in terms of cost we're we're still at about a thousand times last inning **if we ran the synergy support system in AWS today or Azure you're talking about I'm the business would not exist** and so I think that's a that's a interesting thing to look at is is you can be a data scientist on a laptop you can show some great you know things in the cloud but **when you actually go into the field and try and implement it it's either going to not work it's gonna scale or it's gonna be too expensive and any one of those is going to take you out I think that goes back to the slide which is why is 90% of this stuff failing**

70:40

we do have kind of this is kind of our heterogeneous approach if you will we were you know we don't have the money to go build our own hardware but we've decided hey algorithms are gonna get better without us there's no way we can we can dip into that humans are gonna be able to do things that we don't know about in the future we're not gonna dip into that but **what if we created a distributed platform that allowed humans algorithms cloud and everything to work together and to be able to describe that** and so this workflow right here I actually don't remember which one it is it looks similar that the basketball ones about double this so essentially you put your life source in here and it could be VOD it could be a hundred thousand hours or otherwise and then and the decisions are making it's transcoding so it's creating multiple bit rates these orange ones are machine learning the green ones are humans and then there's some analytics in between in order to figure out and filter and and quantify data so something like this is essentially running all of these things that we've been working so the last five years we've been doing just a ton of POC s across a wide spectrum and I'm gonna go through some of those now

kind of how we're using these advancements **first off basketball for you know getting back to location this matters where you did what you did matters now and and not only that we're doing this in near real-time** so when you actually see people the NBA doesn't let you on the court with it with a laptop but you'll see people run down into the into the room in the middle of half-court then they'll run back and give information that's what they're doing is they're allowed to run down and figure this stuff out in real-time so we're actually saying here's the effectiveness here's why you're losing things like that the and we can actually say this area of the court your weak in the last ten minutes here's what you need to do so

just to give you a sense of basketball never thought that there would be about 80,000 games we're in other sports now so we're probably near **150,000 games per year** so if you think about the globe in general and you think about all the courts in the world that's not a lot right it's not you know you think really where we need to get to is all of America all of this with GIS data and everything else but **if you just took all the courts in the world and you were to make those things sensors and real-time streaming components and try and figure out analytics and location within those analytics it's a huge challenge** it's a tiny pinprick on the world and yet it's a huge challenge **any given day on a high day we might be doing 4,000 hours of video and many thousands tens of thousands of games** you know

**interesting other ways that we've seen location coming into this is hedge funds** they've come to us and said look if we look at the Walmart parking lots these are some of these are actually **public data feeds we look at the Walmart parking lots right before Black Friday** before Christmas things like that or maybe even three months before we can start predicting how people are going to spend money at certain times and so they're gonna hedge and use that within their you know algorithms to figure out you know how they're **gonna place bets on the future** in this case it's not only where the cars are located in terms of the globe but literally where they are in the parking lot you can actually make some other analytics that are quite interesting

74:05

**Disney** we're working heavily with both on security and traffic flow so again **you know location matters a lot especially if you throw an alarm of suspicious activity you want to know where the person is not just that they exist and there's a** there's a lot around traffic flow obviously which gates are open and closed these are things that are otherwise kind of hard to figure out and when you start doing bulk and aggregate it's interesting

we are working with the government NGA we've worked with before I know there's someone from the NJ here I saw the acronym and we've included in our system what the government has **the standard of MISB standards around location information within the klv metadata set which is coming in real-time so now you get video and all of the geo-referenced information all at the same time** in that video and then there's some other challenges with that which is like if you have an object in the video how do you determine what where that object is we'll see kind of a fun use of that here soon

**this is NGA workflow it's crazy but it's pretty cool this is real-time video from UAVs and you know over a hundred of them you know figuring out you know everything together** and again it's not just what it's where and when and those analytics are quite interesting **the biggest thing from a location standpoint I think time and location is you're starting to gather patterns of life** and so not just that you have a whole bunch of data points but things that matter in the field and those patterns the actual pattern itself is a level of intelligence that we have not been able to do at scale

another interesting one again location these may not be global locations but Wilhelmsen it turns out that Nissan when Nissan finishes the car they actually sell it to Williamson so you're not buying from Nissan well **Wilhelmsen has to move ten thousand cars per day** so it's a logistics nightmare so they have you know I think it's like thirty acres of parking lot and they're trying to figure out okay what cars are leaving tomorrow get them close to the Train what cars are coming today get them close to this you've got you know 50 drivers their only job 24 by 7 is driving cars it turns out this is **a really interesting you know kind of localized geospatial application which is how do you tell the one guy to go to the one car that he needs and how do you know the car is there** by the way there's that there's a requirement every eight hours they have to audit every car in the parking lot and and so that's a huge challenge one that can be solved we have a **UAV here that's actually zipping along reading all the VIN numbers** and we have all of the **light posts have wide area cameras and the light posts are figuring out what the where the spots are empty and not** and we convert all those to geospatial information essentially GPS that goes back into essentially a **digital twin of the parking lot** and with the digital twin they can figure out where to go and what to do someone else is writing the logistics on you know how to send people to the next five cars what's the least path between the between the cars the bet that guy's gonna move and things like that it's actually a very complicated problem I appreciated it when I went there in a whole new way

77:25

**National Ski Team** for example they also their challenges they have no fixed assets so they go and they put five coaches you know in on the slope you don't know where the coaches are going to be the coaches are all down the slope and what's very important is is while they're moving the camera **you need to know what the camera is looking at from a location standpoint so that you confuse his data with the other data from the other people holding their cameras** and so we were trying to create is a realistic recreation of the skier going down such that you can literally figure out gate 3 is has a little bit more ice than gate 5 and the timing between gates and trying to automate that is an extremely interesting challenge

much more recently this was actually last week a group of us were **at Camp Roberts in California** here we were there with **Boeing and Lockheed and using again their streaming platforms** and the exercise interestingly although it clearly could be applied to other things was **to fly UAVs at a very high level around 4:00 to 5:00 in the morning and figure out heat signatures** turns out there's a lot of elk on at Camp Roberts so that we figure out where are all the hurts then we would send up you know smaller DJI size drones and they would actually go find the hurts based on the information from the first flight then and then all that information is being streamed back in real time and we built essentially a system of **workflow like what you saw there in order to count the elk** it's a it's a it's a big problem but it's a huge advantage to be able to count the elk on a weekly not yearly basis helps Fish and Wildlife figure out the herd size and management and all sorts of other things but what was interesting about it is **you couldn't do any of this without GPS like it's really hard with machine learning** to do tracking unless you know that that elk as long as the elk not running and typically in the morning they're all sitting down or laying down you know that literally the GPS location of that elk and so when you're looking at it through three different drones you can determine that the it's the same elk at the same time and then you don't have to count at ten times if you have ten drones these were pretty impressive programs you can literally just draw a circle in the map in the map where the were the elk were and the algorithms would actually figure out where the drone would stop every X number of feet in order to map out the entire ground space within that system so it was mostly automated we just hit go and we ate cookies and we watched the count it's very cold and I actually I actually had a I do that this was asked in the headquarters figuring it out this is us there's actually a video is gonna show but apparently I put it in the wrong thing it's kind of cool you can actually see the elk in real time as it's tracking the elk and counting it and

80:25

that's kind of the the journey I think you know looking ahead the amount of information that we're getting is huge the fuse **data set is huge everything that we do pretty much is coming with location** we're working with one of the largest malls and in the in California for flow and other things interestingly as you know the hybrid between satellite and inside a building the data are all coming together and so it's creating some really interesting viewpoints and analytics ok as before

81:20

Q&A

we have some time for questions he wants to go first

Question: that's excellent presentation I was just curious like one of the areas that we have a tough time as Eddy is in the response and recovery to major regional disasters and I'm guessing based on the Walmart example in the Nissan example that **we could use these types of analytics pretty effectively to understand the progress of the restoration and recovery** it takes weeks months maybe years we sometimes know very little about in terms of how to move material and support can you comment on that in terms of the use of that

Nils: yeah I think it's essentially **surveying the Delta and the changes rather than being it even every 10 days by satellite you know being every hour because you know especially an emergency response things are happening so quickly** is huge at Camp Roberts and so interestingly they have a whole simulated village and inside that village they have an area of it which is simulated for first responders and there's down helicopters and cars that are crashed and things like that and we were we were looking at how might we track you know one of the things about brittleness is we wouldn't know beforehand as an emergency responder I think one of the hard parts is **you can create all the models you want all day long and then all of a sudden something real happens and you realize that you that the models aren't what you needed** and for example if a certain building came down and there's a certain color and it was in the rubble of another building you might you might be able to look for the color of that building in order to give you a hint of where people might be and in order and to do that by and I've been there in the first responder zone before where they're like hey look for it but you only have 20 people and you're spread out very wide **if you could do this in an automated fashion it would be it would be huge** now so not only for the active right after where you could say hey I'm going to Train something very very quickly on the red brick and if there's an enough concentration of them in a gap or a hole or whatever I'm going to mark that on a map and then also in the humans can start making decisions more educated decisions on what to do immediately and then as you say afterwards as the rebuilding goes on you know are you applying the right resources in the right places essentially its eyes in the sky which is super super powerful

83:45

Question: in terms of **data standards** when I hear 4k or different formats coming through how do you see the data stands of media evolving and how do you prepare for that

Nils: it's a good question you know as you might know in the government there's **the MSIB group I like it because the things that are implemented there are standard and therefore when we implement that we get an interoperate** when we went and played with Lockheed last week we didn't have to do anything we've never used their newer platform before and it integrated so in that part I like it that **the speed at which they move and and you know is very slow** and so the that's a little frustrating I think on the the standards one of the issues I have is that the standards for video were generate I was literally at the table when we were talking about how are we gonna write the mpeg-4 Boxey system and although it's a wonderful system you can put anything in the impact for boxing system has all sorts of issues with it and so I think we need to go back to the drawing board and say look we actually haven't solved what we think we've solved on the video side all the way down to how does it how is it stored hot cold warm and certainly **the standards we have for video do not account for what we need today** so I would I mean my only answer to that is is that **our standards are actually holding us back at the moment** because we're trying to say you need to use the end of the government **saying you need to use the RTSP stream with UDP and everything else those were built for satellite delivery they were built for forward error correction they're absolutely not built for intelligence** and so I think we have a reckoning coming between people saying hey this layer is already done and you know very eloquently put in the previous discussion absolutely not there's a challenge

Question: **what about the ethics of all of this and not just privacy but you know equity and social justice** I mean I imagine there's a lot of things you guys have been thinking about maybe less in basketball games but yeah I mean even in basketball games there's an equity question right it's like so only the teams that have the money to implement something like this are now gonna win if it really does make that much of a difference so I mean have you guys been grappling with any of those issues?

Nils: absolutely and and in basketball they're more academic but they're real and in the other things we do they become much more you know a much more of a human issue in basketball I can tell you just give you give you some hints that are I would have never thought like we just ran ran headlong into them for example the we said hey wouldn't it be great if all the basketball players could get access to our data well we'll do that for free like if the team's already signed up originally it was the team and the coach and then we said hey we're gonna give the players access the coaches flipped they're like you know you know now people are gonna come to me and challenge me with my data and things like that but not only that then the players union flipped because the players in union saying hey you're providing information that is on performance of my player and things like that and and **now you're starting to negotiate with data** and other things well that's still playing out by the way and and then when you get to college it becomes a whole another level of ethics and understanding and we're actually doing high school games and so it's it's pretty crazy and then you know then all the way to you know when you play **a game in virtual reality you know from EA all that data comes from us on a daily basis so if they didn't play well last night they don't play well in the game the next the next nigh**t yeah so the the speed is there the questions are absolutely there the different organizations are asking questions I don't have a lot of answers we do grapple with it a lot some some of the things we've decided as a company to try and fix for example when we went to the NBA and we said hey we want to sell to that women's NBA they were like why and so so we said okay well what if we did a deal with you on the marketing side then we'll provide the service to the women the the women's teams essentially for free as a global service and so **we did work hard to try and balance those scales** and then on the same rate on the college side and the NCAA side we could have an actually a better example this is baseball there are some teams that came to us and said we know you're going to baseball we'll pay you X amount of money if you only go to our team for ten years and it was absolutely beautiful you know it's like hi but we said no and our pricing is the same for every team **and so we have done things that are that we're trying to kind of stop the the inequity component** I there's **clearly there's a better way to analyze that than us just kind of blundering into it** at the college level we made all video movement free regardless of whether you're a customer or not and some level of analytics free whether you're customer or not so with that in mind because Duke came to us and said hey you know we really wanted the same thing right we said no we're gonna we're gonna at least create kind of this minimum bar and then and then people can experiment on top of that and then on the government space it's a whole another whole

Ed: okay brilliant I think that was excellent session thank you very much we've got a quick five-minute break now we're gonna have a panel session

## Panel on Foundations/Motivations :

101:27

something else is gonna go wrong okay guys we could get organized sit down again got another exciting hour before lunch I sent some doubt in the room no really this is going to be really exciting this next hour I think okay so we have a serial panel session coming up before that each speaker is going to probably present four for five minutes or so just their initial thoughts and then we'll dig in much deeper to a Q we haven't actually set the order for this but if **our panel presenters would like to come and join us on the very informal sofa this is very very Google we don't quite use beam I don't know where that came from that Google and beanbags I've never seen a beanbag around the office we tend to have more sofas and nap pods** yes definitely but beanbags not so much okay so actually probably the easiest way to do it is a feature of the panelists when they come up to present just present themselves and then they do their presentation and then as I said hopefully a very interactive Q&A; session after each of the presentation

### Kathleen Stewart, UMCP/CGIS - Perspective: New opportunities through big mobility data analytics

102:25

morning everyone I'm Kathleen Stewart I am director of the Center for geospatial information science I'm at the University of Maryland I'm in geographical Sciences I'm University of Maryland at College Park and so for this foundation's panel I've chosen **big mobility data as the setting for the talk** I've been interested in **space-time modeling** for quite a long time and hopefully some of that will come through in my talk today

so here I'm showing you **vehicle travel patterns from big GPS waypoint data** although you wouldn't know that to look at the image on the slide we're seeing the I'm in Maryland so this is the Washington d.c Baltimore corridor of highway networks and it's a heat map so where we have the brightly lit up areas we have more traffic there this is actually an image from July the fourth a few years ago 2015 and we see the roads really lit up and of course that suggests that we've got a lots of drivers on the roads DC is not well infamous I guess for its traffic would be the way to describe it maybe a bit like here and we see you know Washington DC and it's my cursor where as it does so the Washington DC area and Baltimore really lit up and because **on July the 4th we have a lot of things going on people are driving around with their families and doing social activities** on the 4th and we see the roads all lit up for that reason and **we're interested in those kind of patterns of movement** to understand where people are where they go but **to achieve the this kind of understanding of travel patterns there's a lot of data residing behind those images** and

if we kind of change our scale a little bit and look at something like what we have at the top that's kind of moving around we have travel again on July the 4th so busy social day major holiday in the United States and this is the bridge over to the Eastern Shore in Maryland so a popular place for recreation beautiful beaches place to go and we have lots of drivers heading across to the Eastern Shore all going over that bridge and so again we might be interested in traffic **if we want to manage traffic or understand about travel behaviors of drivers travel patterns those kind of space-time patterns that we've been talking about then looking at significant locations maybe like a key piece of infrastructure and the transportation network is something that we'd be interested in** and the visualization below this image here we go there we go is that same data but now presented from the perspective of kind of origins of those trips and trajectories making their way across the the bridge

105:40

so we have kind of **different story between these two visualizations one captures the amount of vehicles on a road segment at a particular time or bridge and the other shows us kind of the diversity of origins** of where these travelers are coming from and there's lots of reasons why we might want to to understand that better for example to provide signage to drivers who are on their way about what they can expect about their travel route to better manage traffic as drivers are making their their trips and of course all you know we all travel a great deal as part of our everyday lives whether it's in family and social situations or work trips and these **different kinds of semantics associated with movement** are very important and it's something that we're we're interested in for a lot of reasons I'm open I have to click on this to move ahead here we go so

these **space-time patterns are something of interest to us** here in these two images in the top image we see the kinds of patterns on roads by passenger vehicles so mainly cars and they have a very different kind of travel footprint to the image below which is more about truck fleet traffic and here we see more trucks or commercial vehicles really using our main highways again this is in the the DC Baltimore area are kind of the top images is zoomed in a little bit more but still that BC Baltimore area and with passenger vehicles they're going different places and then the trucks they're going into more the residential areas where we live this is what you'd expect but there are differences therefore in the kinds of travel patterns we get and this is the kind of you know important information that we're really interested in extracting and we want to know things for example like **urban - rural differences we might want to know about different kinds of patterns in those kinds of settings risk exposure if we have to plan a you know a major mass evacuation whether it's for flooding or wildfires or any of these kinds of scenarios** how do we move a population quickly then we're really interested in understanding what daily patterns of movement are like to be able to support coming up with different kinds of scenarios that would help us when things turn to too bad from good we need to know how to move in all those kinds of cases and again kind of **multi scale analysis that's such a classic geographic problem really important for those of us interested in geospatial data** is understanding our problems from a multi scale perspective so **whether it's the individual traveler and the trips that someone might take or whether we're looking at traffic as a whole a more aggregate perspective** and more traveled you know patterns of on mass for example these are all kinds of topics that we might be interested in

108:44

now looking at this big whether it's GPS trajectory data as those images were all created using a waypoint **data comes with its own challenges there are irregular sampling intervals that we have to deal with this kind of data there may be gaps in the data there can be trips that seem to start and end in an odd way** that kind of you have to figure out and try to understand am I still talking about the same trip or if we move to a different driver and a different trip so those are some of the kinds of challenges of just trying to get to that end result that we're interested in but again the data gives us a lot still that's complicated to kind of work with and so **we need different kinds of reconstruction or construction algorithms to help us build these trips and trajectories** and there's a whole vocabulary about traces and tracks and trips and trajectories just for mobility related data and so we want to be able to work with the data that we have and then get to something sensible and meaningful and useful and again we need our **we need still our theories and computation algorithms etc to be able to take us there and again when we're working with massive datasets we need new methods** for for dealing with this and

in addition for mobility data we don't just have sort of GPS data that is definitely one very useful source for our data but we have this kind of **data captured through a numerous different kinds of data sources now and each one brings its own special characteristic** for example we have **cell phone data we have location location-based service data from apps we have geo tag tweets or other data from Twitter** and each one of those as you see in my images here has a **different kind of characteristic and gives us a different person affective or view or information about movement** on roads these examples are all taken for the Dupont Circle area in Washington DC and if you go to DC chances are you've you know at least gone through Dupont Circle major you know location popular location in DC and so we notice that our datasets have different numbers of points associated with them and so therefore different granularities of data available for our computations we can derive heat maps as well as actual map spatial maps of the information and we see that depending on the usage of our cellphones or when we access an app **we actually again can procure different perspectives on travel** travel at different times of day apps tended tend to be used at certain times of the day perhaps more than others when and where people tweet as again there's quite a bit you know rich literature now and using Twitter data lots of biases potentially involved in each one of these kinds of data sources so again we get different kinds of perspectives that are very important for us to understand and

here we just see the example of kind of looking at **trips derived from like cell phone data and trips derive from location-based service app data** and we notice that with the location-based app data we have fewer individuals we have data from in this case for our sample that we were working for we had less samples but more detail more spatial detail about the paths and routes and roads etc that we were able to get from the data and with cell phone data were very much it is very much in the context of the cell phone data towers etc that influences kind of the path so **we have to understand more about these different data sources** and make sure we do understand while these different data sources when we use the data for their our final and end products that we're looking at

this is some slides I should say that I sorry I haven't mentioned my colleague Jin Chuang fan in the Center who has been doing some of this work and I'm showing his work here and and we've been looking at this idea of using Twitter data in this example so this is travelers who went to view the **solar eclipse a couple of years ago in 2017** and looking at trips to view the solar eclipses and so using only social media data but then also looking at traffic data to see how well our social media estimates of travel and we're you know how far are people in which parts of the country were people traveling the most to see kind of an event because this can kind of give you some insights again if we're interested in mass evacuations etc and the interesting thing about **Twitter data is it really has some kind of interesting possibilities when we go to kind of a national scale or in the global scale** looking then some of the bias concern because we have we have so many tweets across a big space is may be ameliorated a little bit but again we can get really interesting signals about movement using these kinds of data sources so we have lots of new data coming online that we want to be aware of and kind of take advantage of

a different spin on this mobility data lies with in this case genomic data so we're doing some work with health in the field **of health geography and in this case our genomic data** which is mapped here is those light blue and orange e areas so here the mobility is is not about humans in that case it's about **malaria parasites being carried by vectors so mosquitoes** and so we have genomics of malaria parasites so we're interested to know about genomic parasite gene flow and parasite migration and so barriers and sources of sinks for pair parasites and **so the moving object in this case is the vector and indeed the parasites and using genomic data but then we want to couple that with human movement** so you can see a road networking kind of the the purple there and so we want to couple our understanding of road network and where humans are going to travel to see the sort of nexus at where parasite mobility and human mobility come together because I can **give us a lot of interesting insights into risk and vulnerability and transmission in this case for infectious disease** so again even with genomic data we can think of ways to kind of you know talk about mobility talk about dynamics geo spatial dynamics map these kinds of dynamics and yeah think about these next steps and then everything I've shown you today has been based on data that has already been collected but we have heard in our talks that you know it's kind of real-time is is very interesting too we haven't tackled that yet but that is on our yeah we'd very much like to to kind of start to think about our methods and kind of you know sort of make t**hat next step into kind of real-time** butanyways I hope there was a few things there that can give pause for some thought thank you

obviously well hold on their questions into the pattern so if you have any specifically for caffeine please do you ask then thank you I know you want to come up next okay below salt you know page to your mind that's easy

### Anand Padmanabhan, University of Illinois - Perspective: CyberGIS-Jupyter for Reproducible Geospatial Research and Education

116:50

good morning everyone my name is Anand Padmanabha I'm a senior research scientist at the University of Illinois at the University of Illinois at urbana-champaign and at the cyber GI Centre for advanced digital and spatial studies there today I'll be talking about something about what what we call a **CyberGIS Jupyter with all this big data and analysis coming in how do we ensure that we we can be able to do this kind of analysis at scale and make this results reproducible and make it shareable with people and make it open** I guess in some sense so like

I borrowed this slide and here they're talking about what is the **lifecycle of a scientific of going from a scientific idea to actually a product and our publications and even dissemination and education** and how do we do this in such a way that this is transparent and reproducible can we do this all in a single platform and in some sense a Jupyter I don't know how many of you here know about Jupyter technologies but it kind of provides a platform where you could do these kinds of analysis so

we had couple of papers that we were talking about how to do this analysis in Jupyter and how to scale it across like multiple platforms I had this paper and XSEDE which are in the Perth conference which was talking about how to scale this across exceed that is the national supercomputing environment and our local infrastructure so how can you use it both across our infrastructure as well as burst it out into a national cloud infrastructure so

in all these contexts we have been talking about something called **cyberGIS** and as Mark introduced in his truck this morning cyber GIS we established it as a platform for doing geospatial information science and systems on a large scale on cyber infrastructure by that what we mean is we are focusing on doing high performance computing high throughput computing cloud computing to solve geospatial problems and this kind of bridges the gap between data software and applications initially **we were when we started this journey 10 years ago we did not have Jupyter and so what we were doing at that point was we were creating applications apps and web applications** for everything which is good in some sense it makes like accessibility easy for users **but in some sense it's not scalable that you cannot a single group of people cannot be building applications for everyone with the Jupyter platform** it's kind of it allows that power to be given to scientists and they can do the computation on their own so we have been moving to that angle for for doing our web environment so

what we came up with in the last few years was we have this **Jupyter environment and our focus has now been on providing api's and libraries** which are integrated into the environment and the scaling capabilities and not so much on front-end development for a particular King with particular scientists to do a particular problem but we are providing this geospatial software capabilities and that is all integrated into the platform so everyone can use it to can customize it for their own environment

so some of the key features that we have in this environment is it's it's able **to do data intensive and reproducible research** and it provides kind of a holistic solution that you can build and the other important thing is like everyone can get a similar environment so whenever people are using this environment they are they are getting all the libraries all the api's and being able to access it and it also kind of reduces the barrier which people face to accessing the Sivaji has capabilities like you don't need to know like parallel programming in order to use this now you have these libraries **you can make use of you have Python connectors for them and you can do that from a very simple environment you don't even need to learn shell** and like when I was teaching like class to geography students a significant barrier for them was like if I bring up a terminal and a batch shell and I do like editing in a VI or a nano and then like people are like oh it's it's too tough for me and that kind of is a barrier it's and it's an **unnecessary barrier if you think about it like there are tools which allow them to do this to do the learning without having to do those kinds of things** so capabilities we have that one of the main things is we have like a common platform and we are able to scale it across like the local as well as the national environment

we provide geovisualization and you can make maths you can make you can put layers on the map you can do we have other libraries we open-source libraries as well as other libraries which we can integrate there where you are able to manage the maps you create the maps export the maps and even host it on your website of course

on the backend we have **lot of technologies to be used but these are not getting exposed to people** so that's kind of the the key thing so we we are doing data synchronization we are doing cloud based management with docker containers and images we are using HPC kind of systems but many of these are transparent so one of the key things we did with did this was like one of the case study what

what can we do with this so we had like we work with some hydrologist who they had a national scale problem which is called this how do you do kind of a **flood mapping on a national scale** and this was not done before and this required like lot of computation and you have this NHD like National hydrography data sets which is a huge amount of data set coming from USGS and you have to do this kind of a **calculation which is height about nearest range calculation for every small amount of area within the US** so that's a very intensive computation so we **made this all into a Jupyter environment so we have this whole workflow and coded as a Jupyter notebook** people can take this this is of course as but this is this is kind of a representation showing the whole workflow but we have this workflow available as i Python notebook you can download you can run it so you can reproduce kind of this analysis on your own

so one of the great successes we had for this we ran a summer school this past summer in collaboration with AAG which is the American Society of geographers as well as University Consortium of GIS and we ran a summer school where we brought in like 35 students and paired them with mendes many of these students had no programming experience or very little programming experience and most of them had not done to Jupyter before and the infrastructure we were able to run like 50 people at the same time no problem but **the key outcome of that was students were able to create these notebooks** **which at the end of the week they were able to share these notebooks and they felt confident that they are able to solve some science problems**

so I can show quick example if I I can't get this let's see these kind of yeah this was one of the groups which did like a simulation so they created this whole work book it is almost like a publication that they can take forward with this so all the members with the mentor they are showing their workflow and they have runnable codes here so this is kind of like you have your kind of your results and your background and all the study details as well as runnable codes where there they are running the models they are running they're installing packages they are running simulations they are creating maps and of the outbreak and how it spreads I don't know the details of this but they created this and within a week they were able to do this which was really powerful for us and there were six all the six groups were able to do this and that was that we felt was really amazing and we are hoping to repeat this we are planning to also release this as a community-wide thing for everyone to use shortly

so just jumping ahead and I'm going to make a slight leap we have something called **national geospatial software Institute** we are conceptualizing something so this is an audience which I thought sure I should mention this the **goal is to create like software institute and a long which is a long term hub of excellence for geospatial software** because geospatial software is like distributed all over the basis many people are creating geospatial software people don't know which software to use and like academics are doing some things like open source industry is doing something how do we create a hub for the whole community this is kind of a software this is NSF funded project for conceptualization this George has been involved from the beginning and he is an advisor to the project and we will be we had three workshops from the project many of you were engaged or in the effort and we are probably are creating a report out of this which we will submit to NSF as well as we plan to have a webinar soon probably in December where we will discuss the strategic plan coming out of the workshop so it's not too late if you are not engaged in this process please do let us know what you think is important for the community so with that clock are any questions I guess we can hold it thank you very much

### Mark Korver, AWS

127:54

my name is mark Korver I'm the geospatial lead specialist team which is part of our social architecture team and then I think there's another layer there in between our business development and tech vision team my general role is to you know incubate a team of other solution architects that then support solution architects that we call are either regionally based or account attached have customers with geospatial use cases I lead that team I'm also work with the number of you in the room or here of your firm's in a room I'm generally attached to some of our larger geospatial customers so today mm-hmm I'm gonna I'll make this real quick because I know we want time for the more time for the panel I

I wanted to go back to basic basics my other hat is storage a subject matter expert and it's largely because I've **been working with large geospatial companies for some time now mainly around how to move off of block or file systems to object store** I wanted to speak to that a little bit so you know many of us are aware that a lot of the success right in the open source world is due to you know core projects the **f**oundational projects they PostGIS on top of Postgres GDALand proj and there are many others but these are you know some of the big main core ones that typically have one or two you know like one committer right but the rest of the world depends on them so a lot of the success of open source generally can be attributed to some of these key projects and

I wanted to introduce another idea which is a **you know what's changed in the cloud was different from on-prem i**s that so I say I'm over eight years at AWS before I joined Amazon Web Services I was a customer for about three and a half years I was one of the initial users of EC 2 s 3 and and it was just sqs we didn't have 150 services that took me about half a day to figure out how to fire up the virtual machine now it would take a lot more time to just get a handle on our virtual machine services but it was easier for me when I got when I got started but I was **one of the earliest users of Amazon s3** I think at a national data set level and so so this is so this announcement was about a year after I joined where we first started talking about **a trillion objects in s3** and in less than a year later we're talking two trillion objects and I did a little googling around last night thinking there must be more recent announcement of this how many trillions is it now and I looked at some internal documents and there's really nothing out there which means we haven't publicly announced other than to say that we have trillions of objects so this is everything from you know some enterprise customers you know backup of a server right to operates **well-known for example the Digital Globe moved a whole bunch of sensor data off of tapes right and use the truck to get it to to s3** so and it's everything the genome makes this Alzheimer's image scans just everything that you could possibly imagine is stored in s3 so

I was looking at this and this is **figure** **from the concept documen**t and what's interesting when I look at something like this now is I can't help but think **application silos** right so the tradition of working on data because we you know we didn't have objects for 10 years ago right we had file systems kind of necessitated working on it here on the input on the inside and then go basically going through these silos and impounding getting out the door it's not it's not connected right and what I wanted to point out is

a couple of more recent projects so this is a **earth data NASA cumulus project** and it has anybody her about this before I'm just get a sense of hands right so for those who haven't seen this the general the gist of this is we can't keep up with space in our data centers in our opinion anymore and this is across all of the NASA's distributed active archives DAACs and so this is the software you know the cloudy software implementation of a large NASA class data centers right and what I want to point out is **right smack in the in the middle and it is glacier and s3 you can actually simplify that now and think of think of it as s3 the object storage with different classes of data** so hotter data do colder data and what's interesting with this picture is that there's DAAC a on the Left DAAC B on the right the two completely different you know catalog systems of that store you know petabytes many petabytes of data that are you know storing everything on s3 but **not just storing but using s3 also for content distribution** so it's not so if you can think back you know back in the day when we had to you know get access to some new satellite data typically we'd be **bottlenecked at the ftp server** right it gets announced and then you'd go and then you go okay I'll come back next week just a waste of time right why because it was exactly five FTP servers running right or you get an email from somebody that's angry at you **that is largely solved by object store** this is not just you know simple storage service but it's all the large cloud providers object stores I think we could generalize this and the main thing here is that because the **data is loosely coupled in the object store its horizontally available to anybody's use case** and that's the fundamental paradigm shift that moves us from **shipping data which does not work for streaming big data** right so it gets us out of the world world of having to identify an area or a region of the planet that we care about downloading that and during our own local storage and you know we might have HPC access locally but it took you how long to get access to the data so all that work that we used to do in the past **80 90 percent of the work is around moving data preparing data fixing data is it never that's never right the first time right** is largely alleviated by you know what we now call well we now are accustomed to something called **objects store it's very simple all it is is HTTP put HTTP GET HTTP delete right it does a bunch of other stuff but that's a core concept** it's it's just the URL end of the day

so here's another picture I just saw this I just found this literally yesterday on medium and somebody is talking about you know **a workflow but notice that they're using multiple s3 buckets** along the way why is this important because the object store is not something where for example you store just the raw satellite sensor data the original data coming off the satellite and you're enabled to make that available to anybody or not right pending on your security concerns right but you can make it available all along your production stack so simple things like you know every year we have the USDA's native data acquisition right so we're starting to see customers you know this is not Sat satellites is the aerial aircraft flying around taking state level content where's that data being processed or built where does that production data being built it's being built in the cloud somebody's cloud right might be ours right maybe somewhere else but what's going on now is that instead of moving the data from aircraft to somebody's data center being done in the silos being straight up in the cloud more recently or not announcing ground station which is about allowing you to get out of the business of running your ground station you can just you know stick to your small sets and rent time on our **ground station service** which means that the downlink from the satellite goes immediate to one of our regions were you can do all of this right

a couple of other things I want to point out is so we all know a Tiff's are GeoTIFFS because we're spatial geeks been around for a long time so this is a specific recipe for geoTIFF called **cloud optimized geoTIFF** it's actually quite simple all it is is our **specific arrangement of the TIFF file - to make get requests against the object store more efficient so what we're seeing here is a kind of a maturation a movement of open source tools like GDAL** in this case so that they look like middleware cloudy middleware why is that important why is this happening because people now have access to very large streaming data sets and if you were in the business of moving that data to your compute you'd be out of business it doesn't work it doesn't scale though this is part of the answer to part part of the kind of a recipe to being able to **move compute to data other than rather than the other way around**

a similar thing with the recent announcement with the new features for **HDF5 virtual file drivers** for s Street right why is this happening well they got a you know heads up from work on cloud optimized geo tips the same same underlying reason

another thing I want to point out is it's you know for a few need for a number of years years now we've been sharing data in these **very large object stores on AWS we call them s3 bucket**s so you can make this bucket something that is very private might be monetized it might be you know some university account where you're doing you know very sensitive data like alzheimer is you know mixed data that you share which is one or two other teams or it might be do something that's a you know federal data set that you want to get out of the business of storing data and be just as importantly you want to get out of the business of having to worry about distribution of that data this STAC spec that the further so this is back in 2017 already so a number of companies who were in the business of Earth Observation data you know area data lidar data got together because we realized that we're sharing data out of these buckets and these buckets were **deep streaming buckets that constantly changed we need to have an open way to be able to for example hydrate your own catalog based on the STAC index because I want to do for example push it into something like postGIS** right or I want to put it into redshift or whatever tool that is that I desire on my infrastructure so this is an outgrowth of those kinds of efforts I think that's very important going going forward

this is just a screenshot of our registry of **open data on AWS** so here so this is not us doing open data this is about our customers using best you know working on our advisement about what best practice for open data is and then basically just coming to github and registering their open data and telling the world I have something that I would be happy for you to use and I store my data there but I'm not in the business I've acted to distribute that data the key point here is we for example advise customers especially when these datasets are a large return to **turn a feature called a requester pays on so that they don't get economically ddosed when somebody comes in and decides to move all of the their data from you know Amazon s3 to GCP for example because that might be expensive** the last thing is it's all about sharing right

so please don't think of **object stores as a way to stash data and store it or just archive it forever and never see it again it's a way to share data that petabyte scale** thank you okay

### Jayant Sharma, Oracle

141:43

I'm just going to talk hopefully in less than five minutes about **one simple idea and the concept here is we've talked about things like where something happens when it happens helps you figure out the why** and what you might want to do about it we've talked about at a foundational level at the very basics what you're looking at when you want to manage all this data and manage all this information you're thinking about to is specialization or convergence right so mark talked about looking at either specialized approaches for particular processing the analogy here is the different workloads when you're dealing with your data a workload for dealing with IOT is the workload for dealing with transactions a workload for dealing with aggregations in in-memory pre-computation of certain things and later they will have people talking about representation right so you either have an option of using an infrastructure that is made up of the best of breed of all of these and here's the part we're standing on the shoulders of giants comes as Neil's earlier said you have an alphabet soup of either specialized the homegrown systems that lead to a technical debt

the other option is use converged infrastructure that lets you deal with these multiple workloads multiple representations and the primary of that again as Neil said how do you scale your talent right you've got talent that understands sequel you got for Best of Breed you need talent that understands GQL which is not Google query language but graph so that is the thing when you are dealing with your projects anything these are the options you've got go for something in your data management you have the option of taking the Best of Breed and then dealing with the integration complexities or dealing with the complexities of copies of data when you don't know which one has actually been update and which one is possibly a source of truth or which one is in an insecure backup that gives hackers for example one opportunity of making capital of your data and that's it in a nutshell thank you

### Group Discussion

1:44:40

ed went to check on lunch so he asked me to stand in he had some prepared questions which of course I do not have and so in the moments that he said that until now I've thought of **two things one was the the space and time aspect and the other was kind of the workflow aspect** and so I asked a couple moderated questions and it's really a warm-up because we want to quickly turn to the question that you all have you all collectively have for the audience here

Question: so with respect to **space and time first** so you began with an excellent example of trajectories in space and time is really key I think that's important with respect to where we go with data data science and data analytics you were in fact to mention I think Mark mentioned a book that you wrote some time ago that I'm gonna go look up now do you see we've made progress and where do we go from here and I'm hoping maybe you answer a little bit and then others can think about from your perspective handling time and trajectories in particular what are we able to do now that we weren't able to do before and where do we go with respect to time

1:46

Stewart: what I think has changed now is **we finally have the data that we didn't have all those many years ago** we just we didn't have the you know computing possibilities that we have now and really the kinds of data that we have now so we could imagine and thought about many of the kinds of semantics and situations and applications that we are in fact indeed interested in and now **we were talking about sort of trajectories or trajectory like objects many years ago and now we have them for real and can kind of make them happen and create them and we discover new things that we we didn't really think about before because we didn't understand how the technology was going to deliver these things** I think to us and so in real time space and time real time you know we imagined but but you know now of course we're there so and what's next you know what will be the the next thing to be that we should already be thinking about and it marks presentation this morning he kind of gave us some hints about what **we might be thinking about so moving our mobility questions or space-time dynamics questions into this next era** but still there's so many kind of fundamental questions that kinds of information that we're interested in how do we really get at that that data information kind of discussion again trigger any thoughts from the other panels she obviously did that as her key thing so it's tougher for you guys after that I do

Korver: reminded me of some fascinating applications along those same lines so a few weeks ago there was a another data spatial data science conference where in fact the same use case that you took **Airbnb** used exactly that back particular event not the temporal **but from their listings to prove that they bring benefits to cities** and they're sure the benefit the residual benefit that is they were obviously a lot of bookings but then bookings in those areas within that path stayed up stayed and increased and they were able to show those cities here is the benefit that we bring through a platform to you yeah and another place where this has been used commercially is you've heard about the Waze pool do is essentially just

Korver: a couple of general observations right so uh you know you know Geo sciences in Australia doing some very interesting work on **data cubes** for example which was all wrong you know time series data greeting the planet but you know really sense forty sensation said but at this really more basic level for example I was just on a call yesterday with the well-known engineering firm that has a history of stuff on tape right and I'm asking thank God and asking questions like how many **thousands of tapes** was interesting about that conversation was that there the customer is going to **move off it off of the tape format and it's gonna go like I was talking about earlier to the object store** that we don't want it **we don't want it to be a tape to tape movement** right we don't what what occurred to me though was design goal there is all of a sudden we have the ability for this customer to be able to reach back in time in a way I mean just happy to get the table nobody so you know even for you know just one small project like that was measured in something like 15 petabytes of data right many customers they you know hopefully a year from now they'll be able to go I want to reach back ten years and grab this chunk of data I can do it measured in hours without lifting a finger that's a huge not just know that's imagery lidar data but a whole history that this company had to archive but that doesn't mean they had access to it even themselves yeah

Stewart: so **change I think is just a big driver big important whether its land use land cover change and kind of tracking all those kinds of you know environment development kinds of changes** and just in in general that historic like you might want to understand a history or an archive so that you can understand change and how it's happened and then what you can plan for what you might expect to happen ya know cool

Question: the other one other thought was workflow obviously mark you hit that a lot and on you hit that from the Jupyter notebooks point of view can you read Jupyter on Esther II think we can I think we can Jupyter is part of our ndn machine learning stage maker service so so what do we need to do you know we've been working on workflows for a long time and it always comes around - **we need some workflow language in order to** automate it we saw some great pictures that Middle's put up and the like that are kind of that are obviously useful for them but I've always found it hard to **get traction around people to actually do workflows and any kind of repeatable shareable fashion maybe there's a breakthrough**

Anand: I personally I do think that **Jupyter allows some of that because it is encoding not just the what you do but how you do it as well within the notebook which can be shared** of course there are problems with sharing it's not just completely reproducible in some sense that you need the environment also you need the same libraries for example in order to get the same results but with the technologies we have currently having I think **with the cloud and docker and all those things it is moving in the right direction though I think there is more work to be done** but but I say I see there is at least a way forward not like it's a theoretical concept

Korver: consider something **like Sagemaker which are machine learning product right so you have the familiar notebook metaphor of Jupyter** the difference is that you're working for example was shared earth observation data that you didn't have to voice from here to here to get her going right another simple concept is you know when Jupyter first came out and essentially you're writing you know a web server that can run your Python code and show your output and you could add your comments a great stuff but you're constrained by the cores on that machine right nowadays might be a docker image right but the the **Jupyter going forward is we're sharing code and methods via the notebook metaphor but we're not constrained by storage** **we're not constrained by parallel on the backend so** typically a end user would have you know n number of cores in a Jupyter environment but going forward when they run a job they **could run the whole planet right if you want I mean you have to be careful about the cost** here but they could if they want well cool

why don't we open it up to questions from you all so who's got a question

Question: probably directed to Jaynat and Mark panel - I'm interested their opinions you talked about **specialized versus generalized or centralized** databases do you see any patterns emerging in terms of are they driven by workflows or the kinds of data that they store like spatial temporal versus symmetry and there are some issues in terms of you make copies of the data you make copies of the system of record in many cases **earlier the system of record was just as an instant of time as opposed to a continuous spatial temporal kind of data that's getting stored** and so on what's your perspective in terms of what changes are coming about from a database perspective **does the database drive what's possible or the workflows drive what the databases should be doing?** any thoughts on that

Jayant: so one clarification **I talked about converged in specialist not necessarily centralized so converge doesn't necessarily have to be centralized and is in fact distributed** but the the cop piece over there does the database development of the database advanced drive not necessarily right it's it is the it's not just the workflows it's the infrastructure around it and databases while they may seem to lag the data enable right so for example let's take on the conversion it's not you may need an in-memory Mark had the flow way you may do different things at different layers right you do all the quickly processing filtering in in-memory you then do the aggregations and then you process that somewhere else all right those other things that are also now finding their way in to the data stores and so it's it's not an either/or they go hand-in-hand it is **more driven by the requirements of the processing so therefore workflows or workloads** that's my perspective so

Korver: I don't think this is necessarily geospatial I think you know generally the strong patterns that we're seeing at AWS Ely's does that you know there's unstructured to structured data this is you know going back to kind of OGC early days for me was you know we had these highly siloed systems that no vendors you bought from a vendor we wanted them to speak the same language right across to each other right things like GML or back in Japan was the GXML or something right anyway they had to do with moving data from one system to another we need to have standards to do the otherwise you know what you're gonna buy right and **what's distinctly different now is now that we have loosely coupled architectures and it's and it doesn't cost you money to share data** right in fact you can make money both make money and not worry about the cost on the backend right and that's that's a complete shift and so in that sense there's **less need I think you could argue for standard you know you know implementations like GML to share data back and forth because you're actually exposing raw data** and it's in its rawest form well your exploit you want to do **to give to your customer to your colleague it's whatever the code is to gain value** from that right and so it's unstructured somebody cooked it up somebody literally built some new database thing yesterday and you want to point it at this data what do you want you want it to be **good authoritative well curated content and that's all that really matters that's what we're seeing not just in geospatial but across the board** you know whether it's genomics or Netflix you know logs that's the pattern in satellite space too it's like you know we just want **we want to keep whatever is the most original authoritative data along with the metadata around the sensors because the algorithms we're going to use ten years from now if the gain value from the data** we know talked about the transition to stack an API is part that's why that's my STAC is important you know goes back to basics it's you know we're moving from a world or we ship data so literally one copy of good data it gets reflected a zillion times around the planet storage problem right plus it's like I got this data but I didn't take care of it very well so I don't know if the metadata is correct or not there's probably

Question: a question for mark this is Jay from UC San Diego and just as you mentioned that right now we're reducing the duplication and copy and lot of good data czar AWS hands right now and every time **hitting a good geospatial large data set is it it's like hitting a goldmine for geospatial open researcher** and **is there implementation or plan to connect to the Semantic Web or a knowledge graph** which is not yet Vina touch on today Fran aw standpoint for make the sharing actually easier

Korver: **the closest thing is that the STAC implementation so it's not semantics right** it's different and and you know quite frankly it's amazing because you know the object stores over ten years old now and all the other providers have object stores that look a lot like a stream right but really in Amazonian terms we're still day one it's taken a long time for people to move their code off of you know systems that require you know file systems block storage or whatnot right to being comfortable with using object store and sharing object store so we're just at the beginning right and so you know hopefully in a couple years we'll have missions like like the NISR mission from NASA part of the reason for things like **Cumulus** were we're getting you know eighty **two hundred terabytes dropped on us every day** and people are getting accustomed to that kind of streaming having access to all of it in real time that's new so we have a lot of work to do

Ed: I mean I think **linked data is a bigger thing in Europe for whatever reason** than it has been in in North America I think you know from our perspective there's the academic approach to creating linked data but then there's a lot you can do behind the scenes into the creative markup of web content and and **using methodologies like schema.org to add meaning to otherwise unstructured data and that's a lot of the technology this is behind the knowledge graph** that is a major part of search whether a Google point of view we've long since moved away from trying to present you with ten links on a page or actually answering your question and I think I think the last time we did the analysis over sixty percent of of the queries that someone sends to Google we can answer directly so rather than singing to a webpage that will tell you what Churchill's birthday was we can tell you what Churchill's birthday was and its then **extending that kind of metaphor into spatial data in a pragmatic way that I thought I think is appropriate** and that's you know if you want to find out more **look at the work that we did the joint OGC W3C working group on spatial data might on the web a very kind of pragmatic approach to a more semantic understanding of geospatial**

[lunch]

okay now I want I want to be the guy that says no more questions but I am going to be the guy that says no more questions just because we have a booking of a set of tables over in the restaurant that we're going to use that is from 12 till 1 so I want us to give us the opportunity of taking part of it part of the Google culture but before we do that let's thank everyone from their panel from this morning thank you all these session this morning has been recorded I'll stop the recording over lunch time we will package it up and at some point over the next week or so we'll make that available to you via a YouTube download link ok so the process for lunch is we can go across to Mavericks which is one of the many cafes that Google has across the campus **eating together is a big part of the Google culture so you grab a table sit down and chat to people that perhaps you don't normally see over free food** so that's the whole point so we go there the food's there just grab whatever you fancy by all means make use of the tables that are inside the restaurant it's a bit cooler than it was yesterday but by all means you can sit outside as well but we need to be back in here at one o'clock for the next session but if you want to follow me I'll take you over there