>> Hello, everyone. My name is Amy Walton. I lead the data project in the Office of Advanced Cyberinfrastructure. I have to tell you a quick, funny story that, before that, for a long time, I was with the NASA's Jet Propulsion Laboratory, and it was my privilege at the time to have been at the, working with the Forest Service to try and do some of their first airborne detections of forest fires. The kinds of equipment that they had at the time was basically the size and weight of a refrigerator that they would roll onto the airplane. It would then take overflight infrared pictures, and they would have to drop the infrared film at a location, have it developed and taken to the fire camps overnight. The dramatic changes you are about to hear in how fires and many other data activities are now being done will be incredibly dramatic, so allow me to a take a moment to introduce one of my colleagues, Dr. Ilkay Altintas is the chief data science officer at the San Diego Supercomputer Center at the University of California San Diego, where she is also the founder and director for the Workflows for Data Science Center of Excellence.
In her various projects, she leads collaborative, multidisciplinary activities with research objectives to deliver impactful results by making computational data science work more reusable, programmable, scalable, and reproducible. Since joining the San Diego Supercomputer Center in 2001, she has been a principal investigator and a technical leader in a wide range of cross-disciplinary projects. Her work has been applied to many scientific and societal domains, including bioinformatics, geoinformatics, high-energy physics, multiscale biomedical science, smart cities, and smart manufacturing. She is an initiator of the popular open-source Kepler scientific workflow system and the coauthor of publications related to computational data science at the intersection of workflows, provenance [phonetic], distributed computing, big data reproducibility, and software modeling in many different application areas.
She's also a popular MOOC instructor in the field of big data science and has reached out to hundreds of thousands of learners across the globe. Her PhD degree is from the University of Amsterdam in the Netherlands with an emphasis on provenance of workflow-driven collaborative science. She's an associate research scientist at UC San Diego, and among the many awards she has received are the 2015 IEEE Award of Excellence in Scalable Computing for Early Career Researchers as well as the 2017 ACM Emerging Woman Leader in Technical Computing Award. It is my pleasure and privilege to introduce Ilkay Altintas. Ilkay, thank you.
>> Thank you. Good afternoon, everybody. I'm now presenting you as the PI of the [inaudible] WIFIRE project, and it's a privilege to be here. Thank you, Amy, and [inaudible] for the kind introductions as well. So let me give you a little about me and that I work for so it makes more sense when I talk about how WIFIRE came to being. I've been at the San Diego Supercomputer Center since 2001, and since then, there has been quite a few changes. But really, the transformation happened with the beginning of the center as a NSF center, one of the 3 NSF centers, in 1985. So it's more than 30 years old now, and what started as a center with a deployment of a purchased supercomputer enabling a nation's science then transformed itself into a place that innovation and supercomputing happens, like we have been innovating the first flash-based supercomputer and also the first virtual HPC.
So to say, the [inaudible] computer in the recent past. But also, projects, research projects that are funded by research dollars, both through federal agencies and industry partners, on big data, versatile computing, and life sciences applications. So, and many other different application areas. The center now enjoys quite a wide research program. In addition to the

supercomputing facilities, we have the Western Big Data Hub funded by NSF or cohosted, the [inaudible] Institute, many well-known projects, including Protein Data Bank and the [inaudible] Gateway as was also started here. So there's a lot of activity here that is multidisciplinary. What happened, in the last 10 years, but even before then, is we've seen a huge rise in inclusion of data science techniques, and data management, and data science in the projects we were involved with.

So almost everyone I'd say here does a version of data science along with their research. So in 2016, we thought it would be a good idea to bring together this type of know-how and create, to honor the multidisciplinary aspect of data science activities, to create what we call the Data Science Hub as a virtual organization, a center-wide virtual organization that invites our community to work with us for further innovation, for solution architectures, applications, translation data sciences, platform development, and also working with industry and the workforce to train our partners in that regard in data sciences.

So Data Science Hub definitely has been an active activity that I've been leading since 2015. At the same time, UC San Diego has an exciting new development, relatively new development -- a center called Data Science, Halicioglu Data Science Institute was funded with an endowment from one of our alumni. So the Halicioglu Data Science Institute is now an academic unit with an undergrad program and an in-progress graduate program at the UC San Diego. But also, it organizes these translational expertise clusters to seed and grow the practice of data sciences at UCSD and beyond. So you can see here the clusters are really cross disciplinary. Not just multidisciplinary, but it's really bringing together multiple departments from UC San Diego, and they are growing in numbers as well.

And for my own research, I've been at San Diego Supercomputer Center doing workflow systems research as it relates to distributive computing and provenance for a long time, and our original name was Scientific Workflow Automation Technology, which was more about the coordination of the computational science activities with scalable platforms. But we've also seen the same change over the years that we were facing a lot more combination of data science, big data with HPC architectures and a lot of the exploratory aspects that come with understanding the data and learning from data to influence science. So we've decided there's a new science here, and there's a new type of workflow. So we were rebranded in 2014 as Workflows for Data Science Center of Excellence, but our goal has been not just to put together workflows, but look at the questions and look at how teams, collaborative teams solve these questions.

So instead of maybe focusing on just tools and technologies, really focusing on how problem-solving happens and what technologies, and development, and applications we can bring to fold to enable this work more. So we've been developing the Kepler scientific project workflow tool for a long time now, and we call it that over the years. And Kepler was applied to anything from big data and informatics to IoT applications like the WIFIRE project I'm going to be talking about today and also HPC applications -- for instance, like your dynamics and drug discovery applications. The common theme here is, in these applications, is scientific problem-solving [inaudible], and you have access to data, being able to explore with data, scaling computational analysis.

And then, you have automated workflows and tools for that. It really fuels a lot of different things -- reuse, reproducibility. It saves resources. It helps to formalize and publish standardized

pipelines for the best of the researchers. And also, we know this -- it's really a great tool to train the next generation. And we've seen a huge adoption of [inaudible] notebook-based workflows, for instance, in training activities like that. So what's the common theme in all of these activities? So to say is they spend big data, computational sciences, data science, cyberinfrastructure, and their applications. So their applications, as you see, is a little bit separated on the side because it's really what we are here for -- these transformational and translational applications that help us to come up with research problems rather than having research problems, then we can go find the problems.
^M00:10:14
And in most of these applications, we see now big data needs to be combined with scalable computing. And when you combine these on-demand scalable computing with data, you can solve maybe new types of applications or you can solve all the applications that need approaches that are more dynamic, that are more data driven. And we've seen a huge rise on these names added to existing field -- smart manufacturing, personalized something, computer-aided something, precision something. These are all signs of that data-driven, that dynamic aspect that are being placed into these agile disciplines. So then, you can say every problem, nearly every problem is being transformed by big data today.
And I like this figure because it points to the fact that -- and this is the IBC [phonetic] prediction, so it's just a prediction -- but it tells us that we'll grow by a factor of 44 from 2009 to 2020 in the data we generate. And it's going to bring us to zettabytes, which is 10 to the power 21. And 10 to the power 21 is also what we see in the universe. You know, we see the Milky Way when we scale that much. So we are, in a way, generating a galaxy of data, and the complexity is growing at that scale as well. So when we look at geospatial big data or the geosciences, the [inaudible] data sets that we have that are collected by blood, sweat, and tears in the field still are really important. Any geoscience lab I think has some of that.
But there's an [inaudible] to hear to combine these types of sensing and measurement technologies with these approaches. And any data, geoscience today uses a version of, you know, sensing, or drone imagery, or satellite imagery, or weather forecasts in terms of geoinformatics-related research. And the way to describe this data is maybe too large in volume to fit in a single server. It comes in different varieties, and in streaming rates, in different types of resolution. You know, measuring something remotely has some veracity issues or uncertainty issues related to it, but we all agree, I think, the combination of such data sets with what we know about our research environment until now provides potentially really high value.
And we have to tap into this value. Similarly, in biomedical sciences, we are seeing a huge rise in using instruments or measurement data to drive simulations, to drive modeling. And these type of approaches, they'll only grow in time. And I think what some of the current challenges in sciences is, how do we increase the speed of what we can learn from the data so we can actually inform simulations and modeling activities and close the loop, so to say, between data acquisition, and experimentation, and simulation all together? So there's definitely, then there's the big data, and let's look at this imaging data coming from, you know, microscopic facilities, for instance. There's an opportunity to learn from this data on a regular basis, not just when we need it, but we can actually use big data coming from imaging facilities, for instance,

and combine it with optimized use of HPC to learn, for instance, in the imaging sense, in the biomedical domain, biological entities and models, we can have some metadata aggregation. You know, we can generate domain-specific markups and quantified metric about these data sets. And when we look at that, what I call the high value zone, so if you push the data to that high value zone, we are also commoditizing some activities. We can definitely create consumer visualizations from these data sets or create understanding of these data sets in a more low-resolution form, in a low-resolution form, and we can also use those to teach the science or what we learn from these data sets to different communities.

So when we actually reduce potentially the computational resources that we need to look at these data sets, we are expanding these communities, and commoditizing the digitalization of it, and improving the understanding, our understanding of these data sets by adding the right annotations to it. So they become a lot more valuable, so to say. And based on that value, hopefully, we become also more actionable. And that action calls for benefitting something, right? Their action calls for, how do we then use this value to answer new questions, to improve our understanding of different phenomena, and to benefit, you know, science, engineering, society, education, and things like that?

So the problem is then, how do we get some data, amplify its value, and how do we then use that amplified value to benefit something, to benefit why? And it could be in, again, science, engineering, business, society, education. It depends on the problem. So then, we need to really focus on the problems [inaudible] to reap that benefit. How does the problem-solving happen? Its required application integration. It's often seen as this thing that doesn't matter. It's an engineering approach. You integrate things. It's not research. But there's real research in application integration. For instance, this is a problem that we are trying to solve actively today. We are trying to understand the composition of sea spray aerosol particles, which are small bits of ocean which come off the breaking wave. And they go up into the atmosphere, where they take part in chemical reactions.

And those chemical reactions, in turn, affect the chemistry of our environment, and also our weather, and things like that. And when you look at this solution that was drawn on the board at some point, there's GPU-enabled molecular dynamics in it, which is a really potentially, going forward, [inaudible] scale application. There's large-scale PCA and clustering. There's micromodel optimizations. They can be high-throughput computations. There's continuous access integration, transformation of data. And there's this information that goes back to the experimental design. And we'd like to do this in a way that improves scientific communications or producibility [phonetic] and quality assurance. When you just look at the GPU-enabled molecular dynamics part, it can take more than a million core hours for a system of this size for doing that computation.

So it's a really complex problem. And when you draw that in a nicer way, in a more virtual-like way, you see the application integration quite clearly. We see these experimental data coming in to also combine with computational parameters that we need to also understand and manage as data sets, and prepare the data, and run the [inaudible] molecular dynamics, [inaudible], and Markov State Models. And all of this is connected. And when you solve a problem, you really need to connect all of it. Maybe not all of it at once at times, but as dynamically as possible so we can actually cut the time, optimize the resources based on that integration. And all of these boxes are signs of someone or a group. For instance, the MB

[phonetic] workflow was something that we've done together with my collaborator, Rommie Amaro's, group.
And it is there, but it still needs to be integrated to this problem and improved based on the challenge of this problem. So there's definitely new problems still require some updates, but the existing tools can be used as composable methods into such problem solutions. So we can then say from this application, learning from this application, this really required a multidisciplinary science and technological expertise and required integration of many skills with computing. It requires integration of big and small experimental data sets, which can be historical or real time, depending on the problem and part of the problem. They require individual, usage of individual or community-developed legacy tools. Methods to manage and interpret the data. Modeling and simulation. And when we are communicating the results or running these experiments, monitoring these experiment, they require gateway for communities to access visualizations in dashboards.
^M00:20:11
And once you have that, there's definitely a need for long-term active and passive storage needs. So we can bring some longevity to the findings and the products of the workflows and tools. So, and there's life cycle management, data science, computational science, advanced infrastructure. And it comes together through these years, so to say, to integrate them as a solution. And which is the case for any, in my opinion, data-driven problem. There [inaudible] any systems, data management tools, data-driven methods, scalable tools for research optimization and coordination. Entire disciplinary workforce that know how to work together and understand each other. But maybe last but not the least, maybe the most important is collaborative culture and tools that enable groups to communicate.
I think something that we can improve as we go forward. So there's an ecosystem here that we need to support. That ecosystem, so it's people coming together with the question or some data and goes through these machineries, so just to come up with benefits, right, that x and y are 2 sides of this. But it also requires continuous iteration and programmability. We don't have luxury anymore to do things once because we are actually constantly improving and learning from data in this scenario. And very simplified, of course, if you think of a virtual acquisition preparation analysis reporting and action or story and publishing of that visualization and report. We have 2 phases, 2 entire linked phases on data engineering and computational data science, but every phase today needs to be scalable.
And that scale can be from the computational scale. It could be from geospatial and temporal scales. But, you know, whatever we understand from scale is also changing, and we need to provide programmability and dynamic behavior to enable that. So the ecosystem then, what are the best practices or needs that we should have in such a system? Is, again, data driven, as I talked about, scalability, dynamicness, process-driven collaborativeness, accountability, reproducibility, interactivity, heterogeneity, and definitely including this multidisciplinary expertise for what they are experts for, not to learn everything, but really a team of experts working together and creating a harmonized view over the solution.
So I'm going back to amplifying the data that related to x and benefitting y because I'll ask now, what if x was wildfires? This is the question we asked when we started with the WIFIRE project. How can we benefit something? What's that something? What's that question we are trying to solve? Because we have access to data, big data let's call it, related to wildfires, and we thought

we can definitely monitor this data. At that time, we were able to monitor it. This was about 2008, 2009 we were really pondering this question. We can definitely do more by visualizing it and add value or show the data in ways that are more understandable. Not just streams of data that we monitor, but visual products that make the understandability of these data and learning from this data more.

But maybe even more, we said we can do fire modeling. We can use this data to improve existing fire models. Maybe we can do even fire modeling in real time. Every 15 minutes, we can come up with a fire model. Why don't we have that? Because there was, working with the fire departments at this time, there was a huge need for real-time fire models. The situational awareness -- what happens when a fire happens? The single-most question people ask is, where is the fire and where is that going to be? And there's very little understanding of that during the fire for many different reasons, but wildfires are happening, and they will be happening more. And our preparedness for it heavily relies on our understanding of the behavior of the fire -- its direction, where it's going to be, and its rate of spread.

So that was the question: How do we better predict wildfire behavior using big data? And we know this -- there were tools, there were fire modeling tools, there were real-time information, satellite imagery. Even in the San Diego region had advanced network of [inaudible] stations and a connectivity through this advanced network to emergency command centers. So we had all the bits that I call composable bits there. We had also computing systems that we can scale any model. But there wasn't a dynamic system integration of that. So there's a little bit information on this integration from 2014 in an NSF video, NSF Science Nation video. I put the link here if you'd like to watch later. But [inaudible] to let the WIFIRE system is and how we turned this question into a project that now is being adopted by fire departments.

So in, you know, after some iterations, WIFIRE project was funded by NSF under the [inaudible] program by [inaudible] support. And here are the senior personnel's names, or the names for the project and all the collaborators. But there is a huge number of individuals, actually, who took part in the project, including our advisors from different federal agencies and fire departments. But how did WIFIRE use what I talked about -- life cycle management, data science, advanced infrastructure, computational science -- and brought them together? This picture to the right is actually the original architecture we took on our proposal. This comes straight out of our proposal. We said they'll communicate this data to workflows, and they'll turn these into bits of information for different types of receivers and just ways and portals, gateways [inaudible].

And we'll use models, and coexisting models, and computing facilities, and archives while doing that. So we'll combine historical and real-time data and do this. And at the same time, we'll maintain geospatial databases and visualization services along with the rest of the scientific workflows we are putting together. And data life cycle management, this picture was the data that we have access to and how we created models over them to serve this data to anyone who can use as open services. Those services that are available are through our GitHub site and linked from our website. If anyone's interested, there are also links at the end of my presentation. We had these real-time data sets and historical data sets.

Can we provide curation with action transformations to these data sets with what we know automatically and using the up-and-coming tools and open geospatial standards? We said, you know, we'll then combine these fire modeling tools and data assimilation capabilities to grab

what we learned from data and what we reduced the data to in real time and learn from the dynamics of the existing fires, so to say, to enable data assimilation into modeling tools. And while doing that, we'll use existing infrastructure in computing, data storage, networking, visualization [inaudible] visualization [inaudible] we have access to as a team or we built as a team, I should say. And most of these, if you look at the pictures here, a lot of them are enabled by prior investments from NSF and other agencies.

But if I just list the NSF projects that made Kepler, Comet, PRP, XSEDE, CHASE-CI, and [inaudible] ITR, CBI, SPCI investments, and REAP, GEON, SEEK, RoadNET -- these are all previous NSF projects we as a team worked on. And the NSF [inaudible] is the advanced network that I was talking about that was built for NSF, to supply NSF support for high [inaudible] violence resource in education network. So bringing education and research networking capabilities to the researchers back in the backgrounds, out in the backgrounds of San Diego at the time. And it was then used and sustained by fire department support in the long term. So these are definitely the 4 bits, but again, there is that machinery here in the middle.

^M00:30:00

And what is that? That's this top part of workflow management and enabling interfaces. For the process integration of these bits, we used, since we are more familiar with Kepler reuse, Kepler workflows. And we use maps as an enabling interface to communicate the findings through these workflows and run these workflows behind those map interfaces. And if I were to provide a reference architecture for such a project, this is, indeed, what's going on. We see any computing system as composable and managed through a research, resource manager. And the vision here is then there will be composable services that can utilize these composable systems. And the role of virtuals, in a way, here is moving them to coordinate these composable services, and bring optimization on the use of these composable services and systems together, and then communicate the outputs to different interfaces.

They could be gateways or maps, as it's here, but it could be online tools for research, publication, and things like that. And I'll also show some examples of that towards the end of my presentation. So what happens is then real-time data comes together with historical data and also what we learn about the fire in an ongoing basis goes into these workflows that execute fire models on computational platforms, and the results will be turned into maps and visualizations on these maps. And big data is being used or data is being used in real time to learn from the actual wildfire. So we are doing parameter and state estimation into fire models to adjust models as we go in automated workflows.

And I get asked, what are the curation of the data? What are the machine-learning case studies? There's so many of them, but, for instance, we can do prediction of data on the conditions based on location, not just if, you know, temperature's above a certain and humidity is above a, below a certain value with gusty winds. We have a sense on a condition. But that doesn't mean for each location. Or we use multi-spectra imagery for high-resolution, you know, satellite imagery for detecting vegetation and classifying maybe Southern California fuels. That's something I'm going to talk about. But the problem here is all required dynamic and periodic access to data in a programmatic fashion. So we couldn't just, every time we needed, pull this data. If you didn't build those access services to these data sets, it would be really hard to even think of doing some of this analysis of the data.

So I mentioned the high-resolution satellite imagery. This is some type of workflow that does vegetation classification from this satellite imagery. Our students, for the last 2 years, have been looking into how we can use new network-based approaches and combine it with active storage and GPU-based platforms to really, on a regular basis, to get satellite imagery and turn that into classification of the vegetation in the Southern California area. So it's still in its infancy, so to say, to be useful, but I think we get some really promising results in the sense that what is now updated every 2 years may be this type of approaches, together with the fuel database, experts can be turned into things that we can feed into fire models on a regular basis, like on a weekly basis, we can update fuel databases maybe, if we use these approaches and combine with fire models, which, in turn, provides better fire models.

But then, these workflows, and LA Fire Department Chief, Chief Terrazas, provided us with good leadership, and then he noticed what we were doing through a [inaudible] magazine ad for UPSDI [inaudible] in a funny way. He called me and said, "I would like to support your research, and what can I do?" And at that time, we had a problem with having access to real-time fire data in terms of the fire front. We were told, you'll never get that fire perimeter in real time. So he provided us some helicopter time and, you know, whenever a fire happens in the LA area, we get their perimeters in real time.

And in turn, we thought of, you know, maybe they can use our tools. But then, it was the, how will they use our tools? And we won't just give them workflows to execute on their own. Something needs to enable their access to these tools and their understanding of the fire. Firemap tool, in essence, was born out of our collaboration with LA Fire Department to then bring an interface to access to information related to fire, model real-time fire behavior, maybe analyze what is scenario for certain region, and generate sharable reports of these fires with others. So it's a real-time fire modeling interface. When you do the models, it shows you up to 6-hour prediction of where the fire will be based on real-time data, or forecast data, or data that then will be adjusted by the firefighters.

This was August 2016 Blue Cut Fire, for instance. If you look at this, the nearest other station here that shows temperature, and wind speed, and direction is quite far from the initial point of the fire. What our team did was listening to firefighters' radios to put in the best weather conditions at the initial point and generate this model. And once you see these orange that are here, it's a little behind the model, is the satellite detections of the actual fire 8 hours into it. So we see that actually the model, well, if you give it the, if you curate the information in a good way, it can get mere accurate results with these models. And maybe it's different for each environment, but what the firefighting community always told us is accurate is not the biggest issue for us.

We are after fast. When a fire happens, we want to know, especially these wind-driven fires, we want to know where they will be to get ahead of it because they are really hard to control. But if you can get ahead of it, you can manage it better. So there are definitely many different fire models, and there's room for really accurate models of the fire. But our approach has been to use the far-side model, and our [inaudible] collaborators have a similar data assimilation based model that they are building, actually, themselves. And assimilate the real-time data and learn from the dynamics of the fire to adjust the model themselves. And over time, we also explored that application of the model to different areas to see if we can actually move into different geological, geographical areas and explore with data sets there.

And we have some ongoing work with Tahoe and Nevada groups there doing AlertTAHOE and AlertNEVADA camera systems and sensing systems. And I imagine it takes a village, right. It's a huge team. The leadership is a lot of people come together, including a lot of students, and postdocs, and partners from fire departments, advisory boards with different expertise in the affiliations. And then, what's the impact of such a system? It's a public impact. Actually, surprised us, the interest of the public in this system without us ever [inaudible] it. As you might remember, in fall 2017, we had a huge number of fires in the California area. And throughout these fires, in 2 months, we had 800,000 unique visitors access our service for situational awareness. We are not even making fire models available to the public, but we had 8 million plus hits on our site.

To understand, you know, what the fire perimeters that were announced are, what's the weather looking like, and what are the red flag alerts, and things like that -- there was no integrated map that could provide this information at 1 place in a dynamic way. And 8 million hits means we need on-demand scalability. And the fact that we built it with what we know or the best of what we know about scalability and computational approaches really helped here, that [inaudible] of these workflows and services that we built and being able to actually on demand scale it on the Pacific Research Platform at [inaudible] computers that are also funded by NSF really helped in this regard. Another thing is the collaboration with fire departments. We are collaborating with more than 10, maybe 20 fire departments over the next, last couple of years, and, you know, some of the LA County fires that was, that these products were used and the models for that are here. And recently, actually, LA mayor tweeted a model saying that this is how much was saved. You know, if there was no fire suppression, this is what could have happened. So I think an interesting usage is coming from there. We also started some public-private partnerships in the San Diego area with San Diego Fire Department and General Atomics. So what happens now is when this partnership is called, when a fire happens, SD [inaudible] gets together at the emergency [inaudible] center.

^M00:40:43

[Inaudible Speaker]

^M00:40:44

In December. General Atomics aircraft flies over 18,000 feet. Anything today about the firefighting, you know, activities. And gives us real-time perimeters of the fire for [inaudible] adjustments. And once those models are generated, they are fed into the state SCOUT system. And this is all made possible with this integrated NSF cyberinfrastructure, that which, you know, thanks to the NSF bills, and it inspired this system.

And once you push it into the state system, you can look at, you know, [inaudible] zones and make some decisions based on this predictive capability. The same system was called in the Thomas Fire, but not when the fire started -- way into when the fire started. And here it was used for looking at scenarios. What are scenarios for back-burns to manage the fire? You know, if you burn here, the fuel would go away. Would the fire go further, for instance, by fire behavior analyst there? And if you look at the satellite detection, then after the fact, we can see how this fire actually happened.

It had 3 waves of sentinel events -- 1, 2, 3. And we were actually there right before the third one. And those satellite detections really helped later on seeing how the fire moved, what kind of weather effects happened. Let's come to the educational impact. I mentioned the number of

people involved in this project, and a huge number of them were students. I think having an integrated infrastructure and composable service idea provides students opportunities to be a part of the solution. They see [inaudible] and they can fit their integration, integrate their individual methods into the whole picture, and they can test alternative methods. So it really has provided a big value for us to train students. And also, we had conversations with the firefighting academy and different fire departments. They're up-and-coming firefighters, and they might not be yet fire behavior experts or even their understanding of the fire behavior needs some tools to explore with.

So visual exploration of environmental data and fire behavior really provides value in that [inaudible] as well. And then, we also used the services for using data in our teaching and projects. For instance, we do some MOOCs to hundreds of [inaudible] learners, and they will have access to environmental data through these services. So suddenly, it's programmable and open, and anyone can use it to teach geospatial data processing. And what's the scientific insight? I think it's beyond fire modeling. It's many different areas. So it's the typical, you know, the sum of those parts is bigger than the sum of its parts. You know, the whole is bigger than the sum of its parts. And I think, if you look at the WIFIRE publications, and related work, and news articles even about WIFIRE, you see these, really, the system is built for extension, for diverse geospatial streams in geographical areas.

We are starting to explore with multiple hazards and application to different scientific and societal problems. And some of these have been in, for instance, using the satellite image analysis that we did for vegetation, for instance. We use it for demographic analysis of [inaudible] in Mumbai that you've seen is not here. Or we work with UNICEF to locate schools in rural Liberia. And we also work with different data providers like UNAVCO and NEON. The right, bottom-right image shows that we can now actually stream the data into these lab interfaces and put different workflows behind to access this data and analysis the data. In this case, we just have [inaudible] workflows as prototypes.

And once that's done, it's all programmed, and the workflow itself and the data that was used in the analysis with the results, you can create research object bundles. So it's like the whole analysis as a research student and for this research object that they can now publish through these enabling interfaces. So I think these point us to the importance of integrated cyberinfrastructure and how we can build and sustain them. And I think in any integrated solution architecture, as seeing in my examples, depends on many prior investments. And, but it also increases the impact of these investments. Individual middleware and software really shines in these solution architectures in terms of impact. But the integrated solution itself is research. It's the translational research bit. It's an engineering and research combination, so to say, and it needs to be funded separately.

And I think [inaudible] and NSF has been phenomenal in enabling such systems so far. And long-term sustainability, of course, is still a challenge for these type of systems, but I think it depends on getting users engaged from the beginning, at least at the enabling interfaces part. And what the system needs to be -- our advisory boards really, really influence what we've done in terms of fire modeling. And our collaboration with LA and San Diego Fire Departments really influence how we communicated the information to firefighters. And they've been our advocates to the community in that regard. And now, through their initiatives, we are on our way, well on our

way. We have some of the subscriptions already in place. Well on our way to the Firemap tool being supported by fire department subscriptions.
And our goal is, to back this up, we've continued innovation through research funding to make these tools better and slowly graduate new generations into these platforms for further subscription. I hope this gives some inspiration or, you know, summary of lessons learned in WIFIRE. Of course, we figured out a lot of things, and we are still figuring out a lot of things about all of these points as we go. And we collaborate with many wonderful people, and parts of the research are funded by different agencies. But we are really grateful to the NSF and [inaudible] support in helping us achieve this. Thank you.
>> Thank you very, very much, Dr. Altintas. That was a wonderful presentation. I'd like to let those of you who are online know that if you'd like to ask Dr. Altintas a question, if you would send me an email at awalton -- A-W-A-L-T-O-N -- @nsf.gov while we're on for the last few minutes here, I will be very happy to pass that question along. And so while I'm waiting for all of you to curiously type on your screens, if I could take a moment, Ilkay, and ask, you worked in a large number of cyberinfrastructure areas and obviously across a huge number of application areas of atmospheric modeling, of fire modeling, biofuels, many, many different areas. Let me ask you to take, and step back a second, and give your opinion: Is there a particularly important next step or next direction for the kinds of activities that involve so many different fields and capabilities? And if there is, are there particular challenges or biggest challenges that need to be overcome to reap the fruits of that direction?
>> Great question. I don't know if I'll be able to list one, but I'll start with 4 as we were going through questions.
>> Okay.
>> I really think collaboration is key. We need to create a collaborative culture and back that culture with tools that we can quantify collaborations and enable collaborations. So we are putting some research into this area because I think this is really needed to solve problems and look at different aspects of the problem -- look at scientific aspects of it, performance aspects of it, even our accuracy aspects of it, or even ethical aspects of it. We need to be able to measure the success of collaborations and impact of collaborative work from different perspectives.
^M00:50:01
It's really important to move computational data science forward. The other is this composable systems idea requires a dynamic resource management and tools that can take advantage of this dynamic resource management and utilization of the plethora of scalable computing platforms from lower scales to [inaudible] scale, into utilization by different services. And the other is, of course, typical big data challenges of, how do we even store data and, in the long term, preserve reproducible data sets or reproducible science, is an ongoing problem. And the fourth one is, how do we then understand the impact? You know, once we have these platforms, how do we judge usability of it? How do we actually design them from the beginning to improve that usability of impact?
>> Okay, excellent.
>> Not a complete answer, but.

>> Well, I just asked you to solve the world's problems, [laughs] excellent [inaudible] to that. So I do have a couple of questions that I received from the audience. One of them is, have you published the software stack? Why don't we start there?
>> Yes. If you actually go into the WIFIRE website, there are some links already here at the WIFIRE, that UCSD [inaudible]. Under the Data APIs, you can find the data APIs. And there's a GitHub site for the project that has some of the bits, and the Kepler system that we use for workflows is also available through the Kepler [inaudible] project site. So anything that we've done is openly available, other than, you know, the integrated system, of course, which you have access to.
>> Oh, nice.
>> The, all the composable services are openly available, and that's meant as actually the project that initially [inaudible] just those services. You know, we dreamed of maybe having fire department use it and have the web services to be utilized as the Firemap, but the initial goal was to actually provide these to the fire research and other communities. I think they're all available there.
>> Very nice. Second question that they had here is, does the model take into account how quickly water-dropping helicopters arrive and how much water and fire retardant are dropped?
>> So there's all this room for improvement on those. We now have, through the interface when we do modeling, we can add fire suppression as barriers to the model. We can also disable fuel based on those drops. So we can manually, a fire department, when they are working with the software, they can manually change and put those into account. And that could be, of course, another data set that can be curated and set into the system and can be automated, but that's not something we automated.
>> Very nice. Thank you. I actually have a personal question, and it's based on a couple of comments you made. One that you said when asking question is, collaboration, through all of these different research areas, collaboration is key. And then, you had said earlier in your talk that basically the fire departments learned about you because there was a smaller glim in one of the San Diego magazines.
>> [laughs] Yes.
>> And I will tell you that the very first infrared fire detection system that the Forest Service used and developed was because the researcher that developed it started the program at the Jet Propulsion Laboratory, and one of the firefighting organizations both had children on the same sports team. So there's an awful lot of chance to these kinds of things. Do you have any insights into how we might make those collaborations more robust now that we have--
>> I think--
>> Some options?
>> The human factor is still the first in all of these, that then LA fire chief called me, I was excited. I wasn't like, oh, you know, this is not research. Am I going to talk to him now? So definitely, some excitement in that is needed, an openness to work with the community. It's not something I can publish, all the work we've done, but it's really makes me happy as a human being that we could collaborate. For instance, that was one of the personal impacts, I should call it. But I would like to come back to the importance of marketing -- how that actually happens. When the project happened, it was definitely potentially impactful, and the NSF recognized it. And Science Nation video that I put on this slide were, was created. And through

that, I think it raised awareness at UCSD that this was an important research, and it was used in the UCSD's marketing campaign, so to say. That's how the picture got on the Southwest magazine. And from that, it was, you know, found by the LA Fire Chief Ralph Terrazas. And it went on like that. So I think as scientists, we sometimes underestimate the importance of science communication and how to really communicate the research to public. But a lot of good value can come from those marketing and communication approaches, activities.
>> Excellent. Well, your excitement on this project certainly comes through in your talk, and I want to take just a quick moment to thank you again for taking your time today to provide the overview of all of this excellent research.