BRUNO BASSO:
Hello everyone, my name is Bruno Basso. I have had the privilege to chair the planning committee to organize this workshop 'Exploring a Dynamic Soil Information System'. It is with great pleasure to welcome you here today, and I hope you will enjoy this immense information and data that we will be discussing about one of the most important things that supports our life. I will kick off just by presenting the objectives of the workshop, some of the challenges that we face with these complex systems and present some examples of some complexity again, and how we could use these examples during the workshop.

So, next please. I had the privilege to work together with this elite group of scientists, Ranveer Chandra from Microsoft Azure Global, Alison Marklein from University of California, Charles Rice, Kansas State University, Jim Tiedje, Michigan State University, Kathe Todd-Brown, University of Florida and Rodrigo Vargas, University of Delaware. Obviously, this workshop would not have been possible without the tremendous support and help that we got from the National Academy staffs mainly Kara Laney, Esther Stein, Sarah K'won, Robin Schoen and Deb Glickson.

Next, please. And a special thanks goes to the sponsors of the workshop. The fund after they fund the National Academy of Sciences, the National Corn Growers Association, National Science Foundation, The Nature Conservancy, USDA NIFA, USDA NRCS, and the DOE ARPA-E.

Next please. So, some of the questions as ideally, you know, the objective of the workshop, we really would like to cover, as you may have seen already from the breakouts room and the program, we will discuss, be discussing what should we measure in order to build a soil dynamic information system? Before that, you know, what it's very important to realize that different things have different meanings and different variables for different personas involved in making the decision. So, why should we measure and where should we measure when and how frequently? If we can't measure, what are the proxies? The complexity of harmonizing, storing, retrieving, and also delivering data and also how we use system models to predict and simulate the impacts of soil dynamics.

Next please. So, we needless to say, soil supports our life, but it's so complex and dynamic. And you'll be hearing a lot of this terminology, but why is it dynamic? Because it's highly affected by the weather and the human activities, weather with all the components, affecting processes that occur in the soil as well as affecting the plants that are hosted in the soils affecting the roots dynamic and decomposition, nutrients. And so plants and microbes are highly dynamic over space and time in their behavior. There is a inherent variability in soils. The way they were formed, topography, texture, hydrology, how water flows and how they weather over time. Management is a critical piece. The way we manage or till the soil versus not tilling, if we use a permanent cover that, again, several others management strategy plays a critical component including agrochemicals which have a strong impact both on the plant dynamics as well as on the microbiomes of both plants and soils.

Next please. So, some of the uncertainties that we are very well aware and we'll be discussing. There's obviously spatial variability that will come quite often today I'm sure varies from micrometer to meters

to kilometers. How many samples do we take? What is the volume of the depth of the samples? And how frequently do we take temporal variability in some of the variables in soils is very dynamic process. It could vary from seconds to centuries. And, again, stakeholders involved in making this decision and using information are critical.

What quantity of soils? One particular attention needs to also be dedicated to the method of sampling, whether it's hardcore or the new semiautomatic or fully automated coring. Spectroscopy is very important new sorts of disciplines to be able to collect information from a distance, but they also are not free from uncertainties. Sample, how we sample the soil sample, the processing is critical. Do we go with what level of physical disruptions? Some can be much, much harder and destroying the aggregates, for example, to a point where even the results are affected. So, there is even sieving extents. And so there will be a discussion during the breakouts, hopefully, on how we process samples because it creates a level of uncertainties there.

Chemistry as well interference in the instrumental temperature uses. The reason I say that because as you'll see, next slide, there are aspects in relation to this components. The spatial variability, for example, this is a sampling that's done on a farm and it was kindly shared by Will Barrington. And you can see that the soil carbon, the mean of the soil carbon is a result of the widespread of results across different ecosystems. But if we want to capture a 90% precision, and be within the 10% error, the multiplier using the statistical model, to be able to capture this variability, obviously, makes the number of samples be quite different than what we took. And so, for example, you know, there could be a multiplier of 45 times greater up to reaching 119 samples, if we wanted to capture fully the variability.

Next slide. Laboratories, you hear about that. Even though we're very professional and we're doing the analysis, you still can't help because of the procedures and what I mentioned about the way the samples are processed, we can have range differences. These are, again, an entire farm that was sampled with 96 samples that varied from 1.25 or so to over 1.4, which equates to about five tons of kilograms of carbon kind of variability of error between the different labs, as well as he has now shown that bulk density is a critical component, spatially variable. And so a 0.2 cubic meter of a meter of soils could lead up to eight tons of kilograms of soil organic carbon.

Next slide. This is a wonderful work shared by my friend and colleague in Michigan State Sasha Kravchenko just to show now the scale, the extreme end of the scale of X-ray computed micro-tomography phosphor imaging for radioactive detection as well as enzymes. And so, with this technology, very advanced, I'm sure Dr. Cornelius will cover some of these aspects, we're able to even understand a little bit better the possibility of the addition of new carbon as well as hotspots for indoor productions. And so this is a very promising fields but it's very far from, it's in a mechanistic basis.

And next slide. If we integrate space and time, this is some of my own work, as you see the map on the left corner, just constant change from one year to another about crop yields, soil doesn't change, obviously, in terms of texture, and so it's really the integration of several components affecting that yield change. And so over time, we've learned that there is some very strong components of soil in

explaining the stability in crops, but there are areas that they could just not be explained by soil variability alone, because they vary from one year to the next and so positioning the landscape and weather play a role.

We also learned that the possibility of using thermal imagery from remote sensing is a very nice proxy to be able to detect a variable in soils, which is very complex to measure, soil depth. And so in this case, where you see the map in blue into thermal stability cold and stable, it is really a proxy to show how much water is available to constantly provide the water that is needed in the evening years where it was a little bit drier, so that that area kind of matches the high and stable zone. So, we've been able to scale this work across the whole Midwest, over 80 million acres. And there is a pattern of the majority of a good chunk of the field being high and stable where soils plays a role in others, unfortunately, for several reason, it's constantly underperforming. But it's the unstable zones that they also need to be pay attention because they are a result of the much more complex and dynamic interaction between the system components as you see in the next slide.

So, in this work that we did, we included with Rafael Martinez-Feria and I recently, we showed how much variability over a very large dataset of several hundreds of fields of historical yield over space and time, so soils were able to explain only about 25% of the yield variation. But when you included in the second model there, when you included yield history, then you would learn significantly more because of, again, the stability in soil playing a big role there, as well as understanding what drove this instability as a result of climate. So, soil alone is not sufficient to explain it because it's an interaction between position in the landscape whether in crops. And so as you see here, areas that they are affected by excessive amount of water, if they are in a depression, obviously, they don't produce as much, but if they are, if it is a dry year, they will produce more on those areas.

Next slide. This is something that as really got my attention because if we're investing a significant amount of time in modeling and both whether our crop models or geochemical models, and so if you see, that's on the upper left side, that's a soil map, that's simply replacing the properties of soils to run the model for crops. And the model does a relatively good job, but except in the blue areas where that is a good soil and so, in comparison with the measurements of the actual yield that you see there, that areas it was actually quite a bit underperforming that year. But if you run the model by parameterizing the zones based on this, again, this interaction and moving beyond the static properties of soils where topography plays a role, then the model was able to capture the low yields because of the depression and water standing and be able to affect plant heterogeneity and be able to capture features that just soil alone wouldn't be able to be captured.

Next slide. This is again just a simple indication that we will be covering it and we have a significant interest in spectroscopy and remote sensing actually is a promising tool to be coupled with sampling. This is an example of with hyperspectral handheld radiometer be able to characterize clay properties in a soil just from a spectral signature as you see the different behaviors in the electromagnetic spectrum or either be able to detect soils that are more degraded with this ratio between the calcium carbonate and clay from imagery.

Next slide. Just to conclude the geophysics and underground measurements played a critical role. One thing that has got my attention over time because of this dynamic component of management are tillage effects properties. And so this is an example of electrical resistivity tomography which is approximate the inverse of conductivity.

And so as you see before, a tillage event, following a tillage event, the amount of porosity and so highly resistive it's just that it's very well detected, but there is a very dynamic changing in play basically underline the importance of when measurements are taken. And the maps on the right, this is a profile to show the impact of a, this was a 15 years conventional tillage study where you can see very distinctly the layers of conventional tillage, no tillage, the possibility of having a more uniform and deeper soil where roots can explore deeper layer. But if we do one tillage event, basically you destroy all the structure as well as the CO2 that is emitted. So, there is a particular attention needs to be how we manage this system, both over space and the impact on depth of soils, as well as CO2 evolution.

Next slide. I thought that there was just really a teaser to kick off. And so we're gonna move into some of the core of the workshop with the next sessions being invited speaker. Before we do that, I would really encourage there is a Slack workspace that that you may have seen. So, you are encouraged to pose questions there and follow. I will be reading, there'll be people monitoring the Slack channels. And so we hope to get an interaction from all the participants using the workspace. Next. So, I also would like to point out that on the webpage, there is a soil repository. And we are building it to we encourage people to add additional material there. And so it's you can see that in the web link. The next slide please.

So, it is with my great pleasure that I am going to kick off the official start with the keynote speakers. And I have a real, real honor and pleasure to present the three speakers today. I will be starting with Dr. Cornelius. Dr. Joe Cornelius is the Chief Executive Officer for the Bill and Melinda Gates Agricultural Innovation, known as Gates Ag One, which aims to ensure high quality cutting edge crop innovations are available and accessible to smallholder farmers in developing countries. So, Joe began his career on a small family farm and now he brings more than 30 years of experience and continued dedication to improving the world through agricultural advancement. Before the Bill and Melinda Gates, we probably all know Joe from directing excellent programs at ARPA-E. And I also would like to point out that Joe received this PhD in Plant Physiology and Biochemistry from Michigan State University.

SPEAKER:
And with that, Joe, take it away. Thank you very much indeed.

SPEAKER:
Thank you, Bruno, and welcome everyone in attendance today. And thank you to the National Academies for supporting this exciting workshop. For the next 20 minutes or so, I'd like to provide a quick overview of innovative technologies for managing soils. This will not be an exhaustive list, obviously, but rather a subset of diverse technologies under different stages of development from sensors to computational models. Next.

So, one of my favorite quotes actually came out of nature several years ago. And that quote was Science is informed by what is possible to measure and it takes a great leap forward when we can measure something new. Actually, that's a paraphrase from Lord Kelvin's quote, 1883. When you cannot express it numbers, your knowledge is of a meagre and unsatisfactory kind, which I think is a very inspirational context, especially today, given all of the significant opportunities that are starting to converge between the overlap of physical, biological, chemical and mathematical sciences. Next slide.

As all of us on this call are painfully aware that the hidden half of crop productivity resides right beneath our feet. And for 10,000 years of agricultural development, this has been the proverbial black box. Next slide. And Franklin Roosevelt once said that a nation that destroys its soil destroys itself, which actually seems to be the path that we're on when you look at the amount of carbon that we've lost. In the Sanderman's study, he and his team had estimated over 130 gigatons of soil carbon have been lost globally as a result of land use and agricultural practices. Next slide.

And at the end of the day, soil maintenance is...soil health is just good business and it certainly touches on a broad cross-section of the sustainable development goals that we're trying to achieve that actually are fundamental to the survival of our planet and humanity. So, to start with, as we talk about this particular space, I'd like to provide a general overview of the range of new and emerging tools for soil characterization. And obviously, there are clear tradeoffs between resolution and scalability.

So, a range of sensing tools is important to be able to capture the subtle interactions in the soil that determine its chemical and biological properties and scale-up that insight over large environments. Next slide. As an example of high-resolution measurements, the Department of Energy has supported work at Lawrence Berkeley National Lab to develop controlled lab environments to study plant, soil, microbe interaction. Plants in the EcoFAB platform can be grown on soil and controlled media and samples extracted micro politically for analysis. Researchers have established automated handling systems to grow and analyze multiple EcoFABs in parallel.

This provides a real high throughput platform. Using these systems, researchers have identified minimal microbial communities that are stable and can be used to study changes to plant growth and soil microenvironments. So, this is an example of some of these early-stage technologies and the

potential impact that they can have in this particular case, demonstrating their utility for measuring plant (INAUDIBLE). Next slide.

So, slide ahead already. Alright. So, next slide. So, if we look at understanding root growth and function in the soil, it's important for both understanding the performance of crops in the field and the potential impacts of various crops and soil health. When we actually proposed this program to the Secretary of Energy, Secretary Moniz, he thought it was wildly crazy going from a medical MRI device which weighs tons to a field device. We're now on version 2.0, which is what you're seeing in the field here. And we're on our way to version three, which basically will be a flat plate on the soil surface as a new way for us to actually be able to visualize root-soil interactions and growth and development. Next slide.

On a significantly different scale, we have the route tracker which is using electrical impedance to actually detect root development and can measure genotypic differences across different environments. Here we see a maize breeding population that's looking at root growth and development. The part on the top shows root development and the bifurcating between two different varieties under different moisture regimes. So, we're actually developing soil tools that actually enable selective breeding for specific microenvironments. Next side.

This technology, nitrogen detection, has come out of Iowa State and Nebraska and it's being led by (INAUDIBLE). It's in a process of being commercialized by InGenius AG uses silicon-based microfabrication techniques to produce micro needles that contain nitrate sensor electrodes, which will enable very low-cost production at scale. These particular devices will be able to be deployed in a multiple variety of environments for measuring nitrates and the sensors can be swapped out in the future for other particular endpoints that we would want to be measuring. Next slide.

Researchers at LBNL and Noble have demonstrated the capability to measure soil and plant properties using electrical resistance tomography, ERT. As shown in the bottom left image, electrodes are inserted into the soil and the system is powered by solar cells. The system can collect data without human oversight, monitoring voltages across electrodes at the end of a row of plants.

The researchers, in addition, have developed a new model to correlate this electrical receptivity with soil moisture to a depth of six meters, which was sensitive enough to distinguish between 64 individual breeding pots of wheat plants. Greater resistivity is correlated with decreased soil moisture and data collected in 2020 indicated a number of wheat genotypes that had deeper roots and increased water uptake from the soil. I mean, think about this. We've never bred for crops based on root characteristics such as depth and penetration, which allows us to actually start measuring carbon partitioning in ways never before ever imagined. Next slide.

So, let's pivot to something a little bit different. Several years ago, at DARPA under the direction of Dr Blake Bextine, started a program called Advanced Plant Technologies. I remember when I first heard about this, I thought, this is really crazy, but that's exactly why we have the authors. And the primary objective here of that team was to harness plants and mechanisms for sensing and responding to environmental stimuli and extending that to sensitivities that we can actually

measure.

In this particular example, we've got Interplant, which was not part of that original DARPA program, but it's an example of how these technologies then find their way into the commercial space. Using fluorescent proteins that can be engineered into the plants that have been targeted to actually respond to specific stimuli. We're able to create entire new mechanisms for detecting different soil properties and characteristics and correlating those with crop growth and development. And never mind the fact that these plants carry their own power systems.

So, it allows us to actually be able to deploy them in a multitude of ways that we had not previously been able to do with conventional sensors. And if we take this to the next level, thank you to Bruno and his colleagues, moving to landscape observations actually now takes us into the space where having these sensors actually provide us with site-specific…understanding site-specific factors such as the landscape position and dynamic soil characteristics which correlate crop yield response to weather. This is the end game that we're striving for with these particular technologies.

And it's the integration of the chemistry, the biology, the engineering, the computational science that ultimately creates the impact. And it's at the end of the day what's going to drive success for us in this particular space. Digital and geospatial technologies to monitor, assess and manage soil, climatic and genetic resources illustrates how to meet this challenge. Next slide.

Measuring soil carbon has been a real challenge and it's absolutely imperative that we create…invent new scalable tools in order to be able to break the old paradigm of dry combustion, which required a significant amount of sweat equity. And there is a significant time lag in taking those carbon measurements. Here we have three examples. One yardstick is a hand-held rapid visual sensor that can assess soil carbon and soil bulk density. It uses spectral analysis, resistance sensors, machine learning and statistics to measure and calculate the amount of carbon in an area of soil. Meanwhile, Berkeley is working on a non-invasive neutron-based system that will provide real-time soil carbon concentrations.

The Holy Grail here is to be able to actually have significant, tighter spatial-temporal observations in soil carbon that can directly impact management systems rather than waiting for a decade before we can measure differences in management practices. Being able to do that on a time scale that is measured in years or less. And then Impossible Sensing is a firm in St. Louis that actually has developed a sensor that's on the Perseverance Mars rover. How ironic that we're actually taking a lot of these technologies to other planets, yet we're not fully deploying them yet here in the US.

It says a lot about our…some of our priorities. And in many respects this capability to resolve different forms of carbon in soil using time-domain fluorescents is particularly exciting and a new tool that's on the horizon. Next slide. As the technologies described in the previous slides are not amenable to landscape-level assessments. Models are essential for us to scale up these insights from high resolution sensing systems. Data from these systems are used to parameterize and validate models that can be scaled up or applied to different geographies and ecosystems.

There are a number of biogeochemical models for soils such as COMET, which is used to estimate carbon and greenhouse gas impact in agriculture. MEMs is building on existing biogeochemical models to make to better take into account the impacts of plant growth above and below ground a number of soil parameters. Next slide.

Ideally, models will be able to be overlaid on remote sensing data to provide insight on soils across large acreages or entire countries. The University of Illinois (INAUDIBLE) and researchers are taking satellite imagery across the Midwest and in conjunction with existing technologies, are calculated in soil organic carbon concentrations. The SOC values allow the determination of the carbon stocks in the soil, similar to what the model presented in earlier estimates. The SOC maps can also indicate areas of uncertainty, where focus sampling will provide greatest benefit to producing accurate data. Models such as these in Illinois and elsewhere will continue to improve as more ground truth from many of these sensors that we just discussed will actually provide additional information to inform those models. Next slide.

And a program that has been funded by the Bill and Melinda Gates Foundation ISDA has a primary objective of being able to combine current science and low-cost field level agronomic diagnostics with advanced geospatial analytics to achieve increased cost-benefit of agronomic advisory for small scale farmers in Africa, usually an underserved segment of the private sector. This is an area that actually is even in more critical need of being able to improve and manage their soils.

The ISDA Soil Program is a digital soil map of Africa at a 30-meter resolution, incorporating soil samples from over 100,000 locations plus high-resolution satellite data and cutting-edge machine learning approaches. This creates soil properties predicted for over 24 million individual locations, incorporating data from other projects such as offsets and other datasets. This is an open-access data platform and we're excited that it actually is already creating significant utility. Next slide.

This is a horrible example of a soil fertility advisory tool looking at different constraints. You can go to the website, invites you to do that and you can actually see where now you get down to the granularity on a hectare basis where farmers can actually get direct advisory as it relates to the soil characteristics and health of their particular farm. Next slide.

So, to go back to my opening remarks about scale and resolution, the capacity to make measurements at scale allows research to inform policy, assess impact and influence market adoption. Large scale measurements can create a feedback loop to direct these measurements beyond the capability of research programs such as ARPA.E and NSF signals to the soils to provide insights on soil properties. There are socioeconomic and behavioral considerations that affect policy decisions. At the end of the day, the sensor technologies and the research that many of you are doing inform these models, these policies and actually translate ultimately into a significant impact on economic development. Next slide.

So successful innovation requires three critical ingredients, compelling technology, receptive markets and enabling policies. This is a very robust ecosystem. If you were to have looked at, done a landscape map in this area a decade ago, we would have been lucky to see a handful of entities up

there. But with the convergence of technologies and the hard work of the scientists and the funding coming from many of these public programs, we're actually seeing a profusion of innovation. And by working together, we can actually see a path to combine success. This is truly a robust innovation landscape and it needs to continue to grow and flourish. Next slide.

And that's why I'm particularly excited. Later today, you'll be hearing from other speakers, among them David Babson. ARPA-E has a new program on smart farms. And this is an example of taking these technologies and moving it to the next level.

And I'm really excited to actually hear the panel discussion later today as we bring in experts not only today but over the next several days to actually share with us some of the really exciting opportunities to have impact as it relates to soil, health and climate. So, with that, I want to say thank you again to the organizers, the hard work of the team that put this all together. And to iterate Martin Luther's statement many years ago, great science is compassionate science. Thank you.

SPEAKER:
Excellent. Thank you very much, Joe. We will have questions in the time allocated after all the speakers.

BRUNO:

So it is with great pleasure that 'll call Jerry Hatfield. Jerry I've privilege to call a good friend of mine, over the years we've been working together. He's the director of USDA ARS national lab for agriculture, just retired in Ames, Iowa. He has worked at university of California Davis before joining USDA ARS first in Texas and then in 1989 in Ames.

Jerry's research focuses on the interaction among the components of the salt plant atmosphere continuum and their linkage to air, water and soil quality. But he goes beyond that because he takes scales and salt functions very much close through his heart and the evaluation of farming system and responses to input across the soils, capturing field variation is also a critical piece of his research. He has published extensively nearly 498 papers and without any further ado, your service to long Jerry I, the floor is yours. Thank you very much for accepting our invitation.

JERRY HATFIELD:

Bruno thank you and thanks for this invitation to present this and, you know, Joe gave us a good overview of what's going on and we'll talk a little bit about the soils and agricultural systems and all of you've had basic soil courses, you know, that 16 weeks. And so we're going to cover a lot of material in a short period of time, but my view of soil and agricultural systems is, ensuring how we enhance ecosystems in human existence.

And when we really think about all of this next slide is, it really gets down to what I think that we want to understand and what producers want to understand is really what are the functions of soil. And if you look at this diagram from FAO, it has all these different pieces that we want soil to do. It's got everything from carbon sequestration, to climate mitigation, to production and flood regulation, all these different aspects, but when it comes down in the next slide.

When it comes down to agriculture we can kind of boil this down, because what we want is; we want to provide support for plants, we want those plants to be standing so that we can do our cultural operation. We want it to be a very efficient water reservoir and as Bruno pointed out, a lot of this variation we see across landscapes is really due to inability to supply water and all of this, we'll talk a little bit about that.

We wanted also to supply all the nutrients that plant needs, not only to grow but also to produce that high-quality grain or forage or fruit or anything else. Recently, we see all of this aspect in terms of carbon cycling. What can we do to use soil as a carbon reservoir and improve its cycling? And then we really want it to function as a way of decomposing pesticides, antibiotics, things that we do as part of our cultural operations, so that it doesn't have an environmental impact.

So, when you look at the functions soil and we boil them down to these five things. We have to understand that we really wanna look at the functionality and we wanna look at these different dynamics. So let's go to the next slide. And we'll talk about the current state of our soil. Where we are today in all of this, because if we start being honest with ourselves and we're looking at soil, is

that we have had an impact ever since we began to cultivate our soils. Let's go to them next one. In all of this and I just use the examples from the morrow plots in the Sanborn plots, because those are, those are 110 years of history. You look at the plots from the morrow plots in Illinois, if we've had a corn, oats, hay rotation, we've lost 35%, but if we've had continuous corn, we've lost almost 60% of that organic matter. And in Sanborn plots out in Missouri with continuous corn, we've lost 70% and we see the impacts of this and all of this. Let's go to the next slide.

And really think about how agricultural systems have changed our soils. We've removed organic matter through tillage and we look at all this. As we till the soil, we oxidize a lot of that organic matter back in there to, tillage does what it's supposed to do, it dries the soil out then it spurs activity and we see this change going on in terms of this. We also, if we're honest with ourselves that we really have adopted carbon or cropping practices that limit a return of carbon through the soil, we have very monoculture or limited rotation systems that, basically only put carbon into the soil or for particular time of the year. As a result we have reduced the functionality of our soils and we've increased the reliance on external inputs. Is that we, we have soils that have very limited infiltration rates, we have soils that have limited nutrient capacity, so we end up with supplying nutrients, we supply water to those.

And we also have increased our erosion rates and increased soil degradation. If we look across the landscape, we see this changing piece of these dynamics where we have taken higher organic matter soils and we've eroded them down into the slope. We've gone from the A horizon to the B horizon in a lot of our slopes. So we have as humans and when we have farmed, our soils over time is that we have changed our landscape, we've changed our soils, we've changed the functionality of our soils. Let's go to the next slide.

So if you think about agriculture and a lot of this, and when we start looking at these functionalities, the primary factor affecting agricultural systems is water. You see a lot of these upper left the scenes in across the Island in the spring, we've got potholes, even though we're intensively drained, we were standing in water on that. We've got a lot of fields across the United States in which the lower left slide is that we, we have limited new infiltration rates, we have run off and so we're not capturing that water in terms of using it for agriculture, we're putting it downstream. So next slide.

But our views of soil and agriculture is scale dependent. When we really look at this, let's go to the next slide. Is that in agriculture and when we talk about throughout this whole workshop is that, I think we need to understand that sometimes we think about that profile of soil, we think about the A horizon and B horizon. Sometimes we think about what it is in the field scale, where do we sample that, where do we see that variation? And then really what happens at a landscape scale?

And so a lot of our, our sampling, our interpretation and our management is really scale dependent and it's how we look at that system. And I think we've got to look at the system from a way it says, you know, how do we take that information we understand at a profile and begin to translate it into all the variation that goes on to the field and then take that field into landscape and say, what is the

best way in which we can enhance that functionality from all those different pieces. Let's go to the next slide.

Also have to realize that soils are diverse. We've got a history of the soil forming factors. A lot of those covers depend upon what the parent material are and sometimes we have soils that have, very consolidated parent material and rocky soils in there. And then we've got this really good profile of soil that's extremely deep, our organic matter all the way down through that. So, soils are diverse and so we have to realize that when we talk about these different dynamics, is that we are dealing with a lot of history of management, a lot of history in terms of change, all these different pieces that go together. Let's go to the next slide.

But we'll also have to realize that there's an independence, interdependence of soil functions that, you really can't separate soil water from nutrient availability and you can't really separate support for plants from soil water variability. All of these things that we talk about for functions are not silos, they are really interdependent when we start talking about soil functionality. Let's go to the next slide. And if you just look at this, this is a study that we did. This is, we related mean County yields across the Kentucky, Iowa and Nebraska and soybeans in Kentucky and Iowa and corn, this is related to the national commodity, crop commodity productivity index.

It's one of the databases out of NRCS. You look and say, you know, the better the soils the higher the average County yield and we see a lot of terms in this, the Nebraska data is because we only selected irrigated counties. And so if you could manage the water, the quality of the soil is not important, but if your rain fed agriculture, quality of the soil and how it's influencing this becomes very important. Let's go to the next one.

And there's just the variation of the NCCPI across the Midwest. So we start looking at this at scale and saying, you know, at the County level, here's the good soils and we'd move our agricultural systems up and down, is that we're not gonna always move them into high quality soil. So we're gonna see a lot of variation due to that weather component, due to that climate component as well. Let's go to the next one. And so what do we need to know about soils for agricultural systems? How do we enhance the functionality of soils in all of this? So let's go to the next slide.

And if you go back and here's what we know about soils and water. We know that all the textures have different water-holding capacities. That center graph says, you know, what's available water if you've got a sand, I mean, there's a little difference between the wilting point and, and feel capacity. Get that silt loam, there's a lot of, a lot of available water. Hudson showed that as we improve organic matter, we improve that water storage on the different soils in all of this. So when we think about these dynamics, you know, what do we need to understand and carbon becomes a major part of that as we go to the next slide. You know, we have an impact on productivity, you know, we see these low yielding parts of that field that are the poor quality soils in that, with limited water and then you've got other parts of that field that have high yields because they have that available water during grain filling. Let's go to the next one.

Bruno this is just a repeat of, Bruno this, I think that we need to spend a lot more time looking at how these unstable and stable zones are created, in terms of their parent material, but how do they relate to the functionality of supplying nutrients, supplying water and supplying support for all of this and then how can we look at our management practices to change that. Just go to the next one. Here's just another field variation, we see all these different parts coming through in all of this and so when we look at this, just go to the next slide, a lot of this really comes down to the variation of water holding capacity.

And so when we begin to think about this, you know, these (INAUDIBLE) variations and is a temporal component. And so it's not only what we have as the capacity, is what we have as that weather component and as well as the crop growth component. Because if we would look at wheat for example across these fields, we wouldn't see the same level of yield variation that we see in corn, just because of the different dynamics of the season and rainfall patterns. Let's go to the next one.

So the central question is what can we do to change soil-water availability? Let's just talk about that function for a little bit. And then what do we need to know for information to evaluate the effect of these changes? Go back to that slide in terms of texture, we know we can change it with organic matter, how rapid is that change all these different parts of that soil. Let's go to the next one. And so we have to face reality of our agricultural systems and that reality is that we are dealing with a carbon cycle, a water cycle, a nitrogen cycle, for example, and all the other nutrients cycles that go with that.

So agricultural systems are comprised of a number of processes and cycles. And so if we really wanna look at how we change our system, we have to realize that this interconnectivity and these cycles overlap, they all work together and we need to be thinking much more about how do we put our science to begin to understand these dynamics and interdependencies. Just go to the next one.

If we just look at this in terms of just changing our soil out there, you know, just an example of soil organic carbon change in that is that, bigger impact is how we're managing our micro organisms and fauna within that and then we get down to clay, then you work our way down this. And so we can't separate any of these, but I think we need to understand what may be causing the bigger change in our overall system and how do we measure that as part of that overall component. Let's go to the next one. Just click through those.

Here's a system of these we've been working in, Northern Iowa, this is three different fields. The organic matters started out quite low and then we switched that field to no till, strip till, just click onto the next one. There's been about a two and a half percent increase, you can see in a lot of cases that we sampled fence rows, now we're at 69% on our fence row.

So, we're slowly improving this and the major impact of this has been just reducing tillage within the system and we've been adding cover crops lately, but you look at all this, is we can change our soils. And what we see within these fields is we have changed those stable-unstable zones, we've made those soils that had a lot of variation to have little variation over time, made them much more weather resilient, just because of how we were managing water. Let's go to the next one.

So if we wanna evaluate changes in our soil is that we need to understand that, we have a lot of interactions or processes, that we also need the history of soil management with it. Not only is, do we look at what we're happening now, but what has happened in the past. We need to understand what information is required about all the soil response to management as well. And if we go to the next slide.

So why do we need to have the soil's information? Is because our efficient production, whether we talk about a water use efficient system or a nutrient use or light use efficiency, requires that we understand the functionality of soils. And that the functionality of soils are linked with climate and management in order to produce crops and livestock. 'Cause that, now we get this history impact, we get the current management, we're gonna have to figure out how do we put these technologies together, but the challenge is going to be, how can we simultaneously increase our functionality and production efficiency, through the use of information at different scales and Joe, provided a excellent overview of how we can begin to look at this. So with that Bruno, thank you for this opportunity, give you things to think about throughout this, this workshop.

BRUNO:

Thank you very much Jerry, that was fantastic.

BRUNO BASSO:

It is also a fantastic, great pleasure to have Alison Hoyt. She's a post-doctoral research had the max Planck Institute for biogeochemistry and then also affiliated with Lawrence Berkeley National Lab. Have work addresses, how biogeochemical cycles respond to human impacts. With a particular focus on most vulnerable and least understood carbon stocks in the tropics and the Arctic. She completed a PhD from MIT in 2007, and she will be starting in Assistant Professor position in the Earth System Science department at Stanford University. Alison, thank you so much for accepting invitation. The floor is yours.

ALISON HOYT:

Thanks so much. So today I'll talk about the importance of data archiving and data integration. Which cuts across a lot of the different themes that we've talked about today. But I'll particularly be focusing on the role of soils in the global carbon cycle and implications for climate change. Next slide please.

One of the central questions that were interesting in addressing is, how our land management and climate change impacting soils? So, these changes can be driven by human impacts. For example, here in the case of soil erosion. Our next slide, they can also be driven by climate impacts, for example, in this this is a thermokarst pond where you can see the impacts of permafrost in Siberia. Next slide.

So, in both of these cases, we're interested in what's happening to vulnerable soil carbon and whether it might be released as greenhouse gases to the atmosphere. So, we're really interested in seeing the impacts of management and climate change on soils. Next.

However, there's also many feedbacks to take into account. So, it's not only that soils are being impacted by management and climate change, but they also have the potential to in turn influence climate change, for example, through management practices that can help us mitigate the impacts of climate change. And in all of these cases, we're really interested in understanding these changes and seeing how quickly they might occur. Next slide.

So, some of the central themes that we'd like to understand with respect to soil carbon. To get at these challenges are what are the current soil carbon stocks, and then how might they change in the future and on what timescales can we expect these changes to take place. Next.

So today I'm going to be talking about soil databases and how we can leverage soil data to answer these questions. So, there's two major ways that this is currently happening. The first is through continental scale sampling efforts and data organization efforts. And the second is through grassroots efforts to organize a past data around certain particular scientific questions and understanding of processes. And hopefully you'll see today through my talk that both of these efforts are really needed and fit very nicely together in complimentary ways. Next.

So, the first I'll talk about continental scale sampling and data organization efforts. And these are often really top-down efforts that are very systematically structured. They give us a very excellent snapshot of the current state of soils. I'll share some examples and then talk about the strengths and weaknesses of this approach. Next. For example, in the United States, the rapid carbon assessment has provided

systematic sampling of soils at high resolutions across the US. And this is really helpful in enabling us to reduce our uncertainty and soil carbon stocks. Next

NEON takes a very different approach. And this allows us to have a better understanding of processes at a much more smaller, more limited number of sites. Because in this case, we have systematic sampling sites that represent different ecological regions and this soil archives and soil information can now be used in conjunction with other ecological information that we need to understand the role of soils in the terrestrial carbon cycle. Next.

One other program that I'd like to highlight from Europe is the EU soil observatory that was just recently launched in 2020 and promises to provide a really dynamic and exciting platform for soil knowledge and data. And this works in combination with the European soil data center. A center for data sets of maps, documents, and ongoing projects and events related to soil. Next slide.

It's not only in Europe and North America, that these efforts are evolving. As mentioned earlier in the keynote, you also heard about how there's increasing efforts to sample systematically across Africa through the African Soil Information System. In this case over 20, around 20,000 samples were collected systematically and analyzed, and these are enabling a much better mapping of soil carbon, as well as soil properties across the continent. Next.

So, these continental scale efforts have major benefits. They allow systematic sampling and data organization making the data organized, keeping the data organized and making it available to a wide number of players. They also employ careful sampling design and consistent standards. So, in the case of analytical efforts measurements are often made by the same lab or with shared standards. And in terms of sampling design, sampling schemes are repeated consistently. And brings much better cross comparison. Unfortunately, the con of this approach is that it only gives us a very good snapshot of time currently. Because a lot of these programs that are much more recent. And measuring current is not always enough to answer key questions because we can't go back in time. Next slide.

So, to illustrate this. If we come back to one of the key questions that we're trying to answer, it's not only what are the current carbon stocks, which is extremely important, but also how our soils changing in response to management and climate change. So, this just shows an example of how much more we can learn from looking into the past. Next. So, if we're able to leverage data from the past, it can sometimes even change our interpretation of what we're seeing in the present and may inform our understanding of the future. Next.

This is where some of the grassroots efforts to organize data around central scientific questions come in. Next. Since these databases not only build on current sampling efforts and systematic sampling that's going on, but they also tap into the published literature and archive data and samples from the past. And they try to put this knowledge together to fundamentally focus on understanding mechanisms and processes that might help us better predict the future response. Next.

One strength of these small grassroots efforts is because they're unable to make new systematic, continental scale measurements 'cause they lack the resources. They're really resourceful in terms of re-purposing past datasets and really leveraging the most out of past investments. So here you can see the

long tail of data with most grant awards being relatively small. And most of these studies that are actually conducted are focusing on one question at one place at one time. The great thing about these grassroots datasets is now they're pulling together data from this long tail and repurposing it to answer centralized questions by pulling data from many studies that were originally collected for different things. Next.

This shows that it's really important to archive data because it can be repurposed in the future. But unfortunately, data is being lost extremely quickly. So here you can see the results of a study, where they actually got in touch with authors to try to access their datasets. And they found that due to a huge range of different problems, like emails no longer working people having lost their hard drives, et cetera, that even when people wanted to share their data, they often didn't have it anymore. So, this loss is tapering off really quickly, and it emphasizes that we need people to focus on archiving their own data, because they may not be able to reliably store it for 10 or 20 years in the future. But also, it emphasizes the role of many grassroots databases, which are emerging to compile these different datasets and pull them together thematically to answer core scientific question. Next.

There's many of these small databases. But today I'm going to talk about one effort that I've been involved with as a case study to show some of the common challenges that these databases are facing. And some of the things that they're able to accomplish. So ISRaD, is the international soil radiocarbon database. And it's a large collaborative effort to compile soil radiocarbon and related data. To better understand rates of soil, carbon cycling, and also mechanisms of stabilization in soils. Next.

And all of these different grassroots are databases have come together in very different ways and have had different paths to come into existence. But for ISRaD it's been the product of a really dedicated community. Plus, some course support from the USGS Powell Center to organize these workshops. To bring people together, as well as scientifically oriented funding to focus on the importance of radiocarbon and soils. Next. We decided to, or we have focused on radiocarbon within ISRaD because it provides a strong constraint on the global rates of soil carbon cycling. And that allows us to address some of our key questions such as how might climate change impact carbon stocks and what timescales could these changes take place. Next.

So globally radiocarbon is naturally occurring in the atmosphere at very low levels. But in the 1960s with weapons testing the Thermo nuclear weapons testing increased the concentrations of radiocarbon in the atmosphere substantially. And this led to effectively a global tracer experiment where we can see how this radiocarbon for bomb testing was incorporated into vegetation and then soils, and then respired back to the atmosphere. Next.

And we can use this labeling in effect to see how long it takes carbon to cycle through soils. So, by tracing this radiocarbon bomb holes through the vegetation soils, and then respiration, we can see how it takes anywhere from years to decades for carbon to cycle through soils. And we can really quantify this, which can then help us understand how fast oils might respond to climate change in the future. So here you can see some of our recent efforts and where we find that carbon is cycling with Mean Transit Time of a few years in the tropics up to decades to centuries at the higher latitudes. Next.

Although ISRaD is on one database and really focuses on radiocarbon, it's been able to answer many different questions. So, these are just some examples of other efforts that are underway that have used ISRaD as a platform. So, we're also using ISRaD to look at how old is soil carbon. How is carbon distributed in soil fractions? And can we benchmark earth system models with radiocarbon? And what's really specific to the international soil radiocarbon database is this emphasis on using radiocarbon and soils to understand timescales on the rates of cycling. But what's really in common with many other grassroots databases is that we're pulling data from many, many different published studies that were often intended to answer a range of different questions at different sites in the past. And now we're able to pull them together to answer our scientific question. And that's a fundamental principle that many of these grassroots databases are using to make the most out of past datasets. Next.

There's been many other success stories. As you can see here, these grassroots datasets are answering a lot of fundamental science questions in this space. So, for example just to list off a few here, you can see SoDaH, SOC-DRaHR, COSORE and stuff like synthesis, and many, many more. So, there's been a real pilferration of these efforts and a lot of successes. Next. Another way that we can quantify the impact of these databases is to look at the citations that they're generating. This is of course, metric, but it's useful to think about the impact that they're having. So, this is the example of SRDB. Which is a global database of soil respiration data that was launched in 2010. And you can see that because of all the synthesis efforts that we're able to build on this database. And it's now generating hundreds of direct citations per year and thousands of indirect citations of work that builds on those direct citations. And this is impacting fields as diverse as agriculture, ecology, forestry et cetera. So, it's not limited to a particular sector that impact this work can have. Next.

One question that I have for this group is the thing about given how successful these grassroots efforts have been. And that's really exciting, but one thing is, should we actually be fully reliant on these grassroots efforts to pull together this understanding about mechanisms and past knowledge for every database that's been successful, there's probably many more that have failed or not made it to fruition because of the many challenges that these small databases and grassroots efforts can face. So, to give an example of some of these challenges, they rely heavily on individuals.

So, if funding priorities shift or jobs change, then that can totally derail a database overnight. They also have a lack of standardization which leads to reinventing the wheel every time. They can be really fun to think about science and focus the database around certain scientific questions. But then there's also a lot of inefficiencies when coming up with structures that make sense for each question, and yet can tap into these broader objectives. There's also logistical hurdles, such as the need for programmer time. When a lot of the people working in these spaces may not have the database experience necessary to build these systems. Next.

Even making it past all these hurdles and logistics. Another key limitation of these grassroots efforts is the global distribution of data. Because they fundamentally rely on the published literature. They tend to incorporate the same biases that past work has had. So, most site level studies or many of them are really concentrated in North American Europe and more recently in China as well. And this is not necessarily a problem for a management oriented understanding the impacts of management because those tend to be regional or local.

But when we moved to trying to understand the climate change impacts, these impacts are global. And if we want to understand the processes that are driving them and potential feedbacks, then we really need to fill these data gaps globally to make sure that our global conclusions are rigorous. So, you can see on this is a map of the data distribution in the SRDB database that I mentioned earlier. But most ground-based observational networks have the same biases with a real lack of data in the tropics. And particularly in Africa and South America. Next.

We do have the potential to make new measurements on both through the large-scale networks that we talked about earlier, but also through archive soils. So, this is if we manage to keep around these archives' soils. So, on the left, you can see the aircraft world agroforestry archives in Nairobi, and probably many of you wish your soil collections looked like this. But on the right is what might be more likely to be happening in your lab. So, it's really important that we support on retiring scientists and people who have individuals who have built up large collections of valuable samples. And that these resources become a resource for our community going forward, because we can actually make new measurements given the new technologies and new sampling approaches on past samples, if we can manage to keep them around. And that past knowledge and past soil samples are extremely important.

Since, as we talked about earlier, we can't go back in time. There's also challenges even for well-organized archives in terms of space and resource limitations. So, for example, that archive there in Nairobi actually had to reduce its capacity by one giant room. And that led to them disposing of huge amounts of soil that I'm sure many scientists in the US or Europe would have been extremely excited to get their hands on. But unfortunately, because the network wasn't well connected it didn't make it to those people who might have been able to do further analysis with it. Next.

So, this brings us to another challenge, which is the lack of connectivity and integration, or in many cases, there are foundations of this, but we can do much better integrating both the people who work on these networks as well as the data themselves. And this does not mean that every database needs to be connected with every other database. And in fact, many of the grassroots databases have their strength in that they're very specific and targeted, but it does mean that where we see potential synergy, an ability to answer questions better, the potential for those connections has to exist. And in many cases right now that's not the case that the databases are not really able to talk to each other efficiently due to lack of resources or lack of standardization.

ISCN, International Soil Carbon Network has served as a hub for many of these grassroots database efforts today. But so new databases are constantly emerging and many of them are not connected. So, there's like a much bigger challenge here than ISCN alone has been able to tackle. Next. So, in summary we looked at both continental scale efforts as well as grassroots database efforts. And I hope you've seen that these two efforts really need to work together to be able to understand the impacts of climate change and management on soil carbon.

In particular, the large-scale continental efforts are extremely valuable and giving us a snapshot of the current conditions or those that have occurred in recent years and hopefully will do a great job tracking the changes going forward. But we also have really relied on grassroots database efforts to better

understand the changes by leveraging past data, and to think about the timescales on which these changes are occurring based on mechanisms and processes. Next.

So, given the successes of these efforts, I hope that this week we'll be able to think about how we can better support and improve upon these efforts. And I know there'll be lots of discussion and ideas, but I wanted to get this started by just throwing out a few that I think both individually support, large scale efforts and grassroots efforts, and many of them which are can help both of them. So, for example, we could use more support for cross sector centers to make sure that these data are integrated and that both in terms of people and topics, as well as the databases themselves. These could also be better supported by long-term funding and systems for archiving soil samples. So that individuals who have really fantastic collections of samples also have a place to put them when they retire or to keep them organized over the course of their careers.

More generally we need a stronger mandate, both in terms of a community norm for archiving soil data, so that we don't have a continuous loss of data in the future. And we're actually making the most of all the samples that have been collected and analysis that have been run. More focusing on the grassroots, I had we'd need support for new efforts that can use existing framework so that every new group that comes along interested in a synthesis to address a scientific question, doesn't have to reinvent the wheel, but can instead build on past efforts. And we also, I think a challenge that's unique to these smaller grassroots efforts is the need for a more dynamic part-time resources that can be accessible, for example, tapping into program or time or guidance on databases. Just occasionally here and there over the course of their development and also for more long-term maintenance.

So those are some of the ideas, and I hope that the discussion this week will help us prioritize and come up with other ideas as well. Thanks so much. And I look forward to the discussion and questions.

BRUNO:
We do have a time allocated for Q&A for the keynote speakers, we have roughly a little less than 30 minutes before we break at 12:45. You... Chuck has a question, please, Chuck.

CHUCK:
Yeah, thanks, Bruno. Great presentations, really set the stage for the next several days. I guess a question maybe targeted for Jerry, but then Joe might answer as well is, Jerry, you had in your drivers for carbon and I would argue nitrogen as well, one of the top drivers was microbial activity. So, I guess the question for you and maybe Joe is how do we measure that activity, and what would be the appropriate temporal or spatial scale?

JERRY HATFIELD:
Yeah, Chuck, that's an interesting question and, you know, I pointed back to you, you spent your whole career working in microbial ecology, and obviously, you haven't come up with the answer yet, but...

CHUCK:
I figured an outside person could figure it out.

JERRY HATFIELD:
But in all seriousness, I think that we look at this, the microbial systems or what's transforming residues, root exudates into this. And I think that that is one of the major questions that we have is how do we quantify that? I mean, you go to back to the slide that Joe has, I mean, we've got the microbes clear at that one end of the scale that really working in micro sites. And yet we want to look at this as the field scale. And you look at this. I still think that the CO2 evolution out of this, because we are seeing an increase in respiration because of the increase in biological activity, but then you got to look at it from another way and saying, what's the outcomes of this?

JERRY HATFIELD:
I mean, in terms of maybe aggregates or nutrient cycling or even changes in color that we see in organic matter. And I think that throughout this workshop that we need to... we need to come in and say, what are some of the measurements that we need to be making and start thinking about how do we... how do we get those measurements made? So it's not overwhelming that we're not transporting all the field into the lab to determine the sequences that are going on.

So, I think we need to wrestle with this. And I think when we talk about dynamics, all the information that that the title of this workshop, that that microbial system is probably one of the most dynamic of all of this. And how do we characterize it? And maybe Joe has got some thoughts. But I think that, you know, collectively and I've put this out to people today that this is going to take disciplinary science, and I think it really is a result of almost being a wicked problem that is at that scale that we need to be thinking about it at.

BRUNO:
Joe, would you like to add something else?

JOE CORNELIUS:
So let me just add on top on top of that. So, this is the Holy Grail in the context of being able to actually create measurements, particularly in the microbial space, which are so complicated. And I'm really excited about the Berkeley activity that they're doing today with the eco fab as an example of miniaturization. And there are and that's just one example of where we're actually to get... we're able to actually create new systems, model systems that are more indicative of real world systems and scale them up in a way that they actually become high throughput. Also in the space which I didn't have in the slide deck, primarily because these are things that have been under development for a number of years, but actually are now getting quite close, the nanotechnology and microtechnologies. So as we look at Micro GC's as

JOE CORNELIUS:
And when you start looking at microbial being able to use things like micro devices GC's actually can pick up different respiratory endpoints, et cetera. So, I'm super excited about the progress that's being made on the physical science side. I think the real challenge for us is, as biologists and agronomists, is to look at the telescope through the other end and actually think of creative ways to actually use those technologies that address our specific needs.

BRUNO:
Thanks very much. I would like to invite Alfred (UNKNOWN), has a hand raised, then Michelle.

ALFRED:
Thank you. So maybe a question for the first two speakers and maybe for you, Bruno. So, is this... is this the idea that we do this only for soils in our agriculture, or we do this for all soils? So maybe a yes or no would be sufficient in this case.

JERRY HATFIELD:
To me, it's all soils.

ALFRED:
Very good. I like that idea. Then maybe if I can have another question for Alison. So, Alison, the grassroots or the bottom up approach, isn't much of that work needed to develop methodology rather than to collect data that could be harvested by a bigger database?

ALISON HOYT:
Yeah, definitely, I think those bottom up efforts are really important, both to tackle particular questions because they can be really focused, whereas top down efforts, since we're investing all this resources need to address a much wider range of potential questions. And then they could also be used, as you mentioned, to develop methodology, especially by providing comparisons between different particular approaches. So, one example of what we're doing now with the Israel database is we're making comparisons of different soil fractionation methods because a lot of this data has not all been combined in the same place because it's been done so differently.

So, actually aggregating it all systematically has been really challenging. But by pulling it all together, then we'll be able to compare what the results of different methods are giving us and then make recommendations to the community for more standardized practices going forward.

BRUNO:
Thanks, Ali. There seems to be another question from Michelle Wonder, please ask your question.

MICHELLE:
I had a question for Alison. I was curious about your cross-sector consortia or groups. Who are the sectors and how will that and what will that look like? Or do we have any models?

ALISON HOYT:
I think I mean, even at a most basic level, what you saw today, the discussion of soils and agriculture and then a lot of my work and a lot of these databases that I was talking about in terms of mechanisms have focused much more on the role of soils for climate change. And I think there's really important data that both communities are generating. But even in terms of where they overlap and understanding soil processes and mechanisms, there's not really adequate data sharing and not adequate knowledge between those two groups of what things are happening and how they might be able to support each other.

MICHELLE:
So, can I just sort of maybe... maybe Joe has thought about this with industry, you know, I think we see like FAO and people, you know, trying to unify, make carbon maps and sort of in the government sector and academia. But I think the model for how we would engage industry while maintaining transparency and openness is kind of, you know, I think, you know, our only opportunity to get the money that you the implications. But I think it's... it's really going to be some new kinds of relationships have to get forged.

ALISON HOYT:
Yeah, I definitely agree with that, and hopefully this week will provide some great opportunities for discussion, for creative ways that we might be able to approach that and how we can make the most of data that's even been collected in a proprietary way to answer some of these questions even if we don't have the full access, what parts of that data might might be possible to share and how could that help us?

BRUNO:
I'm going to read the question that came in select channels, what kind of efforts or collaboration are there to connect the scientific public data to the private data collection by companies? That's a very relevant question because they seem to be lots of efforts going on both in the private and obviously in the public. Joe, would you like to start describing a little bit, maybe the experience in the private that you've seen over the years?

JOE CORNELIUS:
Definitely. And actually, I think this is a very germane topic, especially as we're working towards creating more efficient and better open access models. And I think as we look at basically the data

generation side of the equation, we give a lot of emphasis on the private side. But actually, the more interesting data is really coming in from the public sector, both in the context of diversity of data and also being on on the most cutting edge. Industry Private sector data has a tendency to be much more narrow, much more narrow in its domain, primarily because it's product driven. That doesn't mean that it's not valuable and important. It certainly is. But when you start looking at the space that actually will stimulate the most creativity, it's really coming from the public sector domain. Actually, when I was at the DOE, we used to have... IT teams would actually joke saying that the lights would dim around midnight because they knew that's when the private sector was pulling data out of the open source systems so that they could then actually start or try to mine it.

JOE CORNELIUS:
But even in that context, a lot of the private sector does not necessarily have the capabilities to adequately query, query the data. And on a slightly separate note is that the private sector is right as we speak, is very receptive to creating new open innovation models. And we've been engaging... the foundation has been engaged in industry to actually be able to stimulate that. And I think that's a significant opportunity for public sector. Again, the public sector brings a very strong competency that the private sector doesn't necessarily have. The private sector is really good at scaling it. But at the end of the day, what we really need to do is, is develop these open innovation models that bring the two together and allow us to actually be able to capitalize on our strengths.

BRUNO:
Spot on. There is a question from Phil Robertson, Michigan State. Please, Phil.

PHIL ROBERTSON:
Thanks, Bruno. This is a question for Joe and perhaps Jerry and Alison as well. I'm wondering if there are efforts underway or perhaps another way to put it, do funders recognize the need for a set of calibration sites for comparing and calibrating different instruments and different methodological approaches? And, you know, I emphasize sites rather than samples, because so often the greatest source of variability is not at the gravimetric or archive sample scale, but at the field scale for important properties like carbon. And I think this is probably where the intercalibration needs may be the greatest. And I don't know if this is, you know, if this has been recognized by funders, or developers yet or not.

JOE CORNELIUS:
So I'll take a shot at this first. Phil, I love this question because actually I firmly believe that creating these testbeds or these sites are so imperative, especially when we start talking about integrating all the different disciplines. And DARPA has actually been quite good at this, creating testbeds, maybe not at the scale that we need from a from a biology perspective, but this is very clearly directionally a space that we need to be driving toward.

When I was at ARPA-e, it was the first time that DOE actually created a test bed with the terra site, which was basically the large robotic gantry system in Arizona as a way for the agency to actually test the test bed concept. And it's been hugely successful. And I think going forward, it's something that hopefully through workshops like this, we can start encouraging the agencies, USDA, DOE, NSF, et cetera, to actually be able to actually build those type of test locations out.

I know from a foundation standpoint, we fully recognize that having those type of long term investments actually create significant value and the private sector benefits from that long term because innovation gets a chance to actually be direst. And I hope David Babson touches on that. Someone, he talks about the smart farm concept, which I think is an example of how that system could actually be scaled.

BRUNO:
Jerry, do you have anything you would like to add on the testbeds?

JERRY HATFIELD:
I think that, you know, Phil you raise a good question. And I think that this is one of the things that we do need in terms of inter comparison is some standard way in which we know the history. And I think a lot of this is that we often don't know the subtleties that what happened at the site. And then we get all confused. And so we you know, you think about that. And I am a firm believer in the inter comparison of methodologies because when... experiences when we did work in water quality across the Midwest is that, you know, every lab was doing its own thing relative to nitrate. And so we had to come up with inter comparison to technology and sharing samples. And I think the same thing is going to happen on our soil. So it, you know, I think here is an opportunity for Neon.

Here's an opportunity for eLTER to really set aside some places within those systems that become those testbeds, that really devote the effort of saying we know what's happened here, so that when we sample these three different methods, we... we know the history of them.

BRUNO:
Thanks. Alison, anything else you'd like to add.

ALISON HOYT:
I'll just mention that informally, the scientific community has certain sites where there's large concentrations of measurements that have been made through time, for example, Harvard Forest, and also supported through some more formal networks. And one thing from the database and data integration side is to think about how we can make the measurements from these sites a point for connecting different databases.

So, for example, we have database focuses on root traits and databases focused on soil carbon and database focused on radiocarbon and rates of turnover. And all of these different databases do have measurements from these key sites that the community has already been focusing on and making a lot more long-term measurements. And so I think hopefully those sites could also serve as a linchpin to connect these databases that are really siloed into focusing on different... on particular aspects and be a focal point for connection.

BRUNO:
Very good. Thanks. Mark Bradford from Yale, please. Thank you so much.

MARK BRADFORD:

Thanks to all three of you for the talks. So I'm going to ask an unpopular question, given the title for this workshop where people might be unpopular and that is, what constraints should we place around innovation and measurement in terms of thinking it's going to get us toward understanding? And I'll back up just a little bit. So.

So, Joe put out a quote that I put out quite a lot as well. The Lord Kelvin quote about, you don't know unless you measure it. A bunch of economists, when that, quote, came out pushed back and said even when you do measure, you don't know. And Jerry's talk then really highlighted that so much in soil within it is multi causal and it's interdependent. And so being able to pick it apart with more and more measurements we could still be left with some pretty poor understandings. And I think National Academy report from 2010 as a great example when we said we don't know how much it's changing. And so carbon, even as a key metric, relates to many of the outcomes.

So I'm like, how much effort should we put towards innovation for measurement and how much should we put forward toward actually understanding what those measurements mean?

JOE CORNELIUS:
Well, since Mark you called me out, I'll go ahead and take the first shot and then let the rest of the team pile on. But at the end of the day, it is about knowledge. So, being able to measure something doesn't guarantee anything. And that's, you know, that's where it becomes particularly important to actually have as many eyes and as many brains actually digesting the information that's coming out of these systems, because invariably it's going to be that public debate that actually is going to tease out, you know, the insights that we need to have.

I don't know what the right ratio should be between the two, but if we're not, you know, and this actually brings up the kind of going back to my earlier comment, if you're in the private sector and you're downloading a ton of data and it's just sitting in your database, but you're not mining it intellectually, then what is it versus actually being able to advance the science. So. Jerry or Alison?

BRUNO:
Yeah, I just want to add one piece to that. In the brief introduction, I mentioned that some of the challenges is even in getting different results from the same measurement. So, that poses a completely, you know, a step back. And the other thing is this, even though if , hopefully, in this workshop we'll really try to convene our thoughts of how we're going to standardized to some procedure and trying to minimize that. I know because we've been exposed to Lucas (UNKNOWN) presentation about the EU that they have gone in the direction of having one laboratory that obviously has eliminated that uncertainty but has also slowed down the delivery of the results.

And so be able to standardize could solve that. And when I say I mean, even from the simplicity of sieving methods of temperature, that that causes that. And the other thing is, despite the fact of how well we measure it with less uncertainty, we still have to scale it. And even though I've brought up the component of modelling that plays a paramount importance in scaling and understanding this interaction, because even the measurement dimension does not capture the system.

You know, the interaction that as you mentioned, Mark, at what cost that level of results. So, I think we shouldn't... we should bring the two always together, both the sampling as well, the

measurements as well as modelling, because they have to just be almost one thing.

ALISON HOYT:
And building on what Bruno was saying, I think even though there's a lot of uncertainties that we might not be able to resolve with more measurements, there are places where we definitely know that we can do better and where we're missing fundamental knowledge. And I think that goes back to the distribution of data that I showed globally where we are really doing a lot of our sampling in North America and Europe and we're really under sampling in the tropics and the Arctic and Siberia, for example.

And then we're making global models and intuition about processes that are really focused on calibrations from sites that are extremely geographically co-located. So I think that's one opportunity where we don't need necessarily I mean, new innovation and technology would be great. But that's a place where we just know that we could do a much better job and have a deeper understanding with the tools that we already have at our disposal.

BRUNO:
Very good. We will see questions from Slack, then I'm going to present first and then go back to the Zoom questions. And so the question on Slack says, how much are the emerging soil carbon measurement technologies thinking about integrating with existing data systems? So that's... that's really a question from some additional volunteers in tackling people working on emerging technologies or anyone that wants to take a standard but it's critical to integrate the two. So.

JOE CORNELIUS:
I can add.

BRUNO:
Please, Joe.

JOE CORNELIUS:
So a lot of that integration in large part really depends upon how the programs are actually designed at the very front end. And certainly programs that are coming out of most of the federal agencies like NS, excuse me, NSF, DOE, USDA, etc., actually make that a requirement building on top of existing analytic pathways. You know, there's always room for improvement. And there, you know, it's an evolving landscape which in many, many cases are actually being driven by limited resources in the space. But fundamentally, whenever we design a program, we should certainly be thinking about how is the data ported into systems that actually can not only sustain it, but nurture it. And beyond that. I think a lot of times it's left up to the individuals.

BRUNO:
Any other comments before I move on to the next question by Vanessa.

JERRY HATFIELD:
I think Joe gave a good answer to that one. So.

BRUNO:
OK. Please, Vanessa.

VANESSA:
Thanks. I really perked up with the comment about measuring what we model. That's been a huge part of what I do. I'm not a modeler, but models guide all of my measurements. But I do think for a dynamic database it'll be really good for us to keep an eye on what is the next bleeding edge measurement that we currently may not know how to do, but if we had been so static about driving forward with just measuring what we are archiving, cultivating whatever we want to, curating the data that goes.. belongs in models, 20, 30 years ago, we would have missed the advent of a lot of new molecular microbial data sets that I think it's really important for us to look ahead and say, well, this is new. We don't know how, you know, high risk mass back data fits into our current understanding of soils, but let's embrace that. And here we are now within just a few years of that kind of instrument being democratized, we're getting nominal oxidation states of carbon.

We're getting whole new sets of reaction networks that are turning the soil carbon cycle from simply photosynthesis in respiration out into something where we're actually getting metabolic models of how microbes are operating different soils. So, I just want to not diminish the model informed measurement guidance, but really extend a strong vote for what's next.

BRUNO:
Very good point.

JERRY HATFIELD:
Yeah, Bruno I'd just say that that's really the heart of this workshop when we start thinking about what's emerging technologies, what do we know, what don't we know and what do we need to know and how can we be more efficient at the... at our science? And that's why I still argue for the fact that this is a transdisciplinary problem and we need to be casting a bigger tent in terms of bringing technologies together to tackle some of these problems.

BRUNO:
Yeah. Before we go on a break, I just want to end that. You know, why are we modeling a system? You know, the reason, obviously plays the most important role on the type of inputs that go in. The example that I made in introduction is that, you know, be able to understand the feedbacks between the system and soil alone and not be able to be a proper input to be able to capture that level of variability. It's a very important point to make that despite the fact we have good crop models, the input going into these models and depending on the type of question if we're modeling yield or if we're modeling at a biogeochemistry system is one thing, even though the biogeochemistry models will still need to account for the amount of residues and roots that they are returned to the soil. So, I think they are not longer off the hook in be able to model complex system without accounting for plants and management and weather and landscape interaction. So, I just want to say the complexity it's also in response to the objectives of what we're modeling anyway.

Well, I would sincerely like to thank the keynote speakers, you did a fantastic, tremendous job, beautiful presentation, fantastic discussion. And as a schedule we are due to go on a break and we

will resume in 1:05 Eastern Time zone for the next panel discussion. Thank you very much indeed. I would also suggest to just remind and remain on connected on Zoom and just turn camera and microphone off. So we'll be ready to start the next panel discussion on why do we need the dynamics on information system? We're trying to get the point of view of the agencies and some of the sponsors that allowed us to have this workshop. Thanks again.

SPEAKER:
Welcome back, everyone, for this second part of the workshop and exploring the Soil Dynamic Information System. This panel is composed by the agencies that have also supported this workshop, the agency that played a dynamic role with the planning committee, because we met frequently and they also were able to attend a lot of the presentation of the different soil agencies that share that information.

So, it's really that turn to describe to us in the order that I will call them a vision, a view of the agency or in general, why obviously they have a big interest in being part of this workshop today and help support it. And so the title again is Why Do We Need a Dynamic, Solid Information System? And without further ado, I would like to start with David Babson is the DOE Advanced Research Projects Agency-Energy, ARPA-E. So, David, please go ahead and I'll introduce all the others. You have eight minutes to share some of your thoughts. If you have slides, please go right ahead.

DAVID BABSON:
I do have some slides and we can just jump right to the next one. Thanks for having me. I'll start quickly because I don't have that long, with who is ARPA-E? ARPA-E is the Department of Energy's Advanced Projects Funding Agency. We're kind of our moon shot research funding agency. We tend to fund really high risk, high reward types of endeavours. We were modeled after DARPA and the types of challenges that we are trying to address are exactly what you would expect from the Department of Energy's Advanced Research Projects Agency.

So, if you go to the next slide, I can outline what some of those are. These include resilient energy infrastructure, affordable sustainable energy, US economic development and leadership and science and technology. And the thing that I think about, the thing that the programs that I'm working on are working to do are to address climate change, to offer climate change mitigation solutions. If we go to the next slide, we can see the context for why it is that we're pursuing, you know, programs related to carbon farming and leveraging soil for carbon management.

And that is because, as it is now, all of the paths to two degrees of warming go through zero. Our ability to just reduce emissions, even down to zero and still avoid more than two degrees of temperature rise passes by in the 90s. We did not do enough. You can see below the zero line in this chart. The blue shaded area are the net negative emissions that we need to achieve and the contribution of net negative emissions to net total emissions is quite substantial even before the crossover point. So, this is to say that we need to build a very large negative emissions industry very quickly to be able to give us options to maintain our path to two degrees.

And, you know, from the Department of Energy standpoint, because we know that we need to build a very large negative emissions industry, we know that we need to be able to service this industry with technologies that are low cost and energy efficient. And that's why we're interested here. So, if you go to the next slide.

As this group is well aware, agricultural ecosystems can play a large role in carbon dioxide removal with significant reductions through the implementation of best practices and the potential for net negative agriculture with broad-scale implementation of cutting edge technologies. So, we want,

you know, at ARPA-E, push towards developing new technologies that can even make kind of these frontier technologies are outlined here that would get the ag sector to carbon negative to make that even greater.

So, if you go to the next slide. You know, ARPA-E is playing an active role in funding the development of those technologies. As Joe Cornelius mentioned, there are several ARPA-E programs that aim to enhance the role of agriculture in carbon drawdown. And we are actively exploring new opportunities for enhancing terrestrial carbon reduction potential, many of which are rendered ineffective without accurate and accessible soil data. Next slide.

The Smart Farm Program was motivated by the need for highly scalable soil measurement systems to inform market incentives for improved carbon management, with an emphasis on nitrous oxide as the primary driver of positive emissions and the soil carbon as the potential driver of net negative emissions strategies. Next slide.

As mentioned earlier today, the phase one teams are tasked with capturing high-resolution soil data and making the data available to the public. As such, ARPA-E is keenly interested in the discussions and outputs of this workshop so that we can help to ensure the phase one data are the greatest utility to the R&D community and private sector stakeholders. Next slide.

Looking to phase two of the Smart Farm Program, teams are charged with developing the next generation of data collection and analysis tools for agricultural carbon accounting, with approaches including perimeter and drone-based nitrous oxide monitoring, institute sensors for soil carbon and highly scalable model-based approaches for coming to a net emissions estimate. For these teams, harmonised soil information system offers a clear framework within which to capture the data, and offers a broader audience for its use. Next slide.

Lastly, the ability to measure soil, measure data accurately in soil opens up entirely new possibilities for technology development to enrich its benefits, whether it's, you know, through new methods or fixing carbon and nitrogen and soil, or enabling new, highly scalable means of measurement. And we are interested in funding even more technologies in this space to really open up the possibilities for new strategies to do carbon management ecosystem services. And so next slide.

I will put in a quick plug for ARPA-E's open solicitation. It is open now. We are interested in all kinds of new technologies that would help us achieve our aims in this space. And so with that next slide, thanks for having me. And I'll turn it over to the next speaker and I look forward to the panel discussion.

SPEAKER:
Our next speaker is Dr. Jim Dobrowolski from the USDA National Institute for Food and Agriculture. Jim has been in charge of the CAP program at NIFA and several others. So, he's really been highly involved in working and dealing with soil-dynamic process. Jim, the floor is yours.

SPEAKER:
Thank you, Bruno, appreciate it. And thanks for inviting me to speak today. Next slide, please. One more. There you go.

SPEAKER:
So, first I'd like to talk about the National Institute of Food and Agriculture. We're a small agency with a big budget. Right now we're at about 248 folks and we have a budget of up to about $1.7 billion. We send that money out back to the states, part of which is through capacity, where we fund the land-grant institutions to support risky and long-term research extension and outreach, plus education. And that represents about 40% of our portfolio. And then competitive, which is discovery and applied research focused on agricultural production, quality and sustainability.

>:
Now, the study of soils remains an important part of the NIFA portfolio with consistent representation over the 15-year career that I've been with NIFA. More than 1,183 soil sustainability projects, focused on soil erosion, nutrient management and microbial activity, have been funded. And we're going to continue to fund soils into the future, and particularly with our partnership with NSF, through the Signals in the Soil program. Next slide.

>:
So, NIFA awards require data management. We often emphasize connections with existing inventories or networks that include, but are not limited to, training the next generation of scientists in soil science and management, the development of minimum standards and methods for data collection and integration of datasets, and plans for long-term data management, storage and sharing. And also linkages with publicly accessible databases for collection, information, tool development, sampling methods and data curation plans. Next slide.

>:
But we struggled to send folks to the right data repository. So, the potential first stop is our own National Agriculture Library site, their Scientific Data Services. It offers data management, policy and planning, repository management, data and metadata curation and consultation, and preservation.

>:
But although the beta tag is gone from this activity and the site is serviced by the Ag Data Commons, current domain and related informatics expertise is limited to biological sciences, geospatial and biophysical sciences, genomics, federal open data policy, and life cycle assessment. So, there's still some missing links there. And you can submit or link your data to the Ag Data Commons to meet FAIR data

requirements of journals that you submit, both from manuscripts or public access requirements of funders such as NIFA. Next slide.

>:
So, it's been just over a year since the federal call came out for data repository specifications through the Office of Science and Technology Policy. It was about January 21, I think, 2020, when that came out. It made a ripple through the repository folks across the country. It was seeking public comments on a draft set of desirable characteristics of data repositories used to locate, manage, share and use data resulting from federally funded research. And so, a lot of folks use, with the results of that, as a template for developing some of those grassroots databases that Alison spoke of this morning.

>:
And here's how how we can help together as this workshop. By participating fully, we can improve the current systems in place for widely monitoring soils -physical, chemical and biological. We can better understand and document and manage the effects of land-use and cover changes on soils. Next slide.

>:
But here's the thing. OK, so developing and implementing a dataset that encompasses the chemical, physical and biological attributes within the context of environmental and land-use conditions with a network of suppliers is a challenging task. It requires careful identification of known and innovative sampling methods, with suitable metrics and attributes focused on the appropriate users.

>:
And here's the kicker. NIFA has funded the collection of millions of soils-related data points. But where are they now? And I can't answer that. We need to have information about where our data that we fund is going, what data already exists and where, to help us identify new priorities, both national and regional, into the future to assist with our efforts to synthesize some of this work, and to decrease the potential for any kind of duplicate studies that might crop up. So, for NIFA a lot of our data is spread out over a number of different repositories and this is the reason that we need the dynamic soil information system.

>:
Thanks very much, folks, and thanks, Bruno.

SPEAKER:
That was excellent, Jim. Thank you for all these critical points. We'll try to address some of them as we continue today.

BRUNO:

Our next speaker is Matt Kane from the National Science Foundation. Floor is yours, Matt. Thanks again for accepting the invitation to speak to us today.

MATT KANE:

Thanks very much, Bruno. So NSF is a little bit different from other agencies in that we don't have intramural laboratories. We really mainly do only one thing, and that is we review and fund science and engineering proposals. If you could go to the next slide, please. When it comes to soil interaction with the environment, with freshwater, with the atmosphere, with other terrestrial processes, there's actually quite a number of directorates at NSF that fund relevant research to soils and would benefit from soil information system. Of course, there's the biological sciences and the Geosciences Directorate. We also have programs in environmental engineering increasingly interacting with all of these activities across directorate aspects with the social sciences and with computational science and engineering. Next slide.

NSF also has these ten big ideas, which our previous outgoing director, France Cordova, sort of marshaled the forces of the foundation to produce. And four of these, I think, really, the notion of a soil information system, a dynamic soil information system, maps well on to. One would be harnessing data for the 21st century, understanding the rules of life, navigating the new Arctic, and then, finally, growing convergent research. These are all sort of umbrella activities at NSF that involve a number of different directorates and program areas. Next slide.

We also have a number of core and special program areas, really, that will benefit from a soil information system science funded by our engineering directorate and programs in environmental engineering and sustainability. We've had a special activity called Signals in the Soil, which, like some of our other activities, although it's headquartered in engineering, it's been a collaborative one involving program officers and other directorates. We also have programs in the Geosciences Directorate, geobiology and low-temperature geochemistry, and geomorphology and land-use dynamics. I mentioned navigating the new Arctic. We have programs both in the Arctic and Antarctic. And then a variety of programs, of course, in the Biology Directorate from plant genome research and ecosystem science to our more long term programs, and these long term programs like long term ecological research sites and macro's system biology and NEON enabled science are really what I'm going to focus on in relationship to a dynamic soil information system. Next slide.

Two years ago, the National Science Foundation completed construction of the National Ecological Observatory Network, or NEON. This is the largest single investment that the Biology Directorate has ever made.` And the goal of NEON is to enable regional to continental scale, biological and other, research. It's really a force for team science. But most importantly, NEON produces from some 80 sites around the country and 21 different geographical regions. It produces 180 different data products, and this is open data available to anyone for use. A major theme at NSF, going forward, is

open data on the environment to help us address environmental questions and solve environmental problems. Next slide, please.

Open environmental data is not only available made available through NEON, it's also made available through other NSF program areas like the LTER network, Long-term Ecological Research, the Critical Zone Observatories Network, or CZOs, and a variety of other programs, some of which are more focused, for example, on the ocean or on biology. And these all interact with other open environmental data opportunities, such as satellite data from NASA. And a major interest of NSF, going forward, is enabling the scientific community to make full use of all of this open environmental data to address questions, to identify questions, to really understand the earth and its biota as never before. Next slide, please.

And so, right now, we have a new open competition for Center for Advancement and Synthesis of Open Environmental Data and Sciences. And this initiative rests on four pillars. First, it's going to be an incubator for team science for analyzing and synthesizing open environmental data. And that's where, really, this soil information system comes in, is it's a form of open environmental data. The center also is going to be involved in developing creative and innovative cyberinfrastructure to drive the science and support the science. This center is going to be an inclusive and equitable enterprise like really never before because open data has a democratizing force. It enables democratization of science and engineering like nothing ever before because anybody can access it. You don't have to be at a large institution. You don't have to be part of a large team. You can become part of a team

no matter where you are and work on this. And then the fourth pillar of the center will be data science training. Our letters of information for this solicitation are due April 1, preliminary proposals, April 29, and then by invitation only, full proposals are due September 15th. The budget for the first five years of the center will be $20 million, ramping up from $2 and $3 million in the first two years to $5 million a year for the last three years of the first award period. And that's really, I think, where NSF interest is going forward. And this is an area that soil information system will really be but one part of the open environmental data that's brought to bear on future science and problem-solving. That's it. Thanks.

BRUNO:

That was excellent, Matt. Thanks for sharing the NSF programs and the one going forward, very, very interesting, and also the role of soils within, especially, the new center.

BRUNO BASSO:

Our next speaker is David Lindbo. David currently serves as director of the Soil and Plant Science Division at the NRCS. David has been really an active member on our planning committee, and I look forward to hearing his thoughts on this. The floor is yours, David. Thank you.

DAVID LINDBO:

Alright. Thank you, Bruno. Hopefully everybody can hear me alright. So I wanted to start by really saying we don't know what we don't know, and that's really how this particular workshop, I think, really got started. There's a lot of information, but we don't always know what other people have out there and we could use it. I think that's where when I think of the why we're doing this, that's the why, so that we can start talking to each other. And as we start talking, we're going to start producing things.

So if you'll go to the next slide. So wouldn't it be great if and just click through, please, here, if we could get soil information on a ten meter grid with properties, interpret it for land use, use real time water and climate information, soil moisture, water table depth, irrigation needs, forecast the effects of conservation practices, including dynamic soil properties of soil health, water based resource concerns, erosion, and then determine the effective practices for desired land management goals, the state and transition? And if we could have all of that in one place.

We've heard a lot about folks collecting data and it's scattered. What do you do with it? One of our roles within within the division and within NRCS is to provide the producer, the landowner, the land manager with the information that they need in order to manage their land more, more effectively. Lots of technology can be brought to bear here. So if we go to the next slide, so what we are looking at is developing a dynamic soil survey and there are five parts to this.

The first four parts you see here, soil maps, including our traditional vector-based products that you're mostly familiar with soil survey, but also raster based products that will look at properties with depth on varying scales. Tie that to ecological sites which talk about what plants are there, what is the ecological community and how do you move from one ecological state to another in a given landscape, and then pull in dynamic soil properties. Those properties that change on a human scale, that change due to management. These are the soil health properties.

But they could be more than that as well as we start to think about urban effects and other land use effects, and then finally tie in the climate and hydrologic data. Since all of the all of the parts we've talked about so far are so highly related to climate and the hydrology, it all needs to be brought together into a dynamic soil survey that uses the static data that we've collected as well as temporal data. The fifth part of the Dynamic Soil Survey is the people and the information. I combined them together because we need the people in the field collecting the information, synthesizing it, and then curating that information so that we can use get it back to our dynamic soil survey, so it continues to grow. It continues to produce information that is of value to more and more people.

So if we go to the next slide. A dynamic soil survey can incorporate a number of things. We need field data, we need scientists, and we need your information. And this is a critical part of this particular workshop, is to recognize that there's a lot of folks who have collected information, but we don't always have access to it. We're getting better at it. But then even when we have access to

it, how do we truly incorporate that data into a unified system or unified, in this case dynamic soil survey. So we do need everybody's input here.

Next slide. So really that for us, the dynamic social service, the future of where we are going within the division and within NCRS so that we can use the soils information as the foundation, the backbone for resource management, for truly working with soil health and incorporating those practices, looking at conservation planning in order to make and collect or get the best use on the land. And then for various initiatives, whether they be climate related, water quality related, urban agriculture related, they're all going to need that soil information so the decisions can be made properly. From this workshop, we hope to be able to see what's out there. Start talking to people and move us forward to the future. So, with that, Bruno, thank you.

BRUNO BASSO:

That was wonderful, Dave. Thanks for sharing the overview of the agency and share every thought you mentioned in there. So hopefully we'll move the needle by quite a bit during this workshop.

SPEAKER:
Our next presenter is John Mesko. John is the senior director at the Soil Health Partnership at the National Corn Growers Association. John, thanks for talking to us today.

SPEAKER:
Thank you, Bruno. And thank you, everybody, for attending. This is a great conversation and I'm enjoying listening to everybody's perspectives. My slides are really meant to answer questions when we get to that point so we can leave this one up or we can take the slides down while I share my comments. The Soil Health Partnership is a program of the National Corn Growers Association. We work in 16 states and we have over 200 sites on real working farms where we are testing and evaluating the impact of various soil health changing practices on the soils themselves, but also on the economics of the farm. My understanding of the community that we work in is that there are a few organizations...a few projects that have the combination of soils data, management data, economic data that we do at Soil Health Partnership. Our funding does not come primarily from the National Corn Growers currently, we receive a little bit less than 10% of our funds from the National Corn Growers.

So, if you hear National Corn Growers and you thinking checkoff dollars, we don't get very many. And so, we are out there writing grants to some of the agencies that have been already speaking. We are out there working with supply chain partners to raise funds to continue this work and continue to develop some of these answers that we're working towards. And really the basis of our program is founded on the understanding that, you know, we all want to see changes. We all want to see climate change mitigation, as we've been talking about here. And within agriculture, to the extent that we can affect climate change, it rests on the shoulders of farmers who are going to make changes to the way they farm. And very few...the farmers that are already implementing some of these practices and I'm talking about cover crops, reduced tillage, nutrient management. The farmers that are already implementing those practices are what we would call early adopters. These folks are probably going to do it with or without the involvement of a program like ours.

Many of them don't require an incentive payment, although they'll happily take one. A cost-share payment to implement one of these practices. But certainly, these folks are interested in these new techniques or new technologies because they have an interest in that. What we have been targeting here in the last two or three years is trying to reach the larger portion of that bell curve. The middle adopters, those folks who have, by virtue of being part of the farming community, observed what's going on. They've read the articles in all the magazines. They're aware of what we're doing and other groups are doing it with on-farm analysis and so forth. And yet they still have not adopted some of the most important climate change affecting practices like cover crops or reduce tillage. We have 170 million acres of commodity crops produced in the United States. And depending on who you ask where somewhere around 10 to 14 acres of cover crops currently, even though cover crops are probably the hottest topic in agriculture right now.

So, we have a long ways to go. And our program has been developing the tools that we think can and our experience shows make help to make...help farmers who want to make good decisions about their practices. And that is data. For example, what is the impact of a cover crop on soil indicators? What is the impact of a cover crop, for example, on yield not just in one year, but in multiple years? We've got now seven years of data in our dataset. So, we provide data, we provide soils data and outcome data in terms of yield and changes in soils. We also provide an opportunity for farmers to have peer to peer networking. Our community would say that that might be the single biggest barrier to adoption of practices because each farm is different, each farmer is different, they

have different set of equipment. They have different access to, for example, cover crop seed or fertilizer, whatever the case may be.

Regionally, those components vary from place to place. And we are aware of the need that is in place. And our data shows that early in the change cycle for farmers, there is typically either an increase in costs or in some cases a decrease in yield. While farmers are adopting this learning curve, going through this learning curve if a farmer has never planted a cover crop before their first time out, it often doesn't go well, has a lot to do with timing and seed application rate and technology that they're using. And does it fit into what they're already doing on a year in, year out basis? So, I guess my message to you all is that I think the Dynamic Soils database is fantastic. We would love to get our hands-on information that is easy to reach. We currently are sending soil samples out with a probe on our 200 sites and collecting soils data, as you can imagine, that's very expensive. And we would love to be able to access this remotely or access it from a database that is pulled everything together.

But I would maybe to throw a wrinkle into the discussion, having just that, in my view, does not get us to where we want to go as a community, as a society, in terms of helping large numbers of farms change their practices. We have to be able to match that up with the management data that farmers using or whether it's planting date or the amount that they spend on their cover crops seed or the change in equipment that they are using as they change their practices so we can help them understand what the impact up or down is on their bottom line as they make these changes. There was a time in our farming history many years ago when farmers were more diverse. Farms mainly all included livestock. And so, there was quite a bit of diversity in farming. Now, on those 170 million acres of commodity crops, it's pretty much the same process on every farm. Corn, soybeans, corn, soybeans. In most cases we're affecting, we're promoting and asking farmers to make a very big change when we ask them to adopt a new practice, even something that may sound as simple as adding a cover crop to the crop rotation.

So, with that, I guess I would just say that our work is really, really focused on bringing not only all the soils data we've been talking about here together but the management data, the information that will help farmers make good decisions to help them achieve some of those practice changes we're all looking for. And that, quite frankly, it has been quite a challenge, particularly with the idea that farmers, most of them, believe that that information belongs to them. It's their business choices. It's their purchasing choices that they make. And they're probably not that interested without a strong connection with an organization. They're not just that interested in giving that up. So, I'll leave it at that and maybe we can get to some questions later.

SPEAKER:
Very good. John, thanks for your perspective into the critical role that farmers play in the challenges and difficulties. I share the thought of linking soil data with management because that's really a big part of the system, how dynamic that feedback is.

SPEAKER:
Our next speaker is Stephen Wood. Stephen is a Senior Scientist for Agriculture and Food Systems in the Global Science and Climate Change team at The Nature Conservancy. Thanks for speaking to us, Steve.

STEPHEN WOOD:
The nice part about going last is you can just agree with everything that's been said before you and the vision that Dave put up in his 'Wouldn't it Be Great If?' slide, I think, really matches our own vision of what we would love to see out of a dynamic soil information system. And perhaps, just a little bit of organizational history and background would help you understand how we've gotten to that point.

We, as an organization, I think, are often best known for the work that we do around land managements and land protection and oftentimes, most so in the United States, where we have chapters in all of our 50 states that acquire and protect properties, amazing properties all around the country. And that is still absolutely central to who we are as an organization.

And over the years, since we were founded in the mid-1950s, I think we've started to appreciate more and more that the biggest environmental challenges are not just local challenges, but they're also regional, national and global challenges that we can't just solve by protecting individual properties. So, that's led us to appreciate more the importance of integrating working lands into our vision of conservation and protection, and that new approach has led us to prioritize soil as a core part of our working land strategies. And so, that's something that's relatively new for us and it's something that's led us to start to work in some new ways. So, next slide, please.

This is a theory of change or a strategy for our soil and agriculture work in North America, mainly in the United States. And what I wanted to highlight here is that this strategy for us is really what we would call an impact strategy. It's something that builds on our decades of state-by-state and chapter-by-chapter level work, doing amazing work on the ground related to agricultural systems and it's about thinking about how do those things come together, those chapters come together into a broader strategy where we can have a broader impact.

So, when we first formalized this, probably about five or six years ago, we set a very ambitious target of having 50% of corn, soya acres in soil-health practices by 2025. And that's certainly a target that we're maybe not likely to achieve, but the purpose of setting such an ambitious target was to get us to think about what types of new partnerships need to exist and come together, in order to have impact at that scale. And so, you'll see in this cartoon here, that includes things like working with private-sector policy, scaling demonstration, working with farm advisors, new financial incentives, et cetera.

And one of the things I think that is interesting that's not on this chart is what we've been talking about already today, which is empirical soil measurements. And so, in a lot of our works, we don't take a lot of direct soil samples to measure whether, say, soil carbon is changing over time as a result of what we're doing. And part of that is because the scale that we're working at is quite broad. It's not a field-by-field type strategy and program like SHP, as John described. And so, that's put us in a situation where what we need are systems that allow us to assess the adoption of practices at broad scales, and to assess and evaluate how soil might be changing at broad scales over time - so, again,

the type of vision that Dave laid out in his slide.

And to just end, I guess I would echo what John said, but also a question that was asked in the plenary session by Mark Bradford, that not just about how do we understand changes in soil properties through time for us, but it's how can we take that knowledge of soil property change and convert that to knowledge and insight about agronomic and environmental outcomes that are the sorts of things that we create strategies around, and at the scale in which we work, from regional to national to even global scales. So, thank you.

SPEAKER:
Excellent, Stephen. Brilliant.

BRUNO:
That concludes the panelists, and we have about 20 minutes or so we could address questions to the panelists. I do have already one that is for John. And the question is, is the soil health partnership data available in any of this data repository? At what stage is the data processes and possibility of using them?

JOHN:
We are working with the University of Minnesota on a GEMS program to bring our data into one place and make it so that it can be shared with other partnering organizations, folks who want to work with us to achieve some of these goals. It's shareable now, but it's not very easily available just yet. We have really over the last year, one of our major areas of emphasis is trying to bring this together so that it can be shared and used by other partners.

BRUNO:
Right. Any other question from the attendees here while we navigate the Slack? Yes, Alfred, please go ahead.

ALFRED:
Yes, thank you very much. Maybe for David or maybe for all of you. So, there seems to be, but I may be wrong, there seems to be quite a few parallel efforts, and should we worry about that or should we say those parallel efforts are good, because people are going to obtain different types of data from different organisations anyway?

DAVID LINDBO:
So, Alfred, I'm assuming you're talking about David Lindbo. Problem with too many Daves on the call. So, yeah, there are parallel efforts. That's great. Most of the parallel efforts that I'm aware of, we talk to each other and we share information. There's always ways that you can take information, tweak it a certain way, one way or the other, or perhaps for a different purpose. But yes, we talk to each other. I think it's great to have people move in parallel as long as you understand that if you look at railroad tracks that are in parallel, they eventually cross in the future if you look down the track and that we have to keep that up. So, thanks.

BRUNO:
Stephen, did you have something... Your hand came up. But Stephen would... OK.

STEPHEN WOOD:
No, I didn't have anything.

BRUNO:
OK, sorry.

JOHN:
One thing I would add on this question. There are a lot of parallel efforts. There's no question about it. And from our perspective, it boils down to what the stated goals and outcomes are. When soil

health partnership started seven years ago, we received quite a bit of investment from the supply chain community to help us get started and to help us learn about and understand the impact of these practice changes on soils and on agriculture. In the ensuing years, many of those supply chain partners have developed their own variant of what we're doing. And sometimes I think in most cases it's not as robust and it's not as maybe scientifically viable, but it meets their needs.

JOHN:
So, thinking about a food company that wants to demonstrate to consumers that they're trying to invest in regenerative agriculture or agriculture that is promoting and advancing climate solutions, they may not need the finely tuned data that this audience expects. They may need just to be able to say our farmers are doing A, B and C, here's the proof that our farmers are doing that, buy our products. So, there is a lot going on, but not everybody is working towards the same ends. And I think that's really important. The other thing I would say is that funding I mean, I don't need (UNKNOWN) anybody hear this, but funding drives access to information, funding drives access to priorities and setting the agenda for what those expectations are. And I think that in our cases is a very, very important factor as we think about how we navigate through learning these practices and helping farmers understand them.

BRUNO:
Right. Thanks, John. Yes, Chuck, please go ahead.

CHUCK:
Yeah, I'm just gonna add on to that. I think that's one of the challenges for this workshop. And Dave Lindbo and I were talking about trying to get this workshop pulled together. The question is there is a lot of parallel efforts and at different scales, temporal and spatial scales, a farmer, land managers, maybe managing a 10 acre GPS sampling points or whatever, and managing at that scale, but then NRCS, in some cases, are working at a national scale.

So, I guess that's the challenge. And the opportunity is how do you take all these different sources of information at different scales and put it into a more dynamic information system. Dynamic not only in time, but dynamic in the sense of chunks of data are coming in at different times of year or different decades in that sense, but that's the opportunity. And I saw on Slack that somebody asked about citizen science so, again, there's an opportunity to add a robustness to our data set. The other challenge, though, is make sure that there's quality control of the data.

(CROSSTALK)

JIM DOBROWOLSKI:
I would like to certainly open communication lines with my colleague David Babson too, because we had David over for detail, I do believe last year. Didn't we, David? Over (UNKNOWN).

DAVID:
Yeah. I was over, I guess, two years ago now. But yeah, I was over there for a couple of years.

JIM DOBROWOLSKI:

Yeah. And so working with you, particularly on not only this new soil initiative, but the (UNKNOWN) initiative, which is the water piece that goes along with it, the (UNKNOWN) is pushing forward. We're very excited about being co-partners in those. And so, hopefully we'll be able to work together into the future, just like we collaborate with the National Science Foundation now.

DAVID:
We should be in touch.

JIM DOBROWOLSKI:
Absolutely.

SPEAKER:
I think one of the reason, in the introductory remark this morning, I pointed out obviously the personas, the stakeholders, play a critical role because in the type of information that they're needed for, one of the stakeholders that needs to make a decision are different from another and that's very important. Both Dave and Stephen made that point, that sometimes the measurements are so critical and valuable, but it is more about delivering information that will drive to a change that can be obviously verified through different systems.

SPEAKER:
So, I know we will continue to tackle this point and even addressing some of the potential proxies that we could use to get at these final kind of outcome that we all look forward to help stakeholders make better decisions. There is a question from Slack. What role does the standardization play for sampling and compiling the data in databases or soil information system? How could open geospatial consortium standards be applied, expanded for this purpose? Good question. So, standardization for sampling, very important. Anyone wants to take...

DAVID LINDBO:
So, since nobody's jumping on that one, I'll jump on that and say we've got a couple of folks within my division talking later this afternoon who can perhaps address that better than I can. But standardization is indeed critical so that we can compare. But the critical part, and I think somebody had said this earlier, is knowing that methodologies change over time, we have to be able to look backwards and make those comparisons as well. So, having those plots, those locations that we can test, again, using different methods and compare is going to be important. Again, I think the more we want to rely on or be informed by the science, the more we do have to have some standardization, some testing, etcetera. And I believe that FAO, through the global social partnership, is working towards that with some of their work with harmonization and the (UNKNOWN) global Soyland Laboratories. So, standardization, yes, we've got to consider that as we move forward.

SPEAKER:
Surely. There's a question from Kathy first and then Matt.

KATHY:
So, I don't have as much of a question as additional comment that when we can start working with

standards, we need to remember that that's not a static thing, that that's something that's sort of an ongoing and evolving work. I think oftentimes there's the tendency to just do a one and done, but as new methods evolve and new measurements start coming online, we need to have some way to extend and revise the standards.

BRUNO:
Well thinking. Yeah, it's a good point. Matt, I saw you with your hand raised, please.

MATT:
So, I just wanted to say that the standardization issue is part of the reason, it's one of the reasons behind why NSF constructed NEON, the National Observatory, ecological observatory network, is we have 81 sites that are collecting soil samples in an identical way under very strict quality control, storing them in a single biorepository now available for a variety of different analysis. Both soil and water and other samples obviously are collected, but the focus on standardization in NEON is really a unique opportunity to have that kind of central control over a continental scale.

BRUNO:
Any other questions for the panelists? An opportunity to address some questions to the agencies. I guess I'll make a point that even though we have all agreed that we help farmers make a change in practices - that's the ultimate goal -, often these practices are driven by an economic profits that they often don't see immediately, and they are willing to change if there are policies and incentives. So, we haven't discussed much about how we could influence policies.

We're all waiting for the new administration to, as we hear from the media in general, that there could be ways of incentivizing even further the change of practices, the (UNKNOWN) fields and, again, reducing greenhouse gas. There's the climate bill (UNKNOWN). So, policy for me plays obviously a critical role and I think we should consider the way science impacts policy as well as policy impacts the final decision. So, without necessarily commenting on my comment, there is Stephen Orgo from Colorado State. Please, Stephen.

STEPHEN ORGO:
Yeah, I have a question for the panelists. So, I'd like to hear a little more about maybe what you see as some of the challenges beyond just the data. I think John Mesko, you brought this up very well, that you have a lot of farms out there growing corn and there's different equipment, there's different soils, different conditions. How do you take this information? What are the challenges you see taking the data and then using it to give advice, maybe the farming community? And any of the panelists who might have some to add along those lines.

JOHN:
Well, thank you, Stephen. What I would say is that what goes into a farmer making a decision is very complex set of points of data. And farmers are not just scientists, they're members of a community. They live and work in a community of people. And deciding to change a practice, a major change in practice, means they might have to stop doing business with somebody and start doing business with somebody else. It might mean they have to stop a regular association with somebody who they get advice from and start an association with somebody who they now are gonna take advice from.

It's much more complex than presenting a set of data to a farmer and saying, look, if you start planning a cover crop in three years, you will save X amount of money and you will make X amount more money in terms of yield and you'll be doing good by the environment too. It's just so much more complex than that.

And if you talk to anybody that's trying to sell anything to a farmer, they'll tell you there's a reason why ag businesses really specialize in building a community around farmers. That's why farmers wear ball caps that have the name of the logo of the product that they're buying or the color of the tractor, they're the product their buying. If you really want to change the hearts and minds of farmers, we have to do all of this data, we have to do the incentive payments, we have to do the peer to peer network, but we also have to make it very cool to be a farmer that farms in this way. And right now, the momentum as to what kind of community I want to belong to for most farmers is I want a big tractor, and if you tell me that my farming practices are gonna require me to get a smaller tractor, I might not be as interested in that. I like a big tractor. I like to be seen as the one who's out there pushing for the maximum yield. And even if it costs me a little bit, I wanna do that.

So, there's all these narratives that have been typically pretty entrenched, but they're starting to change and that may take some time, that's a generational change. The next generation of folks that's coming in agriculture, all of them probably went through environmental science in high school as opposed to just FFA. They probably learned something about recycling. They might have learned a lot more about the impact of individual choices on the environment, those people are now in their 20s and 30s and they're starting to take over these farms and they're starting to become more open to changing practices. But it's really not just throw some information out there, 'can't you see this is an obvious choice? You should do it this way', it takes much more than that.

DAVID:
I think one of the things to follow on that point that would be very helpful to get farmers to move towards adopting more sustainable practices and new technologies is the connection of those farm practices to new types of carbon markets, which is really what the focus or what is the intention of the smart farm program is. And at the root of that is data. If you're going to actually connect farm practices to expensive carbon markets like those that exist in biofuels, and that could be established for bio products, for food, for other sorts of things like that, you need to have very highly accurate data available at a low cost so that you can consistently attribute value to the individual practices that farmers implement for their production system. So, that's one of the important needs of the data.

And actually to follow up on a point that John made earlier when talking about why it's important that there's a bunch of different simultaneous activities occurring in parallel and for different things is because there are different markets, there are different sized carbon markets. If you're connecting farm practices to a $15 a tonne carbon market, the resolution of your data can be a lot lower than when you're trying to connect them to a market where the carbon price is $200 a tonne. And so, there's this opportunity space where Bruno was mentioning about how to science inform policy. If we had the technology to offer better data, more data with higher resolution at lower cost, more consistently, that would give policymakers better tools to establish more robust carbon markets and ecosystem services markets that we could then connect to farm practices and that would drive more

sustainability. I would just counter John's point, I don't know that we need to make sustainable farming seem so cool as much as we can help make it profitable, and that will make it happen.

BRUNO:
There is a question from Slack that I'd like to post to, I guess, the panelists to start with. Is how does training fit into a dynamic solar information system, or more specifically, thinking about standardized database training for students, network for early career scientist, outside partners like industry and non-profits, extension programs that would translate into optimizing on the ground management. So, maybe Jim, because (UNKNOWN) pays so much attention on the extension correctly, you had some points to share on that.

JIM DOBROWOLSKI:
Yeah, I think that it is important that we do provide students with some insight into how they're going to have to translate information to the folks on the ground that are actually going to use it, because it's not the same as it used to be, in particular with delivering pamphlets or having having traditional field days or get togethers where you have a large group of farmers that are coming to listen to you speak about the newest technologies. We have to incorporate the social science aspects of this. We need them to focus on sociology and psychology and trust building so that we can promote behavior change and adoption in other folks. I mean, we've spent a lot of time looking at how we get people to do the right thing from an environmental management perspective and we need to do the same thing with the agricultural communities.

We have to become part of those communities and build a trust level so that extension, in particular, and outreach is not like 13th out of the list of people that they trust. Maybe their seed person or their fertilizer person are the number one and number two. And oftentimes those folks have been trained by the extension folks. And so, we've got to start thinking about a new way of providing that service and we need to do it earlier in their career. We often graduate a lot of students out of the university system and they might end up as extension specialists or extension faculty with very little training in that activity. They've been trained as hard core scientists, not as extension folks that are going to attempt to translate the information that they're given and improve either the bottom lines of farmers or improve the environmental conservation aspects of it.

So, there is a lot to do and it's very difficult in a lot of cases if you're going to even go younger than that, which we probably should be doing as well, because the state curricula are so packed in, it's difficult to get pieces in, wedged in there, that can provide them with a wide range of information that might help them later on.

BRUNO:
Very good, Jim. Dave, please.

DAVID LINDBO:
Yeah, I agree with Jim as far as getting your extension folks trained so that they can interact with people as well. That's critical. But it's also critical for the folks that are getting their BS degrees now in soils or agronomy or environmental science to really get more computer information, GIS training and programming. Our employees that we're hiring now, we're looking for that as part of their

training so that they can do things in Python or R or fill in the blank. I don't know it, but I know some of you that are training students that we'll be hiring. The good recommendation is take those courses so we can really use that expertise.

BRUNO:
That's great.

JOHN:
Those are great comments. I think there is a perception in communities like what we have here that the extension services is relied upon by farmers for new technology and new information education. And as a former extension agent myself and someone who's been a farmer and has 30 years experience in agriculture, I hope I don't step on anybody's toes here. But the reality is that the extension service, by and large, there are pockets of great extension outreach individuals, but by and large, the extension service does not reach farmers in the way that they used to, certainly, and it doesn't reach the types of farmers that are making the kinds of changes or that we want to make those kinds of changes going forward.

BRUNO:
There are two more questions, and I have one from Slack. Phil, please go ahead.

PHIL:
Thanks very much. So, this is a question for John and Stephen and perhaps others, but particularly you guys, because you've been working with farmers. And in your time working with farmers, do you find attitudes towards healthy soils changing among those that John described as the middle part of the bell shaped curve? I mean, clearly the ones on the tail, early adopters and innovative farmers, they have a very healthy respect for healthy soils, but the middle part of the curve. And I'm wondering what role, if any, are retailers playing in this change or lack of change?

JOHN:
I'll share a couple of points from my experience, and then Stephen certainly can add to it. Starting with your question about the retailers, retailers are in the business to sell things. And if there's a practice that's gonna require less chemicals, they're not interested in promoting it. I'm not trying to be difficult, I'm just trying to be efficient in what I say here. That's really what it boils down to. They're also not really equipped with personnel or with the motivation to help farmers adopt the new practice unless it's gonna involve them purchasing something that that retailer sales. And frankly, there's not a lot of retailers who sell cover crop seed. There's not a lot of retailers that sell (UNKNOWN) tools for their equipment.

And so, what we're finding is that a lot of retailers are trying to partner with folks like NRCS or Soil and Water Conservation district people to provide some of that technical assistance related to some of these environmentally sustainable practices that farmers want to employ. That isn't a ringing endorsement. If the business that you rely on for all of your inputs and all of your crop decision making tools isn't really helping you with those decisions, it's really hard to get over that hump. I think, in short answer to your questions, I think the press, farm press, has done a great job of making awareness around soil health and some of the practices here, but it's not following through just yet.

That's my opinion.

BRUNO:
Question from Alfred and then the Slack one. And then we will have to close this panel to continue with our program. Alfred, don't see you. You had a hand up for me.

ALFRED:
I had a hand up. But maybe a quick question, and I don't know who to address it to. So, we keep coming back to this largely agrarian focus of these databases. I mean, are we talking two million farmers in the US? Are we talking develop databases for the world population? Are we talking...? That's one question. The other thing is maybe foe Stephen and is whether the nature of conservation and NEON, maybe would like to speak up a little bit more what these databases will do for you.

STEPHEN WOOD:
I think for us, what they would do it's mainly in the insights, as I kind of suggested towards the end of my presentation. We don't do a lot of pure research at TNC. A lot of what we're interested in is insights that we get from soil knowledge about, say, projecting regional changes in water quality, water availability, greenhouse gas cycling. So, I think that there's gonna be an inbetween between the databases and our actual final use of it that would come through the interpretation of that. And then to your other question, I think we are speaking to a broader dynamic soil information system that's not just agrarian.

BRUNO:
Let me read this question from Slack, and this is relevant for the agencies. Do any of the panelists have an example of a data repository information system outside soil science that they could consider a successful model that will be worth mimicking or building off? Maybe the NSF, Matt, if there is anything you've been exposed to or Jim or David.

MATT:
So, the ocean sciences has a centralized database they call BCO-DMO, that has all kinds of biological, chemical and physical data for ocean science. That's one potential model. But I think... I'm not sure that there's another model that soil information system per say should follow. I don't see it following the ocean sciences model or GenBank as a data repository for genetic sequences. I think you're just talking about heterogeneous data and it's a huge challenge in understanding the soil environment.

BRUNO:
For sure. Perhaps the climate science has taught us a little bit more also on storing information, but it's hard, yes. I also didn't think of an easy example. Anything from (UNKNOWN), Jim?

JIM DOBROWOLSKI:
Well, not that's gonna have the breadth of data sources that are gonna have to be in the repository. I think you can think about the... Well, let's see how many years have been now. Matt, 15 years that NSF has funded (UNKNOWN), which was the consortium with the universities for hydrology data, time series data, those kinds of things. And that was where we from (UNKNOWN) would send all of

our water related data sets to. But it's, again, fairly limited relative to the number and types of data that they can receive. And so, I think from the soil perspective we saw some pretty amazing looking repositories, though during some of our meetings that we had prior to the workshop. And so, I think we might lean towards some of those newer versions rather than the ones that have been established for a long time.

BRUNO:
Thanks, Jim. Well, with that we're just at time, we need to close this panel. would like to thank all the panelists for attending and providing these valuable insights.

BRUNO:

Our next speaker on the program is Alison. Alison has an interesting job is really, to synthesize this eating they think at, in an agency's or a group of people that have shared the information with us or with the course, we have met for over a year longer because this workshop was to be held in person only last year, but they really gave us an amazing opportunity to learn so much. And so, I pass the floor to Alison, to synthesize and share with us what we learned during this year. Thanks, Alison.

ALISON MARKLEIN:

Hi everyone. Thank you, Bruno for the introduction. I do have a very interesting job today, which is to summarize our listening sessions as Bruno mentioned. Next.

So, our, the past, over the past year, our goal has been to understand what exists in terms of a dynamic soil information system. So, we met with 18 different organizations, including US agencies, International Agencies and members of the Private Sector. And as a result, we've come to a few different conclusions, which are there's high potential for increased inter-agency communication and collaboration. There are a few dynamics soil data, and we have identified needs and gaps and funding for monitoring is needed for long-term dynamic Understanding. Next.

So, we met with several different organizations, including many who we've heard from today, including the NRCS NEON, US Forest Service, Soil Health Partnership as well as International Agencies and Private Sector. Next.

We asked each of these different organizations, a variety of questions. First, what is your vision, and what do you want to do with the soil data? What is working well with your current database or data collecting effort? What are the roadblocks? How are data curated, transferred, analyzed, and shared? What are the drivers of change? What do you want to be able to do? How is your data being used and who uses it? What infrastructure is needed to capture and store the data? And at what spatial and temporal scales are different variables measured? I'm gonna go through and first mentioned these US-based and global data products that cover including the United States. And then I'm going to go through two examples of how we organize data from some of our listening sessions. So, the NRCS has the NASA database as well as gNATSO so, and SSURGE. NOAA has the national integrated drought information system and national coordinated soil moisture monitoring network. Soil Health Partnership has a database. NEON has 15 active soil data products. The US Forest Service has the forest inventory analysis. There's also dataONE, the global soil information system, ISRaD, Global Soil Health Partnership and ISRIC. Next.

And overall, what we've learned is we've synthesized a variety of different challenges. One is continuous funding for monitoring, and this one is particularly important. A lot of the agencies are really excited about funding science driven questions that are a few years in scope. Whereas we really need to have this long-term monitoring as several other presenters already mentioned to get long-term effects. A lot of the organizations are understaffed for soil science and data analytics. There's several issues with data, privacy and security. There are different naming conventions for data that provide present challenges for data harmonization between datasets. There's also issues with common methodologies, including sampling and analysis procedures. There's also errors associated with the methodology analysis and

facilities, as Bruno mentioned in his introductory talk today. It's really difficult to capture spatial and temporal data at varying scales. And there's the challenge of re sampling destructive samples. So next.

So, my first example that I'm going to talk to is the European Joint Research Center. We had a couple different presenters from this organization and this screenshot is a template of how we organized a lot of the information from our sessions. So, the goal of the joint research center is to provide scientific evidence and data for policies on soil. It includes soil data information from 27 countries, and the soil data is used for agriculture environment, climate change, biodiversity, and human health policy. They have a soil information system, that's harmonized and free. There's also the land use cover area frame survey, and survey in collaboration with Eurostat. And then here in the table shows the data collected, which includes soil physical and chemical properties. The spatial resolution is roughly at a two-kilometer squared basis for Lucas', but sampling's done on roughly a 14 kilometer by 14-kilometer grid. It is irregular. Their temporal resolution is dynamic every three years. The same sites are re sampled for land use and land cover change. And it is regularly updated as new data becomes available.

Some of their challenges are they've had laboratory issues in the past, and now they only have soil data analyzed from one lab to reduce the interlab uncertainty. There's still very large uncertainty in spectral and remote sensing. There's diversity in interest from different UPU member States, some of which don't actually have monitoring systems and certainties and scaling. And there's a delay in reporting from analysis. And this is primarily due to the fact that all of the analyses are performed by one lab. There are several assesses. They do have a harmonized dataset for soils for Europe. The data are used by different stakeholders, including scientists, modelers, policy makers, and farmers, and there's strong collaborations between groups and organizations in the space. So next.

This is a map showing an example of their database. This is the top soil organic carbon content, and you can see the different levels of soil carbon in Eurasia. Next.

So, to summarize a lot of the information from this organization, their goals are ag, environment, health, human health, climate change, biodiversity, and ecosystem services, and all are driven by parliament with the 27 member states. It's part of the EU budget, which is approved by parliament. Resampled every three years with 2009, 12 and 15 available online already. 2018 will be available soon as it takes the 18 months for analysis. They measure physical properties, including texture, book density, moisture, depth, topsoil, chemical properties, including carbon nitrogen, phosphorous, micronutrients, pH, and soil contamination properties. And they do DNA sequencing as a biological properties. There's also some information on land use, including crop type and management systems. And they have over 300,000 samples to 20 centimeters depth. And it's an irregular grid for policy relevant locations. And their main challenge is data privacy. As many of the other speakers have mentioned. The lands are often private and the heavy metal data is also it says not private, but it is private. Next.

So, I'm gonna talk about one of the organizations in the United States that has soil information system. And this is NOAA. They have a national coordinated soil moisture monitoring network. Their goal is multi-platform soil moisture, including grid, project products, measuring in situ remote sensing and numerical model output. They partner with different agencies at both the federal and state level, and there's over 150 end users. Their data includes 21 different mesonets. And then they have a NASA product and a NOAA product for soil moisture. Their volume metric water content is measured in percentiles, and they do spatial interpolation using SSURGO soil and PRISM precipitation data at the four-kilometer grid. And this is near real time and regularly updated. So next.

There are several challenges associated with this product, for example, how best to represent soil moisture. Whether these are percentages anomalies the drought monitoring category, volumetric water content, or millimeters. And how do you communicate uncertainty for each of these types? There's also the underlying data maps are not currently available due to funding, but the goal is for this to become accessible and map data is currently available. They also recognize the uncertainty in sensors and the validity of them, and there's challenges with data integration, including spatial distribution issues and the representativeness of each of the points, soil depth, the record period, and data gaps and sensor performance, metadata, and data formal variability. They have established proof of concept and have been operational for greater than one year now. And so, I'm really excited to see how the dataset evolves over as more time is collected. This is one example of their map. This is the blended soil moisture product at five centimeters. And it's the high-resolution gritted soil moisture map with combined Institute model generated and satellite data.

So, to summarize the NOAA data set, this is a multi-platform soil moisture gridded product, managing in situ, remote sense and numerical model output, including state agencies and federal. They are developing a cyber infrastructure. And they have several different next steps, which are really exciting, including interpolating soil, moisture data with space, time, and depth. Blending, these different data sources, further validation and quality control. Goal is to make the underlying data accessible. And one of the things that was really interesting about this presentation is recognizing that we all do better if we coordinate with each other. And this is one of the motivations behind our workshop is to coordinate these different soil data products. One of their other future goals is productions. Next.

So, our group synthesized the different properties in some of the data sets that we discussed. On the left column are the different organizations, including the European Commission, NEON, CSIRO, The University of Minnesota, Rothamsted, Viresco, the Soil Health Partnership and ISRaD. And the top column is different soil, physical properties, soil texture has been defined or has been collected for each of these different organizations. Bulk density for almost all of them. There's also quite a few that study soil moisture, aggregate stability and depth of top soil. And then soil temperature and available water capacity are also in some of these. Next.

We also looked at the soil, chemical properties. Carbon and nitrogen are in, are used in many of the, or in all of the products that we have shown here. Phosphorous micronutrients and pH are also shown by almost all of the organizations. Metals, nitrogen transformations, carbon isotopes, and nitrogen isotopes are also included in some of these. And for biological processes, there are a lot fewer data for this. There are some microbiome omiox, there's PLFA, root traits, carbon respiration, soil health indicators, including the Cornell Comprehensive Assessment, pathogens and microbiome phenotyping are some of the reported properties.

We also created a timeline of soil data products showing when each of these have started. NCSS GLOSIS were established in 1900. The forest inventory analysis, ISRaD and SSURGO were set up in 1930. The legends of this have shifted. So, the years don't correspond with where they're pointed, but they are in chronological order. So, 1966 was, ISRIC. And then starting in 2006, we had a lot more products developed, including NEON, DataONE, NIDIS, FAO and NMDC as well as the SMAP. Next.

And we have a variety of recommendations for developing a soil moisture, sorry for developing a dynamic soil information system based on our 18 conversations, including funding monitoring for long-term, understanding. More repeated measurements in the same locations, more communication

between agencies, which I think we're doing today. Increased clarity on nomenclature. More metadata on the sampling methodologies and processing. Data at a scale relevant to farmers. And archives soils for future analyses. And I would like to thank all of the presenters who have spoken with us over the past year. It's been really informative and yeah, thank you very much.

BRUNO:

Excellent, Alison. You pulled it together. It was really a challenging job because unfortunately, you know, it's just so much information and it was truly a privilege for us. I would like to open, we still have about little less than 10 minutes to, before we go on a break and start the next panel. So, if there are any questions in relation to what you have heard so far, and what were you surprised or anything else that didn't come clear, please raise your hand. I can say that the next panel does have, a some of the representative that they spoke with us, so they can go in a deeper, deeper dive on that. And we could also use this time in case there are points that were really covered and we could discuss for a few minutes. I hear that, you know, Tom Hengl, had you had a point about the OpenGeoHub foundation you would like to share and talk to us about that.

TOMISLAV HENGL:

Well just yeah, I just noticed the list though. I was just pointing to some datasets.

BRUNO:

Right?

TOMISLAV HENGL:

Yeah. And we actually exposed all the, I was appointed to the compilation of pulling data which would put the Geo club, so you can see all the import steps and how the binding was done. So, it's a really transparent system and we would like to convert it into a community system. So that people can just pull modify ads and you contribute, but we are a non for profit organization. We promote really open-source data sharing and collaboration, so really simple.

BRUNO:

Really good. Thanks Tom.

TOMISLAV HENGL:

Thank you.

BRUNO:

We have a large group of attendees and we could only cover so many. And so, I would really invite anyone else that is willing to share other systems that they've been working weird or aware that that could benefit this conversation. Yes. Michael, there is a hand raised and I would welcome your question, please.

MICHAEL:

Thank you. I do have a question for the previous speaker. And it seems to me that there's sort of three pools of problems that we're dealing with. One is the standardization of the collection of the data. Two is the standardization of how it's stored and then three, one of the techniques to harmonize the data. So, we can, you know, we can combine data of different formats, scales, and types. And I'm sort of wondering if Alison in your conversations with your stakeholders, whether they will the people are sort of this as an issue, certainly standardization of data collection is important. We don't want to overstay standardize 'cause then we lose innovation, the metadata. So there seems to be a pretty good history of how to, what metadata people need to store. So that we can understand the providence of the data and how it was collected. But it seems that there's an ongoing effort to try and harmonize the data as well. Am just sort of wondering what people's opinions were in terms of the sticking points.

ALISON MARKLEIN:

Yeah, that's a really great question. People were definitely very interested in increasing the harmonization of datasets and we had several conversations about this. There are a lot of challenges as you've mentioned. And it's really hard to create a standardization, especially when not everybody is in the room at the same time. So, I think that's one of the goals of this workshop and getting this conversation started is to actually have people recognize how, like, not just what other data are available because that wasn't known like between organizations, but also to figure out ways to make the datasets interoperable.

MICHAEL:

Thank you.

BRUNO:

Right. Thanks. Alison, Chuck please.

CHUCK:

Sorry. I was just on mute. Yeah, I guess one of the things that kinda help with the discussion is what surprised me in our conversations was the lack of ability to use remote sensing far as ancillary data to make the soils information more interpretable or that that seemed to be kind of a consistent message as well as not using that kind of data set that wouldn't again, expand maybe utility. And then the other thing, I don't think Alison, you know, I don't know, I can't remember now it's been a long year. But again, some of the timescales or in frequency of some of the measurements, particularly on the biological side. Anybody can comment on that.

BRUNO:

Yeah, I guess I'll take a quick stab on remote sensing since I'm pretty heavily involved. And as you know, we shared the same thought. That, they play a critical role in helping at least guiding through, you know, the target sampling, but even using information directly, you've gotta separate, you've got the optical sensing, which have a significant limitation 'cause they don't go through the ground. So, you'll be able to see you know, different zones and areas radar and microwave have a possibility to go through clouds and penetrate the soil. I think a very valuable tool the way I've used as you know, Chuck is that plants cover that soil and the plants we really talked about what they're seeing in depth soil. With the thermal imagery that I was referring I think that that's a such a promising tool because it just reflected soul depth and then the possibility of having available water that you would need necessarily to go and poke holes or everywhere, just to learn that these plans were doing better because there were just simply able to satisfying the evaporative demand.

So, I think we have to be creative in how we use remote sensing, especially with the fact that now the cover and role in verifying, you know, practices. And so, there is quite a bit of components since we have gone into direction that the changes of practices will affect soils as we know they do. And so, there is quite a bit of a role of remote sensing in there whether it's optical or remote. We are able now to detect whether a soil has been tilled or half of her has cover crop, pretty well obviously chlorophyll content. So, there are all proxies about what the soil is, but directly underneath the soils, you can only really booster our knowledge if we have a significant amount of information where you could train the system by knowing, you know, the detection and scale it by correlating reflectance with observation on the ground. And I think the future will go much more in that direction. Given the fact that now there are fusion products coming actually from Europe, where they provide daily coverage of biomass, soils temperature, soil moisture by just simply fusing the different sensor. So, I see a promising role and our share all of your thoughts as you know, we haven't had this conversation. So that will be a topic for tomorrow in the breakout.

CHUCK:

Yeah. And I was kind of placing the thought, you know, and, and it's not necessarily sensing into the soil is even like, you know elevation or landscape possession, things like that that would help in turn, make the soil data more interoperable. You know, we've talked a little bit about some microbial data using the wetness index and the radar data greatly helps expand, you know, the understanding of the carbon or

microbial data. And I think that's what surprised me on a lot of these data information sites. That they weren't using accessing group remote sensing with NOAA, or, you know, whatever Australian agency or whatever to help better interpret the data. That's all I'm saying.

BRUNO:

Excellent. We're do for a break, but allow Alfred to ask the question if he's, if it can be brief.

ALFRED HARTEMINK:

I'll try to be brief. Thanks, you. So, the, I think in reply to Chuck, there's a lot of, there's a lot of bottom-up studies that in digital, so mapping that use all three in parameters, but maybe they have not entered into the database. But my question is how are we going to guarantee that are going to be more studies with depth beyond the 20 centimeters within this framework?

BRUNO:

Yeah, that's a complex question there. I think some of this will really shoot. It's our hope to go in a deep dive in the breakouts because they're separated by that. So, unless there is anyone that wants to comment back on Alfred's point about solve that, which puts into the fire here, 'cause it's a big driver of variability within field, but anyone, if, if not, I know people are patiently waiting for this break coffee and bio break. So, without further ado, we adjourn. We really, again at 2:55 with another in-depth session chaired by Chuck with the different representative of the agencies that we have listened to. Thanks for listening. And please tune back at 2:55 PM Eastern time.

CHARLES RICE:
Alison gave a good summary of the last year, the bad thing about COVID, we were supposed to have this a year ago, last June. The good thing is that we spent the last year listening to different information systems around the world as Alison, 8 teams. So, it was about twice a month we were having meetings. So, we were able to get a lot of information. And so, what Alison presented was a summary. But what we have this afternoon is we picked on a few soil information systems to explain what's being done now and some of their thoughts on the process. So, we have a line up here. You can see on the slide. Mark Farrell is with CSIRO in Australia. For some reason, he didn't want to get up in the middle of the night to present this. So he provided a pre-recorded message presentation and then we'll move to the other presenter. So, go ahead and start Mark's presentation.

MARK FARRELL:
Hello and thank you for the invitation to speak today. Unfortunately, the time difference has meant I'm unable to participate live. But I have tried to address the questions present to discussion with regard to Australia's soil information systems, and I hope that this will be helpful to you. The collection curation and delivery mechanisms for Australia's soils are not legal in and of themselves. Our overall aim through initiatives like the Soil and Landscape Grid of Australia, the SLGA is to develop data products that are accessible to many types of users.

It is important that the data is harmonised and the system established in such a way as to be continually added to and upgraded. There have been a number of historic activities over the years in Australia, including the Australian Solar Resource Information System known as SRIS and others. There has also been substantial work undertaken by individual state level were viewed this whole survey and data availability. And of course, there is also the multitude of individual research projects that generate data on soils in some way, but out of the universities and research agencies.

We've been able to make substantial progress in terms of delivery through the Soil and Landscape Grid of Australia Web portal. And this allows access through the browser, Google Earth or data pool for use in GIS or other special analysis workflows. Consequently, this allows access for users ranging from casual interest all the way through to major research and land management projects. When it comes to the users of the data and their applications for it, the short answer is that the user base is diverse, widespread and growing.

At present, the main users would, of course, be other researchers and policy professionals within government departments at both state and federal levels and with goals of either furthering understanding in agricultural or environmental research or working on Land-Use policy. Increasingly, however, there are other users, particularly facilitated by more user-friendly interfaces that enable agronomic consultants and even landholders to learn more about the soils that they manage. A growing area is also the increased interest in valuing natural capital in which the quality and health of soils can feature heavily.

At present, most of the special data in the Soil and Landscape Grid of Australia is from a single point in time and the major harmonization challenge exists to reconcile soil sampling dates that may differ by decades. Data are currently available at 3 odd seconds on the 90 by 90-metre pixel resolution when aspiration to be at least 30 by 30 metres, if not finer. At this stage, primarily due to data

availability and demand, the majority of variables captured are the more traditional form. For example, pH, soil depth, landscape attributes. But also products including mineralogy, plant available and fragments can come in line.

Looking to the future, there is demand for time-series dates to be derived or delivered, particularly to enable a better understanding of the trends at the state of Australia's soils. Increasingly, there is also demand for information on soil biological variables. Both function and community structure. So by this I mean a database of microbial sequences that is special resolved, but also things like nutrient cycling and carbon turnover rates.

Another potential utility for soil data is to integrate with food and produce traceability systems and being able to link with those as they are developed, whilst also capturing other variables, such as isotopic data which would broaden their application. This is proving particularly important as Australia moves to improve its understanding of being able to trace food and protect its high-value export markets. And with that, I'd like to thank you for your time, and I hope you find this useful.

CHARLES RICE:
Alright, thanks, Mark, virtually for that presentation you heard a little bit about the Australian setup.

SPEAKER:
So, next up, we're gonna have a tag team from NRCS with Drew Kinney and Skye Wills. Drew is with NRCS... And I lost my bio, and then Skye is also with NRCS. And they're gonna talk about the soil information system that's in place for the USDA NRCS system. So with that, I guess I think Drew, you can go ahead and start, please.

SPEAKER:
OK, next slide. I just kind of want to give a brief overview of some of the data sets that are available right now from the NRCS website and through the National Cooperative Soil Survey effort. If you look at some of the data on our site, this is our traditional vector-based model data that we provide. Probably the one that's most people are accustomed to is our Soil Survey Geographic Database, or SSURGO, a county-based soil survey that we've been conducting or working on actively for well, since 1935 as a soil conservation service.

The data is structured so that it is really to scale. So, whatever scale that your information you're looking at to do any kind of analysis, whether it be a land resource by region map which would be more of a continental scale type data set, all the way down to our SSURGO data set, which is what we consider our field-level data set. Most of that data, the SSURGO data set was conducted at a scale around one to 24,000. We had a couple of other higher scales, one to 12,000. Some have done it one to 63,360 scale. But those are, if you look at our product that we deliver through Web Soil Survey, that is all pretty much done to a one to 24,000 scale.

We've always known for years that there was an issue with our vector-based data set. You know, it's been a tremendous data set, but there's some limitations. And one of those limitations of recent years was it doesn't lend itself well to modelling. So, we looked to create our data more into a raster data set, which is more conducive to the modelling world.

We produce a number of data sets in the aggregated format or raster format. We produce a gSSURGO product, which is essentially the same as our SSURGO product but in a raster-based format. We deliver that in a ten-meter state data set as well as a 30-meter raster CONUS. Within the last couple of years, we started producing a gNATSCO product, which we call our best available soils product. For whatever regions that we don't have SSURGO data, we have our general soils maps, STATSGO data. So, in the gNATSGO product, we've combined those two, so where we don't have SSURGO, we still have some soils information in there based off the STATSGO data sets. It's proving to be quite popular.

And then within the last few years, we started conducting raster soil surveys using a lot of remote sensing and digital elevation models, wetness indices, satellite imagery, and using that in an inference engine to create soil surveys. And we're starting to produce that information and that also shows up in our gNATSGO data set. All of the raster data sets are available through the USDA Geodata Gateway site. Next slide.

We also provide a number of point information that we've collected during the course of the soil survey work. Most of that laboratory data that we've collected in the field is available through our soil characterisation database, which is available through our website as well. And we have about 65,000, almost 66,000 pedon, individual pedon website information in that database. And we also have another 24,000 official series that we've identified. We have that information on that site as

well. We are also working on a pedon database. All the pedons that we've described in the field that we have, including those that we have not sampled, exist in our database and we're looking to deliver that information here in the near future.

We have some hurdles with that as far as privacy information, so we have to do quite a bit of cleanup on those pedons. But we have about 456,000 full pedon descriptions. If you look into our pedon database as it is now, there is actually about 700,000 pedons, but they're not all complete pedon description's. And next slide.

This is actually a representation of our NASIS database. And its National Soils Information System that is not only what we export our SSURGO product from, which is the SSURGO is our public available data. This is also internal to NRCS. It is a very robust database and as you can see, a very complex database. It encompasses just about every aspect of our day to day operations, including our planning that we also do within this NASIS database schema all the way out to our delivery mechanisms, which are primarily our two predominant delivery mechanisms. Our Web Soil Survey and a web service called Soil Data Access, where you can actually put queries to the Soil SSURGO database itself. Next slide.

You can get access to all of our soils information through this website. And this is actually the tools page on the NRCS Soils web page. You click on any one of those nine buttons and it will take you to any of our data sets, including a lot of our applications and some of the tools that we've developed for working with our data, including our raster data sets. Those tools, they're available for download. A lot of them are produced in our script and some of them are actually ESRI applications. But really, that's all I have for today.

SPEAKER:
So, I'm Skye Wells, I'm the National Leader for Research. Oh, here we go. So, now, we're gonna go back one. Sorry. There we go. So, I started thinking about what I could add to Drew's presentation, and I'm thinking about it in terms of research and when I started with the NRSC what I wanted to know. So, Drew showed you kind of an overview of like, the finished products we have and then the really complicated stuff that goes into NASIS. But really, we need kind of an intermediate level, I think, to make sense of it. So, I try to put a process diagram here.

SPEAKER:
So, we go all the way from collecting samples. Stuff goes into NASIS like Drew talked about. Some samples are sent to the Kellogg Soil Survey Lab, and there are a lot of analysis and other things that are done there. But the way they analyze and store the data isn't the way the public or even other soil scientists need to interact with it. So, they have to export it, and then all of the databases we work with, we can aggregate together. There's some R scripts and some NASIS queries and some cool things that are online, if you're interested in that. But essentially, it's all these behind the scenes databases are summarized and pushed through to components and data map units that go into our published map. And then we can take those published maps and then we can do scripts and queries and all that stuff on them again. Next slide.

>:
So, this is a representation of just the Kelloggs Soil Survey Lab samples. So, the way it's organized is it's focused very internally in terms of managing samples and analytes, methods, procedures, equipment. And then we take that and we integrate it with how we think it should connect to the NASIS or an NCSS database that we produce. And so, I know that's a lot of words, I don't expect you to read it just to know that there's like a nested hierarchy of complicated systems that leads to us have having and providing soil data. Next slide.

>:
So the thing that allows us to do this is a series of guidance documents. These are very much the standard. And as Kathy mentioned earlier, notice that they're each version and we try to store this information in our databases. So, go ahead and hit next a few times and cycle through these. This includes how we sample the soils and includes the National Soil Service Handbook, which is actually statutory and says how we have to do things. We have two different lab manuals and then we have some documents that go into what is in our information system. What's in NASIS altogether that's as specific as what is the meaning of each column? How is it displayed? What type of data is it? What are the allowed entrances into that column? So, we need guidance documents that take us all the way from the very basic stuff and relatively simple to the very precise, how do you actually put information into the system? Next slide.

>:
So, I wanted to highlight what I think is really important and is maybe overlooked when we just talk about the things we produce is that it's not just the information system is what all these people do in the information system. And from NRCS standpoint, this includes soil scientists that are gathering data.

It includes lab analysts that are making measurements. That includes a review of regional staff at a couple of different points in the process. And so that interaction of individual people and that QAQC process that not a lot of small scale data sets can keep up with, really is at the heart of us keeping on track with this vast array of data we have. OK, next slide.

>:
So, I thought since the title of this is Dynamic Soil Information Systems, I'd highlight the two ways we use dynamic in our databases. One is in the data hierarchy. We actually have a way to record when we visited the same site multiple times on different dates. So that's really strictly dynamic. And we also try to put information into infer dynamics and land use and management information. We haven't always collected that over time, but we like to get into space for time comparisons or maybe the ecological site descriptions that Dave mentioned so that we can infer what's happening over time. So next slide.

>:
And so one thing you guys might have heard of is the Rapid Carbon Assessment or RaCA. So, in order to make that happen, we had to take our standardized guidance and our standardized databases and lay a new level on top of it. And I'm going to tell you about this, because all of this data has been collected. It all goes into our new SSURGO products. However, if you want the individual point measurements, they are on a giant Excel file on my desktop computer which is not ideal. It's not the way we want to do things, but sometimes for special projects, we have to do it that way. And so I thought quite a lot about how we might want to do this. And this is in the future, but we're still working on it. So one more slide.

>:
So the methodology and guidance documents are collected, they're put on websites, and then the raw data lives on my computer in a very large Excel file, I should say. And we can share it but there's some steps we have to go through and that part isn't automated. And so maybe we want it to be in the future or maybe we don't. I'm not sure. OK, one more slide.

>:
We are working on the next step in this process, and I'm including it, even though it's not quite existing, because we actually do have contractors and developers working on this and we're calling this the DSP Hub Dynamica Soil Properties. So, we're gonna link data from multiple USDA sources. We're gonna apply models and interpretations with data. So that should allow us to bring in things like remotely sense data, geographic data sets that don't to our normal conceptions of point samples. And then it will also allow us to embed metadata. That will include all the versioning of the standards, versioning of data and models. And that way, whenever we make predictions or do anything for interpretation, it's stamped into the system. And so it's pretty ambitious, but we will at least have a prototype within a few months. So, I just wanted to bring that up. And that should be my last slide. Thank you.

SPEAKER:
OK, thanks, Skye. So that gives you a little overview or an overview of the NRCS data set and their efforts.

SPEAKER:
We'll switch now and head over to Europe and Dr Luca Montanarella is from the European Commission Joint Research Center and they've been doing a soils database. So, Luca, if you want to give an overview? Thank you.

LUCA MONTANARELLA:
Yes. Hello, everybody, and thanks, Jack, for introducing me. And thanks to the organizer for inviting me, actually. Yes, I work for the European Commission. For the ones who are not familiar, the European Commission is the Executive Body of the European Union. And so let me talk on how we started to deal with soils in the European Union.

We wanted, since it's our mandate at the European Commission to build on European Union, we wanted to build on the European Union for soils and for soil data. And so in the early 90s, we started the process, which was quite lengthy and painful, I must say, in bringing together into one single system and into a harmonised database, data available in our EU member states.

That required a long and quite tedious work of harmonisation, of harmonisation, especially of data sets that you must be aware they are collected in very different time, under very different conditions and using very different methodologies and standards.

Many European countries have a long tradition in soil science, have a long tradition in soil survey, some dating back even to the 19th century. And so you have different classification systems, different standards being adopted and used at national scale over many years. So, it took us quite a while to come to a common understanding of European soils. And we take it as a success story that we could build a common European Union for soils and that's the picture you see here.

At the very end even our candidate countries, so countries who were candidating to enter the European Union, usually have joined the European Soil Information System before even joining the European Union. And one effort that then came out of this was also an effort to make this data public and particularly make them understandable to the general public. That's why we invested a lot after completing our Common Soil Information System and also preparing a series of documents, particularly in the series of atlases.

'The Soil Atlas of Europe' was the first one, and now we have many others. By the way, we then embarked also in doing the same exercise in other continents. We completed 'The Soil Atlas of Africa', 'The Soil Atlas of Latin America'. Currently, we are completing together with our Asian colleagues, 'The Soil Atlas of Asia'. So, there is been something initiated through this exercise that is still ongoing.

The other thing that you should be aware when we talk about developing common soil information systems is that we work at the interface between science and policymaking. We are not a research organization. We are a science policy interface within the European Commission. Our our is to support EU policies related to soils with relevant evidence, data and science-based in order to make science-based decisions, so using hard evidence for EU policymaking related to soils. And that's

why... next slide, please.

That's why, having completed the first common understanding of European soils, allowed us then to initiate a process that led to having all the common legal framework for soils in the EU. And this is crucial for us because that justifies then the further investment in more sophisticated data collection exercises in more detailed information systems that are coming up then in the later years of this process.

So, in 2006, we presented a package that is known as the EU Soil Thematic Strategy. I don't go into the details of that because this is not about data it's about how we want to manage, in a sustainable way, soil resources in the European Union. But what is important in this strategy and next slide, please, is to understand the core understanding and concepts that we embedded in our EU vision concerning soils.

We embarked in a vision that is centred around the vision of multifunctionality of soils. So, we don't look to soils only for one function, which is what I heard, by the way, in most presentations until now in this workshop. So, essentially, the function of producing biomass, so, essentially, the agricultural activities that are based on soils, deforestry activity, all the traditional soil science-based soil data that have been developed over many years, mostly geared towards improving our agricultural production and productivity of sorts.

The European Union adopted in the Soil Thematic Strategy, a multifunctional view of soils, and this is crucial for us because these seven functions that you see listed here, it took us quite a while to negotiate them in the Parliament and in the Council, are these seven functions that we have recognised as delivering services beyond property rights. So, one crucial element of our legislation is that we don't want to protect soils, we want to protect soil functions that are relevant to all EU citizens. So, these seven functions deliver services that are going beyond private property rights, and that allows us then to legislate about those functions.

I mention this because this, of course, has been driving all our data collection efforts. So, we collect data in order to document progress made in protecting those seven functions. I just list them here, we can quickly go through them. Beside the biomass function there's the storing, filtering, transforming nutrient functions or producing, having good drinking water, for example, the biodiversity functions or being habitat to be protected, the physical cultural environments, all the issue about urbanization, infrastructure, the source of raw material function. So, we have still quite an interest in having soils that deliver to us some raw materials.

Very important to us in the climate change debate is, of course, the function to act as a carbon pool that was singled out exactly for the purpose to be then used within the climate change legislative package. And finally, a very important thing for European citizens, which is the historical archaeological heritage, geological heritage function. Europeans are very much attached to it. So, all this explains a little bit our efforts to then document this with data. Next slide, please.

So, we have now operational, a system which is called LUCAS. It was already mentioned in some of the previous interventions. We want to detect changes in soil properties. To detect changes you

must monitor. You must go regularly on the same spot and measure again parameters that you would like to detect improvements or changes in some way over time.

So, we have devised a system that goes under the name of LUCAS, which stands for Land Use Land Cover Arial Survey, where we have on a regular one by one - sorry, two by two kilometer grid, which corresponds roughly for the European Union to one million hundred thousand points. A stratification of sampling points, which allows us then to go on a subset of roughly 25,000 points, which are highly referenced. So, very well documented where these points are and what type of land use is happening there, so we have surveyors sent by us on these points on a regular basis.

And we have also a systematic approach for collecting those data and samples. One key element of all these systems is, of course, the archive of soil samples. So, having a long-term storage of samples over time, where you can go back to previous exercises so that you can also reanalyze stored samples for some parameters which are maybe stable over time. And the second big element that we introduced was to abandon the idea to have several laboratories doing the analysis in the different EU member states. And instead, we decided to select one single laboratory at EU scale that would do all the analytical work.

This for the simple reason that our experience with the many interlaboratory calibrations we have been doing over the years, showed to us that interlaboratory variance of results. So, if you give one sample to 50 laboratories in the European Union, you get 50 very different results, even for very simple measurements like pH or things like that.

So, this system is now an operational system. So, this is not a research exercise. It's run by our statistical service, which is AeroStat and is part of the official statistics of the European Union. We are now currently doing the first survey, the 2021 2022 sampling survey. We already completed 2009, 2015, 2018. And we have all the firm plan for expanding this system in the near future. We will probably go up to 250,000 sites to be surveyed on each survey. So, ten times more this because we have a firm plan to use this tool, and I will talk a little bit about this in the implementation phase of our new European Green Deal, which is very much centred on soils. Next slide.

So, what do we measure there? We a list of parameters that is evolving over time, by the way, because it's trying to be adapting to policy priorities that are emerging over time. These parameters get measured on points, but of course, then we do some special interpolation exercises to produce maps that once you give, example, organic carbon or phosphorus content or fungicides or some contaminants or whatever, and so you get regular reporting about these parameters with this system. Next slide.

So, let me conclude with what is the future vision. It was already mentioned, we launched on the 4th December last year, 2020, a new initiative of the Commission, which is called the EU Soil Observatory. This is embedded into our new policy framework, which is called European Green Deal. I don't know if you heard about this, but this is a very large program that our new President, Mrs von der Leyen, has been launching. Soils play a central role in the European Green Deal because they are linked to many of the strategies of the deal, particularly to the climate change, the biodiversity and the farm to fork strategy. And so we will have a much more structured monitoring system even

further strengthened from what you have seen so far. And all data we hope we will then deliver as always over our European Soil Data Center that you can freely access and all data freely available.

So, this is just to give you a very brief overview of what we are doing and what our plans are. But I'm more than happy to respond to any questions or further details you would like to know.

SPEAKER:
Thanks, Luca. So, we'll have questions and the panel discussion afterwards.

SPEAKER:
So next up, I will stay in the (INAUDIBLE) and Rik Van den Bosch from the...from ISRIC and we'll talk about the world's information network that they have. So, Rik.

SPEAKER:
Thank you, Jack. Thank you to the organizers to have me. Can we go one slide up, please? The other way round up, the first slide. Yes, thank you very much. So, one slide about our institute. ISRIC is a small institute in the Netherlands. And our vision is a world where reliable and relevant soil information is freely available and properly used to address environmental and societal challenges. So, we are working with 25 people every day on gathering soil information, harmonizing, standardizing and provisioning soil information, but also helping us to do the same for their own territories. We have four workstreams at ISRIC. One is about standard-setting and standard-setting is typically something you do with a team of people, with a community. So, we do that basically within the global partnership of the FAO. And our standards are set for measuring soil, for analyzing soils, for setting up databases, for interoperability and for serving data.

Our second workstream is global data provisioning. I'll have a couple of slides on that later in my presentation. We do quite a bit of capacity building and capacity building is geared towards the National Soul Information Institutes around the globe, predominantly in Africa, and we also work on applications of soil information. Next slide, please. So, one of the products that we have is the world soil information system or WOSIS. WOSIS is our point data repository. Actually, ISRIC is acting as a custodian of soil information. Sorry, thank you. As a custodian of soil information. So, everybody always has datasets and is not sure whether he or she can gather that data for himself and also serve it to the public. They can send it to us and we put it in our repository. We serve it so everybody can find it. It doesn't get lost. And at this point in time, we have about 450,000 profiles registered in our data repository. And we work almost daily to standardize those datasets into a standardized environment and that is then called the WOSIS database. Next slide, please.

So, the standardized data from all those data that we have in our repository are now to about 200,000 profiles that we serve openly on our website with a CC BY license. And in addition, we have 24,000 profiles that we cannot share, but we can use for our own applications. Next slide. So based on that point dataset, we produce a global grid on 250 by 250 meter through predictive soil mapping or digital soil mapping using also, of course, the covariance from Earth observations. And that global product is, as I said, on 250 by 250 meter. It provides both physical, so physical and so chemical parameters. And we predict up to a depth of 2m. It is a reproducible workflow. So, every now and then when we get additional datasets, substantial amounts of additional data, we can run the whole workflow and update our predictions for the globe. Next slide, please.

So, if you want to access to those point datasets or to those grid products, you can go to www.soilgrids.org. Both the point data and the grids are available there, but we also have uploaded them to Google Earth Engine. So, you can also go there to get access to it. Next slide. A couple of other remarks. We are also working towards trying to quantify...soil quantity indicators. So, at this point in time, we only have the basic soil parameters. But our next step would be to also quantify in a more routine way soil quality indicators. The users of these products are clearly global users. Very many...most of them are academia global modelers, but also UN organizations are using these global products. And we believe that for real impact on the ground, these global products are not very suited. It's much more effective if we help National Soil Information Institutes to produce soil information systems for their own territories, using their own datasets, using their own knowledge of the soils and for their own clients and users. So that's why we have increase in capacity building

program where we work with National Soil Information Institutes around the globe to help them to produce those systems.

And last remark that we would like to make is that the GSP, Global Solar Partnership managed by the FAO is working on a global scale information system which has a federated character and we are also participating in that. And the objective here is that we can provide a sort of a standardized package to national institutions where they can start with to build up their own national soil information system relatively quickly under the umbrella of the GLOSIS methodology. And ultimately, those nodes, those national nodes will be connected to each other through common standards and a global…and a discovery hub so that users can explore these different nodes in the different countries and get information from the entire system. The big advantage here is that it can really help national institutions to get going quickly, especially those in the South, with setting up their own national soil information systems. I think I will leave it here. I have made a couple of notes based on the presentation that I heard, but I might probably introduce them when we get to the discussion. Thank you, Jack.

SPEAKER:
OK. Thanks, Rik.

SPEAKER:

So, alright. Well, and then finally, we're gonna come back to the US and Dr Samantha Weintraub is gonna talk about NEON. You've already been introduced to NEON. Matt King talked about it from the NSF perspective, but we'll get a little more detail on the data collection, soil data collection from Samantha. Samantha?

SAMANTHA WEINTRAUB:

Great. Thanks. It's wonderful to be here. Next slide. So, I think I don't need to dwell too long on this because we've already heard about NEON, in a couple of different sessions today. But we are a continental-scale US-based observatory. There are 81 field sites, although 47 of those are terrestrial. So those are the sites that we're monitoring soil plants and things that grow on land. We do produce 181 data products with the key goal of shedding light into how our ecosystems are changing. So those data products are very diverse from eddy-covariance, micrometeorology disease, ecology, biodiversity monitoring. So, soils is a small but important piece of the broader suite of things that are going on at the national ecological observatory network. We do have 15 soil data products. And I'll explain a little bit more about what those are in the next slide.

So, our soil monitoring and measurements are varied in both temporal and spatial scale. In terms of time, we have some products that are one-time characterization efforts, and we did partner with NRCS. We've heard a lot from today. So, using a lot of their standard methods for both fields and lab procedures for doing soil taxonomy, pathology, geochemistry, and then moving through kind of the inter-annual seasonal timescale.

We sample for biogeochemical processes. We measure soil, carbon, soil, and organic nitrogen soil microbes, both through PLFA and genomic efforts. And then going to our sensor data, which are streaming at, you know, on the minute timescales. So very broad temporal types of data. And in terms of spatial coverage, we're going from single points in the landscape up through maybe one Hectare of, with a dense sensor array, and then using stratified random sampling approaches to characterize entire sites that are multiple kilometers squared large. There's also a depth component here. So, depending on the type of soil sampling, we're either focusing on the surface or we're going all the way down to either one- or two-meter depths. Next.

So, we, NEON has very standardized procedures for how all of our data are ingested, created and shared. This is just a graphic on left showing the general workflow. This would be for an instrumented system, a data product such as soil moisture coming from a sensor, but it's a similar much of the workflow is overlapping between observational data. Where let's say we have technicians going out and measuring things in the field or data going to labs, and then chemical and physical data coming back.

So, the data are adjusted. We have some initial quality control to even ingest that data into what we would call the level zero, the base data. But then there's quite a bit more processing that happens in it. There's some degree of algorithms calibration, we're doing quality assurance. Some of the sensor data products actually need quite a bit of calculation and higher-level data algorithms to go from, you know, voltage from a sensor to unit of soil moisture let's say, or temperature.

So once all that's done, it goes into our process data repository. Gets published and then is available free and open on the NEON data portal. So, this is where you could go to get our soil data products,

but really to get any of our data products. The soil data portal is, or the NEON data portal is also the hub where you could download the protocols and the algorithm documents and data product, user guide documents that explain how the data are collected, curated and published. And you can also download our free and open data either from the portal, or I would say probably the majority of users are actually using an API to access our data. Next.

We have a wide variety of data users and data collaborators. I would say probably our bread and butter to date are academic researchers, professors, their students, post-docs. People who are seeking to just understand fundamental questions in ecology, soil, ecology, biogeochemistry. But lots of other people are involved. We know that local land managers we have, we don't actually own the land where any of our sites are.

So, a lot of that is owned and managed by the forest service. Several sites are managed by the Nature Conservancy, and we know those land management partners are interested in trying to see how they can use our data to inform land management at the sites. We have been working with lots of different data repositories to promote the NEON data showing up in synthesis efforts like the SODA database, ISO bank, that's getting off the ground, which is a stable isotope database to improve discoverability and integration with other data efforts.

There are at least a couple instances I know of where industry groups are kind of NEON curious and seeing if they can use our data and partner with us. Certainly, modelers are very excited about the standardized and continental distributed data sets that we have. And they're active partnerships, for example, folks at NCAR. Educators, we have of kinds of tutorials and lesson plans and modules for how we can teach students how to use big, open ecological data, including soil data and there some partnerships going on with different national labs.

So, I think I'll just leave it there. You know, we can talk more in the panel and I'm happy to answer questions about kind of pros and cons and what NEON can offer. But what is more challenging with the kind of standardized top-down system that we have. I think that Alison articulated really well some of the trade-offs this morning. So, thanks very much.

SPEAKER:

OK, thank you, Samantha.

CHUCK:
We have about 10 minutes, 12 minutes or so to open discussion. I haven't... Does any... Is there any hands raised up there? No. OK, any questions? And again, we've had... There was a question from Slack. And Luca, and you presented this to the group, to the committee, but kind of surprised about the quality of the testing between labs, comparison and the issues related to that.

LUCA:
It's a big issue, it's a big problem, at least in Europe, I don't know in other parts of the world. We had a previous experience. We were running over many years forest soil monitoring system, maybe you were aware of it. It was only on forest area, on soils, run within the ICP Forest Initiative, where we were doing extensive inter-laboratory calibrations and we were using the national labs for doing the analytical work. And it turned out that suddenly PH between Germany and France was changing across borders, or things like that, simply because measurements were done differently, let's put it this way.

So, there is an issue here of having comparative data when you use multiple laboratories, especially if you do it on a national basis. To build a common European Union, you cannot have that suddenly some soil properties change when you cross the border for any specific reason, unless there is a real natural change in the landscape on the properties, but on average some parameters that change suddenly, that's due to the analytical part or maybe to the sampling part. That's also why we centralize the sampling strategy so much and we use teams that are completely trained and managed by us directly, not on national level, because otherwise we would end up every time with a patchwork of different parameters, giving different responses according to national boundaries. This is what we wanted to bypass, is this system.

CHUCK:
OK, thanks, Luca. Tomislav, you had a question? Raised hand.

TOMISLAV:
Yes, I have a question just for Luka. So, we use LUCAS for a couple of projects now, it's really amazing data set and I think it really shows that Europe is... Maybe I'm biased, I'm from Europe, but it shows that really advanced in the monitoring campaigns. And I just want to ask Luka just to give us an idea, it's about 20,000, 30,000 locations every three years. What's the cost of that?

LUCA:
It's costing us, every survey, around eight million euros, which in dollars is probably, I don't know, $10 million? I don't know, euros, I don't know the change rate. But I mean, in euros, it's eight million roughly.

TOMISLAV:
But I would still say it's a low cost considering the benefit of having the whole continent covered. And do you think the cost in the future likely are going to drop? Because most of data is based on soil spectroscopy, right?

LUCA:
Well, let's say, first of all, concerning funding of such an exercise, it's not an easy thing to convince the parliament because, keep in mind, this is an operational system. So, it goes into the normal

funding of the European Union. So, we go through a discussion in the European Parliament for the funding of this exercise. And we have really to have good arguments to convince that taxpayers' money of European citizens should be used for soil monitoring.

Concerning reducing the cost, it's true what you say. From the very beginning we introduced the issue of collecting also spectral reflectance data from the soil samples so that we have now probably one of the biggest spectral libraries in the world. I don't know how many hundred thousand spectra we have there. And our dream was, of course, to get rid of the wet chemistry part for many of the parameters. I must tell you, the results didn't show us that we can really move on towards lower cost, using only spectral reflectance for some parameters still, but maybe we still need to do some research on that. But we really still continue to use a lot of the traditional wet chemistry for most parameters. And this is, of course, adding to the cost.

But I must tell you, the highest percentage of the cost is not the analytical work, it's the sampling. So, having a team, and we have a team of roughly 5,000 surveyors, they need time. And so, it's quite costly. So, it's sampling overall compared to other media.

TOMISLAV:
OK. Just a bit of criticism also, something that also that Alfred (UNKNOWN), I think mentioned, it's a pity it's only zero to 30. It's a real pity. People walk to these places, as you said, this is the most expensive for people to go out and walk, and then they describe the soil zero to 30 or zero to 20. And it's a real pity because the soil is interesting all the way to 2m, almost. So, that's just a bit of criticism. I don't know who decided that, but...

LUCA:
No, I can explain to you. It's a compromise, unfortunately, as all things in the European Union. The compromise is due to the fact that we must decide if we want to have more points or if we want to have more in-depth sampling per point. So, having less points, but spending more time on each point, because at the very end, the cost of sampling for us is simply the cost of the time of the surveys. So, we can have a strategy where we spend more time on single spots, on single sampling sites, and have less points. This is a decision that depends on the purpose of the survey. At the end, we have been striking a sort of balance of that, could beat it in the future, we we will go more in-depth on certain sites or build a tool system with intensive monitoring sites and maybe general monitoring sites on top, let's see. For the moment this is the balance we have been striking, but it could change in the future. Yes.

TOMISLAV:
OK, thank you.

CHUCK:
Alright. I've got a couple of questions, but I'm gonna take moderator choice here and ask one for everybody. I guess for all the different information systems, what do you see in the future as far as your end users changing? Luka, you mentioned you're more for the policymakers. Samantha, it's more for the scientific academic community. I guess, how do you see - and this is for everybody - the challenges and opportunities for expanding the community of users. Anybody can jump in. And as you expand, are there different needs?

SAMANTHA WEINTRAUB:

Yeah, I could start. So, I think that the vision of NEON is certainly to provide ecological information that could be used to set policy, manage land and make decisions, but I will say, and a few people have mentioned this, that you would need a level between the data that we provide, would need to be kind of interpreted, modeled. You would have to have more derived products to have it sort of be relevant and digestible by policy makers and decision makers, I think. And so, I believe that's why I kind of at the phase we are now, academic researchers are the ones who can kind of deal with it and really take data and make it knowledge. But if we can figure out how to do that more regularly and systematically and then make it available for land managers and decision makers, I think that will be a really exciting evolution for NEON hopefully over the years and decades.

CHUCK:
Rik?

RIK VAN DEN BOSCH:
Yeah, we had quite long debates about this this week. In the past we were always focused on higher resolution, higher accuracy on global level, but now we discovered or we didn't discover, we realized that global users don't need higher resolution. If you look at IPCC, they want information on 1km by 1km or 10km by 10km. So, we are focusing more on those global uses and package the information that we generate (UNKNOWN) towards their needs and probably aggregate that information. So, that is a movement we make, which is a bit surprising for ourselves. And then on a more landscape level, we rather not do it by ourselves, but work with the national institutions to get proper products for national territories. So for us, it's really also focusing on the user. And depending on the type of information that they need, we've got to repackage what we have according to their needs.

CHUCK:
Sky?

SKY:
I can jump in here. From my perspective with soil survey, soil plant science division in NRCS, we're really interested in sort of a two-pronged approach. In that dynamic sense, we want to have more specific information, both spatially and temporally. That's really helpful for things like PrecisionAG really, we hope, helpful for lots of other land uses and decision makers. But we're also just interested in providing new properties that we don't think about in these older information systems, so that things like Greg was talking about with the soil quality indicators, we have some DSP for soil health project. We're doing things to try to figure out how we can incorporate the information into a system. So, we will need to make individual measurements, but we'll also need to figure out rules for aggregating measurements because just giving people a lot of raw data isn't really our business.

SKY:
And so, we are having a lot of conversations about how to maintain our current user base and our current products that people rely on and also expand into these new areas and it's not straightforward.

SPEAKER:
I think I'll probably expand on that too and say that a lot of our future delivery will probably be cloud based because of some of the size limitations that our current systems have and just the size of information we're trying to deliver to the public.

CHUCK:
Anybody else? There was one question specific for Samantha. They want to know what the difference was between LTER and NEON.

SAMANTHA WEINTRAUB:
(LAUGHS) So, I would say LTER is very question driven. Each LTER site has this theme and they have specific questions and hypotheses that they are testing using their own methods and setting up the management of their LTER how they want to. NEON is really different, it's to provide continental scale data sets, observations of ecological and environmental properties using standardized methods, standardized lab, standardized sensors. So, again, harkening back to kind of Alyson's points this morning, I think they're well suited to approach different kinds of problems. I think they're complementary and they both have an important role to play in our ecological data ecosystems, if you will, but very, very different approaches.

CHUCK:
And I'll add to that, since I'm situated here at Konza, which is the Tallgrass Prairie LTER. As Samantha said, it's objective driven. It's what's the effect of fire on the prairie or grazing on the prairie and prairie health. And a lot of times are co-located, but not necessarily, but they don't have the... They're different missions. So, Alfred, I think you had a question.

ALFRED:
Thank you, Chuck. So, maybe a question for whoever wants to answer. So many of these databases have been developed with maybe the idea of the databases, 20 or 30 or even older and have been developed with a particular user in mind and a particular community in mind, perhaps particular, as I mentioned earlier, an agrarian field. But how do we serve the rest of the community? How do we serve the urban community who wants to have information on contaminants or hydrocarbons or (UNKNOWN) or anything that is really maybe affecting many more people than when we currently serve with the current databases?

CHUCK:
Anyone wanna jump in? I don't think NEON has any in the the urban environments, do they?

SAMANTHA WEINTRAUB:
I'm gonna say, unfortunately, there was a decent urban set of sites, so that would be a theme, to look at everything we look up within the urban and they were de-scoped. And so, we nope, we're wild land and (UNKNOWN). Yeah.

CHUCK:
Yeah. And others LGRs and urbans, there is two or three of them. But Sky, do you have any comments, or Drew?

DREW:
Well, we've wrestled with that for quite a while ourselves, as far as what the customers we deliver to, ultimately it's up to them on how they're going to use our data. And I don't know, because of the diversity of our audiences, I think it's almost impossible to cater it to anyone one specific. So, we try to give as general information that we can that can be broadly used.

SKY:

I would just add to that, I mean, that's part of why we have several different versions of our data out there. It's meant to go to different users, different levels of detail. But I mean, we're working all the time to incorporate new areas like urban areas, urban gardening. It's not a static group, but it's to meet everybody's needs.

CHUCK:
OK. I've got a question from Slack. The question was, I guess Europe, you're looking at combining monitoring systems at different scales like the National and the LUCAS, are there similar challenges in the United States to combine monitoring systems? I guess that refers to maybe NRCS versus NEON versus USGS. Anybody wanna... Is Dave Limbaugh still on?

DAVE LIMBAUGH:
Yeah, Chuck, I'm here. And, yeah, that's gonna be an issue. So, we've doing what we can, I'll leave it at that as nebulous as that can be at this point. But yeah. So, I'll leave it there.

CHUCK:
OK. Alright., thanks. Well, we've used up our time, so I'll turn it back to Bruno to end and put what's on tap for tomorrow.

LUCA:
Thanks a lot.

SPEAKER:
Thank you.

BRUNO:
Well, I just want to conclude this very productive day by quickly going through the agenda tomorrow, is quite different. This a little bit more deep dive. Later in the afternoon we will start with fireside chat chaired by Ramveer Chandra with the industry representative, representative from Land O'Lakes, UPL and Pivot Bio. So, look forward to hearing about that discussion. And then the rest of the day we have to really focus on breaking out in sessions. There will be two parallel session with three different topics. You can see the topics there. We really would like the participants to go much deeper into discussing measurements and sampling and archive from all the way from collecting in the field to processing the samples, how we archive.

The topic B is the collection and duration with all of the harmonization and related issues. And then topic C is data analysis, data analytics and models, machine learning and process-based model. We'll break after that, and then the people which have been preassigned will change and go to the next session. And so, we will have a good mix of opinion and people don't have to stay always within their fields and boundaries. On Friday we'll have a very detailed discussion about all the topics.

Given the time, and especially for our European friends, that it's getting late. Today is about 10 p.m., I would like to conclude today with a special thanks to the speakers, starting from the keynotes and the panelists and people that have contributed throughout the day. I thought it was a very productive day. As you know, it may not seem that we have gone far, but I really think we have really put out a lot of great topics of discussion, which we plan to continue to go deeper, as I mentioned, in the next couple of days. We will have a nice synthesis at the end as well as a workshop report.

So, one of the things about tomorrow with the session, if we could really envision for the note taker and the moderators in the session, see if the discussion can go towards the direction of either consensus or very complete opposite point of view, but at least to see what could be reported out as a potential direction of how we're going to build this sort of dynamic information system and use previous experiences, as we have shown in some of the cases today.

So, with that, I really thank you very much indeed for your attendance, participation. And I will see you tomorrow at 11am Eastern Time. And with that I wish you a very good afternoon or good evening to anyone.

SPEAKER:

Hello, everyone, and welcome back to day two of the workshop, Exploring a Dynamic Soil Information System. We had a wonderful day yesterday, very engaging and learned a lot. Great presentations from the keynote and panelists, wonderful discussions. So, I look forward to another great day and I would just like to pass it on to Ranveer Chandra of Microsoft Azure Global, who will be moderating a fireside chat with the industry representatives. Ranveer, take it away.

SPEAKER:

Thank you, Bruno. And hello, everyone. Good morning and good evening wherever you are in the world. And today I'm joined by three esteemed panelists. So, this is to continue the discussion we were having yesterday. And as many of you heard, the role of private sector came up multiple times. And today, we have three key leaders from the private sector, from both established companies, startups, from a farmer perspective to input companies or the leaders on that space to some of the innovative startups in that space, too. So, the key people we have today, one is Teddy. Teddy serves as the Chief Technology Officer of Land O'Lakes, which is one of the largest farmer-owned cooperatives. And he's leading Land O'Lakes Ag Tech and IT organizations. Teddy is responsible for developing and implementing technology solutions for retail and farmer customers to help them produce more sustainable outputs by leveraging agronomic insights from Answer Plot locations with Winfield United Innovation Center and the knowledge of the organisation. Teddy's application of technology and data to the practice of farming has shaped product offerings, such as WinField's R7, Answer Tech and ATLAS portal, which are all key to some of the things we are discussing today. There's a lot of data that is gathered, a lot of information about farmers, and it's also about giving insights to farmers. Teddy holds an MBA from Indiana University and a Bachelor of Science in Mechanical Engineering from North Carolina State University,

Adrian Percy... he serves as the CTO of UPL Limited, a major crop protection company that is a leader in global food systems. He's also a Venture Partner at Finistere Ventures, a technology and life sciences venture capitalist investor focused on transforming the food value chain. Adrian is an advocate of the need for and benefits of modern agriculture. He's also a strong proponent of the development and adoption of new agricultural and food technologies that support global food security while conserving the environment. So, born and raised in the UK, Adrian holds a Doctorate in Biochemistry from the University of Birmingham. He now resides in the Research Triangle area of North Carolina in the US.

And the third panelist, Dr Karsten Temme, is CEO and co-founder of Pivot Bio, a company that has designed and commercialized a microbial nitrogen solution to displace synthetic fertilizer. He has dedicated his career to improving farmer outcomes while decreasing agriculture's environmental footprint. Dr Temme earned his Bachelor of Science and Master's Degree in Biomedical Engineering

from the University of Iowa and his PhD in Bioengineering from the University of California-Berkeley Joint Graduate Group.

So, welcome Teddy, Adrian and Karsten. It's really an honor to have you in this panel. So, to begin with, I want to start with the theme of the panel, that is, why do you care about soil health... about dynamic soil health? So, let's start by Teddy.

SPEAKER:

Yeah, thank you, Ranveer, and good morning, everyone. Glad to be here with you today. You know, soil health, why don't you care about soil health? That would be the question. But jokes aside, I mean, from myself and for our organization, I mean, it goes back to what Ranveer said a little bit. We're a farmer-owned cooperative, farm-to-fork cooperative at that. And, you know, we look at the three divisions within Land O'Lakes. We have crop inputs, animal nutrition, and the Land O'Lakes dairy business. And crop inputs is the largest group of the three, it makes up nearly 50% of our company. And the way that that organization works is we're a value-added distributor and we buy products from a variety of different agricultural manufactures and we sell that to ag retailers that then work with farmers to actually put these inputs into the ground and then manage the crop throughout the lifecycle.

One of the areas we differentiate ourselves and we continue to differentiate ourselves is to bring key quality recommendations to farmers and how to best manage their fields, how to best manage the crops that are on to the ground, how to care for it, not only in a productive way for their bottom line, but also in an environmental, sustainable fashion. So, soil plays a very, very key role into that. One of the pieces that Ranveer mentioned was this concept of Answer Plots, these research throughout the areas in which we do business, which is majority of the United States where real crops are. We plant different seed varieties. We test them in different soil conditions, soil types, different environmental conditions, different farming practices. And we collect this data. And then this is what's made available to farmers to make the optimal decision on their field. And making sure that we provide the best possible recommendation really goes back to do we understand that soil, do we understand the dynamic properties of it? Now, it's just not a static thing, right? It's over time. It's changing, different. One given field of 60 acres may have different compositions in it.

And so, how do you make that recommendation in such a way that it's most productive for the plant, but also for the health of that soil? Because we want to make sure that that soil kind of goes from one generation to the next. So, I mean, there's a lot more to get into this about this, but it becomes a very key point of who we are and how we serve the farmers who are the owners of our cooperative.

SPEAKER:

Thank you, Teddy. And Karsten, do you want to add to that? Why do you care (CROSSTALK) I know this is fundamental and (CROSSTALK)

SPEAKER:

Maybe I'll start more on a personal note. And thanks for giving an intro to all of our backgrounds. I think I'll note in mind, I'm trained as an engineer, but it's an engineer in the bio... the living part of our world. And so for me, I think that the passion with agriculture and specifically around soil health becomes the dynamic living components. The part that makes soil and its influence on agriculture change on a daily basis and in all the different spatial parts of our world. So, maybe three things I think I'll try to highlight throughout the day is you know, ultimately at Pivot, we're trying to help every grower produce a more bountiful crop in a more efficient way. And we want to be able to move new ideas and innovation into the marketplace fast as possible. And at the end of the day, we'd like to be able to leave things better than we found them. So, make sure that that underlying asset, the land itself, is more productive and more sustainable in a way that's beneficial not just for the grower, but all the rest of us around the planet. So, I think that's where my passion comes in. It's dynamic because this is a living world. And then there's a lot left for us to learn.

SPEAKER:

Wonderful. Thanks, Karsten. And Adrian?

SPEAKER:

You know, I mean, I probably would say the same thing as Teddy to start off is why wouldn't you care about something so fundamental to our production systems and to the sustainability of our planet? But it is a little bit ironic, I'm sure, for many of the folks on the call that may have been in this area for 30 years or their entire career, that this is now seen as kind of the new frontier. But that's really the way, you know, industry is looking at it. I mean, I think the interest in this area has only emerged relatively recently. And I think across the input industry there's now seen as tremendous white space that, you know, we could provide growers with much better support, much better understanding of their soils and then provide inputs that can help them optimize, you know, the yield on their land, while at the same time, conserving resources and improving the soil quality.

So, this is something that's become pretty fundamental to our business and beyond just understanding the soil and using things like a soil mapping system would be so useful for that to help us actually design some of these products. So, providing prescriptions to growers is one thing, but actually understanding how our products are working in the R&D phase is also really important. And obviously, soil is a tremendously complex system. And right now, sometimes we struggle to really understand how some of these products work and to make sure that we have a... we can actually go to growers and say, OK, under these conditions, in this environment, these products are gonna work in this way. And that's why we need so much more research in this area. And so, I'm so happy to see these types of different conferences taking place because the more and more input that we have, the more interest and more research going on, it's going to make us much more effective for making these recommendations to growers that will meet both their needs, but also the needs of society moving forward.

SPEAKER:

Wonderful. Thank you, Adrian. And I think with today among the panelists, we are representing different sides of people involved... private sector involved in the agriculture value chain. And as you all said, soil is key to your business and to what the company and what you as an individual care about as well. The next question was more on the theme of this panel. That is what is... in your opinion, what is that holy grail for a soil map? What are the things you want to be measured? If you could create this dynamic map, what are the different things that you would really want? So Adrian, I'll start with you. That is, from UPL's perspective, if you're thinking of OK, this is the view that we want created, (INAUDIBLE) all in terms of physical, biological or chemical properties, what are the things that you really care about?

SPEAKER:

Yeah, I mean I'm not - I'll say up front, I'm not a soil scientist, so I'm not going to go into tremendous detail. But, you know, as I kind of hinted, we feel we're at the early part of a journey here and we need much, much better understanding. You know, we need basic standardization of measurement systems, which are scalable and affordable to use that go across geographies and countries even. I mean, that's kind of a fundamental need because we're all working or we're working to comparatively different systems right now. But I think integrating a lot of different data, whether it's on the microbiome, on social structure, on water retention, all of these different things, if we could understand this more holistically and then understand also how these individual parameters influence the quality of the soil and the output that we're getting from it, that would be tremendously valuable.

And, of course, we want that for ourselves as an industry company, for our basic R&D purposes to understand that, but we also need to be able to provide that information in a way that's understandable

to growers. That they can actually take actionable things out of it and actually help manage their farms. And I think that's one of the challenges that we have, you know, to get this in a farmer-friendly mode, if you like, that with all the other stuff that they have going on, that we can provide them with advice that makes sense to them, that they can act on or not. That's their choice. But at least they have an understanding of how different activities that they may choose to initiate on the land will have, you know, on their soil health and on the output that they have with their cropping system.

SPEAKER:

Thanks, Adrian. Teddy?

SPEAKER:

Yeah, I think those are some good insights. And part of it, you know, as we just we just heard, I think, making sure the farmers understand. That stuff for me gets to the spot of the the holy grail, right. When it becomes actionable from the perspective of the farmer. I mean, and the other piece is, you know, we've made as Adrian mentioned, it's great that this topic is coming up now and we've made progress over the years. And I look at even seven years ago when I got deeper involved in this space and where I am right now, before, we were just happy that farmers were even using a soil map to make decisions. Like period, like not even... forget the static or dynamic, just the fact that, you know, you are actually looking at a soil map and understanding the soil composition to be able to make decisions. And we do see, I mean, based on the research that I mentioned earlier, we do see differences in how different varieties perform, depending on what soil they're on.

Now, that said, some of them have even pushed the envelope more today now, where they're actually in addition to using a soil map, they do soil samples every year, especially in the fall after the harvest to understand, you know, has the composition changed to a certain extent and then compare that to a yield map and then be able to make decisions that way. So, those are the guys that are more on the sort of the front end of it, I would say, right now. But going forward, I think it becomes critical to even further understand, well, as we look at it, because the soil maps, the soil samples they take, it's not done every year. It's probably done once every three years or so. But that soil is constantly changing. Not only the composition of itself, but also, you know, if you've had some soil erosion, if you put in some conservation practices in place, have you regenerated some of the soil in itself.

Is there areas that require more conservation practices? And that's a big deal for us. I mean, when I talked about environmental sustainability, it does go back down to have you changed your tillage practices? Have you put some terraces in? Are you putting buffer strips? Do you understand where

erosion is happening and how can you avoid that? But again, doing it in such a way that it's not unproductive or unprofitable for the farmer. So, having that dynamic understanding of what's happening with that soil at any given time, not just once a year, but even throughout the season so you make the adjustments. If we have the ability to understand that consistently, then all of a sudden we can - I mean, from our perspective - make better recommendations. The farmers could adopt better practices. It's better for the environment. So, I think the key part for me in that what's the holy grail is the word dynamic. You can actually be dynamic and understand it in such a way that it's consistent. So, I don't know. Karsten, maybe you have a different opinion, but maybe you can expand further on that.

SPEAKER:

Yeah, I love where both of you have kind of taken it so far. And maybe I'll build on that a bit and say, you know, zooming out, I almost think that someday our holy grail in the space is going to be to bridge length scales and time scales, unlike most other domains of science and engineering. And for me, the way... the lens I look through today is almost in thinking about customers. In the sense that there are different types of customers out there that may care about the kind of dynamic mapping we're talking about. And on one hand, from a grower's perspective within each season, how do we make that crop as as bountiful as possible, the way that that plant responds to environmental conditions or stresses throughout the season? That's something that the time scale and length scale is within the local soil structure of the root, it's something that happens on an hourly or minute-by-minute basis across just about 100 days of the year.

And then maybe if you think of a company like Pivot, being a customer of this topic, I'm trying to figure out how do I bring new innovation to the market as fast as possible. And my time scale is really one growing season.... every growing season is as fast as I can ever innovate. So, it's a different type of a time scale. And I care about - I think, Teddy, you mentioned it earlier - the consistency of product performance or maybe the interesting differences that happen across parts of our world, the different spatial implications of that diversity around the world. And then maybe a different customer might be stakeholders who really care about the longevity of every field, whether it's somebody who might care about carbon sequestration or the grower themselves and how we can serve the topsoil on that acre. And now, we're talking about time scales that are a year, length scales that are acre by acre. And so, that challenge means that whoever the customer is, we've got different time scales and length scales that really matter. And how do we bridge all those?

SPEAKER:

Thanks, Karsten, Teddy, and Adrian. I think the responses were - and it was great to see yesterday that most of the discussion was on what data sets exist right now. If we were to create a dynamic map of soil

throughout the world, how do we create it? And we saw some public datasets. We heard from some key speakers and organizations that are building those datasets. Today, from your answers, we're able to bridge that to some of the ways in which you care about these datasets. What kind of applications could be built? And from the things that all three of you mentioned, it's the customer, be it an organization that cares about carbon or a farmer or eventually those are the use cases which will determine how much dynamism do you need? What kind of things do you need to measure that, what depth, what properties you need to measure? So, that's wonderful.

The next question was around right now, when you think of soils, you have a lot of data that you collect, that you already are doing to map the soil within your organization. I know some of that might be proprietary, but at a high level, what are the things that you bring together to map soils for your customers? And this could be things like even farm management practices that could have an impact, soil samples, geospatial data, or even genomic data. But I wanted you to elaborate a little bit more on that. What are the different datasets you have? So, Teddy if you could start?

SPEAKER:

Yeah, absolutely. So, you know, again, like I said, we've kind of been working on this for a little while. You know, we started with some of the public data that's available from the USDA. I think that's kind of that was a starting point for us. And, you know, obviously that's not as accurate as we'd like it to be, but it was better than nothing. And definitely - and I don't want to minimize it because getting that information in itself was a big feat. So, kudos to the folks that were able to provide that. I think in addition to that, I did mention a little bit about some of the soil sampling that happens, especially after harvest. So, we're trying to get the number of farmers that will adopt those practices because then you can create a lot of different - I mean, the whole momentum behind that was variable rate, right?

So, where we get variable rate fertilizer, variable rate seeding, variable rat applications in the future, a lot of that. Not soil as the only component, but it's a big component of how you create those variable maps. So, that has an input as well. And then you take yield and yield maps and things like that to be able to get to that spot. But again, like I said, the soil is constantly changing. And so, how do you get that information in a much more - not in real time, it's a little bit further out, but as closer to real time as possible. That would be the next frontier in my mind. Now, you asked about what are some of the challenges in probably getting to that spot, right? Number one today, I would say if we were to go to that sort of dynamic type model, I mean, you'd have to be able to get that data quickly back to somewhere that could be analyzed. So, you know, some of the basic basic challenges we have today of just getting some of these systems out is broadband. Like how do you actually get connectivity on the field, so you can make information go back and forth?

It sounds - you know, we can talk about all the futuristic things, but if it sits there on the edge but the edge never goes back to the cloud, I can tell you that that whole system breaks down. So, kind of getting broadband connectivity not just across, you know, like kind of to the farm itself, but on the agricultural lands, that stuff becomes a bigger... the one thing that we have to overcome. And we're working on a variety of things, you know, both Ranveer, you and I are on the FCC Commission on Precision Agriculture. We created something called American Connection Project. So, we're trying to advocate that we need to do more of that. Another challenge, I would say, is the availability of data. I mean, it's in a lot of different places. And one of things you mentioned is proprietary. You know, understanding the soil, I'm not so sure that that's a proprietary thing. Like, what we do about it could be interesting.

So, is there a better way if information is collected, for everyone to get access to understand at any given point or in every field, what is that composition, what are the changes that have happened over time? I mean, making that available. And today if you want to have that information, you got to go get it somehow or you got to partner with somebody to go get it. And we're just slowing down, I think, some of the more innovative things we can do going forward because we don't have sort of the base starting point, that's why. And that talks to the third problem, which is this idea of collaboration. Like this kind of open source type mentality, like not in the software space but that thinking, that's something in the ag space we could use more of. And we're still a little bit sort of in the protective of our like here's what we have and we think it's critical. So, these are the things we have to overcome over time, I think. And we are. I'm not saying we're not, but that could expedite more of what we could do. But I'd love to hear from my panelists here if they're running into the same challenges or not.

SPEAKER:

I'm happy to maybe go next, if that's OK, Karsten.

SPEAKER:

Go for it.

SPEAKER:

So, maybe it would help, you know, with the audience just to kind of walk you through very kind of high level how we're developing some of these products. I mean, we start off in, you know, with a new biostimulant or a new soil enhancing product in the laboratory, move quickly to the greenhouse. And, of course, we can do a lot of testing, under very, very controlled conditions to start to understand how

different nutrient levels or different water regimes may affect favorably or unfavorably the activity of that particular molecule. And then we'll go on to our research farms, which we're in the process of - if we've not already done it - of really mapping to a very, very high degree, you know, the soil premises on those farms. But then we have an enormous leap, so we go to kind of open field trials and actually get the products into the hands of growers. And as I said before, you know, we really want to be in a situation where we can very confidently predict the outcome of using those products under different conditions. And I think that is a frontier right now in terms of how it really works with some of these products.

And of course, we need better soil mapping, we need dynamic soil mapping because we want to follow these over season over season, how the quality of different soils improve hopefully over time. So, that dynamic portion is really, really important. I mean, we're also using this for regulatory purposes, you know, to look at runoff and dissipation of products in the soil. And those soil maps are very, very important from that point of view. Again, with the environmental protection angle, that we have to make sure that whatever we're doing on the farm is not harming the soil or the environment in any particular way. So, there's a whole lot that we can do with these. And I think, you know, we've got to a certain point where they're useful but they're certainly not perfect. And I think the proposal that we have here to create something that's truly, you know, universal for the United States will be tremendously valuable for industry and then ultimately, for growers moving forward. Karsten, over to you then.

SPEAKER:

Sure, Ad. You know, and at Pivot, I think the contribution we want to bring to the basic science is around maybe the smallest of time scales and the shortest of length scales. And kind of by way of explaining some of the inspiration for our company, is when I was doing graduate research in the field... the nascent field of synthetic biology, one of the big innovations at the time was to move from doing bulk fluorescent measurements on a population of cells to using a cytometer where we could get a fluorescent measurement on each individual cells that grow in a culture. And it was a new window into how cells behave because you could see different populations emerging. And at the same time, the ability to write DNA meant instead of taking a very specific and targeted approach with genetic engineering to make a change in a genome, we could write thousands of combinations into DNA and have a new order of magnitude of information to test hypotheses. So. I think we want to take a similar approach when it comes to what some of how Pivot operates on understanding the dynamic soil properties that influence products we design.

We try to collaborate with as many research institutions as possible. One to highlight is we're doing some really interesting work with mesocosms, with teams at Iowa State University, to be able to have a

very controlled environment but in a real world environment and looking at different interactions between microbes, crops and maybe the non-living components of soil. And then at the same time, Pivot can bring resources to bear to take that kind of science to a different scale. So, we have done tons of shovelomics over the years to dig tens of thousands of root samples across the world and digitize the microbiome in that root structure. And it becomes a plant by plant look at how the plant, the soil microbiome, the chemistry of the soil, the structure of the soil, all interact to be able to lead to an outcome. So, that's really where we spend a lot of time trying to push the boundaries of science. And then we try to work with as many folks who collect other forms of data from satellite all the way down to sensors in the ground and figure out ways to integrate them together.

SPEAKER:

Thanks, Karsten. These are great answers. I was busy taking notes and I thought each one of you made really awesome points. Thanks. Thank you. So, the next question which came up yesterday as well was, you know, if we have to build a dynamic soil map right now, all of this data doesn't exist in one central place. There are different organizations that have the data, some of them in the public sector, some in the private sector. Part of the problem is around discoverability. Who even has what data? I was just learning something that the agrimetrics team in the UK had built, which was really cool. So, one part of it is how do you discover data but the other part is how do you incentivize different organizations to share data so that others can use it?

Now, part of the reason - I'm not saying it's with like your companies, but it could be the different organizations - that are using some of this data is not shared could be either due to privacy reasons. For example, farm management data. This is something which could be private data, which couldn't be shared. Or it could also be... the other reason to not share, it could be IP reasons. That is, you just want to protect the IP. So, this is something - if you leak that data, you're compromising that. So, I wanted to get your thoughts on how do you enable more of this public-private sector collaboration where different entities might have datasets which are unique? And one of the datasets that, for example, could be farm management data because that comes up and that's something which usually the private sector has much better data on. And on the other hand, some of the research data, like detailed data from these long-term research sites, that's with the public sector.

So, the question here is more on how do you think we could encourage more collaboration, encourage and facilitate more collaboration with the public sector in your case? And in your case as well, that is how do you... you could also give examples of how do you work with universities? Karsten, you mentioned you're working with Iowa State, but broadly, how could this be extended to the research community, like for the people in the audience today? So, we could start with Karsten.

SPEAKER:

Sure. You know, I think the more incentives we can have that show how there's a a better goal or a better outcome by working together, I think that always gets people's creative juices flowing. And sometimes it feels like there are more disincentives to making data available than there are incentives. And that might just be a challenge of communication. And one of the things that's benefited Pivot well in the past is there have been a few either incumbents or stakeholders within the broader agriculture community who've been very open in telling us the challenges they face. Telling us the problems that they have a difficult time solving. And it's created an opportunity for us to find ways to work together. And I think the more everybody is willing to talk about the things that are hard for us to solve, the more it helps bridge some of these gaps. And that's been one of the things that's really helped any of the collaborations we do or any of the joint research we've done with academic partners is to begin by talking about where we have difficulty operating and then finding how we can complement each other because of different resources or skill sets.

SPEAKER:

Yeah, Adrian, do you want to go next?

SPEAKER:

Yeah, I mean, building on what Karsten said, I mean, I think there's an enormous scope, but there's also an enormous appetite for folks to come together in this space. I mean, this is a space which is new. And again, I know that folks have been working on this for a long time, but I think the focus or this degree of focus is quite new on soil health. And, you know, I think all of us recognize, whether you're in the industry side of things or in academics or maybe in government policy as well, that there are so many challenges that we need to solve in a relatively short space of time. So, I believe that we should be coming together and form collaborations that go across industry with academia, with government, policymakers as well, certainly around what I would say as kind of pre-competitive activities. Trying to create a foundational level of knowledge and science that we can all share and all benefit from. And then, of course, there are always, you know, IP and as you've alluded to, confidentiality issues that many industries will hold on tight to because they are for profit and they have... with all the investments they make, they want to see a return on that investment.

But we can layer on top of those some of our product offerings. But you need that foundational part to start, whether it's working in a carbon-type market or working with a crop input. I think we have to be working to similar standards and with similar tools. Otherwise, it's going to get absolutely confusing for

growers who we're trying to advise and help. So, you know, I do think and I see opportunities here. And as you're aware Ranveer, I'm sure Teddy, as well and others, there are these consortia where there was active discussions ongoing. And I hope that those will actually mature into something meaningful where you've got different industry players working together on common themes, common topics to create a foundation that we can then all benefit from moving forward. Teddy, what do you think?

SPEAKER:

Yeah, those are... I mean, I wholeheartedly agree with both Adrian's and Karsten's point. I mean, the collaboration is going to be key going forward. And I think the conversations like this one we're having and this whole session over the last two days are going to continue to have the discussion, understand where maybe some of the challenges might be and then see if there's a way through that. That said, I mean, the IP piece, the intellectual property, I don't want to minimize that. Because obviously, companies, as Adrian mentioned, have invested quite a bit, you know. And so, they don't want to see those investments go out the window. Or even if they invested in something that is useful for everyone and is not really proprietary, again, they made that investment. How are they going to recover that in the future? On the other hand, the privacy issue on the farmer side, that's very real. I can tell you that from a farmer-owned cooperative piece where they go, "OK, sure, like you as a company are collecting my data and we have an agreement that you're not going to share my information."

Now, the reason why they don't want the information shared is not that they see any malicious activities happening with their data or somebody doing something different, but they could easily see where a decade later, all this information that's been shared could be used against them. All of a sudden, you know, you were doing some modeling and analysis and you're like, oh, this farmer hasn't followed XY practices. And so therefore, this is why they're having so much runoff, et cetera, and they're getting penalized for it. So they go, "I shared the data for the good of all and I'm getting penalized for it as a result of it." And that fear is real. Whether it's going to happen or not. So, we have to be conscious of that.

So, the value in my mind has to be understood by all parties involved. Like what is the value of having a shared soil... a dynamic soil map that we could all leverage? What's the value for the farmer? How are they going to benefit from it in the future? And here's what they get if they share. And what are the incentives they can reap from it if they do share? What about the companies that are collecting that data? What about academia that's working and going deeper in an analysis in areas where companies maybe don't have the ability to go really deep into things? And how do we make sure the universities are comfortable sharing all that knowledge that comes out and it's not sort of caught up in the whole like, well, who owns the IP for this and all those types of things?

And then you have the public sector on the government side. I mean, USDA has done a lot of work and we leverage that quite a bit today. But again, you know, they ask for data in some cases like, well, how are you going to use it? And this gets back into that whole, OK what is the government going to do with this information? But at the same time, from their perspective, if they're going to be doing more, they will need some funding that comes from the administration at that time to be able to give dollars in that effort. So, which means the rest of us probably have to advocate and say that's important enough so that those funds get allocated to do this type of work. So, you know, it's the whole... as a whole, we have to come together to be able to enable this.

And then you also have companies that maybe are not in the agricultural space, like a technology company, like Microsoft (INAUDIBLE) and maybe has a bigger interest in we should come in because we want to do... we want to enable companies, we want to enable individuals, we want to enable farmers to do better. So, how can they contribute? And again, in some ways, they're a third party that is not involved in the ag industry in itself. So, you know, you can have these other players that come in and have sort of a different view on things. So, you know, it'll be interesting, but it all starts with the conversation and coming together and realizing that there is value for all of us in doing this.

SPEAKER:

Those are great points. And I really like this coming together. The suggestion of the private sector and this was just mentioned in the Slack channel as well, where all the stakeholders exactly to your point, Teddy, Adrian, and Karsten, could come together and talk about what are the incentives, what are the shared challenges? And that could be a great starting point because everyone would benefit by sharing the data. It's getting the right incentives in place. And I wanted to add from a technology company's perspective too, one of the things we are doing in research - and the computer science research community is doing a lot of work in this space - is around multi-party compute. That is, how can you share data, encrypted data, that is, how do you do AI, for example - artificial intelligence - on encrypted data? So, I could be sharing data with you, you don't see the data, but you could be doing aggregate analytics on top of it.

So, these are some of the primitives that could also help and assure growers and assure stakeholders that you can share the data and then improves along with it that what information could be gathered from it and what is not leaked, which could be one of the tools that we have. Things like multi-party computer. Homomorphic encryption is one of the new things that is there in the computer science world. Which could help with this as well. But yeah, this is one of the - we're still looking at various ways in which we could make it happen. And some of the learnings through this workshop and as you

mentioned, this is the starting point. We need more of these forums where different stakeholders can come together in one forum to talk about the challenges.

So, we're coming towards the end of the panel. One question is around new business models or new scenarios that could be enabled once you have this kind of data. That is, from your company's perspective, assume we were successful, like ten years down the line, we had a dynamic soil map, how would your business change or what new businesses could arise if we had such a such a dynamic map? So, we could start with Karsten.

SPEAKER:

You know what, a tough question that to make sure we get right, I think we'll have plenty of ways we can suggest things and we'll be proven wrong over time. For me, it always comes back to who's the customer and what makes that customer willing to part with cash in exchange for some benefit? And kind of going back to my intro, the three types of customers I think about today are helping every grower be a lot more productive with that crop each season, helping companies like Pivot be able to bring innovation to market faster, and then thinking about stakeholders who care about the underlying value and sustainability of each and every acre. And I think a more complete and powerful dynamic map is going to be useful to each of those types of potential customers for different reasons. And maybe it leads to different types of businesses or scales of businesses that can grow out of solving challenges that we face. But I think anything from inputs companies to the companies in ag themselves to growers and other industries, there's a lot of opportunity there.

SPEAKER:

Thanks, Karsten. Adrian?

SPEAKER:

It's a lot and this will not be a complete list, but I mean, as Karsten just mentioned, soil health inputs, I mean, the ability to research and then provide concrete recommendations on their use, so they're really effective and we really get the full benefit of using them and not over-using them or under-using the, you know, every kind of drop of whatever we're putting in the ground actually counts. And coming with that, I think more advisory services that we can as an industry provide to growers. And by the way, ten years is far too long. So, we need to do this a lot quicker than that. You know, the carbon market is something that many companies are looking at right now. Can we support growers to get a cheap value from good sustainability practices on their land? I mean, can we get better crop prices? Can we get

carbon credits, these types of things for those types of things, which are obviously being pushed from many different directions and are very necessary.

So, I think this is something that these maps can really help open up. And then things like precision planting, which, of course, is already the reality now. But can we improve our ability to do precision planting and precision application of different products? So, there's many, many different things that this would enable, some of which are already occurring. But this would add a lot more validity and power, I think, in some of these different business areas.

SPEAKER:

Thanks, Adrian. And Teddy?

SPEAKER:

Yeah, so I think they've covered quite a quite a bit of ground and I agree. I mean, you know, ten years out is a long, long ways away. It's only ten - I mean, if you're in row crops, that's only ten cropping seasons. Ten chances, so that doesn't seem a lot. But what we've seen even in the last three to five years, there's been a lot of changes. So, it will be interesting where it goes forward. You know, Adrian mentioned the carbon markets, carbon sequestration, there's a lot of interest in that area. We're doing quite a bit in that space. But again, understanding, fully understanding what sequestration means and how much carbon is being sequestered and how much is being emitted and what are the things that are driving emissions.

If we kind of got a better handle on that, which goes back to this idea of understanding the soil to the fullest extent, that could even stabilize some of the pricing. I mean, how do you actually price it today? It's a little bit all over the place. And, you know, there's some people that talk about, you know, it's three dollars a ton, up to 150 dollars a ton, whatever. So, not sure what that is. But so, understanding that, that could open up new avenues. And again, it's a great revenue stream for farmers with the right incentive of like, here's the right practices to follow and you actually will get compensated for that. So, I think that's a big one that if we can unlock that, lots of possibilities there.

The other one is you know, obviously, we thrive on making a lot of insights today, giving a lot of recommendations to farmers. I think but it's still more we're still product driven and the insights sort of come along with the products. You could get to a spot where the insights are more driving the value ultimately, whether it's from the companies like UPL or Pivot Bio or Land O'Lakes, kind of aggregating

that together and saying, "Here's the actual insight, here's what you should do and that's what you should pay for. And then obviously, these are the things you should you should do along the way. And these products you should use." I think, you know, is there a switch to more of an inside driven subscription type model could be some of the avenues that could get opened up in the future, which I think could be super interesting.

And also, you know, as we talk about sequestration, the other flipside is reduction and carbon reduction. And I think there's value there that we can unlock. Meaning, like just not doing it is not the answer either. Or not doing what you're doing today. It's more like are there more value added products that you could be putting on the field in specific spots? And those are for the providers in the space, those are higher value margin products. But at the same time, it drives to better sustainability as well as better productivity for the farmer. So, there's maybe even new products that come into the market that I think could open up a lot of channels and avenues there.

And then finally, the last one I'll say is - we've talked about this in some futuristic kind of discussions - is more biodiversity. I mean, today in row crops, I mean, it's corn and soybeans. And yes, there's potentially thoughts of - there's a lot of farmers trying, "Can I get into different crops?" You know, peas and lentils and things of that nature. But I mean, you have to change a lot of the practices and you don't know if you're going to be successful in the climate you're in. Will the soil be able to support what you're trying to plant? They need the equipment, et cetera. However, I think, as consumer diets change in the future or different needs across whether it's fuel or fiber changes in the future, are there more diverse crops that we don't even know about today that we could be planting once we understand our soil better? So again, that's a little bit further out and we really don't have a crystal ball for it but there's lots of opportunities there, too.

SPEAKER:

Thank you. These are amazing viewpoints. Thanks so much for all the discussion. And to summarize the last point, I think a dynamic soil map helps build better soils. It's better for the environment, it's better for the farmers. It's also better for the world because you can grow much better food if you know the soils better. So, thank you again for the discussion. And with that, we'll move to the next session. So, people on Zoom, please wait a minute to be placed in the breakout room. And those who are on the webcast, they can view any room by switching in the tabs at the bottom of the webcast screen. There is also a Slack channel that corresponds to every room. So, see you all there in breakout sessions. Thanks.

BRUNO BASSO:
Good morning, everyone. Good afternoon for people connecting from different places. It is with great pleasure that I welcome you again to the third day of exploring soil dynamic information system. We have had another very productive day yesterday. We had the opportunity to interact among each other, as you've seen from the program. There were three breakout sessions in terms of, you know topics. One that really looked at data collection and duration measurements and data analytics and modeling. So today I really feel, it's a critical day to synthesize what we have learned. And we aim to have a discussion an open discussion after the presentation of each of the reports that are coming in the next minute or so. And the open discussion is for people to engage again and be able to share points that may not have been covered during the different breakouts.

And so that's very important to hear from you, especially if you were either more focused on, you know, you miss the, some of the other breakouts, I can guarantee that the report I've seen what the other people ever put together from their notes comprehensive, and but there is always additional topics to be discussed on.

BRUNO BASSO:
So, without much further ado, I want to share the program. The next activity is the report back from the breakout. And we will go for that until the break then Kathe Todd- Brown will share with us some of the continuing engagement opportunities. Very useful sets of links on things happening and not put new opportunities. We'll have a synthesis session by Doctor Jim TG and Rodrigo Vargas will also support additional comments from the Slack channels that we have paid careful attention to it. But often we may have missed some of the points. And so, Rodrigo will have the goal of even synthesizing a little bit further thing that we may have not discussed in details. I'll just add a few comments as a concluding, and we will adjourn again at 2:30. So with that, we will start with the first report back from the breakouts on the first topic, which was called measurements and sampling and archiving. You can see from the agenda that there were lots of bullet points to go through. Sampling and measurements are critical.

So, with that, I would like to invite Doctor Chuck Rice, who's going to report a breakout session to us. Thank you very much, Chuck, take it away.

CHUCK RICE:
Thank you, Bruno. So yes, we had a robust discussion yesterday in the two different sessions. And so, I'm just going to try to provide the bullet points and summaries of what we discussed. Next slide. So, the question one of the questions is what should be measured in soils, and we didn't want to get bogged down into the detail 'cause there's a lot of debates on what should be measured and then even how to measure by particular methods. But the consensus of the group was really what is the question that a project or a group is trying to solve. And that really determines what's the key measurements. And so, it needs to be project or go specific and you know, whether it's an academic or a scientific exercise. Whether it's related policy like in the EU was discussed earlier in this week. Is Atlanta management, a

producer or land manager may need much different types of measurements. And then just kind of the broader what's the state of the solid resources for a region or for a country. And then the other question is that, you know, if we want to advance the science of soils.

CHUCK RICE:
Then measurements may be guided by developing new theories. But also, the data that we collect and if it's a robust resource, it could help drive the development of new theories. And each agency we've heard some discussion and over the past year, each agency has a specific goal as well. And so not only what methods, but what at what temporal and spatial scales is important.

CHUCK RICE:
Again, are we looking at continental, regional, or down to the acre or hectare size for management. And again, and go back into the measurements at what temporal and spatial scales depends on the measurements, sorry, the objectives if we're looking at erosion, you know, that's at the decadal scale possibly. Or if we're looking at greenhouse gas flux, we're talking minutes, or even seconds as things change dramatically and what kind and amount of gasses are emitted from the soil. And then we need to think about the soil information system as what it can provide for models. And so, the models need provide input on what kind of measurements they need. So there needs to be a two-way interaction, however, for a dynamic solar information system, the challenge then is how do we integrate different goals and objectives from these different projects or agencies in order to have a more robust, soil information system? So, we can look at, you know second or a scale that seconds to decades. Can we look at microns to kilometers in that sense? Next slide.

CHUCK RICE:
So, we did have a dose discussion on what should be measured. We kinda separated out into physical the, I guess the three key properties on the physical side that was consistently measured was texture. And often a lot of studies don't include texture surprisingly. Aggregate stability, or some measure of soil structure. And then bulk density. And the key bullet point here is that bulk density is a measurement that has to be taken onsite and can't be archived. Some of these other measurements can be the soil sampled and then put into storage or analyze later. But bulk density is something that has be taken on site. So, water content, color and aeration status, you know, color is an indicator of anaerobic or aerobic conditions. And so that's really important for the later attributes. So, a poor distribution possibly mineralogy and horizon designation.

CHUCK RICE:
On the chemical side, organic carbon most consistently measured and also pH isn't not surprisingly, but also nutrients. And there is pretty good discussion that we need to have organic matter fractions or carbon and nitrogen fractions of the organic matter. The biological piece, those methods are really still evolving. The biology started evolving surprisingly. But you know, some measure of diversity genomics is becoming more available both in different labs and costs, phospholipid analysis. But it was also mentioned don't, you know, we shouldn't forget about some of the more simpler techniques, chloroform fumigation, because it's convenient. It can measure microbial biomass, but it also relates to

the act of fraction and fits into a lot of the carbon models, soil models and then enzymes. And then the other question is at what depth, and again, it goes back to what's the questions that we're trying to answer obviously be nice to have a one meter, if you're in South America, two or three meters of profile, but you know, that may or may not be practical in every particular collection system. Next slide.

CHUCK RICE:
We didn't spend a lot of time talking about temporal and spatial scales. I did know, you know, again, it depends on the, the objectives and or I mentioned some of those attributes. A key point, again consistently came out as we need the ancillary, the meta data is probably the biggest need and deficiency in datasets. We need simple latitude, longitude. What's the location of those sampling sites? Of course, climate, weather data be useful. And then the other thing that's harder to get as land use history, a vegetation type, whether you're talking about forest or grasslands or an ag system of the different kinds of crops. And then it'd be helpful to know productivity biomass and I added here, and there was a little bit of discussion mentioned. It'd be nice to know biomass root distribution depth as well.

CHUCK RICE:
In ag systems, agricultural systems. We need to now tillage, fertility, yield and crop rotations, and other aspects of that agricultural system. And then if it's an urban environment, you know what, again, the history is important. You know, if we're looking at contamination what was the, you know land use or surrounding area? Was it industrial plant or housing and other attributes? Sensing, we talked a little bit about sensing from the remote sensing, aerial sensing. We can get very detailed information on elevation by plant productivity. We can sense remotely tillage the type of vegetation or crop. And then, a little bit of discussion on below ground sensing as new sensors that are being developed to be placed in the soil. We can measure some of the more dynamic properties at, at, at small temporal scales seconds or minutes.

CHUCK RICE:
And that's gonna add a lot of opportunities, but we're gonna have a lot of data. And then we also to be thinking about new techniques, avoid the physical extraction of the soil out of that landscape. Can we sense bulk density or other organic matter in place? Next slide.

CHUCK RICE:
Standardization there's a little bit of discussion that about standardization, you know, a lot of countries labs have a standard method, but even just given the example here is a phosphorus available phosphorus. That method really depends on the soil chemistry, the Bray P, Olsen P or Mehlick P, it depends on the pH of the soil. Archiving the samples. We need to be archiving but again, if it's dry samples, the story conditions, temperature. And then we have issues of accumulate or accumulating samples for a space need. Biological, it'd be nice to have a bunch of minus 80-degree freezers around the world, around the country. But the store, just the sand, the soil itself, or maybe be cheaper to extract DNA and then freeze those samples. The advantage with enzymes, some of those can be assayed from dried samples. But then the issues become a space freezer, space costs. When somebody suggests that we need to have the Svalbard Global Seed Vault in Norway, we need to have one for soils. So that's a big global project. Next slide.

CHUCK RICE:
What came out. The other thing that came out was we need to really have a network of reference landscape sites. One is so that we can have a collection of our methods, but also as we develop new methods that we can reference back to those sites to calibrate new methods or techniques to kind of the standard methods. And then we also need that data could then be used to help color calibrate different models, soil models, carbon models. So, the suggestion was that maybe we should take advantage of NEON, LTER ARS sites, but also Land-Grant University network. What is there about 80 or 90 land grant universities in the US? And so, it would be valuable if each Land Grant University dedicated a few acres with a management practice or a land use practice to serve as reference sites. And that would allow us to look at soil and climate variability. And the thought was maybe we should submit proposal to NSF or USDA like an RCN or some other network that can help support that effort. And I think that's my last slide. So, thank you.

BRUNO BASSO:
Thank you very much, Chuck. Can pretty comprehensive overview. I'm sure there are points to discuss when we open the floor to everyone. So, I would like to invite Ranveer Chandra next to report on the collection and curation recount.

RANVEER CHANDRA:
Yeah. Thanks. Thanks Bruno. So, in our session, we discussed the collection and curation of soil data, mostly looking at once it gets to the cloud, how do you make sure that you're following the fair principles, making it find-able accessible interoperable and reproducible. And we looked at various things. How do you store it? Whether you use it in the cloud or not? How, whether it should be distributed? Whether it should be stored centrally the use of AI on this data. So, if you'd go to the next slide.

RANVEER CHANDRA:
So, one of the first questions we asked was what does it take to make the process fair? And especially for soil data, are we using the right principles it. And then you'll see multiple things we talked about with respect to that. The first one was, how do you incentivize sharing that is the data that we need to build the accurate soil information system requires data from multiple stakeholders. It requires data from academia. It requests data from the government, from private sector, and also very importantly from farmers. So how do you bring all of this data together? So, for the academics, all of the suggestions that were discussed, included journals and incentivizing journals to require data to be shared. And data to be shared in a particular format so that this data is easy to it, it's not just about a credit, it's easy to reproduce it because once you know, which data stream was used for that paper, then it's easier to reproduce.

RANVEER CHANDRA:
It's also easy to find additional data streams the additional databases. Which if there's a new paper and you get some homemade. This repository, this is something that was encouraged to have journalists require citations to the data that was used and that data to be made available. The other interesting discussion was that the younger generation, it seems are more open to sharing data. And but on the

other hand, the flip side was that some people who are more senior who are closer to retirement were like, well, you know what, on the other hand, people who have spent their lives collecting large amounts of data, will be like, well, the best way to put others to make use of this data is to make it public. So, it seems that there is more consistency across the participants, that there is a need to share data.

RANVEER CHANDRA:
And the other discussion that came up with us, when we are doing this data sharing, we should be consistent with industry. That is consistent in terms of both when we shared the methodologies used maybe even the standardizations. Then with respect to business models, that's essentially, how do you encourage people to share data, the different stakeholders, academics, government, private sector, farmers. One of the, some of the new business models, which it seems especially, but agriculture might work out well, is one is carbon markets. Where with carbon markets, farmers might be more willing to talk about what they're doing in the farm, because they have financial incentives to do that. Similarly, for crop insurance, that insurance of any kind where yet again, is another reason why people might want to share data.

RANVEER CHANDRA:
The next thing we discussed was how do you find the data? That's how what's the representation that can help you find the different data sources given that if there are so many data sources and Kathy will talk about some of that after the session. So, we talked about how we could use some of the learnings from the semantic web. You can use things like knowledge graphs to link all of these data together. And along with the data, the meta data as well. And if we do the citation piece correctly, as we talked about previously or journals, then you could create this DOI citation to link all of these datasets with the publications as well. Next slide.

RANVEER CHANDRA:
One of the other questions we discussed was that, yeah, we are collecting lots of data, but how do we train people to use the data? That is a lot of times the soil scientists are not the deep AI scientists. And so how do we get not just soil scientists, but also the farmers, the public sector, the government to use all of this data to make sense of a lot of this data. So, then we discussed various things here, for example, Wolfram's Natural language for non-coders is one way in which we could look at this extracting information from the data and making it easy for people to use the different data sources that exist. Visualization was another one when we were talking this came up when we were talking to the USDA, which is collecting large amounts of data using AI on top of this data.

RANVEER CHANDRA:
But one of the suggestions was, well, how do you visualize this data? How do you present soil data to someone who is either a direct consumer or someone who might be taking decisions based on it, or someone who wants to even do science on top of this data. Visualization is an important problem. And especially if you look at the problems that Chuck pointed out, when you're visualizing soil physical, biological, and chemical properties of soil. How do you present this information? To people in the field and people outside the field? The other idea here was to use some of the new NLP, the natural language processing technologies that are being developed in the computer science world.

RANVEER CHANDRA:
On things like conversational AI or machine teaching web. How do you get, the question here was you have so much data? How do you run AI on top of this data? So, to run AI will have to, one of the discussions was on. So, with conversationally, I, people are talking about using your speech and then converting that speech and con and converting it into a query to the database. There are some recent advances, like what is being done by GPD three, which where you can give it a sentence and it then can convert it same to a more structured query onto your soils' database. So, there are different advances here which can help us train which can help train people. People here corresponds to all stakeholders to use soil data.

RANVEER CHANDRA:
But of the other questions, which also came up in Chuck's presentation was data annotation. That is data by itself is not enough. This data needs to be accompanied by metadata about how that data was collected. Whether there what does it correspond to this metadata is important. And one of the questions was who annotates this metadata. One of the data sources that we looked at was not just the way data is being collected right now, but also citizen science. What if there are this is more crowdsource data? Then who does the annotation, should we, should we be having scientists to it? Is there a mechanical Turk model where you can just have distributed the annotation to other people? Maybe even in some cases outside the field, or in some cases also use artificial intelligence to annotate, to add the metadata, to the different data streams that exist in these, soil datasets. Could you go to the next slide please?

RANVEER CHANDRA:
So, one of the things we discussed was around data fidelity and data veracity. This was basically a very interesting trade-off that we discussed. That should we be sharing the entire. When someone collects the data, should you be sharing the entire dataset or like, including everything without cleaning, or you should have strict cleaning procedures. So, you do the calibration and you do strict curation of the data and only upload this curated dataset. So, we discussed both of these, about the pros and cons of both. And there are organizations doing both. Like, for example, when we talk to researchers in the US and Europe, they have very strict procedures typically of cleaning the data, curating the data, making sure that the data is clean.

RANVEER CHANDRA:
On the other hand, when from Australia, Andrew (UNKNOWN), he was talking about, well, you could share the entire data. And then figure out how do you make sense of that data in the cloud, then you can filter things out. And then Jim, who was in a panel as well, he mentioned about GEMSdata, the genedata, the DataBank, that Jim will talk more in his synthesis. But he talked about, you know, in the end, this was a big debate. And in the end, they decided to share the entire data. And it actually was the right decision, as opposed to trying to have strict cleaning, curating procedures on top of it. And the point here would be that once you've put all the data, then you could have others, like for example, scientists or users, then filter that filter, add their filters on top of the entire dataset, which was interesting. This is a debate which I think we didn't get a conclusion on, but these are just two points of view of on how do you handle the data fidelity?

RANVEER CHANDRA:
The other question we had discussed was what is the level of data that is being shared, whether it's a temporal and spatial, this could be shared when ontologies. We also need a level of privacy. That is how do you make sure that PII data, personal and private data is not shared as part of this data sharing process. And in our discussion, it came out that this is also limiting to some extent, the amount of data that is shared. That is farmers, for example, or other stakeholders might be careful about what PII data might get revealed. If you start sharing the entire data. We discussed some of the latest advances in cryptography. Using which you are able to share encrypted data. That is, you can share data in an encrypted format. So that you can still run some of the AI on top of it without sharing the raw data itself, these advances, for example, confidential computer, homomorphic encryption, these are all technologies which are still very new, very recent.

Some of them are being used in a few areas of a few industries, but they might be relevant to building the soil information system as well, especially when we are looking at data that is private and needs to be carefully handled. Could you go to the next slide please?

RANVEER CHANDRA:
So, then we also looked at another question which was on data storage. That was, how do you store the data? Should you store it in the cloud? Should you store it in local servers? Should you store it in one entity worldwide that hosts all the data centralized or should it be distributed? And we have a very interesting discussion around it. But people realize that the cloud has its benefits. That's if you share it in the cloud, you get all the benefits of easy sharing. You get large amounts of compute. You get all the benefits of the cloud. You don't have people managing the servers. One of the concerns that was highlighted with super pricey for downloads that's once you upload the data, it's great. But if you have to download all the data, then that can get pricey. Where I think a lot of the cloud economics are making it more and more easy to start sharing data.

RANVEER CHANDRA:
Like, for example, with some of the open data schemes that are being addressed and there are some incentives making it easier to host all that in the cloud. The other discussion was around with It should be centralized or distributed. That is some of the soil data could have national security implications. And there are regulations around it yesterday. We learned that there are some countries where if you're caught carrying a USB stick with your geo coordinates and some information about it, you can get arrested for that. So that would prevent a centralized data. So, what that means is that we would have a distributed dataset where probably need solving country in your own solving cloud. You would probably host some of this data. But then we need to have metadata or some way to share some of this knowledge graph that we discussed earlier of mapping the relationship between all of these datasets. So that we believe the future of soil datasets would be distributed, especially across countries.

RANVEER CHANDRA:
And then we'd have to come up with a way to query this distributed dataset. The last thing we discussed it was how do you learn AI on soil data? When we talk to practitioners yesterday to scientists. Not many people are doing AI on soils data. The thing people said was, we're doing it a lot to understand what's happening now, but can we use AI to predict the state of soil in the future? And the challenges here of

course, are around data fidelity, plays a role, the data quality. How do you make sure that the data is correct? Those all need to be handled. But the other thing that came up also was, well, we, because it's in an earlier stage in soil science compared to many other fields, we have an opportunity to make sure that we use AI the right way.

RANVEER CHANDRA:
What that means is that we avoid any bias in the data. We make sure that the data that doesn't have bias, they need AI we use uses the reproducibility AI methods, the responsibility AI methods, so that you, you have more explainability. That is whatever the models predict, you should be able to say, why did it predict what it predicted? We also need to communicate uncertainty when we are using AI. And finally, we should be doing all of this, using the right ethics, making sure that we are not using AI the wrong way or so insights. So that was it from our session.

BRUNO BASSO:
That's wonderful. Ranveer. Thanks very much. Excellent summary. I will be reporting the summary of the data.

BRUNO BASSO:
Analysis and models. And this was a team effort with Alison, Katherine, Todd Brown, and Raphael Martinez, who was the note-taker, who did an excellent job in writing nearly 12 pages of notes that we had to distill down. Next slide, please. We wanted to start adding some color and this word clouds from the report of the notes. And it's interesting to see some you know, the words that obviously came more frequently than other little bit outside the remote sensing was a topic that was discussed a lot, in addition, to obviously models AI and machine learning and uncertainties. Next slide, please.

BRUNO BASSO:
So, we started with the question of the positive things, you know, what is the promise of the current machine learning, AI methods? Well, the first thing that we all concur, we had two very engaging groups for quite different. The conversation was quite different because of different people attending with different expertise. So, the two groups was very nice for us to basically capture the range breadth of possibilities coming more from both representing in agricultural systems and ecological biogeochemistry, statistical modeling with more process-based models. But we all agree that all depends on what we're trying to model.

BRUNO BASSO:
OK, so any modeling approach is objective-driven. Machine learning can help generate new hypotheses. They came up a few times. They could actually be then verified experimentally or with process-based model. I started the session by basically showing that we are going in to direction of using not just one model, but following the Acme approach, which we learned a lot, and I explained some. I may go a little bit later on covering more details of that. But ensembling using more models both from machine learning process-based model, it really helps to correct characterize the uncertainties from different sources, including model structure. And so we learned an extensive amount by running the ensemble models within Acme throughout the world. And then it's important to have different methods on top of this to use that as a triangulation of methods that could allow to better verify what the AI is generating.

Next slide.

BRUNO BASSO:
So, an interesting point to that machine learning is it started to be using to predict management practices based on detailed farmer surveys. So, from characterize, qualitative converted into quantitative information and trying to predict management across a larger landscape. And machine learning certainly can help find missing data and filling data gaps. They can also determine where and when we need to take measurements based on what we know, it's currently missing from the current datasets that we have. They play a critical role in augmenting optimization because of better exploration of the parameter space to be done more efficiently.

BRUNO BASSO:
ARPA-E and SMARTFARM well, both came as an interesting and appealing possibility of using this data from sensors that there're being designed and implemented. One of the examples of very interesting actually is one thing I'm working on as well using AI and remote sensing to model observed (N2O) data and trying to get away from self-reporting information globally. Next slide. So, the concerns were probably the list is a little bit longer. The machine learning are really seen as a black box. They are black boxes. It's really hard to fully understand how machine learning makes decision. It's critical to match methods with the right question that will help with answering. Machine learning, the data, the results cannot be extrapolated beyond the training data. That's certainly a big thing to consider.

BRUNO BASSO:
There was conversation on frequently suffering for overfitting. How do we use machine learning and AI to derive insight? That's a very important point. How we go as even mentioned from machine learning to machine teaching or machine knowing. So, can we really flip the thing and be able to use what we have learned to guide us in making decision. We can't do that until we feel really good about, so domain knowledge to me, it's still a relevant and very important and that's why that calls for true integration just having different disciplines in a team. It doesn't mean things are fully integrated and so be able to derive insight it's something that I like to really underline to do eventually learn about the outcome of different practices and changes been made.

BRUNO BASSO:
Data availability for machine learning is limited, validated process-based model can help, but I underlined the validated. We need to be sure that we are understanding the process and be an independent validation. Models have a stronger capacity depending on the types. Let's say the models are not used to be more robust across and be able to simulate in the case and say above productivity without extensive calibration because some of these inputs and may not besides the inputs and ensuring on the quality of being puts additional information may not necessarily be available. But if the model is robust enough, could provide ancillary data to improve the machine learning as we recently did with some publications. So, we are data starved, but also starved for good quality data that often the quality component is not fully tackled. Next slide.

BRUNO BASSO:
We should not use machine learning alone for sparse data stream. It was in the questions and it was

actually very correctly pointed out because it behaves very unpredictably. You basically suffer of modeling the noise rather than the data. It's hard to decouple the two. And so there is more potential to do this sorts of things by having hybrid approach linking machine learning with process-based model. Machine learning is another tool, but it doesn't replace other tools. So, it's like a new advances in medicine, it was a very good analogy. They don't just come and replace everything else there has worked to or it covers different components.

BRUNO BASSO:
So, modeling prediction without knowing the breakdown is very important for experimentalists. They pointed out a very strong weakness in communicating this with modelers that modelers embark into lots of modeling activities, but completely forgetting about legacies and site histories and previous management, water table, in-season heterogeneity. And so that's something that it's critical to consider. And we discussed about some ways to overcome that. And I'll talk in the next slides. How do you harness the power of machine learning while preventing abuses? It's just not a panacea or (INAUDIBLE) it's just we have to be careful on that. Next slide.

BRUNO BASSO:
What are the challenges when we integrate models with data, and how did it come together, can we improve that parameterization, validation, critical piece of benchmarking? As it was pointed out to this seems to be common word across for both sampling, but also using this data to validate models. So, from matching with measurements, Bayesian is certainly, I think all the community heavily involved in models as exposed to this. But (INAUDIBLE) methods is a really nice way to set up priors and be able to get better parameter estimation in models, with characterizing some of the distribution and uncertainties and knowledge about the past.

BRUNO BASSO:
Process-based model I guess indicated to be better at predicting the future because of by definition being more system-oriented, system approach to be able to capture fully the interaction between soil, plant, climate management interaction whether it's managed or unmanaged ecosystem, but especially when AI may not have all of the data for training. We really depend on the amount of data that we train the system. And that's often a limiting factor.

BRUNO BASSO:
One of the ways of overcoming the point about you don't know, the knowledge histories is to use plants as indicators of their variability and not just over space, but also in times. And so, as I pointed out in the introductory talks and was also represented both by Jerry Hatfield and Joe Cornelius about the possibility of using previous knowledge about stability, how different zones behave as a reflection of the integration of the systems that could really help understanding why areas that are always highly productive could be parameterized by capturing their level of objectives, both in soils and some of the features that otherwise wouldn't be able. And so, be able to capture, for example, soil depth which is often kind of given into for granted or used as an assumptions.

BRUNO BASSO:
And so, possibility that what we learned as I mentioned the stability, the thermal analysis be able to be a

proxy for soil depth and be able to satisfy evaporative demands versus low zones, more prone to compactions and you know, shading from trees or other situation that ever been clearly kind of identifying they're detected. And so, parameterize the models to capture those as a proxies could be a way where data is known available about the previous. So, using plants as tell us which bean are you in and how stable you are all the time. So, the instability, it's obviously very dynamics. And so, that puts different weights and understanding the ups and downs and variation through cline and a landscape as also shown in the initial introductory talk. Next slide.

BRUNO BASSO:
So, modeling soil organic carbon, nutrient, and water dynamic requires proper simulation of crop yields. This is certainly within the agricultural systems. I just kind of make this a request in general, that often purely geomechanical models are used to go beyond what they really were designed for. And assumptions are made beyond the knowledge of capturing the true system. So, you wanna a model soil kind of (INAUDIBLE). You gotta be able to model yields correctly, how much nutrients and roots are petitioned back, and how much water and nutrients uptake are taken up. So, that's a critical piece.

BRUNO BASSO:
We capture the variability using crop histories, yield maps, and remote sensing. There was initial discussion again, where remote sensing starts to play a role in capturing the heterogeneity of plants. And one points again back on the agricultural systems heterogeneity of plants that have emerged. That is one of the main factors driving productivity and variation also from one year to the next step, as affected by an appropriate management, for example, or planting and tilling in wrong times awfully not tilling at all. Learning from AgMIP, I've been involved since 10 years.

BRUNO BASSO:
This is a program started (INAUDIBLE) by Cynthia Rosenzweig and Jim Jones and John Anton. It was really all the community of modelers coming together, and that taught us a lot. It should teach this community a lot about using ensemble and what we learn about the median of the models for how many, what's the minimum number of models that we should be using to capture the model structures? We compare the uncertainties whether you run one model with 20 climate models or one climate models with 20 crop models. Believe it or not, we had more vulnerability coming from the results of the crop models. And we use very interesting approaches of having uncalibrated, completely uncalibrated seed test the robustness of the models. And many models were able to capture the yields across times and sites and many other models were needing to have additional input to be able to capture that.

BRUNO BASSO:
So, I'll make a point in the end about, you know, trying to capture the complexity and tradeoffs between model complexity and to be justified versus simplicity, but to be justifying the reality of capturing the system well enough. So, scales as is critical scales for any measurements may not be the same as some of the people. The experimentalists, I guess, we're taking measurements at a scale different from what models are make predictions. So, you go all the way from genes to microbial respiration and flux towers and how do we link those? And there is a significant amount of new data coming in metagenomics. We start using QTL for gene modeling to be embedded into crop simulation models.

BRUNO BASSO:
Spectroscopy is (INAUDIBLE) and models are actually unable to use that fully. And we need to work on that. That was a very critical point here. Next slide. Dynamic soil information system needs to have scientists three metadata that go with it. So, soil enough is not sufficient. It'd be good to record what happened to that site. Ideally, we would like to see a nested hierarchy of models this was a point that someone made to upscale predictions. We have to see how that is gonna work when the models are designed differently. But that's certainly an interesting approach of capturing the level of details at different scales. Information system needs to have repeated measure panels and to reduce the uncertainties. What about underground sensors? That's again, came up N20 and volatile compounds, organic compounds, and then development of new sensors.

BRUNO BASSO:
So, we kinda excited on that, how they're gonna play a role, certainly useful in testing models, but also we hope to change the knowledge in some of these models to be advanced, honestly, being so much into this crop modeling world and some of the sciences not necessarily improved over the years. But we certainly we've been doing a better job in capturing inputs better which has always been a limiting factors in models. We feel that seldom we don't do a good prediction because we don't understand what's happening. Rather the model is really not driven with often the high quality of inputs representing the system, or it comes from a place that is not even close to the point that you're trying to scale to. There is untapped potential in thinking remote sensing with process-based model. That's certainly true, even though this is it's also a process that has improved. I mean, this is data simulation has been around since the '90s.

BRUNO BASSO:
And there is uncertainties that remote sensing as well as measurements don't necessarily capture what is really true that it's uncertainty in doing that. Remote sensing do not provide a direct measure of LAI. There is a significant amount of uncertainty there. We can work improving because new resolutions are counting spatial, temporal, spectral, fusion, the sensors of optical with radar is coming together thermal imagery from different sources. So, that's all coming towards this possibility of linking the two by being able to predict what the next image is going to be. If we understand the feedback between soil, climate interaction. Next slide.

BRUNO BASSO:
We need something like LUCAS that came up, we really thought that the European union is doing a great job in basically imposing through a legislative approaches, to sampling a high resolution in three times with a long list of variables. We wish that we'll go deeper as, but let's look at Montana (INAUDIBLE) that's trade-off that's costing and time of the people. Even though we believe that 8 million euros, I mean, $10 million a year it's really peanuts for the value that what you will get out of you know, basically the life under our feet that feeds us, and all the other support systems that gives to our lives. Organize the soil database to make it easy, usable to modelers. That's very important. So, that was covered quite extensively in the previous sessions.

BRUNO BASSO:
And so we need the system that helps us synthesize from different sources. Often data come when

different data stream different resolutions. So, I know we have several activities ongoing need for facts, and the USDA Ag library, National Ag library, we haven't been working on our monetization of data. So, that's awfully going to help in the near future work with global assessment. It's hard because again, different sources and resolution. How do we select important covariates from different sources? That's how we select that. We need better ontology. We need to communicate, and certainly the measure risk and be transparent about. Next slide.

BRUNO BASSO:
Small sample, large uncertainty, we need to be more transparent on that. Increasing the sample data for machine learning doesn't make a more accurate necessarily if the population has a bias. I think I just have a conclusion slide. Setting up scales and the objective of any modeling approach is critical, as this small intended as direction. So, scales and objective of the models need to be clearly specified and upfront and be able to use the proper system to tackle that. ML have a great I've process and promises as well as concerns you heard that.

BRUNO BASSO:
Tradeoffs between model complexity versus simply the often models are not balanced. They suffer from very detailed mechanisms, highly represented depending on the developer and having very unrepresented mechanisms, and the final result is really predominant on what is not well-represented. Integration and fusion of a domain knowledge to make some assumptions, machine learning, remote sensing, observed data model is key to integrating and fusion dose. We need to develop better soil ontology, FAIR and file format standards. And we need to improve the resolution of data both the spatial and temporal. I think that's the last slide.

BRUNO BASSO:
So, that concludes the report of the sessions. As you see, this is like drinking water out of a fire hydrant, and it's almost too much, but that's what you get when you work with soil. And so I would like to... We have a little less than an hour. We will break at 12:45, but I would like to open the floor for discussions and things that including, you know, other members that help each of the wrapper tour report. If there are things that were missed from the reports you could start by raising those if we missed. And so, thanks to all of you for attending those session, providing these comments, and let's open the floor for discussion. Please raise your hand. Yes. Phil Robertson, please.

BRUNO BASSO:
Yeah, Bruno, thanks. You and the other laboratories have done a really impressive job of synthesizing comments and discussion. That was I don't wanna say chaotic, but disorganized in many cases. So, it's great to see this come together. I've got a couple of comments on breakout A summary and discussions. First of all, I think it's worth calling out LTAR as the ARS sites of interest. These are 20 sites selected for the representation of cropping grazing land sites across the US and that's just a minor comment. More substantively, I think maybe we should consider whether it would be worth identifying a minimum dataset expectation. Of course, as Jack pointed out very articulately different stakeholders have different needs and different expectations, but you know, the list that the group put together included a number of soil properties that are or should be common to all interested parties.

BRUNO BASSO:
And I think, perhaps a valuable outcome of this workshop could be identifying, for example, a tier-one set of variants to all database contributors should strive to provide. You know, for example, taxonomic information, book density, texture, PH, SOC, and so on. With a second-tier two set perhaps, of desired but not crucial properties. And this could become known as the NASM 2021, a minimum dataset, or the rise Basso at a minimum dataset for soil characterizations.

BRUNO BASSO:
And I know there are other efforts that have tried to do this but largely they, as far as I know, they haven't gotten a lot of attraction despite the fact that it could be super useful. And it could also provide guidance for what actually goes into the databases. And as the database start reaching limits and considering the real transaction cost of maintaining them they may want to, or those who organizing them may want to limit the data coming in to data that meet minimum expectations, not just for metadata, but for actual properties that have been characterized.

BRUNO BASSO:
(CROSSTALK) Yeah, go ahead.

BRUNO BASSO:
So, yeah, Phil, that's a great suggestion. I guess purposes of the report, we can only document things that were stated or said during the workshop. So, if you can send or anybody else send your ideas to us Alfred, I had a publication, I think he had a list of 10 or 15. I think we probably all have the same or pretty similar, but I would encourage you all to the audience, the participants to send down. And then we can maybe refine that. I think that's probably legal from a national Academy perspective.

BRUNO BASSO:
I think the fact that Phil reported this great suggestion both about using (INAUDIBLE) site as ARS, as well as you guys, suggested like during the conversation we had yesterday about potentially including experimental station in the Land-grant university, there're I don't know the real number maybe 80 or so that will provide (INAUDIBLE) historical sites and that means there are lots of information that could really also help. But the minimum dataset, I just wanted to comment that that was the approach that we actually took many years ago. I know probably Jim Jones is connected and he was in that meeting several years ago in India, about selecting minimum dataset to be able to start modeling agricultural system globally through, and was the beginning of the snap project that led to the (INAUDIBLE) system.

BRUNO BASSO:
So, having a minimum dataset would serve both decision-makers and about real data information coming from different sites, but also modeling because of the different systems. So, that goes very much hand to hand. Thanks very much for raising that Phil. Again, I'm glad to see several hands I'll go in order. I think there was Kathy and then I'll call Steven (INAUDIBLE) then Mark.

BRUNO BASSO:
So, I was just gonna report out on some of the chat that's been going on the Slack room. So, someone suggested that we remember that the experimental forest and rangelands are another set of sites that

we can draw on. (INAUDIBLE) soil carbonates, so inorganic carbon is another variable to add to our list of desired soil properties. Somebody reminded us that open science framework has been stood up for to facilitate reproducibility in science and that there might be some lessons we can learn from that community in building out a soil informatics center. And Steve Wood, who I'm not seeing his hand up. There's a thread going on Twitter right now around a comment about valuing insights, not just big data. And so what do we need to do to remember that when we're drawing together these datasets?

BRUNO BASSO:
I think Mark Bradford is here raising his hand on that topic, probably.

BRUNO BASSO:
OK, lemme go just in all the way. But I also would like to comment back on Steve Wood comment because it's critical and Kathy, you have more points or?

BRUNO BASSO:
Let's see. Steve Wood also reminded us that there were several tiered soil indicators lists that are already developed. So, we need to remember that they exist and not necessarily reinvent the wheel.

BRUNO BASSO:
Steven, please. Steve (INAUDIBLE) almost.

BRUNO BASSO:
OK, finally unmuted. Yeah, so I've been thinking back, particularly presentations on Monday or Tuesday, sorry. How Alison (INAUDIBLE) brought up different types of datasets out this long tail of research. And I wanna second make sure that's emphasized. I think we need to bring that into the soil information system.

STEVE :
As well as the sampler, it was nice to hear you bring, Lucas up. You know, I think that that type of systematic sampling is incredibly valuable and it's nice that you set up that with statistical design so they can draw inferences on the entire population across Europe. You know, we talked briefly about that in group A yesterday, when I was in the discussion. I think Chuck (INAUDIBLE), was thinking more about the existing data out there. And we surely, we want to make sure that existing data is in this sole information system, but I also think a statistically based sample of the country in a way, like they've done with Lucas, is valuable. And I would think of that, that even beyond the country, they get about the forest assessment work that that FAO does where they have sampling around the world. And another thing Alison brought out in her presentation was the lack of data in a lot of places around the world like Africa, for example.

STEVE :
And I think if we start thinking bigger, you know, globally here, we can get involved with organizations like the UN, which could promote a global assessment of soils, and that would be extremely valuable. You know soils are very important resources we know for society, and monitoring them in this way is critical, and I would argue so. That was just something I wanted to add here, make sure that's not lost in

this report.

BRUNO:
Well, thank you very much, Steve, Mark Bradford.

MARK:
Just make it quick. Michael Young was actually ahead of me. I know it's hard to (INAUDIBLE).

BRUNO:
OK. You have to the (INAUDIBLE) way they come. So please, so sorry for that, Michael, go ahead.

MICHAEL:
Yeah. So, I'm sorry if I'm jumping the line here. I think, first of all, it was a really terrific role, Bruno and others. Thanks very much for putting that together. I do wanna make a point about the ontology. It is obviously very important if we're gonna create federated data sets, is that mapping, you know, the language from dataset, the dataset is a challenge. The geoscience community has tackled this a little bit.

MICHAEL:
Scott Peckham at CU Boulder has developed a geoscience ontology for a program that DARPA is funding, that's known as (UNKNOWN) or model integration. And I'll put in the chat or in the Slack, just a link to some of the resources that they've created. And it's important just so that we're not recreating something that the framework that they have used or the schema may be readily applicable to the soils community and they may have already included quite a few of that, those variables. Just wanted to let people know that that's out there.

BRUNO:
That's very good. Thank you, Michael. Mark.

MARK:
MARK: Yeah. Yeah. So I asked you a question for Chuck and (UNKNOWN) and it relates to sort of elevating this conversation that Kathy just raised some, Steve, what about insights. So when (INAUDIBLE)discussion your breakout groups came around what we should measure, and I'll just give a little bit of background as to what I mean by that. So you know, there's different reasons for measuring.

MARK:
I look a lot of them, the omics data as trying to work out, to if you like identify causes of effects, but there are many variables we listed that things like solar, organic carbon and others that we already know have effects. And what we want to do with policy and practice people is actually quantify the effects on say water storage or something like yield. So when you take those kinds…that lot of focus on quantifying…if you like quantifying when you change, one of those causes the effect that it has, that raises a completely different set of questions about how we do things like analysis and what we actually measure. So it doesn't lend itself to ML and AI particularly well, right? Lends us to causative inference, and we haven't been doing anything on that since like world war two as soils and act people it's all being medicine and economics. So I'd love to know if there was a much of a rich discussion around that in

terms of how we provide usable information, especially given the other conversations that we've had over the last few days about wanting to link change to actual outcome

MICHELLE:
Yeah, I could start and (INAUDIBLE) and others as well, who were in that, who were breaking a breakout session yesterday. So Mark, those are great points. And I think one of the things that, we didn't discuss this yesterday, by the way, this was known as one of the blockers (UNKNOWN), but we didn't have a solution to that. One of the approaches that is promising is what you referred to the causal inference work. So there's a new stream of work in AI on causal AI and causal ML using these causal inference techniques and bringing in other deep learning techniques to do causal inference. And that is something that we are looking at in the context of, but that's more we don't know if many people working in this space applying causal ML to make simulations better.

MICHELLE:
But that is one of the ways in which you could address the kind of concerns, but this is a complete gap right now, as far as soil types is concerned in closing the loop, looking at... But causal inference is a more recent tool, it's been applied, causal ML has been applied for in different industries in particular, as you said, in finance and travel industry, there has been some interesting work using causal ML, but I think a similar application of that in soils would be very interesting.

BRUNO:
Chuck.

CHUCK:
I guess. Yeah, that was just part of the discussion, I guess, in one sense Alfred brought up, you know, we need to have new theories in soil science. And so, he was suggesting new theories and develop what should be measured, but also if we didn't have data that can be interpolated into developing new theory. So I guess it goes both ways. And the other part was then, you know, are there mechanistic models that can infer and then say, you know, we should be measuring such and such. And so yes, that was mentioned briefly in one of the breakout sessions...Good point though.

BRUNO:
So just one addition to Steve's point about being scientists, very true that there are several indicators going around. So what it really misses is that these indicators are not necessarily scalable. And so I think that's where things have to come together, that the measurements and the modeling by reproducing these observation. And in order to get insights, you need the long-term effect. So at least the knowledge of that particular practices being in place for a system extent of time. But there requires also to know some level of initial conditions on where they started. And so to me, that's kind of the limitation and that's where again, models could, even though the data may not be present, but discussing this whether to a land manager or an operator or a farmer is itself saying, we change this system on this day. That's where models could come and the soil indicators could be in the end, something that shows that the impact of that practices.

BRUNO:

The other point, so that leads to basically long-term, but also making domain knowledge and in survey basically conversation to be able to initialize the models in testing this hypothesis and pull apart the system. Because realistically, not just because I work more with crop models, those are the ones that will provide the level of details that you're interested in given the robustness again of the models and the possibility of validating them. So the point is what we need to have a better knowledge when we ask insight is like, can we identify initial conditions and places where the practice has been on in order to infer about that, and can models understand and reproduce that, such that if we go to another place, we could extract that. And so I see that component very important to link to your point, because I value that enough. And you know, this is a lot of applied science and we need to solve problems and getting insight is problem.

BRUNO:
It's very important, it's not just, I know it's on TNC top list, almost in many others mind, but the limitations again is long-term data and knowledge of when things started and spatial heterogeneity of components of these indicators. Any other points? Yeah. Colin (INAUDIBLE), please.

COLIN:
Hi. I can't emphasize enough that, you know, the biology of these systems is a thing we're finally discovering.

BRUNO:
Yeah.

COLIN:
We know who does what for the first time, sort of ever. And, you know, in plant biology, if you didn't know whether you were looking at a tree or at moss, that'd be a big problem and you learn a lot by gaining that information. So I think there's huge potential here. I know not everyone agrees with that and that's fine. But like we keep drawing analogies to Lucas too. So Lucas has invested big time in this. They're sequencing right now, 1500 sites, multiple horizons ends up being thousands of samples. This stuff like work, I don't know, my group is finding it, we can't get enough of this data. And the data we haven't had yet is really this macro scale data, which is exactly what we're talking about with this sort of sampling. We haven't had large spatially distributed data paired with the functions. We really care about, like, you're talking about soil fractions. That will be amazing. You're talking about productivity. That will be amazing. And these are also the things that industry cares about, right?

COLIN:
Like if you look at the regenerative ag market right now, if you look at what's happening in those places, people are developing microbial products left and right, half the people I did my PhD with went to work at Monsanto that's on Bayer, to go look at fungal endophytes, to go look at soil microbial communities, because those people see those things as huge levers in the system. And I think if we miss that, we'll miss an opportunity and that's it.

BRUNO:
Spot on Colin, very critical. I wanna add something actually. Last night I had a conversation with a

colleague of mine, he's a biologist. And he said you know soils are involved in genes and species ecosystem, climate landscape, management interaction. So it's all of the above. And you just raised an incredible point and the point of Lucas came several times. It's just hard to see how the agencies don't come together to design things in combination with everything that we said, you know, the benchmarking and the data collections. But the biologists have certainly been underestimated. And so, very good point. Thanks again... Maybe one, I have a comment for the group A.

BRUNO:
What I didn't see Chuck, you know, you and I talked about it (INAUDIBLE), it's two things that I know the community is a little bit uneasy about, and one is the possibility of capturing heterogeneity and be able to, in light of all the system going in terms of interest for carbons and climate benefits and so on of you know, be able to capture the spatial heterogeneity with the number of samples, that they really described the system, rather than having a narrower, significantly larger, or what you're trying to estimate. That's one point I'll let you speak for a while. And the other is, again, the standardization of procedures in the lab.

BRUNO:
My point in the introduction, it wasn't about there is a good lab or bad lab, potentially is there are. All those measurements may be very well correct. It's just about agreeing on procedures for a particular things. And so this probably requires obviously a different type of meetings with chemists, analytical people. But those two points didn't seem to get the attention that the community instead kind of calls for.

CHUCK:
That's good question, Bruno. I think on the standardization. Well, we just got focused on some of the methods and then the ancillary data, and there we spent most of our time. You know, there are inner lab comparisons, there are standard soils that people, particularly salt testing labs, set around for the chemical procedures in there. And you know, there's different soil science side has their methods manual. So those are pretty standard. But they do vary as I was trying to point out with the phosphorus. It depends on the soil conditions, you know, whether you have a high pH or low pH, so which soil available test, phosphorus tests you run.

CHUCK:
So that's one of the complexities and mineralogy, and, you know, if we go down to Brazil, it's completely different because (INAUDIBLE). So that's, the matrix is really important. But yeah, I think there needs to be (INAUDIBLE), because of funding, the inner laboratory comparisons have dropped out, there was kind of a standard. There is a North American, I forgot association for standardization on that. But yeah, it is an issue Lucas (UNKNOWN) pointed that out yesterday or Monday was I guess, or Tuesday. Sorry. and so we need to kind of keep on that. But I guess the other thing is, depending on the project, what method, you know, the methods are chosen for a particular objective. So the main thing that we need to do is then, or the communities to do is then figure out how are they comparable or, you know, make some comparisons in that way.

BRUNO:

Any comment on spatial heterogeneity.

CHUCK:
The spatial. Yeah. We didn't even approach, and maybe nobody wanted to tackle that, but that is an issue, you brought up, you know, just carbon measurements. Actually I think carbon measurements are easier than bulk density. Bulk density is the one that drives me crazy. And is affected more by people than even the technique itself.

BRUNO:
It can only lead to 5000 kilograms of carbon air.

CHUCK:
Yeah. So yeah, there's been a lot of work, you, you know, the geospatial analysis that others have done. We probably need to have some recommendations or, you know, some comprehensive studies to look at detection. And there's different, you know, I guess it depends on what you're looking for. If you're looking for a carbon market per se pick on that, you have to have higher resolution, particularly if it's money involved, you know, financial markets, it's in setting it's different, but then, you know, if you're looking at long-term trends, say, carbon change across the continent, then maybe, you know, well, it's still an important issue, but the detection limit or the variability is less, how I say impact. You know, it's less, it's still important, but it's less important than when you got dollars and cents tied to it.

BRUNO:
There were just two points that, you know, the community's kind of debating, and it's important to keep in mind, just because the Lucas decided to do, just pick one sample and you know, the spatial variability, I believe that in addition to soils other features have to be captured, you know, positioning the landscape and land data heterogeneity. So when you sample you can capture a lot of that heterogeneity, by having stratifying sorts of layers of information that could help capture that.

CHUCK:
And as Colin and my way early work with (UNKNOWN), you know microbial activity, you know, we're talking spatial variability of microns or millimeters in size. So go anaerobic aerobic, and just.

SPEAKER:
So, I wanted to break in and bring to the front that Michelle has her hand up.

BRUNO:
That's right, yes, Michelle, please go ahead.

MICHELLE:
I posted in the Slack channel a paper by (UNKNOWN) and others, and I kind of, you started out, Bruno, with your presentation with (UNKNOWN). And I think some of the points that would maybe really help us deal with problems with say enzyme interpretation and when, and where, and how to spend the money on microbial community, or even the way we interpret information about biological response. He really is raising that and proposing an approach to do it. So that paper's worth a look. And I think for

us really thinking about marrying theory with root for statistical fitting, where we're, you know, using remote sensing or remote sense lists, you know, I think we need theory around that, and we need to be really cognizant of the real big problem of marrying, you know, cores that are averaging lots of microbial environments. And if we really wanted to understand nutrient use efficiency, for example, and get a good understanding of the co-location of the microbes, the enzymes and the roots, they for breeding, which is, you know, sort of people's vision now to get it right.

MICHELLE:
We have to just be honest that a lot of, you know, scaling our marrying information is just really, you know, leading us down a rabbit hole or a total waste of time and energy.

MICHELLE:
So, you know, I was trying to, you know, talking to Felipe 'cause he's pretty, he has another thing that I could try to find where it's a really critical of digital soils approaches, because I think of that fundamental feeling and he makes a kind of hilarious critique of detective that goes and finds everything and wants to pull everything into the answer. And then how Crusoe who happens to also be a Belgian like Felipe, you know, follows the theory. Right? And so that comment that was made earlier, I don't remember whether it was by you Bruno or who you know, these sort of tiers of minimum data sets and objectives would vary tremendously. And so then that sort of costs an application. And again, this is a point that was made in the early soil quality work you know, with summaries of sort of, you know, when you spend the money and when you go deep. And so I think, you know, if we're gonna have a discussion of this just clear communication, that would be nice for people newly coming in.

MICHELLE:
And so they don't have to repeat the mistakes that we've continued to make over and over. I'll get off my seal box.

BRUNO:
Well said, Michelle, thanks very much, fully the agree on that. There is a comment from Slack. Next gen biology data needs to be linked to micro-scale responses to be relevant for partition policymakers and others. Anyone wants to comment on that. I mean, that goes back to understanding the data and the functions, and we help make better decisions. There is a separation of how much is basic science and advancing knowledge versus how much we can directly apply some of that in that conversion. They may require additional steps of space and time in between, and that's where these hierarchical approach could help. Anyway, just wanted to share that point with you. I (INAUDIBLE), but I'd like to address Colin's point, please Colin.

COLIN:
Yeah. Yeah, I totally agree. It needs to address the things you're saying, it needs to address crop productivity or force productivity or something like that. How many data sets do you know where those observations are paired? I know very few. My team has worked to generate one. And also, when you look at those macro scales, you need minimum, you know, 70 to 100 observations. And even that is light, right? But these are not huge data sets at the end of the day, right? Like we can meaningfully characterize a site with actually one sequence profile if you sample distributively. And if you pull course

and do things like that, you can generate a microbiome profile that's useful. We've shown that using the neon data, that publications coming out very soon. And those are the things that are going to give us those breakthroughs. We haven't had those pair datasets. And so we haven't been able to answer that question.

COLIN:
So I know it feels like you've been hearing about managing omics too much and too long and not seeing a breakthrough, but I think that breakthrough is about to come because it's being rolled out at these macro scales and that's where we're going to see it. And that's certainly my opinion, but I think, you know, we're finding those things.

BRUNO:
Colin, one thing you may be aware, and I don't know how much you relate into this, but in the private sector, there are several companies doing, maybe Trace Genomics. You probably know they're doing lots of analysis in that, trying to relate that to yields. My point there, is there too much of the leap between relating the functions of, and you know, the community of the microbes with the yield, having such a deep knowledge of what affects yields. And how pulled that apart. Do we need things in between, how would you go and kind of tackle that?

COLIN:
Yeah, I mean, ultimately, what you need to do is take that metagenome profile, whether it's amplicon, you know, barcode regions of 16S or ITS or fungi, we find the fungi are actually often the ones that are the players that are exciting and pulling these levers. And then you need to translate that into features, features that you can put into whether it's a simulation model or even, you know, a statistical model of productivity, and that's where you find it.

COLIN:
And so there's lots of ways to do that. You know, we've had success mapping these individual organisms to pull genomes that are available through JGI, and now they've sequenced so much that we can finally do this in a meaningful way, especially with fungi. And also think about the organisms you're focusing on, right? Like some organisms are ephemeral and I get it bacteria, for sure. Some fungi lived there in the same spot, in the same place for decades. Like we know that from even just the natural history of fungi, like people go out to the same spot in the forest year after year, they pass it down for families because they know that spot makes those fungi because that fungus and that mycelium lives there. That stuff might really matter. And that's how I think we need to start approaching these things. But again, you can't see them if you don't build these pair datasets.

BRUNO:
That makes sense. Before I pass it on to Rodrigo, I wanna say (INAUDIBLE) that to possibility of catching these features in models, we are doing a little better job in modeling genes to reproduce genetic coefficient you know, to go into models. But I think it's about time that we do the same using you know, metagenomic and features in soils to be able to see the functions and the behavior, because to me, honestly, Collin, it's too much a lip of a fate to use just statistics and saying, OK, this level of gene has affected the yields, because they are finding very good correlation and you guys, you gotta be careful

with the correlation and (UNKNOWN). And so having an intermediate step by the models, capturing everything else, and what does this group of microbes or fungi do in order to have an equivalent of the coefficients that we suddenly ignored. Now, we only have proxies as active pools, you know, decomposing with (UNKNOWN) and so on. We don't do anything like that. And I think we can do much better job in doing that. So that's something good point, Rodrigo.

RODRIGO:
Yeah. So I just wanted to bring out a side the discussion that has happened in the Slack. But I see that Mark Radford is...

BRUNO:
Yeah let's go with Mark and then you can read the comments on Slack.

RODRIGO:
Yeah. We'll keep probably bring everything, but it was related to soil indicators about soil health and how can we use them? So Mark, you probably want to jump in.

MARK:
Thanks for the Rodrigo. I realize you raised it in the chat. That's one or other I should speak. So Steve would actually raise a question originally about microbial indicators and enzymes and said that, you know, often in (UNKNOWN) and we see an increase in enzymes, we think that's a good thing, 'cause that's increased microbial activity. Whereas in, you know, many other soils, we see an increase in enzymes, we think it's an indicator nutrient limitation. So therefore, how should we interpret these indicators without being, you know, with the risk of being accused of someone who's only got a single horse in the ring? I think this comes back to this broad question around insights, information causation, and how do we use it.

MARK:
In that, if you were trying to tell a manager change the following, so you can get the following outcome in terms of yield or water, do we tell them to change the indicator or do we tell them to change the course? And, and so then the indicator becomes what we call like statistics and attribute of the course. And isn't actually something we can either measure or quantify effectively. So that's why I was raising these questions previously about other areas of statistical causative inference and philosophy that we haven't touched on a lot. It really comes back down to this practical aspect that I feel it was really (INAUDIBLE) by the Guinness brewing company back with students, teachers back in the day, there's a rich literature there. But I'd love some thoughts around this, this question. Should we be measuring indicators or should we be measuring course?

BRUNO:
Well, just to respond immediately to that I would really like to see if Mary Firestone could say something about. I know she's connected about microbial indicators...shed some light on this, as well as Dr. TG.

MARY:
I'm trying. I'm sorry. I would love to say something I've been (CROSSTALK). OK. So there's a number of

things. First I want to start with Colin's comment that fungi are permanent and bacterial are ephemeral, and I completely agree with you in some ways, and that is Tom, Dick and Harry are ephemeral, and those are the, let's say the 16S taxonomic. What isn't ephemeral is the functionality, and the functionality is fundamental. Then the next question is how do we establish that functionality? And so some people talk about enzymes but those could be pretty ephemeral as well. What then is the cause and the indicator? What is the cause of those enzymes? What can we hope will be semi-permanent and deterministic, and that's going to be the genetic capacity. And to get to that, we've got to go to metagenomics. I'm sorry, but we do, and we're not going to get there with 16S. 16S is surficial, it's ephemeral. Thank you, Colin. So that's one. What was the other thing? I almost couldn't resist saying something.

COLIN:
Mary. I think you just made a great case for measuring soil biology.

COLIN:
I think you made a great case for measuring soil biology right there.

RODRIGO:
(CHUCKLE) OK. But what we also know from history is if we go out and measure (UNKNOWN), we can't predict de-nitrification. Well, we know that there's a metabolic capability of it. So, what we're gonna have to learn, and this is still evolving, is how we put pieces of information together. And it maybe you need to combine genes that determine the capacity to (UNKNOWN) with genes that are indicative of (UNKNOWN). And there certainly are those genes too. So, unfortunately, I don't know that we're ready to use (UNKNOWN) genomics as a fundamental soil characteristic, but we're getting there very fast.

STEVE :
That's right. Especially if we link the functionality with the responses to what they drive as a change in the system, whether it's productivity to building those data sets of integration. I always like to bring it to the system within the smaller system itself of this whole biology, but how that feeds back to the bigger picture. That's something we should never forget. Vanessa.

MICHELLE:
Thanks. Mary, I love everything you said. I would like to add just a couple of things. One is I think it's important for a dynamic information system to think about this tension we have. I spent a lot of time yesterday talking about time series, that I think a lot of these data are useless as single points in time. But I think we need to think about measurements of our potential and even our old school enzyme assays, our potential measurement. Same with metagenomes, that's the potential for the soil.

MARY:
To go down deep the rabbit hole, we would love to have a library of time directed (UNKNOWN) profiles. We may not be able to have that, but what may be more tractable is developing models that take that potential genomic potential, enzyme optimization conditions and link that with site history and link that with maybe some of the physical and chemical characteristics of the soil so that we can understand that, OK, this is a soil, maybe we see a lot of these genes for the near genes or whatever, but it's a super

coarse textured soil, so it's highly unlikely to saturate unless it's under really extreme conditions. And so, I think novel models that bridge biology and physics, biology and chemistry all together, and that's something that's really near and dear to me, would be a way to maybe give us more robust information that's useful for future predictions.

SPEAKER:
I add onto that, Bruno.

STEVE :
Yes, you can. I was trying to see the order of the (UNKNOWN). Go ahead now and then I'd like to invite someone else to comment on.

SPEAKER:
Yeah, I'll just make a quick comment. I think Mary and Vanessa are right spot on. And I think what we need is... I think Steve Orgo mentioned the (UNKNOWN) model, where it starts to look at microbial aggregation interactions, there's some other models like the NDC that are more (UNKNOWN), that kind of model aeration status, and that would be helpful then to (UNKNOWN) the potential activities with the soil environment. And that's where kind of the gap is in helping to understand those system models. Thanks. I'll shut up.

STEVE :
No, thank you, Chuck. Steve, I know you... Steve Wood, TNC, I know you've been active in...

KATHY:
I'm actually gonna break in here, Bruno. Kristen has had her hand up for a little while.

STEVE :
I saw Kristen and then all of a sudden disappear from my screen as a hand, they usually pop up. So, Kristen, please go ahead and then I'll invite Steve to comment on insights

KRISTEN:
Yeah, I just wanted to build on (OVERLAPPING CONVERSATION) Mary and Vanessa. That's awesome. And thinking about some of the conversations that we had about below ground sensors where we could develop some of this high resolution data. So, if we think about de-nitrification where we've got a little bit of traction with AI, we've got some really great process models, and thinking about that same example that Mary brought up from a mechanistic perspective, thinking about these co-expression networks where we might be able to think not just about one function, but like multiple indicators that those suite of genes may be strong indicators (INAUDIBLE) great path forward. I think by using some of these sensor technologies that are on the brink of coming out, if we could get some of these gas samples in real time, then we would know where to sample within the soil in order for us to start rolling up some of these traits, we need a place to understand where to target those at the landscape scale and then thinking about site specific within a farm, within a farming landscape.

MARY:

And that's what we can do with the satellite and (UNKNOWN), say, OK, (UNKNOWN) this part of the topography, that's where we need to be measuring in order to get it, these specific mechanisms, and then we can roll it back up to these traits into something that's a bigger level.

STEVE :
Thanks very much, Kristen. Steve, would you...? Yes, thank you.

STEVE:
Yeah, thanks, Bruno. I just, I guess, articulate or echo what Mark's been saying, too, which is that I think a lot of the conversation, understandably, has been focusing on how do we understand changes in soil properties across space and time. And I'm totally in support of that, but just echoing kind of where TNC is coming from, we also need more of a conversation on what do we do with these measures. And I think generally it's been our experience that even if you can measure something across space and time, we just don't know how to interpret that and what to do with that. So, what is a change in a soil protein index mean or change in, like Mark mentioned, enzyme potential, even microbial biomass. Some of these things that we've been measuring for long, long periods of time. We just feel like there's not solid data that tell us about what a change in those properties mean for a change in agronomic outcomes, environmental outcomes.

MARY:
And so, even if we go all in on a dynamic soil system that tells us fine scale detail about how these things change, it's not gonna be hugely practical and useful until we can say what those changes mean. So, just wanted to reiterate that, I know that's been something...

STEVE :
(UNKNOWN) I tried to support back about the scalability and converting into insights. So, I'm glad you reiterated that. Any comments on both Steve. Yeah, Steve Vogel.

STEVE VOGEL:
Yeah, I completely agree with what Steve was saying there. And I really think... And I think, Bruno, you brought this up earlier. I think this comes back to put it into a process-based representation and seeing what we learned from that and where we have that knowledge gap or where everything does seem to work together as we think it should. I think that's really one of the main values of our process-based modeling. So, I would just, like I said, I agree completely with what Steve would have seen there, but then I think we need to take it to the process-based modeling and see if we can represent that in our mechanistic models.

STEVE :
Spot on. Yes, Melissa, good to see you.

MELISSA:
I just wanted to reiterate my colleague from TNC, Steven. There's so many things to measure, there are so many timescales and considerations of dimensions, but I just also couldn't reiterate more how we need to have the objectives and the why frame the discussion and then set the measurement and the

narrative we're trying to achieve over time at the appropriate time scale for whatever that question is we're trying to answer. And that's what has been a little bit challenging. I've been in and out of this meeting, but I just feel like it would be great to frame out that first. Maybe there is a set of whys from the research community and what research is trying to track, similar to the analogy I used in one of the breakouts I was in on session A was the climate data collection on long term trends and climate change and atmospheric shifts versus day-to-day predictions and modeling for weather and decision making, whether it's daily or weekly or seasonal, versus long term trends happening.

MARY:
And so, there's infinite things you can measure, but you need to first determine what it is that you're trying to look at. And then there's a whole other market based set of questions that clients and users need on that end, and what is the appropriate data cost effectiveness, etcetera, for policymakers, market driven actors, etcetera. And I think they're different and we're not talking about that, and so it's like a big mess of every metric that we could be using in limits. So, I would frame the conversation on the why and then talk about the appropriate metrics. Thanks.

STEVE :
Very good. Thank you, Melissa. So, if we want to be proactive on this points, which are very good, how would we structure a dynamic soil information system here with this community? Do we, instead of talking what do we measure and obviously start from one, we measure it, what would be the drivers? Would that be the questions and then have the samples underneath? That seems what it came out and it clearly serves that objective, but we need to help the agencies or even ourselves, if we get together and design such a dynamic system, it may be beyond what we think it should be, even beyond (UNKNOWN) that they have questions embedded in, but they still have a list and they measure frequently.

MARY:
So, I wonder how do we structure, if we structure based on insights and objectives and then see the design, the scales and the parameters and the variables to measure. So, to be proactive, how would we envision a potential system that we can go forward? Yes, Melissa.

MELISSA:
I think it might be subsets of teams working on the different questions. That's where I think this has been amazing for me to revisit all of the updates and technology, all of the conundrums and methodologies. I appreciated people saying how even now we haven't settled on common methodologies probably for measuring nitrogen or nitrification or any of basic things because labs are using different methodologies and others have said it's not that one is better than the other, but they just have different underlying assumptions and processes. So, you can't compare them. They're apples and oranges.

MARY:
I think I would recommend we organized by what are the questions we're trying to ask and then what sort of soil data can we collect. And then, someone else mentioned in another session how there should be some connection between them and that we we need to be cost efficient and mindful that you're not

gonna be sending out postdocs to far reaches of the sites and then soil, unlike atmosphere, you do need to collect wide spatial variability over time over different dimensions. So, where we can leverage and continue to use, whether it's citizen science or existing field sites with long term datasets, making sure there is some minimum data collection, but not everybody is working on the same sets of soil questions.

MARY:
And then, how do we organize? I don't know if it's the Soil Science Society of America, I don't know if Nick is on or is it an academic consortium that gets an NSF grant, but let's organize and then create different parts of people working on it and have some connection between them. I feel like otherwise we're not all working on the same thing or answering the same question...

STEVE :
Yeah, this is one of the, I think, very valuable thing we're getting out by talking together, maybe one of the first time more recently about these things. And so, connecting the dots between... One thing is teams have to be working within the disciplines, but they have to be part of a bigger system, so their dot is connected to the next, both (UNKNOWN) and objectives. So, that's for sure. Steven.

STEVE VOGEL:
I would just add to Bruno comment there. I think also we should be thinking about the environmental challenges out there. I think this was brought up earlier in the meeting. But food security, climate change, all of these bigger issues. And then, what are the gaps there of understanding, and then the research questions that we need to be asking around those from pushing our theory to pushing applications to address those issues. It seems like the dynamic soil information system would be supporting all of that in working towards ultimately addressing these grand challenges that we have.

STEVE :
For sure. Thank you, Steve. Catherine.

CATHERINE:
I just want to say that I absolutely agree with everything that Melissa and Steven just said. Driving with the core research questions... Well, I'd say starting with the overarching objectives of what types of claims or actions are we hoping to be able to make or take as an outcome of this work and then focusing on what are the core research questions that we can ask to better get us to that point is key. The only other thing that I would add would be to really, once we identify those questions, try to put out as broad calls as possible to say these are the questions, who has data already that might help us answer these questions? At what scale? And then, what are the gaps that exist? Because I think there's a lot out there that we can work with and it might not necessarily be at the scales that we're looking for, the geographies that we're looking for.

MARY:
But, again, there's been going back to day one when there was that presentation about all of the grassroots work, I think it's really important that we amplify these calls for specific research questions tied to unique use cases and make sure that we're being inclusive of all of the data that might be captured on the ground.

STEVE :
Very good. Thanks, Catherine. Kathy, any comment from (UNKNOWN), just not paying attention, knowing you are looking at that?

KATHY:
So, we've got an interesting little side conversation, Mark and I, going on on whether or not there is a fundamental divide in approach between a process driven approach versus sort of a statistical correlation assessment. Those tend to be two very different communities and models and to think about the systems differently. So, that might be interesting to capture and report out. We're seeing lots of excitement around sort of genetic and omic potential for data that seems to get a lot of folks fired up here. A lot of thoughts about sort of keeping in mind big questions. There are some folks dropping in some historical context. So, these are questions that we've been chewing over as a community for a while and there's some literature around them that needs to be recognized. So, folks are dropping in citations for those.

STEVE :
Wonderful.

KATHY:
I see Joe has a comment.

STEVE :
Yes, exactly. I saw Joe. Thanks, Kathy. Joe?

JOE RUDEK:
Yes, so Joe Rudik. So, I know there are many questions to be asked, but if the question is what is the climate benefit of various different practices, I go back to the (UNKNOWN), one of the breakouts about looking at the integrated impact through looking at gas fluxes. And that way you deal with a lot of the heterogeneity both in space and time, depending upon how the measurement system is set out. But I think given all the variability that's out there, we may really only be able to answer what the impact on the atmosphere is by sort of measuring in the atmosphere.

STEVE :
Anyone wants to add anything to Joe's comment? OK, well, we covered a lot. We do have about ten more minutes, I was wrong about the time, I read it wrong. We break at 1:00 and reconvene at 1:30. So, other request? Any additional comments that we could capture in these last ten minutes before we break? And then we have a small presentation by Kathy and a synthesis by Jim (UNKNOWN) and we'll kind of close after that. There is a question from Slack. Carbon fingerprint of the food chain, household's information system provide information to this question. A very good one. I mean, when you make the food system in the chain bigger and kind of help by bringing a lot of the components on the circularity of the system, how much of things we take from the soil and return back to the soil, so carbon, it plays obviously a critical role there. And I know there are groups working on the emissions from the food system and food chain. So, if anything, one has more, please head there. Yeah, Kathy.

KATHY:
Well, I was just gonna ask Julie to unmute and talk about her traceable approach, I don't think we've heard anybody use this particular language yet and ask if she wanted to clarify that a little bit.

JULIE:
I was thinking of a tractable approach to a widespread dynamic system would be to have it be hierarchical so that if you had a system that was systematic, I mean, you could even base this of NRCS measurements or something like what the USGS was trying to do a few years ago, where you have a very basic set of measurements that you conduct in a systematic way with some level of funding. It's georeferenced, and then any other studies that people are doing, as long as it's georeferenced, that basic data set could be used as a contextual thing, it could be used for modelers. A regional study, a site specific study could be added to that system, built over time if you wanted to have additional measurements in a large scale, systematic way, because now there's new types of measurements or we think something else is important and you find the funding for it, you could go back to those locations.

MARY:
Just think of it as we all have different questions, we have different agency missions, we have different objectives, but if we could all link it and anchor it in this sort of basic system that has just really fundamental stuff like texture and whatever you want to include in it, that gets away from the issue of all the different kinds of things that people might want to study. And so, you think of it as something that you could build over a long period of time, and instead of just having data in data repositories, now you have it all sort of in a georeferenced repository that's all together.

STEVE :
Thank you. Thank you, Julie. Any additional comments (UNKNOWN). Getting close to the end, I just want to give this opportunity a little more, use these last few minutes. Maybe some of the other committee members that haven't shared some of their thoughts. Maybe Ramveer.

RAMVEER:
Yeah. so, Bruno, it's been a great journey since we started planning this event.

STEVE :
It is.

RAMVEER:
To learn a lot about soil science for me. I'm a computer scientist and to learn from the experts on soil science. And what I've learned is the opportunities that exist for people in computer science and soil science to work together. And some of it came up today as well. For example, (UNKNOWN), something new that's happening in computer science that could be very relevant for some of the modeling work going on in this community.

MARY:
The other work around connectivity and IoT, the Internet of Things, is something where... So, essentially

what the Internet of Things does for people in the call is it allows you to streamline data from things all the way to the cloud, so you don't have someone going taking notes, all the data is usually put in the common data models and all the data is stored in the cloud, and then you push it through these things called EPL pipelines. So, essentially streamlining the entire process of data collection to driving insights from that, these are things that are taking off in other industries, and I think soil science could benefit a lot from some of the work in that space. And the other one, which I mentioned in the (UNKNOWN) of my breakout session around data sharing, around encryption, and some of the work that's happening there on how do you do AI on encrypted data, that could, again, alleviate some of the concerns we've raised around privacy and sharing of private data.

MARY:
And also, the other thing is that on (UNKNOWN) we had these discussions around genomics, and how do you sequence at scale. That, again, some of the advances around high-performance computing, and how do you run these things at scale, would benefit some of the discussions we are having here. So, that's been one of the learnings, I've been taking extensive notes and I'll be bringing in a lot of computer scientists to this problem. So, really excited about the discussion, and I wanted to thank everyone for it.

STEVE :
That's excellent, Ramveer. Yeah, very important to see different perspective and also both the industry you represent and the advancement that you can make parallel to the university, the research is paramount. So, it's very, very well said. And we hope... I think I'm a strong believer in a public private partnership, we can advance, I guess, faster on that. Thanks again. (UNKNOWN).

ALLISON:
I would also add that I think it's really good that we have so many different questions and so many different streams for collecting data. And it is a challenge for combining the data and making different cases useful, but having these different frontiers is really gonna push the science forward. And then figuring out how to share the data between different projects. Well, yeah, that will be the challenge.

STEVE :
Thanks, Allison. Yes.

RODRIGO:
I would like to share some thoughts also based on this journey that we have throughout this year. We talk with many, many stakeholders and many scientists and people representing different agencies from the private sector, different federal agencies, using different techniques. Let's say someone that will just take a shovel and a soil pit all the way to data mining, computer science and also remote sensing. So, to me, one of the most interesting things is the variety of tools that we will need to understand soils, that we are a very broad community, that we can talk about different aspects of soil functionality and how can we describe those different indicators that we have been discussing? But really the challenge of how to bring this together, and not just only in terms of data, but also in terms of concepts and how can we work together.

MARY:

And maybe we will need to come up with advocates from the community. And these advocates should be from different societies, maybe, and the different agencies, because (UNKNOWN) like we're saying, we haven't talked much about remote sensing. There is NASA, there is (UNKNOWN). There are many other products and there are also efforts around the world, and how can we bring this together to favor increasing knowledge in soils for specific applicability. So, finding those advocates and finding people that connect not only the data, but also commit the stakeholders and the providers of the information, I think it will be something important to think as a community.

STEVE :
Wonderful, Rodrigo. Yes, thanks very much. Last two minutes. Any last comment on that or on different topic?

RAMVEER:
And just wanted to add to what Rodrigo said. I think in addition to all of the remote sensing, I think that in combination with sensors, the new sensing methodologies that are coming out, and I think we need more research on breakthrough sensing technologies as well, that can really democratize sensing. These sensors are still expensive, it takes a lot of... It's non-trivial to do measure all the things that, for example, Chuck presented in his readout. Some of the work that we have been doing, for example, on can you use Wi-Fi signals from your phones to measure soil? So, we had it in this paper on using - which I presented at the academies last year - was on using the time of flight of Wi-Fi signals to measure soil moisture and soil electrical conductivity.

MARY:
Now, the vision is that can you use your phone to start measuring soil, some soil properties? Of course, the challenges will come in, what we discussed in the other breakout session on the fidelity of this data, how calibrated is it, especially if you're democratize sensing, everyone starts putting in data, that just becomes a different problem, but that also makes the measurements much easier, much cheaper so that everyone can start sending this kind of data.

MARY:
I think we need more breakthrough research in that space on new sensing methods beyond the things that we've already been looking at, these electrochemical sensors, maybe there is something else. Maybe there are audit sensors, maybe something with remote sensing, what Rodrigo was mentioning. But some of the work that you've been doing there, Bruno, could be very relevant too, how do you combine these sensors with remote sensing.

STEVE :
Yeah, makes great sense. Thanks, Ramveer, for your comment. Kathy, before we close down.

KATHY:
Yeah, I just wanted to draw some of the the points that have been popping up on the Slack here. So, I think that there's still open questions on how to handle complex and dynamic soil biome over different spatial and temporal scales. So, I think that's worth highlighting again. There's also been a note of a lot of criticism recently on the potential for soil carbon to mitigate climate change, and that this is actually

an opportunity space to call for more funding. That we need to know more about soil carbon dynamics. And as a soil carbon modeler, I certainly can't disagree with that sentiment. Yeah. So, I think that's sort of where the Slack is ending up right now.

STEVE :
That's great. Well, that's very nice to see an engaged community on the Slack channel as well. And thanks Kathy for reporting it. I think we have reached the hour and I would like to close and thank you for the active participation, but also to strongly invite to remain connected for next session starting at 1:30 for one hour with some, again, brief presentation of opportunities from Kathy and then syntheses by Jim Jones and additional points by Rodrigo and concluding remarks. So, we'll see you in 30 minutes. Thanks very much again.

SPEAKER:

Hello, everyone. Welcome back from the break. What an exciting sessions we had earlier. I would like to jump right into the next agenda program here. We have a continuing engagement opportunities presentation from Kathe Todd-Brown, University of Florida. So, Kathe, please, take it away.

KATHE TODD-BROWN:

Alright. So we didn't want to leave you guys hanging and we wanted to give folks an opportunity to continue to engage around soil data and soil informatics. So these are some of the opportunities and resources that have come up over this workshop and some of the things that we collectively were able to come to as we stay engaged. So if I could go to the next slide. Alright. So here are some resources from the group. So this slide primarily focuses on U.S. resources. So the NRCS from USDA has a soil database inference, there's the National Microbiome Data Collection, the Agricultural Collaborative Research Outcomes, LTA, the Soil Moisture Active and Passive, National Geochemistry Survey database, LTER, the Consortium of Universities for Advancing Hydrological Science, and NEONs. So these are just a few of the U.S. opportunities and we will be pushing this to Slack, I believe, afterwards. So you don't have to worry about transcribing websites. So, next slide.

So moving slightly a little bit more international, there are some FEO resources. soilrevealed.org was highlighted as one of them and the ISRIC Soil Data Hub, SoilGrids, and the WORLDSOILS project for ESA. Soils 4 Africa has come up a couple of times as part of the Horizon 2020 program for the European Union, and then OpenGeoHub has a couple of projects here that also have sort of larger aggregated data sets. Next slide, please.

So coming back to the Earth a little bit, MESONET. I believe that this is primarily climate data to help contextualize soils. This is a little bit more of a grab bag. So there's a Soil Health Kit created by Colorado state researchers. That might be interesting to people. Open Data Science is a little bit more generally focused. The Soil Spectra 4 Global Good, this is something that John Sanderlin leads that's focused on building an open-source soil spectroscopy library for machine learning and calibration validations. The Global Soil Biodiversity Initiative is still active, and it should be really interesting to folks here who were looking at the OMICS data and the soil biology groups. Call and go check. I'm sure you are well aware of this group. And then the International Soil Moisture Network and the International Soil Carbon Network are both two active groups looking at doing data collections and integrations. So next slide, please.

So some examples of non-soil repository and information systems that we could maybe start thinking of drawing on, the World Resources Institute, Resource Watch, the Global Biodiversity Informatics Facility, Avian Knowledge Network, and the Hydrologic Information System. So next slide, please. So some additional resources from Breakout Group B were the DOE's Joint Genome Institute, and the National Microbiome Data Collaborative, now has a trellis interface, and that's sort

of a data visualization tool to help you explore data collection, and then Zooniverse for citizen science. I'll have to look into that. I'm not sure what that is. So, next slide. So some funding opportunities were highlighted by the group. The first one being the Signals in the Soils, and then the second one being this new Center for Advancement and Synthesis of Open Environmental Data and Science. My personal hope is that when the Center gets funded, there is a large soils component. That would be fantastic. And then some other sort of soil data organizations that are maybe sitting in this space that are sort of one-off projects.

So the Coastal Carbon Research Collaboration Network is out of the Smithsonian Institute and that's specifically focused around coastal soil course, and that's a data harmonization and integration effort. The Ecological Forecasting Initiative, which is something that Mike Dietz's is very involved with, is slightly more broadly focused where they're simulating entire ecosystems and looking at calibrating biogeochemistry and hydrology in an integrated platform. The International Soil Modeling Consortium has what they call a DO-Link science panel. So that's focused on data and observations, linkages to soil modeling. So we were talking a lot about feeding into process models and sort of using modeling to examine the data, and this would be one organization that's actively doing that. And the International Soil Radiocarbon Network. So if you had radiocarbon in soil data, this group is particularly interested in gathering these data sets together in one cohesio database.

Alright. So where are soil informatics being talked about for conferences? So we put our heads together and we came up with a few opportunities to coalesce with colleagues in future space. The American Geophysical Union, it's an obvious first choice, and the European Geophysical Union, both of which have soil informatics groups that meet and talk regularly, the Soil Science Society of America, Ecological Society of America, the Soil Ecology Society, and the Geological Society of America. Maybe slightly smaller though not much groups and conferences where soil informatics is talked about. There are two biannual meetings, so they happen twice a year, put on by the Earth Science Information Partnership and the Research Data Alliance that are maybe slightly more focused on soil data or data specifically, and they're generally soil components in both of those groups that are talked about. And then finally, every year, the international data week is sort of a conference call organized by Chordata and the RDA that just talks about data in general. And there's frequently a soils components associated there as well. So if you were looking for places to submit your next abstract, by all means, take a look at some of these organizations.

And then, finally, so one of the things that I find really exciting about soil data and soil informatics and doing data-centered science, in general, is this really interesting opportunity for collaborative science and working groups. Typically, as researchers, we'll write a proposal and then we'll go off with our lab and collect the data, conduct the study, write the paper, and push the data. The building sort of reanalysis products or aggregating data like this has a really interesting opportunity to do science collaboratively between groups. And so the Earth Science Information Partnership has an interesting cluster's construction where special interest groups will meet regularly and talk about data issues in that particular subset. I co-chair a soil ontology and informatics group that I think has come up a couple of times in the Slack that everyone here would be welcome to join. We're meeting next week. There's also an ag and climate cluster in ESIP that focuses slightly more on agricultural-ish concerns and tying into certain sustainable foods components.

The Research Data Alliance seems to be primarily out of Europe. I'm less familiar with this group, but they also have several agricultural-focused data, they call them working groups, that might be of interest to folks in this community. And then finally, there are a couple of Global Soil Partnership working groups that are currently active, particularly around lab harmonization and quality, improvement, and soil spectral library development.

And so with that, I think I am on time. And if you have any additional opportunities you want to draw to folk's attention, I would encourage you to drop them onto Slack. And I hope to see everyone at one of the conferences.

SPEAKER:

Wonderful. Thank you so much.

BRUNO BASSO:

..Kathe. We'll move to the next point on the agenda, which everybody is looking forward. I would like to invite Jim Tiedje. to share with us a synthesis of what we have learned. Thank you so much. Jim, take it away.

JIM TIEDJE:

OK. Good afternoon. So. I'm going to try to draw together the synthesis of what we have done. So next slide, please.

So I start with what's the value of dynamic soil information system. I abbreviate it, you see, to make it simpler. But what are the values? So we've talked a lot about agricultural values. That was sort of the core beginning of this. So increased productivity, profitability, sustainability, enhanced land quality. There's also the environmental aspect such as improved ecosystem services, improved soil and water quality, contributions to the Greenhouse Gas mitigation. But there's also a third component, and that's fundamental...I will call it planetary science because soils make up one-third of the planet's surface and there must be fundamental science that's part of that. So hypothesis-driven about scale, trajectories, lithosphere, about interactions between the biological, chemical, physical, about ecological and evolutionary change, the mechanisms behind that. But I encourage ones to also think about planetary science. I will have more to say about that later. Next slide.

So these next slides take off from the three keynote speeches. So what are the functions as illustrated by Jerry Hatfield? And these are some of his points. So support for crops, of course, but also for cities, roads, forest, cactus, tundra. Meaning, other land uses. Now, I point out that we don't know the land use in the future. So if we gather information, it's not for 10 years, it can be for a longer term and other uses of that land in the future. So he mentioned also about storing water, recycling water, distributing water, safe water, supplying nutrients, reliable biogeochemical cycles, and I say planetary cleanup. In other words, the products of society and of agriculture are recycled and the soils do that for the terrestrial component. So Jerry also pointed out it's about water and carbon, and to understand dynamics and interdependencies. Next.

so Joe Cornelius talked about measurements. What can we measure? I also add what should we measure? We had a very good breakout summary A session that talked about measuring. We certainly have the standard historical, proven chemical and physical measures, not so much for the biological or historical. He talked about plant roots, plants as sensors, roots as the window below the soil surface. I think scaling is extremely important. It has been for a long time, time and space, from aggregate to ecosystem-level modeling, the sensors. And this certainly is one that needs multidisciplinary science. Next.

So Alison Hoyt then talked about soils, what I would say, at a more global scale. So about management over time, the projections for the future, carbon mining, erosion, the importance of historical data. And I point out, especially, time series is an extremely valuable resource and maybe

we can capture more of that from historical data. So the archives are also extremely valuable. Connectivity and integration of data and efforts. And she pointed out the value and diverse approaches from systematic continental scale, but also grassroots, smaller-scale projects. I would call some boutique projects because that adds diversity of information to what... One adds consistency, the other adds diversity. Next.

So collect lots of data. So this is big data. So what about Big data? And we had breakout group B, and the use of that in models and so forth, and breakout group C. So I just put down a few points from those different sessions, but those are very well summarized just a few hours ago. My estimation is the data science era is...we're already in the data science era. It'll grow in the future. It'll probably be one-third of the effort in this domain. Next.

OK. So that summarizes some of the beginning points. Now, what has been done historically, in terms of soil science data, a lot. We have existing data. So we have archived data and at various status, reflecting different countries' interests, different scientists' interests, and not very easily accessible. We also have data at risk, and that's data between the eras of science, meaning the first part, I would say, between the 1900s, the last century, basically, and the data management. So we've had a lot of change in personnel and in instrumentation, and the data management is completely changed. So that, I put under the category of data at risk. What do we do about that data? Then archive soils are also extremely important because the time series is so important. So we have a lot of air-dried soils. Biologists don't like that, but perhaps we could learn how to make corrections from air-dried soil information to tell us more about the historical biology. Now, time series, I mentioned it provides that information for the time series. One famous example, there was a paper in the SNT by Dutch scientists on stored soil samples on antibiotic resistance genes. And that really led the era of recognizing the environmental side of antibiotic resistance. That paper has been cited many, many, many, many times. So that only could be done because there was stored soil samples in which they could go back and make those kinds of measures. Next slide.

So now, what is being done? And so we've heard about here in the U.S. and some in North America what's been done. We've also learned about what's been done elsewhere. But important point here is what are the lessons learned from what is being done. What are the values, how to communicate those values, the different scales of effort, the resourcing to do those kinds of studies and how that can be improved? What are the users? How can one effectively communicate with them, extend that user base? And what about nimbleness? Because sometimes things get set up and they get very structured, which can be good, but it also can be bad because they don't adapt to the future. So the point here is lessons being done. And we've had these listening sessions, and Alison Marklein summarized those in the first day, and these are from her slides. But in our effort, there's two portions of it. One is these listening sessions over the past year and the other is what we've been doing the past two and a half days. So this summarizes the various U.S. international, and private sector groups that contributed to what we have learned so far. Next slide.

And one thing we've learned is challenges. And the point is, from these challenges that we've learned from the other groups, how similar are they to what we have learned from this session? And

many of them might be supportive of what we have also learned at this session. So there's a coalescing of challenges from both of these inputs of information. Next slide. And then the recommendations coming out listening to these other groups. And I point out here the ones that Alison put forward from that particular group. But all of them are similar to what came up at this particular meeting. So there's a coalescence of information from these diverse international groups and from those attending these meetings. Next slide. Now, one thing that came up several times is how are we interacting? What is our connectivity? And I point this out because it appeared in a number of talks and discussion points. So one we need attention in the future. Next slide.

OK. Is this dynamic soil information system vision? Is it a Big science vision? And if so, we should plan like it is. So, to me, that means several things. That means that we organize at several levels and with several components. So it takes leadership, it takes community building among various different sectors of science and users and funders and countries. These efforts take major guiding scientific questions. Because in the end, it is the science answers that is the biggest driver. So what are those big guiding science questions? And then we need to build a support base in several sectors. That's certainly the other sectors, the science sectors, the public sectors. You know, soil is one-third of the earth's surface. And from the public point of view, they fundamentally care, if communicated well, about that one-third of the surface that they live on and drive their resources from. And Big science like this is heavily multidisciplinary. So we need to gauge those folks from the various disciplines so that the potential is whole and it's balanced. Next slide.

So if this BIG science approach is taken, what are the downsides? So fragmentation of the community. People have their own interests, they're not happy with one direction or the other. And so it fragments, and support is lost. It can be overdesign. It sinks on its own massive weight because there's always a tendency to measure everything everywhere all the time. And that's too big. I was involved in the first startup of NEON, not as the leader, but participated in some of their activities, and that's what happened to NEON. Everybody wanted to measure everything and it sunk. And it was dead for a while and then resurrected later and trimmed down a lot, and that's why it was successful. Sustainable resources, long-term commitment is difficult. The historical structure of funding is short-term, at least from a competitive grants point of view. So sustained resources are important. So that's gonna be a problem. Also, one needs to keep a dynamic and visionary, but also not lose focus, because some BIG science efforts, if they stick only to what they're doing at that period of time, they will become outdated and not adapting to the future. But they can also become too diverse and lose focus and also not be successful. So those are potential downsides. But if an effort is put into a Big science organization, what if the plan fails? Is it a wasted effort? Is it all lost? And I would say no. The planning will provide for better pieces that can be pursued at a smaller scale because more effort was put into the planning. And then, likely, a larger funded effort that if it's bottom-up, if it's totally a bottom-up system with small projects, it won't be as large as one that was planned from a larger scale, even if it was not completely funded. Next slide.

Now, I also think it's important that we learn from other models. And we have done that in our listening sessions with other people working the soil science area, but there are other communities we can learn from. So I think the one that is most similar to us is the marine science oceanography field. I mean, they are dealing with the other two-thirds of the earth's surface and they are parallel

in very many aspects, in the chemistry, ni the biology, the physical aspects of it. A difference is it's cheaper for us to get to our sites than it is for marine scientists. But they have a simpler system, too. They don't have as much heterogeneity, at smaller scale, that we have. But otherwise, they're extremely similar. And I think a lot of things that we can learn from them, because they have to put more into getting to their sample sites, it forces them to be more organized. So their ship time is a big issue. But when they use that ship time, they're extremely well organized. I actually spent one of my sabbaticals at the University of Hawaii with the Marine Science Group. So I learned a lot about how they organize and how things are done effectively. But there are other fields, too, the atmospheric field, the astronomy field, particle and neutrino physics, and they point out those two because they both have decadal plans. In other words, they get together and they plan what are the big questions that they want to answer over the next decade. And then lay out a vision for that, lay out a funding plan, and have generally been successful in getting that funding.

So do we need a decadal plan for this initiative? Then there's the human side. They do it a little differently. Their new initiatives are usually roadmaps for which they cede new major initiatives, and the human microbiome is one. They are funded for 10 years at a big scale. And then if sufficient science is there, the rest of it is funded under their particular directorate's. And geosciences also have some big-scale projects as well. So what can we learn from other models? So we learn about infrastructure needs and how to obtain them, we learn what are the science drivers, we learn about organizational strategies. And learning from other models, I don't mean that we just take, more completely, the whole model, but pieces of that model that adapt to our situation. So that's the way I look at other models, is, no, we don't take everything completely. What pieces of those models best apply to our situation? Next slide.

Want to say a little bit more about learning from others in my experiences. So the first one is the lessons I learned for a restroom. So I was at a conference sponsored by the American Society for Microbiology in 1995, and it was on Metagenomics, the new science that we could learn from the sequencing of environmental samples. And there were about 40 of us at this resort near Jacksonville, Florida, on the Atlantic coast. It's a great meeting, and the great vision came out of that. In hindsight, it was a small vision, but we thought it was big at the time. But my lesson from the restroom was, you know, we were at this ocean resort and we were just in this room all the time. Other people were going to the beach and they were in their swimming suits doing other things.

So we were a strange group. And so in the restroom, somebody asked me, you know, what we're doing. Why are we sitting in that room and not going to the beach? And I said, well, we had great new methodologies to study about all of the microbes in the world, most of which are unknown. And this person thought it was kind of weird. "What money is in that?" He said. So I said, "Well, some of these microbes might produce new drugs for new biotechnology products." And he said, "Oh, then I understand." So the next day in the restroom, because we're still in that room and never going to the beach, somebody else asks what strange group we are. And so I said, "Well, we're scientists and we're studying. We want to learn about discovering new microbes that might produce new drugs or something like that." And he said, "Oh, that's strange. If you're scientists, I thought you'd be wanting to understand very new things about life." And I said, "Oh, yes, we are. We're

understanding about sort of the frontiers of how life can exist, different ways that they can make energy and survive different conditions." And he said, "Oh, that's more interesting."

So my point from that is the different audiences. One was interested in money, how what we study makes money. The other wanted to know fundamental advances in knowledge. Now, we didn't meet a third day, but I think a third question would be, well, what microbes do for the world around us, and that's the environmental perspective. And back to my first slide, then, the components that we need are, what is the value in the monetary sense, what is important for the environmental sustainability of the planet, and third, what contributions can we make to vary fundamental science that advances knowledge overall? So that's my restroom example illustrating the three components that I think are important in these kind of activities.

They also mentioned lessons from physicists. So, I've worked on a couple - several different projects with physicists, but one I'll mention is the deep underground science initiative which was or is neutrino physics. I would just best describe it as neutrino physics, where the goal is to dig very deep in earth to get shielded from the cosmic neutrinos so that other measures can be made about neutrinos, for example, thermobaric decay, a few other kinds of major physics questions. And from that group, they believed to be Nobel Prize-winning activities over the next decade or so. So how do they organize? NSF puts out a proposal, a call for proposal, for the first level, which is to develop the science questions and the other components needed. They get together and said there's gonna be only one proposal. All of the people involved come to one meeting, they determine what the one proposal is, submit one proposal, they get funded. And then they strategically keep the community together, organized around those major goals, because the size of the amount of money they need is huge. And if everybody's together, then they can make it. And they told me, "You, biologists, you'll never get more money because you argue with each other and you criticize over what you do." So they tried to keep the community together, they developed these decadal plans, what everybody agrees is the major component, and go for it. So that's what they've learned from the physicist and how they obtain larger-scale funding.

Now, I admit that the physicists, the astronomers, the oceanographers, they have core instrumental needs that they must come together on to drive them together. And other scientists don't have that so much. So NSF has a program called major research equipment and facilities construction. So NEON came out of that funding. It's not to fund the science, it's to fund the construction. And it was originally done for big physical facilities, not distributed ones like NEON, but there are a few other distributed ones. So IGO and the Daniel, in a way. Solar telescope would be others. Not that that's, necessarily, a source of funding for this activity, probably not, but the principal of organizing around those kind of goals is one that could be applied to this particular goal. Then GenBank, I mentioned and came up in the summary by...in B, and that's part of the International Nucleotide Sequence Database Collaboration, in which GenBank is the U.S. contributor. Then there's a European component and the Japanese component. And each night, those three exchange all the sequences in the database. So whether you access that in Europe or Asia or the U.S., it's all the same information. But strength comes from that agreement and that collaboration.

Now, as was mentioned earlier, when this got started, I know some of the people that were involved in starting GenBank. And there was a big debate over data quality. Do we check this sequence before it's released to the public? Large arguments for that. Don't put junk data out there. It's got to be good data. But at the end, the guy who led it for MIH said, "No, you submit the sequence and we're releasing it to the public. The data, the user beware, of course," which, in the end, turned out to be a great decision because to check that kind of data over would have cost a huge backlog, and it wouldn't have been released for years. It's better to have that out there to be released, many people looking at it, being able to correct it. And as things are going forward, it's certainly the right decision. So those are other examples from learning from others that could be applied to this situation. So next slide.

So this is a figure that I have used. Joke Handelsman and I were co-chairs of the NRC report on the New Science of Metagenomics, which came out in 2007. And I drew this figure at that time because, at that time, it was all about sequencing. And that was before even Alumina came on the market. Most of the people thought all the money would go to sequencing. I didn't see it that way. I was more worried about the data analysis, the computational side, and then we'd get the experimental side. But that's kind of how I viewed this figure. I drew at the time... It's how I view the field of metagenomics which sort of became the microbiome science as it developed over time, so that it would grow in terms of resources and effort, and it would change in its components because it is more of a decadal plan and that one would get to the experimental phase, in which, first, there would be observational outcomes and there would be a hypothesis derived from that, and that would lead to, then, a more experimental hypothesis driven science. And this figure that I drew then is pretty much what happened over that period of time. But now we're past that time. And so what's next? My point here is that I think it's good to have a vision. And so I would challenge people, that's your homework, just to draw your vision for the next 10 years for this topic area we're discussing, what would be the component parts and how it would develop over time and how it would grow. So that's why I show this. It's important to have that vision for the next decade. Next slide.

So to conclude, then, so thank you for your participation. There were over 280 the first day that participated, and to our previous presenters because there's a whole group of people here that spent a lot of time preparing information for our committee and provided good input to that committee. So they and you are important contributors to this activity. So thanks. That's it.

BRUNO BASSO:

Thank you so much, Jim. That was brilliant. That's really paved the road for the future and for us, for sure, to gather our thoughts. As planned, I would like to offer Rodrigo the possibility of synthesizing what has been coming over Slack both in the last few days if we didn't catch. So Rodrigo had this assignment and even more recently in the last two sessions. So, Rodrigo, take it away. Thank you.

RODRIGO VARGAS:

Thank you, Bruno. And I want to thank all the participants for being here with us for the last three days and also for all the participants in Slack. I also want to encourage everyone to continue

communicating with the community. Also, please, continue using Slack now to share different ideas and different opinions. So what I want to do in the next five minutes or so, I want to bring some points that were either discussed in the morning or discussed in Slack as a part of wrapping up this workshop. So I want to start with the challenges, some of the challenges that we discussed. There is a lot of information out there, but how are we going to put it together? We don't need to reinvent the wheel. We need to increase connectivity and for information and knowledge. We do have to incentivize the community from different sectors to share data. And maybe this would need to change the reward system of how we view these.

We also talk about archiving data, but not only the challenges of archiving this information but also the physical samples. Who is going to pay for this and how are we going to do it? We also discussed about data fidelity. When data fidelity from the point measurements from the laboratory, there has been a slight discussion about that as well, but also about the data fidelity of the value-added products. This could be maps, this could be machine learning product, and discussing implications for research, for education, for extension, for management, for commercial applications. And that discussion needs to be continued. And finally, something that is happening right now in Slack is discussing that we need to work, also, towards a unified community. Because as we were hearing just a few minutes ago, this is something that we need to work together to build and work towards big science.

Quickly moving from challenges to some of the opportunities that were discussed, we can talk about the importance of to link space and time information. And this can be seen as a challenge, but I want to say it as an opportunity, an opportunity for models to bridge information from physics and biology. We have an opportunity to get large spatially distributed data and how to put it together, these physical and biological properties for applications in industry, research, etc, based on the stakeholders that we will be interested in. We're talking about microscale responses and how are we going to provide information and knowledge to practitioners, policymakers, again, to different stakeholders. We also had a very nice discussion about how we interpret the indicators and how we measure causes. But there's also been some challenges of how we define them, and topics related to soil health, soil functionality. And we may require different approaches to interpret data for these specific purposes in terms of data collection but also data analysis. And we got some discussion about different statistical approaches and how to do this.

Another important point was the inclusion of new computer science tools such as confidential computing, for example, encrypted data. We want to have data available, but we also have to think about national security issues, policy issues, commercial issues, etc. And we also talk about, in terms of computer science, what is discussed as artificial intelligence responsibility and also accessibility tools of data supported by artificial intelligence. We have a lot of information, but we have to find tools and how to make that information easy to users for knowledge discovery. I think it's important also, in their reward system of the impact of the information, how to improve the traceability of this data. So we talked a little bit about persistent identifiers, PIDs. These cases, we can identify data sets, we can identify publications, but very important, how we link this database with information, with the authors. And there were examples of PDI graphs. This is very important for agencies to

track the impact but also important for the independent researchers or groups to see how the data has been used, and also for applicability purposes.

Another important point of opportunities is define ontologies for soils and important to communicate what we mean by different variables, but also how we define uncertainty. And as an opportunity, we have to not forget the data that we collect, the variables that we're collecting, but also the information that's associated with them in terms of uncertainty. We have to better quantify this. We have to make a better discussion on how to disclaim and interpret it for each one of the stakeholders, which is something that also requires more discussion. But it is clear that the information will be, information in terms of uncertainty, will be useful for different stakeholders and for different value-added products that this community can contribute. And there are some of the things that we discuss about moving forward, and I recognize and I'm not covering everything that has been discussed in this workshop, but just highlighting a few points, are we have to think about the applicability of such a system and we have to keep in mind the grand challenges, what they are going to be. We can decide (UNKNOWN) security is a grand challenge, global environmental change. And that applicability, it is extremely important. But ultimately, it is going to be a community effort. And this community effort will require a subset of teams working on different questions, and can talk a lot about how we can learn from different examples and different communities to move towards developing dynamics of information system.

Finally, a few topics for moving forward. We need to think about education extension and outreach efforts. We talked a lot about how to collect data, but also we cannot forget about training the next generation of students. We have data, but also we need to have answered questions from these data. And we have to train the next generation of students to access that information in terms of knowledge, to gather and knowledge discovery, and data interpretation. So I do think that, in any effort, we have to be thinking in the investment of developing an information system, but also training the next generation of students that will access that. And this is an effort that cannot be done by just one sector of the community. We recognize that there are federal agencies involved in this, there is a private sector, there is research community, there is the opportunity for citizen science, and also there are national and international efforts that we can learn from them or we can work with them.

And we also talked about increasing data sharing. But in principle, we can agree on the data sharing. And we know that, in the terms of our research, it is a requirement to share your data if your research is funded by the federal grants. But there is a lack of enforcing for this to happen. So one discussion that we have is, should there be more enforcing for data sharing from the federal agents perspective or from the journalist perspective? That if you don't share your data, then you cannot publish, or it should be just a better discussion of the reward system, as I was talking just initially, on how to incentivize and promote data sharing, which, in some cases, it is a responsibility and obligation. But recognizing also that, in other cases, it is complicated because of privacy issues or security issues.

And finally, who's going to pay for this moving forward? We need to think about data storage, data discovery repositories, but also education. And it's not only about collecting the data but also how we're going to analyze the data. Maybe, also, we need to think about funding of data mining and the knowledge discovery from this information. So with that, I haven't been able to keep with Slack. Kathe, if you want to jump in to say if anything has been going on. But with this, I want to summarize some of the discussions that we have had early in the morning.

BRUNO BASSO:

Thanks.

KATHE TODD-BROWN:

I think that captures things that have been going on in the Slack for the most part as well.

SPEAKER:
Well, we have come to the end, I again would like to thank, Jim first for having such a nice synthesis in these last few minutes. Obviously, I don't want to summarize even further, but I did have kind of reading Jim's mind that I submitted the slide earlier. And it's far from being the homework that you invited us to see how we envision. But I did want to put some of the concept that and the ideas that have been floating around the last, these three excellent, excellent days. And so if we are going to design a dynamic soil information system, what we kind of agree that it came in needs to be a coordinated interdisciplinary that captured the systems.

The objectives have to be multiple from, you know, monitoring and creating a minimum data set. This is a critical piece that several it's... we need to learn from history. There are several groups that have faced as the example that I've mentioned already about the IBSnet project. We are being successful in modeling crop yields because of an agreement of scientists coming together and decide what is important to measure and for what.

So we have to come together in deciding what to measure. Like Jim said, we can't measure everything everywhere any time. And, so, the minimum data set and the linkage with potentially do that at selected benchmark site is important. Then everything else that comes with it, with sampling possibilities and new sensing and but they're all part of, again, the questions and the objectives all also to serve and modeling serve back the opportunity to understand the systems because we can't measure things everywhere any time. Models will really allow us to build that bridge that we need. And I think the world of insights that has come out, we have to think. So, the idea that Jim DG shared about, yes, ICD is a big science, but we also have to convert into some potentially applicable solution in a shorter-term since there is so much already known on the system.

So, dynamic information system needs to serve also to advance insights, to guide decisions. The scale has come out so many times. I work on very much in spatial and temporal heterogeneity and each of you in your own fields have different scales that then mine. Microbes to the landscape and so on. So, scale needs to be some objectives. Teams have to be sub-teams, but we connect the dots. So, for example, be able to prioritize on some of these themes like soil biology, But see soil biology says that one piece of the puzzle.

As a systems scientist, I always think that things cannot be seen in isolation. Even if they are, you make significant advancement within that science awfully, that science has a domino effect and contributes to something that humanity will benefit both from that breakthrough as well as for the trickle effect on the system. And so the data and the knowledge that evolves from this integrated and coordinated effort then is available to continue to allow us to have healthy soils to support healthy lives.

Funding has come out in terms of, you know, briefly how we going to do that. Obviously, that's a hard call to make. I do see a guiding, an initial public sort of interest and incentives. But private has also... could play a role here. So, public-private partnerships where private sectors could fund particular, you know, either some team, some objectives whether is again, related to the computing sides or other things, because we are in very much need of the dynamic soil information system that is real and that has to happen sooner than later. I want to move on to the next slides with some of

the acknowledgment. My first thanks comes from selecting me to serve as the chair of this committee that has put this program together.

I want to personally thank all of the members and the committee, all of you have done an amazing job, we have met 48 times as official meetings and we have had such a great contribution from the partners that, like Jim said, they put times into it. Without the work of the National Academy staffs, Carolini, Esther, Robin, and all the people in the last three days in organizing have made this a real possibility.

Listening to the session it was really amazing. It was sad not to have this meeting in person because of the pandemic, but we see a little bit of the light now, and hopefully, some meetings along this lines could happen in person sometime soon. But it did give us that opportunity to listen that we wouldn't have maybe come as prepared as we were by seeing all the efforts going on globally. And so I'd like to thank the speakers, both the keynotes, the panelists, but mainly all of you. You have engaged so much in providing great feedbacks, we could never see all the facets and sights of such a complex system.

We have, even when we try not to admit, but strong biases towards our scientists. Like Jim said, you know, we often criticize our own work and between each other. So, that has really been an amazing effort. I think I'm extremely pleased myself how we turned out. And obviously without the sponsors' support that allowed this workshop to be organized. As I mentioned the first day, this was possible to with the support of the National Academy of Sciences after they fund the National Corn Growers Association, National Science Foundation, the Nature Conservancy, USDA, NIFA, USDA, NRCS, and USDOE RPACE.

So with that, I would like to thank you deeply again. And please stay safe and we hope to see you to continue to work and hopefully build a dynamic soil information system. Thank you so much indeed.

CHUCK RICE:
So, welcome everyone, this morning, afternoon, wherever you are. So, this is breakout session to talk about what we should be measuring in data collection. So those are the questions. My post-doc Marcos Sarto is going be taking notes that will be for these two sessions. And then we'll be synthesizing that later this afternoon evening and present that tomorrow. So anyway, I'm glad you volunteered or was recommended for this particular breakout group. So, we're going to spend the next hour plus talking about measurements, sampling and archiving. And we're going to try to go through these questions for measurements AB and C. What should be measured in soils? And the key point here is where, when, and how frequently? And I guess I've put it on a temporal and spatial scale. How should it be contextualized, combined? And then how do you use proximal remote sensing to enhance the data collection? I think reduce the cost of measurement is also a real key here.

CHUCK RICE:
So, I'm gonna, kind of break this out, but spend a fair amount of time on this first question. So just in the context of what measurements. I would caution everyone to not get down into the weeds and specific methods or comparing different methods, like how do you measure arrogance? Or how do you measure microbial biomass? So, let's try to keep this at a medium or high level. And maybe if we can just discuss maybe kind of a key physical, chemical and biological indicators, and then are there some measurements that are more integrative between those and how that relates to soil resources? Soil health was mentioned in the breakout, earlier in the fireside chat, but in some cases it's even broader than that far as documenting, understanding soils for the different customers. So, anybody want to chime in with that kind of context. But let's talk about the physical.

SPEAKER 1:
Chuck. Are you just wanting us to jump in?

CHUCK RICE:
Yeah, I mean, yeah, sure.

SPEAKER 1:
And sorry, I missed the morning session. So, if I say anything that they said, I didn't hear it.

CHUCK RICE:
No, I was more at the higher level and want to innovate with companies.

SPEAKER 1:
OK. So, physical which is kind of what I've spent a lot of time on. I think some measurement of, aggregate stability is key. Some measurement of bulk density. And organic matter can be claimed by the physicists and the chemists and the biologists. So, I'll claim it first. Some measure of carbon or organic matter or whatever is obviously necessary. And I should've said texture. Texture is the first one that you need before you have any need before you can use anything else. Those would probably be the ones that I would start with. There's lots of other, there's lots of others of course.

CHUCK RICE:
Sure.

SPEAKER 1:
Yeah.

CHUCK RICE:
OK. Yeah. Well, you've probably seen a few others on that side. I've given a few talks and talked about the, my Holy Trinity of soils. And that's basically organic carbon, some measure of physical, I like aggregates, but it could be infiltration, whatever. And then some measure of microbial I got my favorite ones, but, some measure of composition or activity. So, but then there's ancillary measurement, I think. Texture is kind of a guiding, to integrate or understand across landscapes textures kind of I don't know why you say a driving force, maybe. Others, can't see everybody.

JONATHAN:
So yeah, I was actually just gonna ask, I have two things. First, just kind of procedural since we actually are a very large breakout group. Do you want us to like raise hands or just shout out? This might get a little out of control 26 people. I'm not sure if you want any organization there.

CHUCK RICE:
Good point, Jonathan. I think we have a National Academy staff person. I will try. Let's raise our hand. And that way I can try to go through and if Kara can help me on that, that'd be beneficial. Unfortunately, the note taking, it's not up on the screen so we'll just have to remember what's being said, but we are capturing all those notes.

JONATHAN:
And so, actually the point of wanted to make, was actually question. In terms of, instead of like maybe diving right into physical, chemical, biological properties, I guess thinking about like, what are our audiences that we're after. Yesterday, there was a lot of discussion that several people made the point that most of the speakers were speaking through an agronomic lens. In terms of what information, they wanted, but there's other communities, the more natural science, ecological science, as far as science communities. And so just kind of keeping these different stakeholders in mind in this discussion would be really helpful.

CHUCK RICE:
So, I think it's agnostic in the sense that, there was a heavy ag focus on that because that's a major user. But we talked about the systems yesterday, you know, NEON, the National Ecological Observatory Networks. So, and there's the urban audience, there's NGOs like Nature Conservancy. They're focused on managed lands, but there's certainly other communities out there in the urban setting. And from a US perspective, NRCS, while they're focused on ag, Dave Limbaugh is on here and he can chime in. But they also work with different community users. And of course, you heard this morning, there are companies that are looking at in what was talked about this morning was ag, but there are companies looking at ag or, sorry, carbon markets or ecosystem service markets that are a little bit broader in that sense. So, the customers are very diverse. So, I guess from say a basic information, how can any of these

measurements pertain to resources. Soil resources in the US and use. Dave, you can correct me if I'm wrong or jams on the line.

DAVE LIMBAUGH:
You're good, Chuck. I just say that from NRCS's perspective we look at all lands for all people. So ag is a big part of it, but urban forest range, it's all there. So, we don't, we don't discriminate.

CHUCK RICE:
OK. Melissa, you have your hand up.

MELISSA:
Hey, I'm Melissa, with the world wildlife fund. I just wanted to build off of Jonathan's question maybe a little bit. I don't disagree at all. And I'm really aware the Holy Trinity, of the organic matter of physical measurements and microbial activity or composition. And all of the things Eileen said in terms of things to measure. But I was also wondering if and why we wouldn't start with the why. So not just the audiences, but what are we trying to determine about what we value or want functioning and soils, right? And so, I wouldn't just say we should depend on what the customer or the client or the audience wants. Because at least in WWF, that is part of what we feel is the problem with an over-focus on carbon, because we think soils and ag land is going to sequester urban and be a sink and all of these things.

MELISSA:
And so, everyone is focusing on carbon and not necessarily focusing on other things that are needed for soils to be a healthy functioning part of ecosystems and supporting biodiversity and landscapes. So, whether it's for ag or urban or whatever, I just think we as soil scientists, and I'm happy to say, I am, I do have a soil science background. Like we should be defining a more holistic notion of what that means. And then within that is I think our role on influencing or trying to influence those that are trying to focus maybe a little bit too much on just one of those metrics instead of the other, because they want to, pay for offsets or whatever through their objective. And try to meet them part way. But I think, I think we, aren't just trying to meet the client's needs. I think we're trying to holistically define what healthy functioning soil should look like.

CHUCK RICE:
And yeah. So, let me give a little bit background. I'll just take a minute or two, I don't want to take up too much time. But what started this whole idea was a conversation back in 2015 when the global soil partnership was producing the state of the world soils report. And what came out of that was when they were writing up the state of soil resources for North America what surprised me was that the Canadians have better information on how their soils were changing than the US did. And I always thought, US had a pretty good system no offense today. But, a lot of when they were producing that report, a lot of things like on, soil erosion and other things, they stopped at the whatever parallel, the Canadian border.

CHUCK RICE:
And so that kind of started it. And then the question is there are so many different groups collecting soil information. How do we collate that information, which is for another breakout group, but how can we

capture all the information, whether it's from government, private, NGO resources to understand the changes that are occurring in the functionality of our soils? I think Luca talked about one of the keynotes also talked about the soil's functions. So, obviously soils have different functions, but we need to, what are the right measurements or metrics to understand how soils are changing upgrading or degrading? I think that's kind of like what Eileen was talking about was key some key measurements that would monitor that change. Does that help?

MELISSA:
Well, I mean, like weather I think there's in the fireside chats. I apologize, I wasn't here yesterday. But the fireside chats, even this morning, we're capturing different challenges of measurement and sort of timescale spatial dimensions. It's like climate, right? There are climate change scientists who are measuring long-term changes in weather patterns and systems that are really important to track and measure. But then there's like daily forecast, weekly, seasonal forecast, and we need them all to understand and use data for decision-making. And so not sure it's exact analogy, but we should be able to walk and chew gum and figure this out. And, having been part, of the Gates foundation in the early days, the ag development team standing up. More than what, 15 years ago, almost now, more than 10 years ago. And soil mapping was a huge part of what they seeded. Like there so many data systems available. And I appreciate the example you bring between Canada and the US. But I feel like what, and how can we align on what we need to do a certain job and task? What is our role as soil scientists to influence that?

And then let's come up with a pre competitive platform. So, we're not talking past each other and competing with each other. Thanks.

CHUCK RICE:
So, Melissa, that's actually, that's part of the question is what to measure, but also when, and maybe not so much how, but when. Because different customers, you know, if it's a land manager, she, or he may need information at daily timescales. If they're trying to measure plant productivity, whether its grasslands or forest or ag crop. But then like with climate change, I like your analogy. We want to look at long-term changes as far as soil erosion. That's not probably difficult to measure on a daily basis, but the timeframe is more decade old. Maybe a year at the, probably at the finest scale. So yeah, I think you're right. Weather and climate. So, I guess that's the question is what scale, temporal and spatial scale is appropriate, as well as the key measurements. Colin?

COLIN:
Hi. Yeah thanks. So, I totally agree with everything Melissa said, and I'm sure I'm not the only one who's going to say this. And it's already been brought up a couple of times, but it's really so important that we capture soil biodiversity in these efforts. So, we finally have the tools at the appropriate, that can be rolled out at scale to identify which fungi and bacteria and soil animals and other biodiversity is living in the soil. And we're finding more and more than that has dramatic effects in the functioning of these systems. But building, identifying those links requires this paired information. You need the microbiome and the measure of aggregate stability or a measure of crop production or something like that, which is often what we're missing to make those jumps. But these are the things that, you can imagine that eventually there's new microbial products.

COLIN:
Some are already happening in agriculture. Most of them are bacteria. There's huge potential in these other organisms to do things. And also sort of capture the biodiversity of soils. In natural landscapes versus agricultural ones. I don't think we don't really have a good handle on the state of soil biodiversity and what may have been lost and what is under threat. And these are really important, natural resources for us. These are things we want to understand and protect the best we can and we can't manage what we don't measure. So, yeah, I'm sure I won't be the last one to emphasize soil biodiversity, but I'm gonna keep saying it.

CHUCK RICE:
OK. Jim, I think you're next. And then Luca.

JIM:
OK. So, my thought in this regard of what we measure is that we have a lot of experiences around the world now about what has been measured. And I'm particularly thinking about the yesterday's presentation about the European dataset. Because from that experience, we know what has been useful? What the issues are? What the costs are? What the values are? So that's where I would start is to look at what has been already been done. And what's been learned from that. And then derive from that, what should be measured. Also without regard is the dynamics. So, some things change more rapidly. Other things like soil texture changes very slowly. So, the timescale in which different things are measured, could be important from a cost perspective. And I'm a microbiologist as some of you know. So, I often look at this from a microbial point of view. What affects the microbial population? And I would put four things down, of course, organic carbon, especially available carbon and pH. But one thing that's often not measured, but it's really important to selection in the microbial community is the aeration status and that's often not measured.

So how would one measure the information that drives the selection for aerobic or anaerobic processes? And the last of course is moisture.

CHUCK RICE:
But Jim, if you knew bulk density, if you knew moisture and you knew texture...

CHUCK RICE:
Do you think you could estimate aeration?

SPEAKER 1:
I would still look for something more direct.

CHUCK RICE:
OK.

SPEAKER 1:
And I suppose, one thing that's simple is the drainage characterization of the soil. Another thing that

relates to that would be color. So, the parameters that you mentioned of course are ones that dry the aeration, but it doesn't necessarily tell the history of that particular site. Where I think drainage characteristics and the color does.

CHUCK RICE:
OK. I'll come back to the diversity measurement back to you Jan, but I wanna move along. Luca.

JONATHAN:
Hi, Chuck. And hi, everybody. From Europe here. I don't want to at all tell how you should do it in US, but maybe I can tell you how we decide what to measure in European union with our LUCAS system. And the main issue is of course, that at least for us at any parameter, it's tremendously to the cost of that exercise. So, each time we add a parameter, we must explain to the parliament why we are doing this. And so, this is very crucial for us that we make a selection of issues that are somehow responding to societal priorities. To priorities that are understandable to the taxpayers. So, I just mentioned few of the new parameters that we have introduced in the recent LUCAS survey. For example, now one of the big emerging issue for European citizens is contamination.

JONATHAN:
So, they wanted us to measure pesticide residues, antibiotics in soils. There's a huge emerging issue in Europe about microplastics in soils. Just to mention few of them. So, these are things that are extremely relevant to the parliament. To the people who decide, but maybe to us soil scientists may be a little bit remote from our backgrounds and from our scientific interests. So, the issue then, an issue that we should clearly define why we measure something. If we measure it for scientific purposes to have new insights into some scientific issue that we are studying is one thing. If we are responding to societal needs, this is a different thing. So, our system is an operational system, so it's not geared to research. It's geared to provide data to parliament and council.

JONATHAN:
So that's just to explain you our background. I was just hearing soil biodiversity. For example, we introduced metagenomic and DNA sequencing in soil samples because apparently now soil biodiversity is a huge issue emotionally because we have a big pressure for many NGOs. We have big pressure from many interest groups on this topic. So yeah, just to give you a little background about how we handle this in EU. But I don't want to teach you how to do it in US. I'm sure that you have tougher drivers and different priorities.

CHUCK RICE:
So, you know, going back to what you said and Jim talked about, I guess what's been done, we've got a long history and database on the physical and chemical. Yeah. We could tweak maybe some things in that. The biological seems to be kind of more recent side, and then the course, then you have to worry about that dynamics. In what happens today may be different than tomorrow or yesterday. And you know, some of the dynamics. I guess, Jim and Luca, you mentioned the bio-diversity and metagenomic. I'm gonna ask and I'm a microbiologist most of you now, I guess the question is, form versus function. If go back to the keynote talks yesterday, the functionality of soils, how can the genomic information relate to current functionality or functionality? And I'm proposing that as a controversial question just

for that purpose. So, anybody want to respondent to that? Jim, go ahead. And then call on Young.

DAVE LIMBAUGH:
So of course, the diversity in soil, the microbial diversity of soil is enormous and probably not really measurable. At the level of genetic variation of different genes. So, there is this paper originally by Al Kanaka which shows this diversity saturate and at least at the functional level, it probably does. At the stability level, that's different, that adds more value to diversity. But overall, I think the question of diversity saturation is one that's relevant for microbes in soil, because it is so extremely high. And as I said, unmeasurable.

MELISSA:
I don't think we should let the fact that it's difficult to measure, let us stop us from measuring it any. Because it's certainly telling us a lot more than nothing already. And can we link diversity to function to know that we need datasets of biodiversity and function, which we're often lacking. And so, it's difficult to answer that in a general way. However, when we build them, we often find them. In Europe, we've linked variation in which mycorrhizal are there. Using the sequencing approaches. We have realizing that we have not saturated diversity curves to threefold variation in tree growth. We then take those soils, bring them into the lab and show that if we inoculate soils with these different fungal communities that we source from the field, they do induce these growth effects on these pine seedlings. And that's just one example of the types of connections people are building. There's a huge opportunity here. It's just, you know, we're only just now rolling out these technologies at the scales with which we really need to answer that question. And so, I think it's a bit premature to suggest maybe it's not going to work.

CHUCK RICE:
I was asking a controversial question. Mike Young, you're admitted.

COLIN:
My background is not in technology. As you can tell. I'm actually in soil physics. And so, we know we really focus on physical hydraulic properties. Which then roll into moisture status, rolls into soil temperature and those kinds of things. And, you know, to me, the parameters that drive the frequency or the scale of measurement depends on the variability of the process you're trying to measure. So, if you had a perfectly homogeneous material, which of course doesn't exist, you'd only need one measurement. And if you're trying to measure climate, which is over decadal timeframes, and you don't need to collect soil temperature or weather patterns on an hourly basis. So, it really depends on what it is we're trying to do. It gets back to Melissa's original comment about its the question we're trying to answer. Now, at the same time, if we want to use interoperability, we need to try to collect the data in a way that other people can use it.

COLIN:
So, it's not just about collecting data monthly, yearly, or daily. It's a collecting at the highest rate that we can because the data should be interoperable for other communities to use. And at the same time, it should be done at some type of a different scale. So, you know, remote sensing is now getting quite good. It's not perfect, but we're almost down to the meter scale with satellite remote sensing, and that's

getting pretty amazing. And that, that includes things like temperature and soil moisture and others. I mean, we're measuring soil moisture down to a kilometer. Possibly even down to 30 meters, essentially scale of typical land classification. So, I think that, in my mind, it's the context of the data that we're collecting at the point scale that we would be able to apply at a broader scale.

COLIN:
And it's not necessarily to say we throw everything and put the kitchen sink at this, but in a sense, that's what it comes down to. I think the challenge is how to blend data from the point scale with the satellite remote sensing, all collected at different timescales. How we go about doing that? And then using the results to answer questions outside of just ag, not at the exclusion of ag, but there's a lot of other places where the information is going to be used. So, I actually think we're, we're doing a pretty darn good job now at collecting most of the data we need. Of course, I'm biased. I'm not a microbiologist. So, I don't really know understand the genomic part of it. But sure, it seems to me that we're collecting a lot of data. The problem is we're not utilizing the data we're collecting very well.

CHUCK RICE:
That's a good point. You know, for the steering committee, as we went around the world and listened to all these data, they saw information that works. One thing that surprised me, is Mike, you brought out is that I didn't believe. I was surprised that they weren't using remote sensing to help integrate and collect other information far as land use. And that wasn't well integrated into the soil information that seemed to be a missing opportunity. That seems to be global in the (UNKNOWN). I think Luca maybe was doing a little bit, but not pointing out any particular information system, but that seemed to be kind of a general observation.

CHUCK RICE:
I guess, you know, the question is Michael is that, you know, I was at a workshop and a department of defense person asked, what do you want to measure? How often? And at what cost? And my glib answer was, I wanna measure everything every second at Pennies. And, you know, but it comes down to them priorities. What's the key information that's most useful in that?

CHUCK RICE:
So, we do have a comment from Slack. Was if skull the workshop is to identify the needs of our dynamics or information broadly, would, it seems like such a system would need to be built from the beginning to accept new and different variables and parameters rather than to prescribe from the beginning, what measurements are allowed? Especially, if the goal is to accept data contributions from our variety of studies, with different goals and objectives. Anybody want to comment on that? I guess my comment would be, we need to build on what we have rather than start a new, we can't throw everything out. And think about that. Julie, your next on line, maybe its Dave, then I'll come back to you, Julie. Sorry.

JIM:
Well, I was gonna defer to Julie, but thank you, Chuck. So, but I think the comments are well-taken so far. I liked what Luca said about thinking about what we measure should relate directly to a function. We also need to consider, is it at the temporal scale and the spatial scale. Example on, chemical

properties on the, on the spatial scale can change dramatically in urban areas. So therefore, should we be measuring lead content in urban areas when, you know, right next to a house that has lead paint, you can have extremely high lead contents, but you go to the yard and it's low. What is that actually gonna mean? So, as we measure, consider why it's being measured? How it's going to be used? And the large continental scale databases have a very different purpose than say, measuring the biodiversity across a field or across an acre.

JIM:
So, we really do need to consider the question, the function. And also, I think we need to consider what is politically cool or scientifically cool at any given time can change. Right now, biodiversity is a great thing. We're going to start looking at it. But will it always be, can we actually use that information? I think we can, but if we can't, we have to be able to admit that, you know, we don't need to measure this anymore, or we've measured enough of this. Let's go to something else. So, there are some parameters that I think we have to do. I think most of the systems look at those parameters for the lot of the folks that are doing the digital mapping. Those 12 or 20 parameters are really key to probably understanding 80, 90% of what goes on. Can we do better? Sure. Maybe on a smaller scale that is both spatial and temporal. And so, let's consider what it is that we're trying to actually use the data for.

And I'd be happy to hear people talk about how we could take more measurements given the time the people and the money that we have. And that may be part of another one of the breakout sessions to talk about integrating these systems. So, thanks.

CHUCK RICE:
OK. Thanks, Julie.

JULIE JASTROW:
Yeah, that's actually a really good lead into what I was thinking about. And that is because we're so dependent on the question in terms of what things we want to measure in so many ways, in terms of temporal and spatial and everything. Could we, if we really want to think big and think about developing computational capabilities, storage capabilities, could we develop a hierarchical system. Where there's something that's really basic. That's measured and it's available and it's geospatially referenced. And so, as researchers or people that are interested in particular things, they could put their data into that larger database. So, it may not have huge coverage everywhere. You have a set of measurements, bill of NRCS that is available to everyone in some sort of easily accessible way that they can use for, let's say their research, but then they could add their data to that, to those points.

JULIE JASTROW:
And then that could be used by other people in not just a database that you can go and grab and create that for yourself. But that it's readily accessible to, to everybody, particularly with the agencies, you know, wanting the data that's produced to be put into databases and things like that. So, if you could somehow reference all of that information, even if it's spotty on this larger system, maybe that would be really good. And then as a separate thing, just talking about function you know, there's this whole idea of carbon sequestration, of course, you know, soil carbon is not forever. It's not stable forever. And there's the trade-off between soil having soil carbon sequestration versus having a dynamic system that

is cycling and providing energy for the microbial community, cycling nutrients and all that kind of thing. So, it seems to me some simple measure of soil organic matter composition that can, would be useful whether that's some kind of simple fractionation or infrared spectroscopy with that you could always predict many other potentially.

So, you know, something like that I think would be very useful to get at sort of the functional side of understanding the health of the system.

CHUCK RICE:
I've got too many screens open here, I think. Well, yeah. Christopher and will hear, comments.

KRISTOFER COVEY:
Yeah, thanks. I would say Julie has... (BACKGROUND CHATTER) Yeah, Julie, I think this idea of a database that people would not only get information from, but contribute to, obviously very compelling. The challenge that comes to mind is how do you filter that data for quality, obviously, or make sure that there's some uniformity in the measurements being taken, what is allowed to be put into it and how do you filter that. I'm kind of thinking about I was really... Luka. your point yesterday about the costs and needing to justify every new thing that gets added. I think is a really important one and narrowing down not only what is it that we really need to get, but also what things can we get more cheaply than we're currently getting them.

KRISTOFER COVEY:
So, I think some things like biodiversity that require the samples to be stored in a certain way, maybe more expensive than... Basically what samples can we get by just anyone taking a core sample or anyone who happens to be in a place grabbing a sample for you. And those sets of things I think we can make much more cheaply through time, so long as the information that we're providing then back to users is actually valuable.

>:
And so, we've run a couple of field campaigns, and consistently what we find is that the expensive parts of those campaigns are the parts of the planning and sort of providing service to the user that are manual and static. So, somebody's laying out plot points, can we ought to automate where we would like samples to be to evaluate a given service. And then returning a valuable product back to a user, that's expensive, making the maps or giving them reports that requires fancy people in front of computer stations.

>:
And then, the actual getting of the dirt, the expensive part, as we've heard pointed out several times, it's all in moving a person around like pieces on a field out to that random location you've chosen in the middle of nowhere in Nebraska and asking them to spend two minutes grabbing dirt from a bag. Well, someone's already there, and so, can we think about measurements which can be taken in a distributed fashion with some kind of either very low cost, easy to use, open source kind of equipment or tools or they can grab that actual physical sample, and then part of the program is in managing the logistics of getting that physical sample to a lab. And obviously, once you get to which lab and the whole you end

up with this, again, something I was really impressed with Luka yesterday, mentioning the fact that you all have gone to a single lab because of these cross lab comparison problems that obviously Will (UNKNOWN) is well familiar with and Bruno pointed out the other day too.

>:
So, sea of consciousness, random thoughts, but somehow can we list a set of things that anyone could throw dirt in a bag and get it to a lab? What are those things, because those things are primed for a really low cost distributed system. And then I'll stop.

CHUCK RICE:
OK. I guess I would also challenge you to think, do we always have to take a soil sample? I mean, I've been involved in a couple of workshops now with sensor development. So, you throw a microchip in the ground, there's some challenges there, but they can measure potentially temperature, moisture, but also nitrate, maybe even CO2 concentration, things like that. Are there things that would help reduce the cost of monitoring rather than just that person physically going out and taking the sample? I just throw it out there as a challenge of thought, don't constrain ourselves. We'll go a little bit more on this and I wanna go talk about standards and reference. Comparisons and things like that. Michael.

MICHAEL YOUNG:
Well, I don't want to unnecessarily jump the gun. I mean, I think that Julie is correct. What we need to do is think about the data sets as being number one, that they had to follow a certain standards, including metadata, so that other people are able to use the data. USGS has been using... They have little quality indicators, level one, level two, level three. For example, for the LiDAR program across the country, they have multiple levels of accuracy. And if you want to use the data, it's a level two accuracy of your LiDAR data for your digital elevation model. There are ways to do this. I think the key is that because the community here is so diverse, we will never have one schema that is going to satisfy all people.

MICHAEL YOUNG:
I think what we need to be thinking about is a federated data set that allows other people to pull the data into the use that they have, that the data is discoverable, that is interoperable. I mean, this is the typical fare standards, but the discoverability is really key and that it be made available. Reproducible research has made a big difference in this. In Texas, we have the Texas data repository, which is a free, persistent repository for any of the data that we're using. I'm sure that your states have that, your universities and others have it. Data is real cheap these days, storage is very cheap. So, I really think that it ultimately comes down to the cyber infrastructure that we can use to pull the data in when we need it to solve the question that we have, because we'll never be able to come up with a schema today that's going to address the kinds of data we can collect ten years from now, and it would be a mistake for us to try to do that.

>:
We really need to focus on the format of the files and the type of metadata we're collecting so that the data can be reused in the future and it doesn't just sort of fall off.

CHUCK RICE:
Yeah, good point. And I think that's also one of the other breakouts, is the kind of data storage in that. I think you made a key point, kind of federated repository, because different end users are gonna have different measurements and customer needs are gonna be different at both temporal and spatial scales. Aileen, you want to have a comment?

EILEEN KLADIVKO:
Yes, this might be your next question, but I really want to get it in. For somebody who works in soil health and agricultural systems, one of the major impediments to being able to really learn from other people's data is that the metadata of some types is nonexistent. What has that farmer done over the last five years regarding tillage, regarding agricultural chemicals, regarding cover crops or no cover crops?

EILEEN KLADIVKO:
And so, when we get data and don't have that... And that's not sending somebody out to get dirt in a bag as one of our commenters just commented. There's a lot of information that we need in order to really be able to interpret it. Soil health improving in an agricultural setting, even in that topsoil, that we don't often get. And it's hard to get. Even if the farmer is cooperative, it's hard to get. So, to me, that's a major issue that needs to be dealt with if we're talking about dynamic measurements that change as a result of farmer management.

CHUCK RICE:
It's a really, really good point. I think you and I've talked about going through met analysis of previous publications. People aren't publishing just what the soil type was or texture or Ph in a lot of their journals now, and it makes it hard to interpret. So, it's all that ancillary data that's really, really key on that. I'll go to Corey, I believe.

COREY LAWRENCE:
Hi, can you hear me OK?

CHUCK RICE:
Yes, apologies, I'm in the lab today. Sometimes the soil data collection can't be halted. I just want to mention one point that I think is maybe useful to this discussion. Having been pretty intimately involved in the ISRAaD database that Alison talked about yesterday, I think at the bare minimum, we should all consider measuring things that allow for tie ins to databases in the future. So, I know we all like to think about our questions and the questions we're asking with whatever funding we have, but if we're collecting samples that we might archive, we should also consider the things that we can't remeasure or measure in the future on those archive samples. When we built is ISRAaD, we attempted to pull in a lot of different data sets, but what we quickly discovered was if we didn't have a coordinate of (UNKNOWN) and long for a sample, we couldn't use that information because a lot of our analyses were based on pulling other information from geospatial maps.

CHUCK RICE:
So, other things that we might not be able to reconstruct from an archive sample, perhaps (UNKNOWN)

density is another one. But I think at the bare minimum, we should look beyond our own purposes, especially for archiving samples, so that in the future the samples we collect can be reused for other analyses and they have the bare minimum information required to do that. That's all. Thanks.

CHUCK RICE:
Excellent point site. So, I hate to do this. Well, I'm gonna ask the group... We got a lot of information, physical and chemical measures, the history in that, biological is one of the things and I don't wanna get down in the weeds, but we talked about for biological, biodiversity, genomic data. I guess what one or two other kinds of information that would quantify biological function? And then we'll go into some of the other questions. We've got all these biologists (CHUCKLE). I'll start... Oh, Jim, go ahead.

JIM:
OK, actually, I see that Gupta is on this session and he's done a lot of (UNKNOWN) relative to biological activity in Australia, and Australia has a lot of experience in (UNKNOWN). So, I wonder if he would comment on your question.

CHUCK RICE:
OK. Gupta? Is it on?

GUPTA:
Yes, I think. (BACKGROUND CHATTER) Functional base biological measurements with diversity measurements are two different aspects of it. With experience, we tried in Australian environments for function based measurements, not just measuring functions, but associated metadata, even with the current knowledge of the drivers to the functions to occur, even with the biology happens to be there is essential for any datasets with values for that. And even simple measurements for microbial optimal days in order to extend spot measurements to crop season best. Many data sets biology measurements are more of a spot based measurements, extending the functionality of that biological property in a measure relevant to crop performance, I'm using crop as an example, is the difficult part. And metadata that can extend the collected data for such short term measurements is essential in order to use that for advising and use it or whatever.

GUPTA:
I mean, (UNKNOWN) the exact question, but the value of biological measurements is only there when it can be extended to the end user. Physical chemical properties, at least physical you can get away with some extent, but biological measurements, because they're so dynamic, or a crop season, the metadata is essential to extend that.

CHUCK RICE:
Well, OK, good. So, whatever the measurements that we've been working with, just to help stimulate the discussion a little bit, but it's phospholipids because it's a measure of composition, maybe not the diversity level as genomic data, but then, if I can determine fungal populations, we've seen a good correlation between fungal composition and aggregate stability. And it seems to be responsive to land management, like less soil disturbance, even crop type. And it seems to relate to the function of physical and chemical as well. So, I'll just throw that out. I worry a little bit about CO2 respiration because it's

pretty dynamic and trying to measure it, again, today and tomorrow. If it rains tonight I'm gonna get a different respiration measurement, and just because is high respiration good or bad. It's high respiration, I mean, maybe are you losing carbon? But yeah. So, just a couple of measurements that's been in debate. I lost track of... Corey did you have a comment or was that Jonathan? Jonathan, go ahead.

JONATHAN:
I think (UNKNOWN) Cory wants to go if he can. But I can say quickly... First, hi Gupta, it's been a while. I was thinking actually, especially since Australia is on the line, it's maybe slightly relevant for like American heartland agriculture, but subsurface constraints to plant growth we haven't really talked about moving back maybe to physical chemical properties, but those are incredibly important in small regions of the United States, but more so globally. And so, making sure that we're capturing those, that we're not just focusing on the top 10 or 20cm of soil in our minds when we're thinking about what types of soil information we want.

CHUCK RICE:
Yeah, that's a good point. We've got about 20 minutes less, so maybe we can get away from the measurements. Corey, did you still have your hand up or...

COREY LAWRENCE:
No, sorry. It's up, but I didn't intend it that way.

CHUCK RICE:
Alright, um. Melanie, you have a quick comment?

MELANIE MAYES:
Yeah, I just wanted to address your point about the the biology. I think one of the issues of this is that technologies are running (UNKNOWN) microbiology are still advancing constantly. So, it's hard to pick a defined measurement, and frankly, we don't know what most of them mean yet, but having said that, something as simple as a chloroform fumigation for microbial biomass is a very standard measurement, and honestly, once you do... And I guess one other thing I would say is like a (UNKNOWN) PCR for just fungi, bacteria and archaea. These are both really straightforward things that are not likely to change.

CHUCK RICE:
Good point. (UNKNOWN) chloroform fumigation is you get a carbon measurement, so it's a fraction of that. (UNKNOWN)

MELANIE MAYES:
Yeah, and I would say that these things are being used in, say, earth system models.

CHUCK RICE:
Yeah, good point.

MELANIE MAYES:

And process model data. We don't know how to use metagenomic data.

SPEAKER:
You could do enzymes. Enzymes are actually not a bad approach either. Just the standard hydrolytic enzymes are also a good choice.

SPEAKER:
Yeah, I agree. We're doing that, and that seems - like glucosidase seems to be pretty good. Dave (CHUCKLES) - and then I want to move on to the next question, as far as I think we talk a little bit about contextualizing it, but drainage, I think Michael talked about elevation, (INAUDIBLE), some things like that are really, really important. Go ahead, Dave.

SPEAKER:
So, I wanted to help you move to the next line, Chuck. So, somebody had pointed out the critical part of any of the measurements is the metadata. Where is that sample taken from? What is the history of that land use? That can tell us a lot, from any one of those physical, biological components.

SPEAKER:
And part of that metadata are the standards that we use to - this is the tie-in, Chuck - are the standards that we use to collect that metadata. We looked at and reviewed a lot of literature related to tillage, terminology changes over time, the actual practices, change in timing - so, a way to not only standardize the analyses, but also standardize the metadata. What is it that we actually need in order to make that point data more useful to everybody?

SPEAKER:
Let's gets into the standardization of methods and that. I guess the question is do we need to have standards or do we need to have - well, it'd be nice to have standards, but are they comparable? Can we make comparisons with different techniques? I'll throw that as a question.

SPEAKER:
But the other thing, I guess, that came out yesterday - and I forgot who brought it up in one of the keynote talks, I believe, but then I think Jim (UNKNOWN) mentioned, as well - we have referenced soils or standard soils, but should we have reference sites? So, as we have new techniques, then we can go back to that... soil, climate, region, whatever, to allow for future development and comparison of new techniques.

>:
Melanie, are you trying to talk?

SPEAKER:
I was going to say it was Phil Robertson yesterday that mentioned it as sort of a third set. One is data, two is measuring a soil sample. But then, how do you put that together? What really happens in the field? Well, then this point of reference sites brought up is sort of, to me a third level of information that helps with the overall interpretation.

SPEAKER:
Who's going to fund that? (CHUCKLES)

SPEAKER:
Funding is always going to be a big part of any kind of database effort.

SPEAKER:
But I would argue if we had a reference site, we got a whole network of LTRs, ARS, university, land-grant universities. If we had one acre of land that we just kept at Purdue - Eileen? - kept just (LAUGHS) for a reference, it'd be a huge asset to the community. Not just soils, but climate change. I think that would be a wonderful asset.

SPEAKER:
So, I think that's actually an important value of the NEON sites. They're not agricultural sites, but they are in different soil and climatic regions and have extensive data over time. So, the idea would be to extend that information to the agronomic parts or forestry parts or other land-use parts of that region - so, to use those kinds of existing things as a base.

SPEAKER:
Any other comments on that?

SPEAKER:
I would just say the idea of these kind of sentinel sites is critical for agroecosystem modeling, which - there's a lot of interest in emerging carbon markets. The models are so data-limited now, and building up this network of high-quality data streams, to really be able to improve our ability to forecast - and these models, there's been multiple papers that have talked about the need for some sort of system like that.

SPEAKER:
Dave, you want to respond?

SPEAKER:
I think it's a great idea. I think it would be wonderful if somebody with the National Cooperative Soil Survey would put an NSF grant together, to try to identify those sites at various locations across the country. Just a subtle hint, Chuck? Very subtle. But because I know it - I mean, I was at a university - I know that there are areas that folks know very clearly what has been done from management for years, and those sites are available. So, I think that would be great, tying it in to other sites - LTR, NEON, we could go on and on.

SPEAKER:
But the one that would be really good to use - and don't laugh too hard - are cemeteries. Cemeteries provide a phenomenal resource. Just be careful where you dig, but a lot of them have not been disturbed for a long time. And they can really help us out, but it would take somebody to organize it.

Chuck, Jim, Eileen - I'll just pick on a few of you that I can see right here. So, thanks.

SPEAKER:
Yeah, well, and cemeteries are great, but even just having a managed site for long-term, like I said, just for baseline, it would be really - and every land-grant university gets... whatever, the Hatch Act, the funding. So, maybe it's - what would it take to manage an agro land? Of course, administrators would love me to use part of their money. Michael Cosh.

SPEAKER:
As far as LTAR goes, there's a soils group there, and specifically there's also a soil biology group. The key there is they all have research questions that need to be addressed. So, if a clearly defined research question could be described and procedures referenced, that network could be engaged. But one of the critical parts of that is at each of the 18-plus sites, you really need to have someone who knows what they're doing to do that collection and to care about the result. And I've had lots of hiring someone to do something or asking someone to do something, and if they're not invested in it, the collection can go awry very easily.

SPEAKER:
So, there's not a soil scientist at every LTAR site, quite frankly. There's maybe a hydrologist or someone. So, engagement across such a large - it's very valuable, everybody understands that. But if you really - I don't know if there's any NEON folks here - if you start to engage with the NEON group, you'll find that it's the locals that really have to do the work. And if it's not their niche, their interests, they're going to do what's written down, and that's it. So, it is a bit of a challenge. Siting this type of collection has to really go with someone with passion about it. And everybody's got a short career in terms of soil - length, right? So, it is a challenge. I think land-grants with good soils groups may be a better option in some cases.

SPEAKER:
OK. For the sake of time, I've got ten minutes left. There is this question at the bottom. Do we have enough physical sample archives? We visited with Rothamsted that has - oh my gosh, what, 180 years' worth or whatever, I forgot what it is now, of archiving samples in Mason jars up on shelves. It's a fantastic resource. I've got a 33-year experiment here at K-State and I've been archiving samples about every two or three years that we sample. We archive that. But I guess the question is do we know what's our... protocol, robustness for archiving samples. Of course, the biological data, you know, you need -80 degree freezers. So, right now we're just doing dry-soil sample. Do we have enough sampling? Or archiving, sorry.

SPEAKER:
And we got just seven minutes before we have a break and they're going to move us. Jim, you got your thumb up. Is that -

SPEAKER:
Yes. So, I think archiving is extremely important. And if you look at other fields of science, archiving samples have been really important. And so, Chuck, as you know, or (UNKNOWN) from the academy is

collecting information on archives, samples that exist from different locations, with the idea of them potentially being more available to people. So, I think archiving is important. I think the question is always about how you store those samples.

SPEAKER:
But I would also say that we met also with the Chinese Academy of Sciences and that Institute of Soil Science has a huge archive on, I think, the 30 different soil types of China, collected over the last 30 years. So, there's some good examples around of how this is being done.

SPEAKER:
My challenge is telling administrators that we need space for Mason jars, (CHUCKES) storage. Richard.

SPEAKER:
Ah, yes. So, as Rothamsted was said, it was name-checked, I thought I'd contribute. So, I work at Rothamsted on the long-term experiments. And yes, the sample archive is a real challenge. So, we have nearly 180 years' worth of soil, grain and herbage samples. In total it's over 300,000 physical samples. It's all dry samples. We do have wet samples, as well, but obviously, as you mentioned, keeping wet samples is a much bigger challenge.

SPEAKER:
One of the real challenges we have, though, is actually maintaining all the records for the sample archive. That's a real challenge in terms of what kind of software databasing you use, and also keeping track of the data that's generated from the sample, so that the sample archive is available to researchers globally to come and use. And we have many people each year come and use samples. But the data that's generated from those samples, actually there's no formal process for tracking what happens to that data. So, there's no kind of accession associated with a sample and data that's generated in another lab from that.

>:
So, it's becoming more of an issue, I think, because the samples that we have got are obviously finite. You can't go back to 1850 and resample the soil. So, they are a very valuable resource and the data that's generated from them is also an equally valuable resource.

SPEAKER:
Eileen, I think you've got the final comment before we get kicked out.

SPEAKER:
OK, great.

SPEAKER:
Words of wisdom. (LAUGHS)

SPEAKER:
OK, (CHUCKLES) so, anybody who's getting closer to retiring. I've got some samples that are archived

well and some that are not. But we always get asked, who might ever use this? So, actually a little bit of guidance about those of us who do agricultural experiments, where we have three or four different treatments, and maybe we sample every other year for a while. We don't want to archive every soil sample that we've ever taken. So, even, how do you go about thinking about what's worth archiving? What somebody might possibly use in the future I think is a valid consideration, because nobody's going to look at all the soil samples that I still have in my lab that they haven't made me kick out yet, right? (CHUCKLES)

SPEAKER:
That's yeah, some priority. OK, any - we do have two or three minutes, any last minute comments, summaries of what you heard or what we didn't discuss? Jonathan.

SPEAKER:
I was just gonna say quickly, Dave Lindbo, if I talk to the soil survey archive in Lincoln, Nebraska, it's really well curated and really extensive. And I'm not sure it meets the needs of the soil survey. I don't know if it meets the needs of this kind of dynamic soil information service we're talking about here. But it is, in terms of federal taxpayer-funded archive, there's obviously been a lot of investment in it and it seems to potentially be the starting place to build off of.

SPEAKER:
Jonathan, thanks for that. Yeah, we have - I can't remember - 200, 300,000 samples in pint containers as well as several 55-gallon drums of standard soil. Folks ask for stuff. We can find it. We'll ship it out. But when it's gone, it's gone. And that's part of the problem with an archive. If you constantly dip back into it - and I think we heard that from Rothamsted - when you dip back into it, eventually it disappears. So, that's - again, I think Eileen mentioned it. What are you gonna use them for? Have a good idea. Do you just want to save certain benchmarks? Do you want to save everything or something in between? And that's a critical question that maybe for at least a mind that is way sharper than mine to answer.

SPEAKER:
OK. Jim, I'm going to put you on the spot. Do you have any final summary? And I got a couple of points I noticed but - well, think about it, then. I guess what I heard was that, you know, we've got a lot of different measurements. I think there's an opportunity for biological to improve, but there's a lot of diversity there - in measurements, not in biodiversity. I guess we need to think about functionality. And that's where multiple groups, user groups - if it's an urban area and worried about contamination, that might be separate from ag production or forest productivity. So, we have to think about, again, different user groups. Time and spatial scales are really critical. And again, that goes back to functionality. If you're measuring or monitoring change, erosion is going to be different than, say, water quality or something else like that? Jim, any other observations?

SPEAKER:
I guess maybe not.

SPEAKER:
Alright. OK, so I guess the procedure now is we're at a break, 15 minutes. You will be moved to another

breakout room you've been assigned. We'll stay here. And then the challenge will be this afternoon we'll be summarizing this and presenting some sort of a summary tomorrow in the morning slash afternoon session. Alright, thanks for your contributions. If you've got any other notes or observations, you send me an email. So, thanks, everyone.

MODERATOR:
You need to bring all the stakeholders together. But let's discuss a little bit of that as well. What is the training gap in just in soil science? And how do you – this is again an interdisciplinary aspect – how do you bring, how do you enable farmers to be trained just so that you can take they can use the data? How do you get academic soil scientists to be trained in computer science aspects? How do you bring all of this together? And the last one is how should this data be stored for…? Data as in we're not talking of the soil archives, but that's, I think, discussed in another breakout session. But here, it could be more around how should this data be stored in the cloud for queries right now or queries in the future for things that might be that additional data that you might get in the future? So, those were some of the key things we are discussing here. So Rodrigo, do you want to add anything to this?

And I would also encourage any participants if you want to, if you want us to be discussing anything else as part of the breakout session. But Rodrigo, I'll let you go first.

RODRIGO:
Yeah, thank you. Well, let me start with the concept of FAIR and at least in my opinion, what are some challenges that we will face here. I would like to share my point of view from the research approach, and also from the editorial approach in the in the peer review process. So, what I have seen is that I feel that the FAIR framework and data sharing within soil science, it's behind many other scientific disciplines. And this may be a cultural issue. And the fact is that, we as a community, we're not used to share that information, yeah? It's like, "Here's my graph." And that's it. But really, the concept of FAIR is very interesting because it's about reproducibility of science. It's not just that the figure is there. And we will talk later about the how can we store this information? But I think that we start from a challenge of a cultural barrier that we're not used to share this information.

RODRIGO:
And in my opinion, and what I have seen with colleagues across the United States and across the world, it is about of a little bit of lack of trust, first of all, that where my data will be how my data will be used. In other contexts and cultural context, sharing your data, putting out data there, you lose power, because data gives you some power of what you have, and not being that open also gives you that maintains that power of information. And also from an economic perspective, publishing your data has not been part of the reward system. And all those things, in my opinion, has been a cultural challenge of why we're not sharing the data. That said, some publications now are asking to put your data in FAIR repositories. So, for example, the American Geophysical Union has a requirement that if you publish in a journal, data has to follow FAIR principles. Therefore, I believe that with these incentives, at least from the publishing aspect of data, we will start breaking those cultural barriers, and hopefully it will be more enable to have a FAIR framework in soil science.

MODERATOR:
Thanks, Rodrigo. So, any other discussions from the party? I would love for thoughts from people how do you encourage people to share data, in this case soil data, or what you might think might be

discouraging people from sharing data as well?

NICK:
I'll jump in. I've seen from the farmer standpoint, historically, different arrangements around safe harbor periods. I know this came up in the fireside chat. I appreciate comments around, especially if you're a farmer, privacy and use agreements, and but the potential for data 10, 15 years down the road to come back and potentially harm individual producers. So, I've seen in Minnesota and other states safe harbor periods. So, if the farmer contributes data to a common good to provide a safe harbor. So, I'd love to maybe explore that here as well. What does that look like? Could it be, not only in a state level, but also at the federal level recognizing differences there and then also from an international level, are there further international conversations around regulatory safe harbor for different agreements that were signed into as part of the United States or otherwise where this framework could be shared?

MODERATOR:
That's a great point, Nick.

MICHELLE:
I'm just curious, are we adding these topics like do we now have a five and a six or we wanted to free for all on these topics?

MODERATOR:
Free for all to share, yeah. Yeah, no, so we wanted to just start with the first question that is how do we incentivize people to share? How do we even provide an infrastructure to share so that we are following the FAIR principles with data, making it more findable, more reproducible, more interoperable, accessible? So, how do we get there, and what are the challenges that we have right now?

MICHELLE:
I think that the tri societies have been working on this to some degree. And I think that in a community, there's a lot of variability in sophistication and where people are. So, I think Rodrigo's comments are really germane to older folks, and I think younger folks are much more familiar with and open to sharing and some of what we heard yesterday about the bottom of publications, and those things getting cited, and people figuring out how that, you know, being put at ease that they will be rewarded for it. So, I do think it's changing, and I think the tri society's sort of intentionality, and these emerging ontologies, and people becoming much more familiar and comfortable.

MICHELLE:
And that kind of connects up, I think, with your later or maybe what we'll get into with sort of the methodologies that are available, and helping people do that workflow and track data going beyond, "Here's my composition book, right?" I mean that's we can do so much more now. So, I think younger folks are gonna be an easy sell. I think some older folks and getting exhausted with changing tape backup to disk to this to that, you know? I think we've had so much innovation that people even trying to do a good job in my generation it's been exhausting, right? But I think that that sort of... Anyway, that's enough said.

MODERATOR:
That's a great point, Michelle.

JIM JONES:
Yeah, this is Jim Jones, this is great discussions, and I liked the earlier panel comments as well as this topic. I thought it might be useful to give a few of my experiences just in this same area, because we're working in modeling crop and soil modeling that I've done most of my career, we have thought and we've tried for a long time to come up with one standard way of doing things. And one thing that I've learned in all these efforts is that, that's probably not going to work. It's hard for everyone to because everyone already has a lot invested in their own systems and so forth.

JIM JONES:
And the second thing I wanted to point out is that harmonization can be done. And there are a number of different groups around the world that are working on this. And AgMIP, for example, we developed a harmonization approach that would recognize different, more localized standards and perform the translation, have these data tables that relate one to another and do this kind of thing. And so that's being done, it's being there's some work in USDA, and then the CGIR for trying to get to better FAIR principles and having data shared across centers or even with across researchers within the same international research center. So, there are these things being done.

JIM JONES:
And then I don't know, I don't have all the ones that are being done right now. Maybe Cheryl Porter could chime in on this. But the other thing that's apparent to me is that these need to be consistent with, even though, with industry, because even though industry may have their own ways of storing things and protection of data for the reasons that have been explained, at least there needs to be some way to go between those industry standards or a particular industry way of doing things, and then the applications in science. Modeling and data analytics, and GIS remote sensing all of these things are going to need access to this to really make and take advantage of the science as it's being developed and even to develop the science or taking a broader look at some of these soil characteristics.

JIM JONES:
And I would include in that, I asked a question earlier, I'm not sure that was the right place to do it, but about soil biology. And I don't know, you know, everyone talks about swabbing a living system, but I don't know of any really good way of representing that life or soil health that's there. So, there those kinds of things that need to be done as well to come up with standards. But I think doing this in a way that involves the industry and USBA, and maybe some international groups in some kind of consortium or something might be a good way to help, at least make sure that the different standards that are being used are identified, and there's ways to harmonize among those. And I'm afraid I'm going to have to leave in about 10 minutes. And so I didn't want to say that though, because I think that this is a really important area of effort. Thank you.

MODERATOR:
Thank you, Jim. Thanks. So, yeah, go with that (INAUDIBLE). So, Cheryl?

CHERYL:
Yeah, thanks. And, Jim, thanks for your comments because you stated a lot of what I've had on my mind too. And related to Michelle's comment about how the culture is changing, younger people are more receptive to sharing their data. I think that's absolutely true. But I've also had experience with research, people who have been collecting data over their research career, who are nearing retirement and are anxious to have their data out there and available so it doesn't get lost. It's like their whole research career is at stake here. So, it goes both ways. I think some of the older people are coming around to that as well.

CHERYL:
So, I've been involved with some of the data interoperability efforts with AgMIP. And AgMIP is the Agricultural Model Intercomparison and Improvement Project. And just some of my thoughts, things I've learned along the way is making data FAIR, the F and the A are easy. Making data findable and accessible, there's hundreds of repositories that you can make your data findable and accessible and open. The interoperability and the reuse part is hard. It's a lot harder. And there's sort of three tiers of making your data interoperable. If you can tag it with standardized vocabularies as it's being captured, right from the start, it's a no brainer. I mean it's simple then. Then people can understand your data. But that's not happening yet in agriculture. So, we have a ways to go there. So, that's one branch that we really need to work on.

CHERYL:
And then the other two are having somebody who collected the data, tagged it with a known vocabulary, so somebody who's very familiar with the data explain what all the terms are, and provide the vocabulary. And hopefully, that's some kind of standardized vocabulary. The third is where we most often find ourselves, and that's where an end user of the data finds a dataset and wants to use it for their purposes, and they have no idea what all these terms mean. And they haven't been well defined, and they aren't tagged with known ontologies or vocabularies. So, that's kind of the area that we've been working in AGMAP is methods of annotating legacy dataset so that they are interoperable with datasets may be sitting behind a DUI, they're out on some repository somewhere and you can't really modify the data, but you can annotate it such that it can be tagged with standard vocabularies. So, anyway, I've been working a lot in that areas and I'm very interested in this is more agronomic research data, but soils is certainly a big component of that.

MODERATOR:
And a great point, Cheryl. And I think this is critical for us to not just have the data but make the data usable. And as part of that it's critical to have some of these data's being able to tag it, being having the right metadata associated with the data. So, that's gonna be key. So, Phil?

PHIL:
Yeah, I just have two comments maybe on the first in the last third question. So, I think in my experience, the private sector is the worst one when it comes to data sharing, but they happily take the data, but they never share. And I think it's worth noting, and I think that's a problem, too that definitely needs to be addressed if we talk about data. It's not the public sector that collects the data and the private sector that uses it and collects its own data but doesn't share it. And we know the reasons for

that, but I think it's worth noting.

PHIL:
The other bit is when it comes to training, I think we don't train data scientists, but all our soil science collect data in their degrees. So, they collect data, the data is in a range of formats often, and we'll have those data. If we have to put it in a repository according to standardization, we need help with that. We have a lot of data. Most of us that work with students and worked for a long time in a place, as Michelle could confirm here, we have a lot of data for students and others that is just sitting there either in boxes, either samples or virtual data, whatever it is, that we could share easily if there was a repository. I think the physicists have been ahead of us for 50 years. So, there's maybe something to learn from other disciplines.

MODERATOR:
A good point, I'm afraid, even on the computer science side, I think we're better at... So, I'm a computer scientist by training and also and we sharing data is, of course, it's also a challenge in computer science. The more of the issues that revolve around privacy, but at least it's further ahead than soil science. But of course, there are other challenges too with soil. Of course, it's a mix of public data and private data. It's just so hard to collect some of this data, making this data very sparse, which adds to some of the challenges as well. So, Phil, you have your hand up?

PHIL:
Yeah, thanks, Ranveer. I think to some extent, I mean, this has been a great discussion. I've loved hearing some of the points being made. But I have this thinking, perhaps not thinking, but I have this feeling of the cart's a little bit before the horse here in that we haven't talked about how we get data into the databases. And this question seems to fall between the cracks of the breakout rooms between A and B. And, to me, having data in the databases is prerequisite to thinking about how they're curated and made accessible. And this requires, of course, buy in from different contributors. And I think there may be three major, I can identify three major contributors based on discussions yesterday and the talks that we've heard. And one of those, of course, is academic contributors, another is government agencies, and then the third is private. And each of these have, and there may be others, but each of these groups have different motivations and I think we'll need to be perhaps incentivized differently.

Many of us are self motivated and sometimes by age, as Cheryl noted and Michelle inferred, and so there's not gonna be a lot of motivation required, but others are gonna require incentives. And sometimes it's gonna be a stick rather than a carrot as, for example, publication requirements, as we've seen in many of our disciplines.

SPEAKER:
Many communities have grappled with this, as I think Alfred was making reference to. The genomics community does this on a regular basis. They went through a maturation stage where it became required that all genomics data be put on one of the major databases and made available, oftentimes immediately upon sequencing, even though the person ordering the sequencing, and doesn't, won't necessarily have access to it first. The other communities have done similar to the Ameriflex and FlexNet community does this with (INAUDIBLE), any covariance data, and so on. So we can learn a lot from them.

But I think that one of our…I would like to pose that one of our recommendations be that we think about some of the, bolstering the incentives and motivation for getting data into the databases and then we could think about harmonizing as Jim pointed out.

SPEAKER:
Thanks, Will. And Catherine?

SPEAKER:
So, I'll pick up on that thread of focusing on the collection and curation aspect, and also pick up on the breadth of asks related to the data 'holy grail,' as it was discussed during the fireside chat. And ultimately, I think, if we want to capture the inherent variability that's found in nature and nuanced understanding of management practices across space, across time, and at scales ranging from the soil profile to the field level, to landscape, to capture everything that we're hoping for, I have to echo Nick's point from earlier on that we really have to have farmers on board. I don't want to understate the value of the private sector and researchers and the public sector, but at the core, at least when it comes to agricultural soils, we have to pay attention to farm operators and ensure that the appropriate incentives are in it for them.

SPEAKER:
We can't forget that digitization of agricultural records is new and it's not always easy or accessible to growers. And they're constantly being asked for data these days, whether it's by agronomists or USDA or crop insurance, and we need to ease that data burden for them while demonstrating that there's a benefit. So for example, is there a way that we can ensure that once we align on standards and interoperability, that we can relay this data back to the growers in a format that's comprehensible for them? Can we make it easy for them to pump this information back into the crop insurance system so that they don't have to experience the same burden that they do on an annual basis?

>:
Of course, there are those additional financial incentives that might come from things like carbon markets, but I also feel like on more of a basic level, there's a way that we can better incentivize and empower growers with their own data and we have to do that if we want the data at the magnitude that was described during the fireside chat.

SPEAKER:
Yeah, I know that's a good point, too, Catherine, about how do we...like what are the...and farmers, I agree, are one of the key stakeholders and I would add that to Will's list. I think academic, government, private sector, unless when you are putting farmers as private sector also, but farmers as maybe another fourth stakeholder. And we need to get all of them on board and provide the incentives to all the different stakeholders to get them to share the data. Yeah. Any other comments from the participants?

SPEAKER:
If I may, I'll jump back in, looking across the fourth question on the screen and thinking about the stakeholders we're talkin' about. And it'd be helpful to recognize that there's quite a bit of variation

within the stakeholder groups we're talkin' about. There are many different personas or many different types of farmers, many different types of scientists, many different types of industry players and that vary, and just recognizing the intense fragment or the incredible fragmentation that exists across agriculture in the United States and this, it is exacerbated on a global basis, is not only a challenge, but potentially creates an opportunity in connecting the groups in creating better value capture propositions to the different stakeholders by recognizing that not all farmers are the same.

SPEAKER:
We talk about the farmer needs to be at the table, the farmer needs to be at the table. Well, there's (LAUGHS) a lot of farmers in this country, a lot of different types of farmers, so let's recognize the difference that is in the different types of models that they're utilizing within their production systems. And similarly, within ag retail, within the food value to the supply chain, and otherwise.

SPEAKER:
And let me just jump in here. I'm a big proponent of sharing information, but I also want to point out that you can't share everything and there are going to be some stakeholders or some agreements that will prevent sharing this information. So we got to think about personal issues, or say, talkin' earlier today, the industry issues, national security issues that are related to soils that we need to think about as a community. So although we want to have data, I mean, we measure everything everywhere all the time, we just simply cannot put all that information out there for multiple reasons.

SPEAKER:
So one thing to think about also is that, what are the incentives that we have been discussing? But also, do we really, really need to share everything? Is it a must, especially if there is going to be some policy about a rewards or an obligation of sharing data? So as part of discussion, I think, it's very important to discuss within the stakeholders and within the type of information, what is needed to be fully shared and what are the things that we as a community know that is important, that may be available there, but you may have to have some lucks in order to get there? And think about the Forest Service, where to get access to the inventory data, you can get it, but then you have to go through some processes.

>:
So this idea of completely fair, completely open, also we need to think about within the community how we're going to incentivize it and how that level of data should be shared if there's going to be any restrictions. And I would like to hear your thoughts about that.

SPEAKER:
Can I respond to that? I think, Rodrigo, that's a great point, and I like to see somewhat a unharmonious database where you can say, this is the type of data for this type of area and it's available. And whether, you know, the metadata need to be there, it needs to be described how to collect it and how it has been analyzed and what they have done and a whole lot. But somewhere that there is a repository for data that is not harmonized, because the harmonization of data is part of the, what shall we say, the delay in availability and also makes some people say, "I'm not gonna do that. I'm not gonna harmonize the data and then add it to the database." That's beyond, I mean, like I said, we have a ton of data that just would take a lot of work to harmonize it with NRCS, for example. But we have it available and anyone can get

it. And of course, it's mentioned in the papers that use it.

SPEAKER:
But you can say, there should be an archive somewhere or a repository where people can, I don't want to use the word, but people can put data or dump data for others to see. And it should be, it should be navigate-able through Google Earth or somewhat that you say, "I'm studying southwest Michigan or New Jersey," and this is the type of data that is available and there's, and you can get it. And then, you figure out whether you can use it or not. But then, harmonized, unharmonized database of sort, we have those already, several of them. They might need a few more properties, particularly in the environmental and maybe microbial world, so that there is a lot available, but there's a lot available that we could use if we know we could find it.

SPEAKER:
And I will add to that comment, Alfred, that we often think about the cost of maintaining this, because on one hand, you want to facilitate the stakeholders to share the data. It will be an ideal that you just say, "Here, here's my data." It's a less cost in terms of preparation for me or in a lower level of data organization. But then on the other side, there is a cost associated for doing that work that is not done by the user. You are facilitating the user to share the data, but on the other hand, on the other side, you have the cost of someone to look, looking after all these things and harmonizing to put it in this repository.

SPEAKER:
So another question would be, who's going to pay for this? If this is going to be paid by the private the company, you know, in private sectors, is it going to pay by the federal agencies because it is also the cost of the cyberinfrastructure behind it for having the availability and the flexibility and the easiness to share this information?

SPEAKER:
I have no answer to that. I think if the idea is good, and if we look what a physicist do or the chemist or some of the biologists, we can perhaps copy their model and see how they fund it. I think we need to look into it. But a repository of soil and environmental data freely accessible and searchable through whatever engine, that would be, that would be quite a nice thing into addition to all the things that we have. And people might be saying, "Oh, OK, I dumped my pertinent data or my EM mapping or all the carbon data that I have, I dumped them there and let people play with it," because everyone knows, maybe it's not an age thing, Michel, but everyone knows if you share, you get more work than if you keep it to yourself.

SPEAKER:
Maybe also, a solution could be to adopt the software model where open source software, basically, it's developed by someone who has a need for it. He puts it, he or she puts it out open into the web, and whoever has a need to expand on that for a specific business case or a use case or whatever, and they will extend on that, put it open again. And in that way, it is progressed. So if I follow the reasoning of Alfred and Rodrigo, maybe, in a repository where data is findable to start with, and maybe also accessible, if that is a starting point and then whoever has the need to further standardize or harmonize

the data, and then puts 'em back into the repository with the correct linkages, that could be a kind of a way.

SPEAKER:
It will be very interesting to see actually which data then is going to get harmonized because it points us directly into the needs of the users. Only that data that is really useful or has a use case would be harmonized. And of course, it needs to be facilitated by standards and the right vocabularies. And there's a lot of work to be done on that end for sure. And we're also involved in that. But it could be a model. I know people are still searching for what will be feasible, but maybe this is a way forward.

SPEAKER:
That's a really interesting idea. I like it a lot, and the data would not even have to be stored in a centralized location, but just accessible through a centralized location, perhaps.

SPEAKER:
I think centralized is really that's... I don't really see that's people really don't want to part with their data and they're also the ones who know most about their data. Sure, you could have one repository where you could get our data that people want to store somewhere else, where they want someone else to curate their data. But I think, like a distributed or federated system, whatever is most appropriate is much more the future.

SPEAKER:
Yeah. Yeah, I agree. And many data sets are already out there with a DOI and available. So just having some registration of, you know, here is a bunch of soil data sets that are accessible.

SPEAKER:
And it should be really, really easy to do that. Otherwise, people are not going to do that. I completely agree with Alfred. It's a lot of work. (LAUGHS) So you wanna make it as easy as possible for a soil scientist, not necessarily data scientist, but anyone should be able to just upload their data, provide some basic metadata, like a minimum required list and just put it out there, see what happen.

SPEAKER:
And perhaps have some infrastructure developed that would make it easy for, or easier for people to annotate these data sets and make them more interoperable to do that value added part for useful data sets.

SPEAKER:
I think now in our universities, the sort of that responsibility to take charge of the collection and make sure it's gonna always be available is a difficult thing. So now I think we see with university collections or people now have a choice where they're gonna put their, get their DOI. So I think, I know at our university and other places that have ag librarians, like Purdue is fantastic, where they do a lot of training of the students and they help you with those keywords. And I guess if we had assumptions about how that data was gonna be reused. For example, is it gonna be in a process model or who is the likely reuser of it?

SPEAKER:
Because they don't think it's just the useful data that gets reused. They think the findable data. So some of that, probably tremendously valuable, really oddball data that we don't wanna reclassify inappropriately. That's unique, but somebody can find it conceptually related, right. So if we had a notion of these different applications and then have those sort of tags pushed out for the likely reuse potential might help. So I'd say, I think through our universities, we have at least one part of this that's probably not really searchable through Google, but very available for us to train students.

SPEAKER:
The National Ag Library is another one in the US and the Guardian CGA, our repository system is also good for international type data sets.

SPEAKER:
I'd like to make a comment, Fanny and Sheryl, the model you were describing where data are in a repository and then can be harmonized and put back into a repository, that actually already exists. We're doing it in the repository that I work with that happens to also house the LTAR data. We're not doin' it with soils data yet, but we would like to. We're doin' it with community surveys like organism, not population level, but community level data of organisms. That's kind of off topic for what the main theme of this group. But I just want you all to know that this does exist already. And that's actually the reason I'm here is because we have an interest in harmonizing soils data as well, but aren't sure of what kind of systems and models are already available.

SPEAKER:
But we have access to a lot of this kind of raw research data from networks like LTAR, MSP, and LTREB's.

SPEAKER:
And Margaret, can I ask what is your experience? Does it work? Do a lot of, well, do a lot of users...

SPEAKER:
Yes, it works.

SPEAKER:
..go through the trouble and do the standardization and they commit it back to your repository?

SPEAKER:
Well, yes, it works. It's not quite simple. Depends how well the data is described in the first place, whether it's described at the measurement level with a fair amount of detail which is possible but not always done. Restructuring data is always a custom activity, every data set that comes in. And this is like the data set that I think Alfred mentioned, he said just data in whatever form that it needs to be in. And that's one of our goals is that as we harmonize, we want there to be a canonical data set that is, describes everything that needs to be told about this particular data set. Because if you haven't done that step, you won't have reuse potential years from now.

SPEAKER:
People, you won't be able to tell what this data actually is meant to do if you haven't got a really good, better data log in the first place. So that needs to be there. And then there's a reformatting step, which, if that is really relatively clean metadata to begin with, the reformatting is much easier. In some cases, we had to contact the owners and ask questions, but typically that, and our goal is that if we create a script for someone who has, especially for a time series data sets. I know, this is kind of a side comment, because we work with ongoing time series like LTAR, data are constantly being, not constantly but regularly being updated. They will add another year's worth to a data set. So if you harmonize data in 2018, you probably gonna wanna do it again in 2019. So the system accommodates that.

F11:
Stories

MODERATOR:
But you have to keep in mind that things happen and methods change. Sometimes the format of the dataset would change, so it might not be exactly the same script that ran on an earlier version runs on a current version. So that's another kind of a kink. So, it is possible. It's not perfectly simple. And actually, we have a paper in review describing our experience with the community survey data. It's not out yet, but it'll...we sent it in. But we'll have something to site on that. This kind of activity pretty soon, I hope.

MODERATOR:
That's a great discussion to this and on Slack, there are a few questions which I think are relevant to this discussion. There are people asking about other examples...good examples from even other disciplines where scientific, public and private data sources are compiled in a common database repository. Things like Wikipedia is one where you have different sets of data. Someone harmonizes it open street map, but there are others. But this is...both of these are mostly driven by the public. So, is there...are there other examples of this kind of data storage repositories that exist?

MODERATOR:
So, Mike is here and he would like to talk. I don't know if you can unmute him, but meanwhile, I've been saying that there are different harmonized data repositories out there. Right. (INAUDIBLE) there are other ones that exist. If Mike is out here, where you can jump in. But what I would say also that in this in this type of repositories, data can be findable, could be accessible, but the metadata might not be very well described. Therefore, the interoperability and the reusability of the data, it's complicated. So, it's one thing that we need to think about. I don't know if Mike is still here.

MODERATOR:
Still here. I've just been joining on my phone for various reasons. So, but I think that was the point I was going to make was just that, that not only do they exist, but many of those you just rattled off are also federated within the data one system of repositories. So, they're searchable across repositories. But I totally agree that they don't require metadata the same way that the more harmonized one, but I think that was just, you know, where this conversation started was, you know, the first step was just getting them archived at all and to make sure they're not lost altogether. And I just wanted to point out that, you know, let's not reinvent the wheel. There's a lot of good options out there for that first step. But I

really like the suggestions of ways to be able to annotate these sorts of historical archive datasets by folks that understand them without actually modifying the original data.

MODERATOR:
Yeah. That's a great point. By the way one of the other things I wanted to add is when we are thinking of data sharing and it came up in some of the discussions. Just wanted to let this set of participants know about some tools that exist. Coming in from the technology side...my background, as I was saying is, I'm a computer scientist, I work on different tools. And some of the latest tools that have been developed are around this concept of multi-party compute. That is, how can you share as you might have blocked in? Those are the big buzzwords. But in addition to that, how can you share data with others without revealing that the content of the data? For example, if I'm sending some...I'm sharing some data with Rodrigo, I can share encrypted data so that Rodrigo doesn't know what that data is. But he will be able to perform some analysis on top of that data, some AI on top of that data. And there are different ways in which people are taught. People are talking about enabling that.

MODERATOR:
One of them is called homomorphic encryption. That is, the data is encrypted using...where this data is always encrypted throughout. Right. Like now, as you know, the way, for example, when you use SSL or (INAUDIBLE). The sender encrypts the data, the receiver decrypts the data and performs operations on it. What this enables is the sender doesn't...the receiver doesn't actually decrypt it even on encrypted data, they can perform some of this analysis. Now, that is one of the tools. I'm not sure how relevant it is here, but that could be something that could enable sharing of sensitive data. For example, farm management practices. I know Catherine and others have brought this up to this point. We need farmers on board. This can provide more assurance to farmers that, hey, you know what? You could also share your data and your data won't be compromised because others won't know the raw data, but they might still be able to do some analysis on it.

MODERATOR:
So, this was one of the tools I wanted to make the community aware of multi-party compute. There's another concept called Federated Learning, where rather than getting all of that data in one place, the data stays distributed, but you can still (INAUDIBLE) the model on a distributed dataset. But in addition to that, if you have to make the data going to work, you still need data models. You still need that data in a particular format. But that's, again, one of those other tools that exist in the computer science world of things which might be beneficial for this kind of a dataset as well. So, one other question that came up from Slack was are there banks of soil microbe cultures derived from soils? Are any of these banks of data that is available for soil microbes, that is available because that is another dataset, which I guess when you're talking of genomic datasets, it kinds of falls in that region. But again, is anyone in this breakout session aware of such soil microbial bank data that's available?

MODERATOR:
I'll offer that there is a new effort just underway funded by DOE to create a database for the microbiome in soils...plant-soil microbiomes that is being led out of, I believe, out of Lawrence Berkeley Laboratory, but is spread throughout the national labs and also open to the broader community. That just started last year. And it's a nascent effort, but it's one that...it's one of the only ones that I know of that's

attempting to accumulate soil microbiology information and genomic information in particular.

MODERATOR:
That's good. Thanks, I've taken note of that.

MODERATOR:
So for clarification, do you mean live cultures or do you mean (INAUDIBLE) frozen samples? Very different answers.

MODERATOR:
This is data, not samples. So.

MODERATOR:
OK. So, the genomic data?

MODERATOR:
(INAUDIBLE) metadata you might expect to be able to interpret it.

MODERATOR:
But yeah, I don't know my memory, but there are metagenomics databases for soil data.

MODERATOR:
Are they public databases?

MODERATOR:
I think like GenBank would have people put entries in with some very limited information and you would know. So, I think you could find in GenBank things, but not like Phil was talking about with probably experimental data beside them. Right. Is that what you mean?

MODERATOR:
I think that's (INAUDIBLE).

MODERATOR:
I'm sorry, I can't get my camera to work, but…

MODERATOR:
Hey, Andrew.

MODERATOR:
Yeah. So, in Australia we have, I guess, a coordinated effort around microbiome, soil microbiome, where we keep a whole bunch of soil environmental data and microbiome data. And there are similar initiatives around (INAUDIBLE), I guess so. So, there are those efforts and then in terms of microbes, particularly so that their genomes of soil microbes kept at places like (INAUDIBLE), I guess in database's with, you know, with soil metadata. And then all of the culture collections have the, you know, the

isolation source listed as part of the metadata. So, I mean, it seems like to us over here that the most of the problem isn't around actually holding the data. It's holding the metadata. So, what really is needed is the database of metadata so that you can find all the (INAUDIBLE) from individual researchers and access data. But that metadata or the metadata around those DOIs needs to be such that the data is findable. And that's the hard thing. I mean, it's easy to find a DOI, but to find what's actually associated with that data is more difficult. So, databases of the metadata is kind of what we were always offering.

Sorry, I can't get my video to work for some reason. I'll keep trying.

MODERATOR:
I would love to make a comment that I think is somewhat related and hasn't been addressed yet, which is point A under the first question about current vocabulary, ontologies and semantic resources that would assist in data harmonization and interoperability. I am a scientist with the National Ecological Observatory Network and I do a lot of our soil-oriented data products, data protocols. And when we were, I mean, NEON is a very new network, so we were building from nothing. And we measure a lot of other things in the course of NEON monitoring biodiversity. We have a lot of spatial data. And so, there were ontologies that we could borrow and use very standardized terms, for example, from the Darwin core or there were other efforts. I know. So where possible, we tried to use the same terms with standard definitions that other people were using and so that you could more easily harmonize NEON data with other similar kinds of biodiversity observations from either other networks or individual researchers.

MODERATOR:
For soil, we...I, we didn't really know of standard ontologies that we could follow and maybe there are people on the call who could have helped us find them if they exist. But I would say it wasn't very clear where we should take our lead from. And we ended up just making up terms that made sense to us in terms of organic carbon %. Did we call it carbon %, carbon % organic? What exactly is the definition of that? And so, I think that we could actually make a good contribution by creating a standard soil ontology. And obviously it's not going to help with past datasets. But I love the comments people had about being able to annotate those and make scripts for harmonization. But there's a lot of current data being collected that we could maybe make it easier on ourselves by using standardized vocabularies and standardize terms where we're essentially measuring the same thing, but just describing it different, slightly differently, which means we have to just do more work to pull these datasets together. So that's just one observation to contribute to the dialogue.

MODERATOR:
So, Samantha, can I comment? Have you been...are you aware of the ESIP on soil ontologies working group?

MODERATOR:
It looks like not. I was at another...I was at maybe an ICN Workshop and someone mentioned that.

MODERATOR:
OK. Well, Kathy Todd Brown runs it. So, I don't think she's here. But the ESIP stands for the Earth science

environment...what is it? Earth Science Information Partners. It's basically a working group of a professional association for people who do data management for mostly it's NASA, NOAA, USGS and a lot of academic type institutions like (INAUDIBLE). So, lots of these kind of common issues across all these domains get discussed there. And there are a couple of groups that are organizing...have organized around describing soil vocabularies. So, anybody who's interested in that topic should get in touch with Kathy.

MODERATOR:
That is fantastic. And I think that will actually go a really long way to see those ontologies. But I'm really excited to hear that Kathy and other brilliant people are working on that.

MODERATOR:
Yeah. Kathy is going to join the next session. So, we have a 15-minute break. 15, 20 minutes at 10:00 Pacific, 1:00 Eastern, we have a 15-minute break. And after that, Kathy will join in. So, yeah, we can ask.

MODERATOR:
May I also add on that. There is also an initiative globally which for now has been joined not so much by US, but it's been going on for quite a couple of years where the Pacific, for instance, in Europe also were involved in really trying to come up with like an ontology, but also (INAUDIBLE) and to harmonize that in a way to really make this easier. So, in Europe, there is it's called the Inspire directive, which is basically a law that all European countries have to follow and they have to share their data according to the Inspire soil requirements or specifications, basically. So, there is a standardization going on over there where at the moment actually making the code lists for that.

MODERATOR:
So, they are (INAUDIBLE). And also, this is being, well, we're taking part in Kathy's ESIP initiative as well to just try to link that. And also, Australia, I don't know if anyone wants to pitch in, but they also they have their own (INAUDIBLE), which I think are already quite extensive. So, there are some harmonization efforts going on globally to really make soil ontologies including the code list and (INAUDIBLE). So, I'm actually based in the Netherlands. So, it's getting night here. Slightly different time zone. We're involved in the Global Soil Partnership, which also really aims to help with this on a global level. But it's very difficult to connect everyone. So, I'm really excited that now also in US, this is really picking up steam. I think Kathy's doing a wonderful job and we really hope to cooperate to make sure that it actually gets really global. So, yeah.

MODERATOR:
Yeah. There are some efforts, I guess, as well. So, there's the biological observation networks, that (INAUDIBLE) in particular, I guess. And there are also data formats that are trying to get together. So, you mentioned Darwin core. So, there's now or there's about to be, I guess, Darwin core extensions for the access data format. So that will cover the environmental metadata for genome sequencing studies, I guess. So, soil, carbon, stuff like that, that are, you know, set formats and set units and ontology site. So, there are some efforts to sort of to bring those things together. I think the Darwin core extensions are happening now with the genome science consortium. So, I'm sure the stuff on their website about. This interestingly enough is also quite a big effort in ocean sciences at the moment with the Ocean

Observation Network to bring all the data format ontologies together. So, it's happening sort of across environments, which is good as well.

MODERATOR:
I think this is great, at least we are making…this is one of the positive things out of in the science community, we are making good progress in this direction, which I think is really good. Could I just go back to one of the questions that Rodrigo had asked before around the level of data that we need to share and what…like right now, the state of the art. It seems as though a lot of these distributed repositories we are trying to come up with a common standard. We are trying to come up with a way in which this data can be shared among these distributed repositories. But the other question Rodrigo had asked was around what level of data sharing is needed even among these different repositories or even getting into the repository? Do we have any thoughts on that? Do we need it on a…

MODERATOR:
Field site, Rodrigo, do you want to just say a little bit more about the particular question that (INAUDIBLE) spatial-temporal resolution to all? Yeah, just expanding on that.

MODERATOR:
I guess it can be interpreted in different ways. One is a way that you are posting that for that involved the spatial and temporal scales that we need to have all the information. Let's just call it higher spatial and temporal resolution. And if so, I'm sure there are some variables that are important for that. And within these efforts of defining ontologies, Et cetera, and also maybe it could be an effort of defining the spatial and temporal resolution of the variable of interest. The other, way I was proposing this question it's more in terms of it goes into the use of the information that may be related to proprietary or some sort of data privacy issues.

MODERATOR:
I do touch a little bit on that now, where data could be there can be sharing in a raw format that could be used in many different ways. But there might be a tool that may filter that or some politic to filter that in different for different purposes. Just one example could be, do we need to carbon data? It probably could be easier to share, although there carbon markets, but other things like phosphorus may be more complicated because you get into fertilizer, non-pollution, non-point source of pollution, et cetera, and that also gets into regulation. So, I guess to summarize what I said there is the level and the time and spatial scales and the level of privacy of the data and how can we access and how we should access potential data repository.

MODERATOR:
I guess, added to that some of the wants that came up during the fireside chat or (INAUDIBLE), if that's what it was called, you know, to understand some of the dynamics or properties in relationship to crop yield or environmental processes, we really need the temporal, you know, when it was collected, how deep it was collected, the analytical method used, because that, you know, and we could harvest and combine data with different analytical methods with some of those details. But very rarely do people share those or gathered those in their metadata when they're doing data mining. But we could automate that and have a recommended protocol If somebody wants to ultimately submit those and

share them for this larger interpretation.

MODERATOR:
And then I'm excited about this idea of encryption because what we really often are using is the long lat and the year, the time to try to stack data and marry it. So, if people could have trust in encryption, then we would be cooking with gas or different kinds of users who are, you know, take some kind of non-sharing or not sharing the model. Although anyone, if I guess some of the fear about regulation is interesting to me. I haven't really found farmers to be worrying about that, but you must have.

MODERATOR:
Yeah, I think that covered the question that I wanted to answer and you're right Michelle, that some of the privacy things might have addressed that as well. That's if you could use encryption to address all of the trust issues, I think that can help us take a step forward. So, the other questions, so among the questions, so by the way, before we go further, I just also wanted to mention that one of the deliverables of our working group, this breakout session would be a slide deck that will be presented tomorrow. So, we can spend towards the latter part of the last 15 minutes of the next session, just talking about how do we present the discussion to the rest of the workshop attendee? So, with that we can go to the next question here about the internet of things, or how do we deal with data privacy and ownership?

MODERATOR:
I guess we've talked a little bit about that. But do you have any additional comments around how do you... So, part of what we're talking about is what's the structure of the data, right, how is this data stored, how is this data shared, how do you annotate the data? Which I think was all great. The other question which I wanted to add, which is related to question 1b is one of the questions, is on the fidelity of the data. How good is the data? We talk about data harmonization, but there's also about the accuracy of the data, especially when we look at emerging markets and data from there. This problem often comes up. Do you trust that data, how do you validate what data is coming in?

MODERATOR:
Do you see these issues coming up in the kind of datasets that you work with that is from, if you're trying to use a different dataset or different sets of datasets, the methodologies that was used, or the accuracy of the datasets, do you see any of that coming up along with that? The other thing also wanted to ask is how do we make sure that any of this work is future proof? That there's any data that is collected would still be relevant five years down the line, especially as we start collecting more and more datasets. But wanted to get your thoughts on these questions.

MODERATOR:
I would like to see a few things about data and what's happened in the past 20 years. I think the first thing I would like to know is that we have more data than ideas. I think there is so much data out there compared to 10 or 20 years ago. And because anyone can now in a very rapid time, collect a lot of data in soil science, it's relatively easy to collect a lot of data. We have instruments. We have tools out there, and we can collect data much faster than we've ever been able to collect before. So, there is a lot of data out there and it's diverse. And some of it is good and some of it, we don't know what it is like.

You're just alluding to. And I think we use data of others. And most of us that use data of others, if you don't trust it, or you can't get a handle of the metadata, or that a group that is coming from might not be meticulous, you just don't use it. I think that's what many people tend to do.

MODERATOR:
So, you have a sort of natural filter like the internet. You filter also what's on the internet. You think if it comes from this, I might trust it. If come from there, I don't. And I think that's the same with data. If it comes from a respectable organization, whatever that is, or when it from USDA or NRCS or from history, you see, I can probably use that data and it's probably very good data. If it comes from other people that you don't know or you don't work with. That's it. But I think the idea is that we'll have more data than ideas. Let's be honest about it. And I think we should stick in that vacuum or that momentum for a while. Let people massively collect data for a couple of decades or maybe a decade, and then see whether we should harmonize or find an infrastructure to share that.

MODERATOR:
Interesting point (INAUDIBLE).

MODERATOR:
I think this is a huge problem, data fidelity or data veracity, anyone who has put together a meta-analysis or a large-scale regional or continent skill synthesis knows that not all data can be used for those. And determining the datasets that one needs or one actually can trust to deliver the information that you need is a huge problem. And it comes back to a comment that I made yesterday about the need for calibration sites or long-term sites, where we have a known and very heavy investment and making contiguous datasets that do have that cross-calibration and internal fidelity and the importance of that.

MODERATOR:
So, we try to do a synthesis of data collected from LTER sites, for example, I would have a lot more confidence and (INAUDIBLE) I just mining data that I've collected off of three or four datasets without a lot of knowledge of the providence in that data. So, I think this is a big issue. I think it's one that the community hasn't yet dealt with and you can deal with it to some extent in the metadata. But even so, it's hard to. It's a difficult issue.

MODERATOR:
I think that the comments yesterday about people not trusting different labs to do the similar analysis, even with the basics, is really alarming. And if we think about pooling data in your question about five years, I think when we're doing long-term trends and a lot of our soils data we think is about the inherent characteristics that don't change and we're drawing on our pedal head-on data, that's not very dynamic. So, we assume that the labs have consistent practices.

MODERATOR:
So, I think that's really interesting to think about when we're pooling things, and if we don't trust in any lab variability, how then we convert that into recommendations for farmers? And so that might in a way, you know when I saw scientists look at the state of advising (INAUDIBLE) talk to farmers about soil test information, I feel really ashamed that I don't think we've done a great job. So, I guess where we are

now when a farmer can work with a single lab and take detailed data over time and their own records, that's the most valuable thing for them.

MODERATOR:
Right. Gutzman, you have your hand up.

MODERATOR:
I did have my hand up. I realized I was about to say exactly what I believe Phil said, which is even though we have a lot of technologies and tools and, the ability to generate data rapidly, we can't forget that the tools don't necessarily solve the issue of overall data quality. So, these issues of calibration, whether it's four different regions or calibration and transfer between devices, those are core issues. And we also can't forget about the manifold operational problems.

MODERATOR:
So, thinking through standards, we're actually pulling, shipping, and storing samples. If we're talking about soil repositories and how deviations in those operational parameters might affect the data outcomes. And so ideally, before data would feed into any sort of storage system or common repository, there would be some element of data QA QC that walks through not only how, what were the methods used in actually accessing these soils, then analyzing them and making sure that we're, you know. This does extend to the standards and interoperability element, but it's just such a critical area to unpack before we actually rely on and trust the data at hand.

MODERATOR:
Yeah. Thanks, Katherine. Phin

MODERATOR:
Yeah, I was triggered by what Michelle also said and it goes, I think, along the lines of what Katherine said, and maybe it's very ignorant because I'm not a US citizen. So, I'm not fully sure of the landscape of labs in US, but actually also on a global level within GLOSOLAN, which stands for the Global Soil Laboratory Network, which is hosted by the Global Soil Partnership, we are really trying to harmonize lab methods to the extent that they're not yet harmonized. So, there's different varieties out there for using, I don't know dry combustion, for instance. So, we're really trying to harmonize that.

MODERATOR:
And at the same time, there's also a proficiency test efforts going on where actually we just exchange samples and see what comes back, see if we can improve quality to train labs and to really improve their capability to do decent measurements because it's true, it's very needed. It's at the basis of everything. And actually, the Kellogg lab of NRCS is the one that is participating on behalf of the states. But also other labs are invited to join GLOSOLAN. They can take part in these efforts.

MODERATOR:
Maybe you already have all that on a national level sorted out. I don't know. But it could be useful and it could be nice. And actually, the Kellogg lab also going to serve as the Gold Standard lab within the spectroscopy initiative that we also have within GLOSOLAN, where actually countries are also invited or

every lab that wants to participate are invited to check the quality and the measurements that they do in their lab against the Kellogg lab because we think it's of high quality. So, you'll have the Gold Standard lab in your country.

MODERATOR:
Looks wonderful.

MODERATOR:
OK, I think with that we're coming close to the break. But Andrew, you had a point too, I think.

MODERATOR:
Yeah, I can't seem to work out how to raise my virtual hands. So, I'm raising the real one. Yes. So, we kind of take the opposite approach to Katherine's suggesting we try and take all the data, no matter how good we think the quality is(INAUDIBLE). So, sometimes we'll have problems with transport or storage with a sample, but we usually just record that in the metadata and take the data in store because some you the data is useful for some things, but maybe not all things. So, we try capture all of that information in the metadata. And we kind of, it's a bit of a double-edged sword because you wanna provide a curated dataset that people can just download and use. But you're also relying on whoever's downloading and using it, actually looking at the metadata and saying that they understand what they've downloaded and used.

MODERATOR:
And so, it's kind of a really tricky one, but we've taken the approach to put as much data as we can and just record as much metadata as we can so that people can sort through it and try and use it. We also record the methods pretty thoroughly and the lab (INAUDIBLE) the place that the method was conducted so that you can actually just get. If you want old data, that's produced from one lab they can do that. And then nationally, we have a National Data Accreditation thing. So, a lot of labs have this accreditation and they should really keep the same result for the same sample. And they will have the same standard samples to run from actually giving the same results. So, that kind of helps be with your confidence.

MODERATOR:
OK. I think what you're saying is similar to what Alfred was saying too. That you could just collect large amounts of data, but to your point, we need to harmonize it and you're saying that data can be captured in the metadata and then...

MODERATOR:
Yeah. So, for me, it's all just an issue about the metadata. And I mean, the problem is you have to rely on who is downloading or accessing the data to access the metadata as well and to make sure they understand what they're getting and not just use everything. But the other danger is that we only put up or I only put up what I think is good data. And just because of the sample hasn't been stored in the same way as all of the other data, doesn't mean it's not useful. If something's in that data it's in that data from like microbiological sense. I guess it's harder to prove absence. It's harder to compare it in absolute terms. But it's not to say that it's not useful for some things. And therefore, we try and put it up. The

same thing with the data that we find is private. We still put the metadata up, so people know the data exists, and then they can go and negotiate with whoever owns the data to use it for their purposes. But they can still know the data exists and find it.

MODERATOR:
Thanks, Andrew (INAUDIBLE) I appreciate you joining it's a late-night there or very early morning, middle of the night, right. So, we have a break right now. So, let's meet again in 15 minutes and then we can continue this. And the last 15 minutes we work on what is the presentation we can make tomorrow. Thank you. So, see you in 15 minutes.

BRUNO BASSO:
I would like to, introduce to you the other committee member, Alison Marklein and the note taker to Dr. Raphael Martinez (INAUDIBLE) is a post Doc at Michigan state university earth and environmental science, department. And thanks very much for joining, this brief conference room. We must obviously share same interests and the idea behind this breakout is that models and the way we analyze data coming in, are a critical component of a dynamic sole information system. And so knowing a lot of the people attending this breakout, that have experiences both with data and AI methods, the machine learning as well as models and so I look forward to an engaging conversation on this.

BRUNO BASSO:
The question says, you can read them. We will go in order and we have roughly, I think, an hoour, at 1:00 we have a break. One recommendation was not to necessarily logoff, to be able to stay connected and allowing other people to join this room. I would like to start with the first question and we kept it relatively general but, it is in the realm of machine learning and AI methods and where are the concerns in where they will, fall short. How do we deal with sparse and diverse data streams in soils? So that's obviously a very important, critical question. I mean, just, start sharing a little bit of experience and then I'll open the floor. Both Raphael and I are heavily involved in using, machine learning methods and crop simulation models and actually we're working on integrating the two very closely, as well as, not just focusing on one models, but rather ensembles of approaches and ensembles of models as well as AI methods. How they can be fused together to improve what we, can't do by using its single system alone.

BRUNO BASSO:
And we are suddenly obtaining very, very promising results, it's really hard work and I'm sure we're share some more details as we speak. So, the limitation that we see obviously is the amount of data to train the systems, to train the machine learning. That's one of the most limiting factor. And so, obviously the idea behind and not too normal anymore is to produce, generate data from the system approach, trying to capture the feedbacks between the soil, the plant, the atmosphere. You heard me in the introduction, you know, landscape positioning, the landscape management, weather, all together to not just have this kind of pedal centric view of the data, but how this interactions affect. And so models have a strong role in capturing the dynamic aspects and the feedbacks between the soils and the plants and so on.

BRUNO BASSO:
So if anyone is willing to, the floor is open, starting with how to deal with this sparse and diverse data streams, what's available, anyone that wants to share experiences, ideas, limitation you're welcome to do so. Thanks again for attending, let's get the conversation going.

JOEL DUDECK:
Hi, this is your Joe Dudeck. I have a different question maybe higher level. My understanding of AI systems is that, they're just the decisions they make it's not as obvious how they're making the decisions and that, you need to, there's a whole new science of going into trying to investigate how the AI made certain decisions and I'm just wondering if that, one presents new opportunities to learn about interactions we don't currently understand, or to if it presents a problem that it could be some kind of

spurious connections that are not really, really true.

BRUNO BASSO:
That's a very, very good point. I just don't want to do the talking all the time, so anyone else commenting on Joe's comments, concern, valid idea?

MARK:
I think Bruno you actually kicked it off quite well and you said that with your group, that you're using a multiple approaches. So, you know, like, you know, in various fields that's called triangulation and you should come up the same question, multiple different ways. I was pleased to hear that's why share Joe's concern as well, that when you rely on any one methodology, especially when something is outcome based, you know, we can imagine this can be grateful within sample prediction. So when we were saying, you know, what is the promise? I think when conditions are quite similar at other localities, or within time, then they're gonna do much better than I ever push for a process based models. But as soon as we start experiencing more extreme drought or with changing conditions remarkably, I think we have to be careful.

MARK:
Now, again, you know, I'm, putting it out here as a very binary definition. I think a lot of the time we choose, the predictors are put into those approaches based on expert knowledge. So it's not completely devoid of cause of inference, but I guess that's the biggest kind of concern I have, is when we move forward with it and thinking about predicting out a sample.

BRUNO BASSO:
Excellent comment Mark, thanks very much. Stephen, please.

STEPHEN:
You know, I was gonna add a similar comment as Mark. We're using, we've used artificial neural networks and random forest to generate management data. We work on the U.S. National greenhouse gas emission with EPA and USDA. And the big thing that we've emphasized is that at a sample set, how well are we matching the outer sample and if we're not, then we look to more, I guess, simpler approaches to dealing with the management data where we don't, we don't know the management, I guess I should've said. So I think that's really critical, but I agree with Joe that we still don't know the relationships necessarily and why those are, you know, why we end up imputing say, no to one farmer's farm and not the other one. But we rely heavily on the outer sample to give us credibility in that or the random forest methodist is working appropriately.

BRUNO BASSO:
Right. So in that context Stephen I think, where are the data going in. Is pre-both sensing one of the layers, producing that, because at least for the example that your describing, it seems to be within the, I mean, the uncertainty, but a little bit better understanding of the spectral signature whether a soil has been tilled or not, that we work on that as well. What is the data stream going in training the random forest, that's or any other obviously machine learning algorithm. That's one of the questions that sometimes, obviously we wanna go away from core, you know, correlations without having any

understanding of, if that is really affecting is, you know, causation of the system. We can run into freakonomics kind of thing.

STEPHEN:
Right. Yeah. So our main data stream is actually survey data from USDA. USDA conducts what they call the seep survey, conservation effects assessment project. And so, that's where they've gone out and actually, I think it's a really, it's like a four hour survey or something it's just quite long. I think one of the longest, longest ones they do with farmers. And so they asked the farmer, you know, this year, previous year and one year before that about how they manage this field. And so we have this information and about, what the practices were, you know. The unfortunate thing for the broader community here is that data is confidential, so you have to be doing directly working with policy to have, to have access to that data.

STEPHEN:
However, you bring up remote sensing. In our latest version we are working with a group that created the Optus product. That's Bill Stylists and Hagen and those folks in Hampshire. We're using that product just additional inference on tillage, in particular in cover crops. But yeah, we have pretty high confidence in our input data streams for this and like I said, we keep a sample out that we're not using to train the model and then we rely heavily on that to see how well we're doing.

BRUNO BASSO:
Right.

ALISON MARKLEIN:
One of the things that I think machine learning and the use of process based models can be most helpful with is, generating hypothesis. So as Mark and others have mentioned, you like often machine learning or artificial intelligence algorithms well, are like, we'll just generate the output and you don't really know how or why they're making different decisions. But humans can look at the results of the, the runs and see, like using our brains and how we understand the system, come up with hypotheses for why the machine learning, generated those results. And go through, and do more experimental tasks in the field, whereas if you hadn't done the machine learning you would maybe have thousands of combinations of experiments to do and then you can narrow it down into a reasonable amount.

BRUNO:
Right. A comment, I mean Linda, please.

LINDA:
Yes, sorry. I guess partly and this is really following up on what Alison said too. When we think about the promise of these. I'm a microbiome, so microbiome person and for us machine learning and AI are hugely compelling, because our data sets are of the size and scale and complexity that realistically, even the best non-parametric statistical approaches we have are insufficient for us to detect, the patterns and all this complexity. But the challenge is what are we, you know, what are we trying to model? I think that becomes a real, a real critical factor in figuring out the promise and the shortfalls. And echoing what Alison said, I do think prediction is still quite problematic because we don't know the rigor or

alternatively the stochasticity, the consistency of predictions with all these co-varying things.

LINDA:
But as far as really helping us refine, you know, for the microbiomes, are there unexpected combinations of microbes that come out of an ML or AI approach, that would've never occurred to us in our systematic limited human brains. I think the power of these is extraordinary, but the limitations in terms of prediction are very real and I think that needs to be something that we as a community, need to take on directly over the next few years, really trying to understand their value of prediction versus hypothesis generation or just building understanding.

BRUNO BASSO:
Right. I agree.

JOEL DUDEK:
So when folks in this field are looking at AI results, it's a black box and so are the capabilities that you kind of delve in there like black, black box, something that is also being generated, I understand that's kind of a whole separate skill set.

LINDA:
Are you asking about microbiomes explicitly or soil?

JOEL DUDEK:
No, just actually that the AI, you know, makes relationships and understanding what parameters the AI system used to make sort of connections, is a black box unless you can, unless you have the capability to go in there and unrepentant, you know, keys out what the computer, how the computer, how the AI program made those connections. And I guess you can see the connections and then you say, oh, why does that happen? And then you can delve into it. That's kind of a backdoor approach, but I'm wondering if more direct assessment of the AI tools or, is something that's being contemplated by the people that are using it.

BRUNO BASSO:
I guess, you know, one of the reasons I have, you know, the say I was never trained, even though I know enough about it, but I wasn't trained as an ML in AI, but rather more of a process-based models. To me the way I would see the validity and trusting the connection and the understanding that goes into a model, at least knowing how the model, you know, is built and so on. The models to me provide a little bit of, a level of validity on what the AI is providing as black boxes. So that's why I think going back to the initial thought of the ensemble of approaches as well, ensemble of models of each, is possibly a little bit better way to understand things, but I would encourage anyone really working strictly on AI trying to help, this question for doing the rest. Mark if you want to, but two marks.

MARK:
Just wanted to pick up on Alison's point about being human and that we get to investigate all of these things. So, in terms of AI it's one of these situations, I feel like it's the spiderman recommendation. It's with great power comes great responsibility. Right. And so we have great power with these ML and AI

methods and we should be using them, but we're also inherently human. And so, you know, one, I like to always bring back when I'm teaching that the American statistical association came out in 2016 and they had a statement on P-values. They didn't have a statement on P-vales, they had a statement on good science. And one of the two reasons that motivated them to come out and make the statement was big data. And they recognize the big data shifts, policy practice and drive science and they were concerned that poor data analysis would lead to the wrong policy and practice.

MARK:
And so I think, you know, what we're talking about here is like a broader societal issue, is like how do you harness these capabilities and put the rings on them, so that we don't take them too far. And because everybody wants to tell a story with data and maps are really powerful, they often have an undue influence. So it's how, you know, we got to let them move forward, but how do we, I think it would be interesting to think as a group as well, how do you also reign them in, so that those other approaches gonna look in and have equal kind of impact.

BRUNO BASSO:
Excellent, Matt. Mark, did you have a point also, please?

JOEL DUDECK:
I did. First disclaimer, I'm not a soil scientist. I have, however, worked at an ecological synthesis center as the director of technologies for a couple of decades. So I've had to deal with a lot of different types of data issues ranging from discovery to integration.

JOEL DUDECK:
I recently heard a talk by one of the principals on the Wikipedia wiki data group, and they use a lot of ML and AI there. And he said, which I thought was really interesting, is that you have to take the ML and turn it into MK, which is machine knowing, which means that this is addressing Joe's Black box concern that when you get the results from your your AI models, you then have to materialize them in terms of something that's human understandable, and ideally comparable and interoperable with what other models are saying. So this is a way that you can really calibrate your models on the same testing and say, we are addressing this particular issue and this is what we're finding. So this gets to the importance for me of data representation and knowledge representation.

JOEL DUDECK:
Bruno, you brought up remote sensing, and it's the same case there with entity recognition and information extraction. When you analyze a particular image and and you say what's in it, it still becomes very idiosyncratic. And we really should be working towards classification of images according to standard types of entities that can be then shared, and searched for, and validated by others. So that I think that the approach that we take to representing what we see in data, what data is about, what models are determining is really important. There are techniques now, and you've probably heard of these oncologies and knowledge representation languages. I think that those are really helpful.

JOEL DUDECK:
Real quickly as well, I wanted to deal with the question, how do we deal with sparse and diverse data

streams? I would aver that some of that is due to the distributed nature of the data and the fact, again, that they're very idiosyncratically and heterogeneously described and collected. And so that's there again. And we've heard it several times yesterday, the importance of standardization. So developing common terminologies that we all map our data to, and then expose them for discovery and reuse through the web, I think can really bring a lot of benefit to addressing that issue.

BRUNO BASSO:
Excellent, Mark. Thanks very much. Wonderful points there. Susan. Susan, you have your hand raised?

MARK:
Hello. Just getting my video up here. Hi. Yes. I'm a CS, yes, soil scientist with the soil science division. I'm just thinking about this idea of kind of the black box and trying to explain not only for ourselves, but to users of our results. In pedimetrics, trying to use the uncertainty that are generated with a lot of these algorithms, and translating that in a meaningful way to not only ourselves, but, again, our users. That is a sort of large focus of research right now, is how to deal with uncertainty, and along with that goes a validation of the results so that there is a piece to support them as well. But I think it's really interesting to think about communicating uncertainty through measures of risk or something along those lines that can really be applied for for management and decision making.

BRUNO BASSO:
Yes, Stephen.

STEPHEN:
Yeah, I actually kind of had the same point as Susan. We spent a lot of time quantifying our uncertainty in our soil carbon changes. I think this is extremely critical all the way down to the farm scale. I work with Keith Paustian's COMET- Farm tool as well. I think understanding the farmer, understanding what the uncertainty is in this is very critical.

STEPHEN:
I think this goes back to Mark's point as well, about low sample sizes, because oftentimes when we do have small sample sizes, we end up with large uncertainties and that we should be honest about that. And that should be reflected through the through the information we're providing out there, whether it be the general public or farmers or whoever it might be. And if it's really important to the farmer, if it's really important to the public, then that'll put emphasis into putting more funding towards that and gathering more data on it in the future. I think this is very critical, as Susan was bringing out there, to quantifying that uncertainty is extremely important.

BRUNO BASSO:
For sure, yes. You may have seen in my introductory remarks yesterday, I showed the number of samples to really capture the variability, and even the same I know may not be the topic here, but the laboratory procedures that different labs provide different results. None of them is necessarily wrong. It's just the way they process the samples, the seedings, and the homogenization and so on.

BRUNO BASSO:

So but you go from a point two, and then just simply plugging in the numbers in the volume and we came up with five tons. We know on a yearly basis, we're lucky if we are sequestering 0.300 kilograms. It's just like just beyond comprehension how we don't push for additional investments and characterize these type of things, which obviously go under this uncertainties umbrella. Mark, please, back to you.

JOEL DUDECK:
OK, just a very minor question. I'm wondering about this idea of filling in data streams and uncertainties, I think it's great. Again, this is matching the methods to the approach. So one thing we do know with ML and AI is if we increase the sample size massively because of the correlative nature of it in terms of which it's working, you can get a much more, if you like precise estimate, but it's still wrong. So you can increase sample size as big as you like, right?

JOEL DUDECK:
But if we want to inform practice and say, you do this, it'll have this big in effect, we'll just give you a more precise estimation around an incorrect coefficient. So I'm wondering, you know, again, does that just reinforce this idea that we must look at other approaches, or maybe there are new techniques coming out, AI, where we can diagnose that, but I'm not familiar with them.

BRUNO BASSO:
Right. Yeah. You can be precise and inaccurate. Back to Will, please.

ALISON MARKLEIN:
I guess just building on Mark's question as somebody who doesn't do much ML and AI work, but I typically think it kind of global scales and I'm always kind of struck by how hard it is. I feel like anytime I'm cobbling data sets together. And so I'm curious for those of you who do work on a farm scale or who work on on a more manageable scales, where maybe there's less uncertainty in all of the different data streams that you're trying to plug into a machine learning or AI algorithm. How data limited are you in the covariates that you have to put in? Do you have to use PRISM precept which kind of constrains you to the continental US, or like how do you come up with all the other covariates that go into the algorithms that you're trying to run?

BRUNO BASSO:
Yeah, that's definitely... Yes, Mark. Mark, did you have your hand raised?

BRUNO:
(INAUDIBLE).

JOEL DUDECK:
So I didn't have my hand raised. Sorry.

BRUNO BASSO:
It's still listed. Sorry, it's still up there.

JOEL DUDECK:

Sorry.

BRUNO BASSO:
No problem. OK, yeah. Yes, the covariates are very limited often. But I'll come back to this. Yes, Skye, please.

LINDA:
First, I have to remember to unmute. OK, speaking from a small survey standpoint, we're really interested in those places where we have sparse data and actually running our models to identify the places with missing data so that we can focus our further efforts there. It's really interesting to hear the talk, but I've always been interested in the idea of ontology. We've talked about it a lot in terms of management systems and how do we have a collective terminology for explaining these things.

LINDA:
And we've been working with co-operators on some social health metrics we don't usually measure and trying to have this hierarchy of measurements. So, for instance, you report soil Pi. Well, that means a lot of different things in different parts of the country. So are there some things that we can ramp those all up in together so we can analyze it across the whole country, or we can dis-aggregate it in places where you actually have different measurement and where it's an active area of not so much research from our standpoint, but just trial and error of what we can get to work with our systems. And even internally, we have problems.

BRUNO BASSO:
For sure. One, I know you focus your comment on the ontology, but initially you started by filling in the gaps and you open another can of worms, which is geostatistics, because that's different from my mail, that you have points correlated. So that's almost even more complex because then the covariates also often missing. So you raised two very important points.

LINDA:
Yeah. So actually, I thought about this example we just had because it's both, right? It's the geospatial autocorrelation instance, and it's the feature space. Right. Like, we want different kinds of conditions represented and maybe samples are very near one another, but they represent different conditions, especially things that might be in our model.

LINDA:
And so we're just working on this black soils map for the global soil partnership, and with the things that were just collected as part of the soil survey, we really weren't getting very good results. Steven Marker, who I think is in a different breakout group, did all the work. But when he added our Rabbit carbon project, which was a statistical sampling scheme with coverage across the country, and based to be dispersed on things like soil order, the results they were really good. It was because of that difference in the spread across the feature space of the model. I don't know how to solve that problem, but when we're dealing with found data, that's always going to be an issue.

BRUNO BASSO:

Right. Mark, you do have the hands now.

JOEL DUDECK:
I do. I just I just wanted to comment. Skye, if you were aware of the ESIP federation's work of developing a soil ontology, and if you're not, I'd really encourage you to join it. ESIP is the Earth Science Information Partners. It's a consortium that involves a lot of the US agencies, including USGS and NASA, and also a lot of academic participation. A recently formed cluster there is focused on developing a soil ontology that's being led by Kathy Todd Brown, who is, I think, a participant here. So I'm close to the URL in the in the Slack for you.

LINDA:
That would be great. I'm not familiar with that group, but I do know Kathy, and we've been working with WoSIS and that network to try to get international and domestic terms linked up.

BRUNO BASSO:
It's hard to think that NCIS is not part of that, right? But maybe someone within the agency is is joining, but not talking as often the case.

LINDA:
It's possible, but also not possible.

BRUNO BASSO:
OK. This is very good. We may have to move on to the next question, I don't know how much more if there are last points. We're trying to aggregate all the comments and concerns. We not necessarily have the solution besides maybe multiple approaches I think it's the way. Mark supported the idea that I kind of started at the beginning. Anyway, then any other thoughts on the first point of AI as black box and diverse data streams? OK. We could start tackling the second question then. What are the challenges with integrating models with data? So that's in this way. We can be creative in defining models.

BRUNO BASSO:
First I was obviously kind of biased towards process models with data. By going next to how we parametrized this model, validate, and the question that Phil raised yesterday, how we benchmark the ideas to start creating a set of permanent benchmark insights, and then how we obviously test validate the model with the measurements and so on. So pretty broad category there. But let's see how each one can share ideas or comments on the rise in models and using data input data into models. Steve, please.

STEPHEN:
Yeah, so we do a lot of this here in our work for the US National Greenhouse Gas Inventory, and we have found these Bayesian methods to be really quite promising in our applications. I think one thing that would be nice, one of the things we really need, though, are more measurements to go into this system. We work on nitrous oxide emissions, as well as soil carbon and methane out of rice cultivation here in the US, and we're really data starved, I would say

BRUNO BASSO:
It's, yeah, it's very difficult to, and we need the date and the context of our (INAUDIBLE) model, I'm sure like you would as well, Bruno. We need to not only know kind of those properties in the soils, like texture, pH, et cetera. And the weather data behind that, which we have good data sets for that but then we need to know the land used to management of those sites, and we're kind of restricted down to the experimental sites available to us. And while that's a really great data set in terms of what the LTR and the agricultural research service, as well as land grant institutions have done over the years. It's still a small data set when you think about the domain that you're trying to predict across if you're doing a whole country, on a region within the US and it's, I think right now, kind of a weakness in our available data out there and what we can do, but now these Bayesian methods allow you to pull in those measurements.

BRUNO BASSO:
You basically then run the model with a range of priors for your parameters, and you can actually do this with alternative models. We've been starting to work with the NDC, and Daysen to look into old predictions from soils and looking at that in terms of how well these models are working, if one's better than the other one by first calibrating them to essentially benchmark sites, like, you brought up, Phil mentioned yesterday studying these sites where we haven't really, parameterized worked a whole lot with the models. Hopefully if we can find those, sometimes it's hard given the limited number of sites we have, but then, basically running those models for those sites. And then having that added sample, I mentioned earlier, having an added a sample in to see how those models are working.

BRUNO BASSO:
I've very powerful tools. I think that have evolved in our community over the past decade or so for this purpose. But I think our real weakness right now is we're a data starved. I mean, we (INAUDIBLE) or benchmark sites where there's monitoring of the soil carbon into or whatever it might be that we could then plug into we're looking at hundreds of sites right now, and really, we'd like to have literally thousands, tens of thousands of sites out there that we could do this with.

JOEL DUDECK:
For sure. Yes. Very great point. I agree a hundred percent. Cesar.

MARK:
Yes. Thank you, Bruno.

JOEL DUDECK:
You're welcome.

MARK:
This is around the university of Maryland. I worked with the Epic model and with my team and we have made contributions on these issues of integrated models with data and also participating in large programs and you as well, Bruno can be part of the AgMIP project, which is an international project with many hundreds of scientists participating, and they might be good for all of us to kind of have a retrospective view of what the AgMIP has been doing in terms of integrating models and producing

ensemble predictions and see to what expect, what we can learn from that experience, including, even running just one model by various groups and then seeing what kind of uncertainties they're producing that type of thing that can be really useful for us to try to answer some of these questions of parameterization and validation and so on.

JOEL DUDECK:
Very good. Yeah. I guess I could do that. If there is a general broad interest, I would agree with you, Cesar they will be useful. So, if you'd like me to do it I can, or you'd rather do it.

STEPHEN:
No, you can do that better. I don't know if this is the right place at the moment, but eventually we can expand a bit more.

JOEL DUDECK:
Yeah.

STEPHEN:
Maybe we just need to do as homework, (INAUDIBLE)

JOEL DUDECK:
That's fine. And I can give a very quick overview. And I know some of you are exposed or learnt what AgMIP, Agricultural Modeling Intercomparison and Improvement. It really took the idea more from the climate, the SIMIP, the climate model into a comparison. And yeah, the, was now trying to come together as a modeling different expertise and disciplines of climate, crops and economics to evaluate the impact of climate change. It was initially started in the developing world. We divided the approaches mainly starting with crops. And so, there was a wheat AgMIP, wheat AgMIP maize. So, there is one for every crop and the idea behind this to run multiple models. And in addition to that is that models were first run with a minimum data set, which is basically equivalent to providing all the inputs needed and not necessarily I mean, I know some of the model may be significantly more mechanistic sometimes there wasn't even all the possibility of providing a lot of the data.

JOEL DUDECK:
And so, to me, that's also, kind of a game of the complexity of some of the models, but then the assumptions that you have to make to, that if Mark and I, you use two different values, we are given experience, we will get different answers. And so, anyway, that was one story. In fact, some of these models were completely off the chart in the uncalibrated versions. Then you start, we had phases, phase one, two, three, and four and so on. And then you would improve until you're just about half getting the answer. And some of these models came much, much closer, and it was interesting enough that when you start running projection, then they would fade off again, disappear because there wasn't much of the future knowledge, the way some of the CO2 I am, obviously very supportive.

JOEL DUDECK:
We're trying to do actually an AgMIP USA trying to get the models together for the US, to tackle this kind of thing. It's actually greenhouse gas emissions, carbon to ecosystem services, and so on. So, the

ensemble was really the main feature. And what we learned from the ensemble was that the median of the models, and it was interesting that there were never all a better model necessarily. Yes, there was a group that was kind of always closer, but when you, we did an immense amount of work on statistical design and picking and selecting different models. And so, we really concluded in the end, if you have five models independently chosen the prediction from the median was always better. And we kind of had people saying, "So, OK, you have to have two sloppy models to have a good prediction?"

JOEL DUDECK:
Well, again, it wasn't necessarily like that at the models. There was some, just very much like statistics, the probability of obtaining the reproduction of the most of yields that we tested. I launched another activity, which was AgMIP soil myself. And one of the reason was that in climate change studies soil is not accounted. So, I wanted to even talk about that. How it's not accounted is basically it's reinitialized. The model is reinitialized every single year. So, if you start at field capacity in the Midwest, let's say the full profile of water, you go back so you don't lose nitrates in the fall, or, and obviously for a climate change study 20 years from now, depending on how you manage your soil, you will end up having a completely different soil.

JOEL DUDECK:
And so, we had that AgMIP soil, Epic was part of it. Then we learned a lot by basically trying to tease out, is it the lower amount the residue, you have lower yield because of higher temperature and a shorter development, faster development, shorter growing cycle, but you get is the carbon going down because of lower residue or is it the plants are also going down, I mean, the yield going down because of poor conditions and lower soil organic carbon in all the properties and so on. So anyway, the other thing is we tested because there were mostly crop models how you parameterize genetics. And so, what is the calibration procedure to basically trying to estimate the lot?

JOEL DUDECK:
So, there is a lot that we'll learn there, and I think I'm very glad Cesar you brought this up because the parameterization, there is a rich literature. If you ever want to search about the number of papers that came out, out of this work about how you parameterize genetic coefficient or properties in soils, but one limitation was that we never scaled it. We always used data from experimental station where we had data, we had stuff that we wanted to validate against. And so, I would say, when we go to uncertain territory and we don't really know. And just to conclude, if you, I were to ask you, we ran over what, 25 climate models and 25 crop models. And we ran each of single one, each single model with 25 climates and twenty-five climates and 25 models with one climate where would you expect more variability to come? And I'll tell you the answer is, it came from the crop models. You know, the broad range of results that we got from the crop models was greater than the climate. So that's a fair summary Cesar?

STEPHEN:
Yes, that was great Bruno. Thank you. Just one more thing. It's also, there's another branch of the AgMIP group that is doing global graded climate crop modeling through comparisons. And also, so there's literature in there as well that might help us to understand that piece where we stand in terms of the global scale modeling as a solution.

JOEL DUDECK:
But even though there, the scale is different, but it's still running a point base, alright. So, you have 250, you have the average soil, the average, and you just trying to run the simulation across. So, the assumptions said obviously for the scale are it's a still a point base model.

STEPHEN:
Really is still very course.

JOEL DUDECK:
Yes, exactly extremely course because I've shown variability within fields to be extremely large and temporarily also. Any other experience in parameterizing models or difficulties and validating them and I know it goes back to what, the point also going into chat with Steven and we are indeed very much, that'd be the start that they are limited. So, I'll come back to it, but I have a point how we have tackled that problem, Steven.

BRUNO BASSO:
Hi, my name is Steven (INAUDIBLE). I guess kind of like you were saying a moment ago, we've been looking at other models all of the models, basically, if we make a (INAUDIBLE) products, we compare that benchmark So that would be like our applier or something along those lines. We similarly, we're also kind of wondering if be kind of developed, digital format models, but things like (INAUDIBLE) model itself to try and incorporate the knowledge that we have provided data. But then of course, the question becomes like, how many, how much of that can you incorporate without still maintaining that independence of the variables themselves and the scale which they relate to one another. So, we don't have the answer to that, but that's what we're thinking about and kind of what we're experimenting with.

JOEL DUDECK:
Right. Thank you, Steve. Mark?

ALISON MARKLEIN:
I know Sky has her hand up before me, I don't know if that's left over.

JOEL DUDECK:
She has had it I think since,

BRUNO:
Oh, yeah that's looked over, I'm sorry, I'll take it down.

JOEL DUDECK:
I kind of monitored that. No problem.

ALISON MARKLEIN:
OK, I'm going to get on with it. That was a great summary. Thank you, Bruno. You raised a number of important points as did Steve (INAUDIBLE) about benchmarking and I wonder if this is to bring it back to

the start of this breakout group. This is coming, come back broadly about how do we even approach these approaches, if we're going to start generating a lot of information. So, Joe kind of started this off with raising questions about ML and AI, and I think it kind of highlights from a modeling perspective, whether it's more statistical or the more processed based, how will we expect to use information that would be coming out from a big soil information network? And so, you all throw out a straw individual that you can all feel free to pull apart. I would argue that the ML and the AI approaches will be excellent if they've got all that information for near term local prediction.

ALISON MARKLEIN:
Whereas the process-based models, I would expect them to be really poor at near term local prediction, but they would be excellent as we get more out of sample for regional scale, longer term projections. So, I completely support, the idea of (INAUDIBLE) to comparisons, but borrowing. So, what would we be doing in biology, mostly from the physical sciences, where we represent different understanding in those different models, I'd have a lot more confidence than in the projection certainty, but then they had (INAUDIBLE) needs? But I think the benchmarking needs are really different. Like the benchmarking needs to the AI and ML. You made it to get away with less sites, more locally, more near term. But for the broad scale models, you need many benchmarks as Stephen was saying, but also you want them across a broader distribution of space because what you're really interested in, I think, is the main regional change and not the variation of the noise that would be huge that the models would be projecting at any local scales.

ALISON MARKLEIN:
So, someone can feel free to shoot me down, but I feel like these approaches actually address different needs of different scientists, policymakers, practitioners. I know they're all very mixed groups, but if we're putting it out, some will be thinking about studying climate projections at national levels. And others will be thinking about on the farm level in terms of how that translates to yields. We might want to use different approaches and assemble the data sets differently for those different modeling approaches.

JOEL DUDECK:
Very, very good Mark. I would agree with that. Here at the scale, this is a critical piece. Back to you, Steven.

BRUNO BASSO:
Yeah, I think that's an interesting comment. I don't know if our models, I think there's a couple of scale issues here, right? There's the spatial scale, very local scale versus a larger scale in space. And then there's the temporal scale. What we can break about something now versus what we might be able to predict about it in the future. And I think I agree with what you're saying that it's, I think it's very challenging to predict for a very local scale, like one field in that, at least that's been our experience because there's a lot of things we just don't know about the previous management in particular, most of the time. And that particularly when our modeling is full carbon, that context is really important. Because what we see today can be quite different for a field, depending on how it was managed the last 20, 30, 40 years. And in both, we have two fields that have no till, but they could be at a very different straw, carbon stock level.

BRUNO BASSO:
Depending on, on that previous management. So I do think that makes it very difficult to predict on a local scale. We found that when we predict across larger scales, like a state or our eco region or something like that, a lot of that is noise that kind of disappears as, as we aggregate up and across space. I'm not sure that our process based models are gonna do better in the future if they're not doing well today though, that's one thing I'm not, I'm not fully sure about. If we can't predict right now what's going on in the system, I don't really have a lot of confidence that we can predict out 50, 75, or 100 years from now what's going on. That one I'm not absolutely sure about, but I do agree with the comment that probably process-based models are better for predicting the future.

BRUNO BASSO:
They should be in theory anyway, because they have the underlying dynamics that they're modeling those processes if you will, that allowed them then to represent that climate change effects in the future or something. That the AI system which is highly reliant on current data or previous historical data may not have any information about, what that system is gonna be like 50 years from now. So I do agree that the process based model may, should do a better job for those projections out in the future. I don't know that if the model is not doing a good job today, that it's gonna do a good job out there in the future.

JOEL DUDECK:
Yeah. I think Stephen that's a, definitely a good comment and I agree with you some, but I guess it's a question to you, wouldn't you agree that the, the majority of the complexity of the system as a scientist, we feel that we have a pretty good understanding how soil organic carbon works in a soil. It's a seldom that we don't do a good prediction, because we just don't know enough about that, but rather the lack of proper input to characterize. Like you just said, you know, it's a blind cat shoots. It's like I have no idea how it is and that the initial carbon could be a number on the way down versus on the way up and how you characterize, are you parameterize the pool sizes and stuff like that.

JOEL DUDECK:
So one of the reason I tackle this problem working at field scale and being a purist, I mean, a modeler I thought, what, how can I do much better in getting these inputs? And so when you say we don't know, you know, the previous management, one way that I've overcome that level of uncertainty was by capturing the level of productivity and stability that you heard me and was referred by Jerry and Joe as well on how stable areas. So this area is a relatively good area versus one that is, you know, unproductive or one that fluctuates and some of the work, you know, with Raphael too we learned the areas within fields that they basically do well in a dry year versus. So knowing that spatial variation now at least within the U.S. that information is available. I mean, we, we have subfield variability for 80 million acres at subfield.

JOEL DUDECK:
And so when I showed up quickly, I don't know if it was good enough to be able to catch it, but if you run the model by, driven by soils, you'll miss lots of the areas. If you run it by parameterizing based on the stability math, you capture a lot. And I think in the context of carbon you have to get, I always say that,

you have to get the yields right to be able to know how much rice you use, you returning and the roots and the, you know, the behavior of that. And so that's to me in a summary the, the improving inputs going into the model is more important than ourselves really making this model much more perfect because, you know, we don't understand. There's one thing that we don't understand as much and there's quite a bit of work as, you know, water movement from water table, you know, across the Midwest or drainage and stuff like that. But there is a lots of (INAUDIBLE), if we tackle that by using remote sensing.

So I think of fusion it should not just be ML and models, but much more coupled remote sensing crop modeling to get better input.

BRUNO BASSO:
Yeah. I think that's true Bruno. If we get the crop production or the grass production, of course, production, whatever we're modeling. If we get that right, you know, we're gonna do a pretty good job with this organic matter, but I think back to a study that sander man Baldacci published a few years ago, they were using a simpler model, they had this IPCC model at the time, but they were showing in Australia that, if they don't know where they are on that curve in terms of loss and of course that's Australia, hasn't had as long a tradition of cultivation and agriculture as we have here in the U.S. but, a lot of times they were saying, well, we may just have broken out this land 20 years ago, so if we go to no till we're not gaining carbon, we're just losing less carbon and they're right and I think our studies show the same thing.

BRUNO BASSO:
So I think you're right the production is very important, but we also need to know where we are on this trajectory and that, and that gets back to the land use and management data of that, that we rely on it. And that's another input to our model, right? That we need. So that was my point. But I do agree, I think getting the crop production or whatever system or mowing into model soils is really important.

JOEL DUDECK:
Very good, thanks. Any other point? So we have a 12 minutes for this breakout, this sessions. You'd be learning also when you move I think to the other... Anyway this remaining time should we, could you share ideas? Do you have experience about new data streams, anything that is exciting? I think I need, maybe it's there where the linkage between remote sensing and crop models and the stability maps, really finding us a great deal of opportunities there both to, to reproduce yield as well as ecosystem services. But anyway, you can read the question, let's, let's focus trying to synthesis or something from this group on the point number three. (CROSS TALK).

MARK:
So, I'm really interested in efforts like the smart farm program where they're looking at gaseous emissions, in sort of an, as an integrating factor. So instead of measuring carbon in the soil, you're measuring CO2 fluxes, nitrous oxide fluxes. And, I just kind of, out of the box thinking I'm wondering whether, you know, sort of using AI might be able to use, might be able to dig into the remote sensing data on carbon dioxide emissions that are happening on a regular basis and focusing on, on, on regions where, you know, agriculture is concentrated, dominates and, and, and using AI to see if it's able to pull

some information out, that we don't currently think is, is there with, with these fluxes.

JOEL DUDECK:
Very good. There are ongoing activities, I'm actually part of one or more on the private sector, a large coalition trying to use proxies and trying to get away from self reported data. And so you, Joe you're completely spot on on the tools, you know, AI and remote sensing and, and databases trying to kind of bypass the, the measurements that we'd need to start this Stephen design. So the proxies could be a way to me as new data streams and possibility of, delivering a little bit more. Yes Steve, you have your hand raised.

BRUNO BASSO:
(INAUDIBLE).

JOEL DUDECK:
Right. Thanks for sharing that., I agree. Steve, please. Yes.

STEPHEN:
So to kind of reframe number, the question three here. One of the things that, things I'm thinking about from a, you know, a TNC perspective is not just the alignment of new data streams, but how do we also like come up with new insights stream so to speak. So how do we, how do we align model outputs as well as model inputs? And I think that's one of the things that we're really looking for at TNC, especially folks at TNC who aren't in science roles, you have a hard time interpreting differences in results or insights from models that have different structural assumptions. So I think, again, to, to echo what I just said, like aligning insight streams as well as data strings, I think could have a lot of value.

JOEL DUDECK:
Very much indeed. Yeah. Thanks for your point. I'm sorry if I didn't see your hand, maybe for Stephanie. Stephanie we could remove, a little late now the, the PowerPoint with the questions in case, I have a better view I don't have, a view of everyone. Anyway the, very good point. New insights it's equivalent to, to see, you know, more outcome based of, things that are happening from the system for sure, rather than a new data. The complexity there see Venice, it's always trying to see how you quantify the new outcomes, you know, the new benefits. There has been a discussion throughout, you know, different groups and people about do you, you just, you know, reward, then you adopt practices and just because we know they are good, or do we still have to quantify what are the benefits that they're bringing.

JOEL DUDECK:
So the new insights it's certainly very important point I think for me to see, but there is, we still have to realize if we wanna measure the outcomes at a certain level, to be able to conclude that those practices were beneficial, or we just happy to know that just by increasing biodiversity which we know is right and then should happen, but is, should ever be a way to quantify that to, to really work towards these insight more?

ALISON MARKLEIN:

Bruno can I ask a question?

JOEL DUDECK:
Yeah. So, of course, anyone.

ALISON MARKLEIN:
Relevant to the, to the new data streams it seems like, there are new remote sensing instruments going up constantly and I've heard recently about some amazing super high resolution and I just wonder how, what is the process by which awareness and use of those new instruments is getting into the hands of soil scientists, you know, for model input.

JOEL DUDECK:
You have to be in the system to know. Unfortunately there's, I think there are faces because I work closely with planet to be need a resolution, daily images and I really don't know how much planning is used to do this and how many people. To me I think, being more of a system scientist I'm always looking for the components of the system and how things integrate. Maybe some people working on soils may still, they're silos so I think. Maybe that can be also a problem of people working on soils not necessarily know how much beneficial. I'm generalizing but... And, I agree. New products all the time including fusion products, both radar and optical coming in together. So I think the future is certainly going to be more, you know, interesting from this perspective of using more remote sensing.

JOEL DUDECK:
And the other thing is I mentioned briefly yesterday, geophysics has some opportunity as well, non-destructive, you know. You basically drive the system and you understand possibly depth in the soil and other characteristics, but it always depends on, you know, why you're doing it and who you're doing it for and what's the question you're trying to answer. We got four minutes left, any other last minute thoughts, comments? I thought it was very nice interaction and.

BRUNO:
I was curious about interest in below ground sensors. So, you know, we've talked a lot about N2O production and thinking about, you know, getting it from above ground it's already fluxed out, we know this isn't microbially mediated and, and there's a lot of below ground conditions, plant microbe interactions that might be regulating that and the development of, of high resolution sensors or VOC's and other gaseous products below ground might be useful.

JOEL DUDECK:
100%. The signal is in the soil is looking at that. I know some projects, but for sure, for sure, that's definitely, Christine that's a you, beneficial, you know, for soils and bombed rows and model and understanding what's happening below ground, the hidden half for sure. Thanks very much to everyone, thank you Stephen I enjoyed talking. I thought, hopefully you learned something and Steve please go ahead last minute point.

BRUNO BASSO:
Yeah. I was just gonna add, maybe this is obvious but I think this, you know, this, this work that ARPA is

doing that is really intriguing about new sensors, new ways to measure soils and, and whether it be soil carbon or whatever they're, you know, they're gonna work on next. I'm hoping that those technologies will become part of a campaign in the future and we'll have, you know, more data sets to deal with, with our data problem. Right. So I'm, I'm very encouraged by what I'm seeing there, hopefully that'll, that'll translate into some products out there in the future.

JOEL DUDECK:
Sure, hope so too, I'm optimistic on that. Good point. Well, thanks again everyone and enjoy your break. We're going to break for 15 minutes and I think, you would probably be sent automatically to the next session. Thanks everyone.

CHUCK:
OK. Well, welcome back everyone. Hopefully, we'd done the card shuffle and we've got a new deck here. So this is a breakout A, we're going to talk about measurements, sampling, and archiving - the physical archiving of data, but of soils, and that we had a pretty good discussion in the first breakout. So these are the questions that we're gonna try to go through. And the rules of engagement are that we're, if you raise your hand then I can, I, or Kara the academy staff person will help identify those people. And then those are in Slack. We have a staff person that's monitoring the Slack questions. So with that, again, the purpose of this particular breakout group is to discuss what should be measured in soils. And there's a second question is, where should they be measured, when, and how frequently? So kind of the spatial and the temporal variability.

CHUCK:
Then we had a good discussion on what other metadata is needed to contextualize that data that would collect briefly can talk a little bit about remote sensing, proximal sensing of the data. And then we'll talk about Sol standards. Do we need it in our compatibility and then the physical archiving of the samples. And as I see fills on here, maybe the physical samples, but also maybe even having permanent reference sites as we do new methods in that. So we had a pretty good discussion in that first session on that. So, alright. So with that in mind, so what should we be measuring in soil? And I guess I would caution us not to get down into the weeds. Catherine suggested maybe we'd stay with the roots but not particular detailed methods, you know, how you do soil respiration or how you do, you know, aggregate stability, but kind of stay what are the critical measures and put it in the context of temporal and spatial scales on that.

CHUCK:
So I guess I'll just open it up for if anybody has some observations and again, realizing in a dynamic, so information database there's, as you heard this morning and yesterday, there's different users, there's the land managers. The farmers are probably wanting information, at least during the growing season, maybe monthly scale or even higher resolution than that and at high special resolution or if we're looking at soil erosion. You know, those are, even carbon is more on a decadal type scale. So, think about that. So I guess I'll throw open to the group any quick comments, and then we'll see how we guide the discussion. Alfred, I can count on you.

ALFRED:
Thank you. Yes, you can kind of (INAUDIBLE). So I think what should be measured are, you don't want specific. So I think in broad terms what should be measured is two things. We should measure things that increased our understanding of soils, and we should measure things that increase how soils are to be managed.

CHUCK:
What was the last part? Sorry.

ALFRED:
Sorry. What?

CHUCK:
What was the last comment, things that should be...

ALFRED:
First one is our understanding of soil. The second one is things, we should measure things that increase our understanding of how soils should be managed.

CHUCK:
OK. So what would you, I guess I will get a little bit deeper dive in. Are there some critical physical, chemical, biological parameters that would be a key to understanding the functionality of soils? And of course their management.

ALFRED:
Yes, yeah. Maybe there's about 10 or 15 of them. No, I think we can all list them. I think the question is what do we measure at the moment that we, that is maybe not so useful? And is there things that we measure that we don't measure but that should become regular? I have ideas about that, but I'm sure there's a different list for different people, but I think if we split up, say, we want to increase the understanding of soils, not just for managing, but also as a part of the earth system. And we want to understand how soils ought to be managed. I think that that breaks it up in a, but there's a set of different properties for both of them, I guess. For example, the issue of depth might be very important for the first set of, but it might not so be important for the second set.

CHUCK:
Sure. Yeah. Well, and I guess part of, if we're thinking of dynamic information system, there's going to be specific measurements taken that could be incorporated into a more robust database. But not everybody would want that information, but again, it contributes to the broader science. And that's the challenge is if you've got, you know, an NGO or a agriculture group or forestry take a measurements, how can we bring in that information to create a more robust database? Catherine.

CATHERINE:
So I think the way that the question was phrased was what should be measured in soils. And I just want to, you know, add a layer and figure out whether or not we're willing to get beyond that because as much as there's a lot that I would love to see measured in soils. I think there's also the element, especially when you bring in that notion of management interventions, to want to understand what the effect is on things like yield, for example, or for also wanting to expand more of an ecosystem lens to better capture things like water runoff and nitrification or other biodiversity benefits that might extend outside of measuring the soils themselves. So that's more of a question about scope of discussion.

CHUCK:
Sure. Well, I will tell you just to add to that and then people can comment, but one of the things that we as a committee over this last year hearing, I was surprised by the lack of the soil information systems of using ancillary data, whether it's remote sense to look at landscape elevation, you know, productivity, as you mentioned to extend and contextualize the database, you know, that's in question B here. The lack

of metadata is really apparent, at least in my mind. And in order to determine the function or the monitoring of soils, you need to have that information, I think.

CATHERINE:
I would agree with that and say that it does get at that element of dynamism that we've been wanting to see.

CHUCK:
So, what would be the top one or two physical measurements that you think are really important? I know Alfred, you had your list of 10.

ALFRED:
I would say texture and pH, but I'm sure people can say carbon or yeah. So, I mean, would you like to lose an arm or a leg? That kind of question it is too. I think there's maybe five or six properties that you, if you want to have quick site characterization. I mean, and then we notice properties that we can with pet or sensor function derive from others when it comes to water, for example. But there might be about 10, a minimum of 10, and for a proper site practices, it all depends on what the question is too, of course.

WOMAN:
And I would add that you need more than just surface data. You need to have soil horizon data included.

CHUCK:
OK. Steven, you got your hand up and then fell. I'll come back to you.

STEVEN:
Thanks, Chuck. Yeah, I guess I have kind of a question, so it seems like some of these soil properties are probably not that dynamic maybe, like soil texture, for example, while others might be very dynamic. And I wonder if we're making or distinguishing between those here in this discussion. 'Cause I know we're thinking about a dynamic soil system overall, right?

CHUCK:
Yeah. Yeah. Good question. Well, when we talk about dynamic, it was apparent that people had different interpretations of dynamic, you know, one was dynamic as measuring like CO2 respiration that would occur, you know, change day to day or whatever hour to hour. But there's also the concept of dynamic being different collectors of data would provide into a common database would be at different, you know, it's the system. So information system be dynamic or it could incorporate data as it comes in at different scales or different frequencies and different users. So yeah, but you know, we talked the other, this morning or the previous session, we got a lot of history on chemistry and physical measurements and, you know, we could use that, but the biology may be a little bit less robust in that, but yeah. Phil, you had a comment/question.

PHIL:
Well, I was just going to, I mean, to answer the question what should be measured, it really depends on

who you ask, who's in the room, right. And agronomist is going to have a different answer or a somewhat different answer than a small chemist who's gonna answer this differently than, you know, (UNKNOWN) a modeler, and it all relates to end use. I think, you know, whether you're trying to understand you know, as Alfred comes back, you know, what you want to understand is you're just trying to understand soil fertility, you're trying to understand carbon sequestration, greenhouse gas production, pesticide transmission, you know, and I would approach this from, in part from asking, from the standpoint of asking what aren't we measuring now or what aren't we measuring well enough now because we have good, you know, good basic measurements for the 10 or 15 things that Alfred could rattle off the top of his head. I'm sure all of us could to some degree.

PHIL:
And one thing we're not measuring and just comes back, I think, to Cheryl's point is depth and that we've got an enormous amount of information on surface soil sample, surface soil characterizations, but so little on depth. And this has really come back to bite us as we've thought about. So a carbon sequestration, for example, under different tillage systems. And we have arguments coming up about soil carbon disappearance at depth. Why, you know, I'm personally a skeptic of that, but there are very few long-term datasets where we have good depth distributions of soil carbon under management. You know, likewise, we have too little information on soil biology, but that's to be expected because that field is changing so rapidly and I'm sure it will.

PHIL:
10 years from now we'll want to have things measured very differently than we're measuring them now, perhaps. So I guess I would, you know, I think trying to develop a laundry list of prioritized measurements may not be the best approach right now. But I do think that, you know, as a group, we might identify those things that could be done better that aren't being, or that aren't being done at all now for further consideration. And I would, you know, throw out those two as two candidates.

CHUCK:
They have them. What was the other one?

PHIL:
Well was soil biology. And there, I think, it's going to be more important to store soils than to measure them right now, you know, to find a way so that will, you know, to find enough minus 80 capacity that we can store things until, you know, until it becomes more practical to sequence things as easily as we can measure soil pH, which I expect we'll get to in 10 years, but yeah.

CHUCK:
Well, that's, yeah, that gets back down to the last question, but yeah. You know, maybe there'll be extra minus 80s after the COVID.

PHIL:
Well, that's one positive spin. I thought of the same thing though. We tried to buy a minus 80 a couple of weeks ago after one broke down and, you know, it used to cost $10,000 now it costs $23,000 because of the demand, but I'm sure what it would've gone down to $6,000 in two years.

CHUCK:
But then that, you know, I get slammed just trying to find storage space for dry samples. Now I have space for minus 80s is even harder to justify. So I guess if biology is, so I would agree. I think what you, you know, and Alfred you know, adapt is really critical. You know, we're finding some things down at depth that I didn't expect. I guess the other question is on the biology, you mentioned the metagenomics. Are there other, I asked this in the previous session, are there other biological type measurements, you know, form versus function, biodiversity's town, who's there. And the potential, I guess, the other is, are there other measurements that we don't have enough of or should be taking to understand the function, the other, that soil?

PHIL:
Well, we're learning an awful lot about the importance of soil pores of different sizes and their distribution and continuity in terms of moving carbon onto mineral surfaces, for example. So, and, you know, at three years we made a wish we had so pore distributions for many soils.

CHUCK:
Are there measurements that we should be taking at the time that that soil was sampled, that can't be archived? So I'm leading the discussion, but I asked session, they talked about bulk density. You know, it's something almost, unless you're going to store and tax all cores, which makes it even harder.

CHUCK:
Are there some things that are key measurements that we can't take or you know, 20 years from now?

ALFRED:
To me, it was funny that you mentioned bulk density because that's probably one of the courses and easier, not easy, but one of the courses to have simple measurements to think about, but it's also one of the hardest to take it turns out it screws up so many soil carbon determinations, it's unbelievable, 10% air bulk density just in spatial variability. For example, can make the difference between a soil gain and a soil carbon loss over 10 years. We don't pay enough attention to it. That's for sure.

CHUCK:
Go ahead Alfred.

CATHERINE:
I would like to ask a bigger question Chuck. About not what we measure and what we don't measure, but maybe what we lack. And apart from depth and proper site characterization and sampling soils, instead of vegetations. We liked sufficient teary a new framework for a data collection. I mean, everyone collects data and under the umbrella of climate change or food production or whatever it is, but we let new theories, I think we would have more new theories and maybe you can answer that for soil microbiology. But for this very few sub-disciplines in soil science and actively work on the development of new theory that prompts data collection. Because theory needs data, as we know. So, I wonder, I wonder how many sub-disciplines in soil science work on new theory and in that light are going to collect new data?

WOMAN:
I know there's actually been a lot of innovation in that, from the perspective of soil carbon, like this increasing appreciation over the last 10 years. That there's actually a lot of, very easy to decompose soil carbon and soil. And it's that it's protected by minerals and it's stabilized and aggregates. And I feel like there has been kind of a paradigm shift. In our theoretical understanding of why organic matters sticks around and how you can sequester it in soil. And then arguably, I think that's contributed to the proliferation of some of the tools and things people were talking about in measurement types. We heard about it in the plenaries.

WOMAN:
And so, use their facility, right? Like all these fancy mass spectrometer and imaging techniques and ways to really try to get at where is carbon sticking in mineral in aggregates to minerals what are those molecular scale interactions. In terms of how we could predict stability or destability? I know that there's questions of how scalable is that, right? If we want to think about like predicting and modeling and managing, you know, are those really nano, micro scale measurements, so integratable into like larger scale frameworks, but I think, I don't know, that's just one example Alfred, like I think you're right. That things like that are what that paradigm shift is, has led to all this new data collection to try to really unpack what's going on in the black box as well from a soil carbon perspective. Anyway.

CATHERINE:
Can I respond to that?

CHUCK:
Sure. Go ahead.

CATHERINE:
I think it's a great example in that, you know, of course I think of the work of Schmitt and Lehman who sort of prompted it But I'm thinking it's one example. It would be great if we had 50 of them so. I think in pathology I can think of maybe one or two, but I think in all fairness, all of us who are aware of the sort of science literature of the 1950s or the 1920s, probably even a better time so the theory development drove the data collection for a large extent, people had ideas and hypothesis. And I think we were short of that. I think there's a ton of data. There's a lack of ideas and there's certainly a lack of theory.

CHUCK:
I would maybe respond the other way is if you had a robust dataset, you could develop new models and theories.

CATHERINE:
Yeah, I agree. That's the way of doing it. I think that's what we're trying to do with the universal soil classification. But in many sciences, the theory drives to data collection. And I don't know whether we do sufficient. You can probably can detail a little bit about it in so microbiology check, whether there is sufficient new development interior, or rather there's a lot of parroting as I would call it.

STEVEN:
I guess, I think I agree out in front of the scientific community, we certainly need to be thinking about theory and moving that theory forward. Right? I think that's obvious to us to work in that area. But I think this database would also serve the purpose of applications right as well. So that might be a little bit different dataset, which maybe is what Chuck is thinking about. That there's a broader need here than just theory. Although I do agree theory is important. As scientists, we need to be moving that forward and hopefully data like this could serve that purpose.

CHUCK:
Yeah. So, I guess the question is, and it goes back to funding. I'll put this in context. What started this whole idea this workshop? ` Alfred was back in 2015. We had the world the status soils resource or the state of soils. And it became apparent when the North America was putting together their state of soils that we didn't and US didn't have the right kind of information to document how our soils were changing. And Canadians, no offense, the Canadians, they actually have better data than we did, which kind of surprised me when I thought we had, you know, we couldn't determine what our erosion rates or document that or some of the other things. So that's really what started this whole idea. And so, there's a lot of end-users that are needing whether it's NRCS or farmers, or, you know soil resources, forestry, whatever that need to know how soils are changing. And that's probably, you know, there's the academic component. But then there's the end-user and that's gonna probably maybe fun or where we collect information is gonna be driven by those questions. And somebody, Catherine, did you have your hand up?

I I'm trying to multipurpose here.

CATHERINE:
I would likely respond to Lisa then Chuck.

CHUCK:
What time?

CATHERINE:
I said, I would like to respond to that point.

PHIL:
Go ahead Alfred.

CATHERINE:
I think what I see a lot here is that there is a lot of data collection by end-users and lent users, so to speak. There's a lot of farmers here that do EMM surveys. There's a lot of sampling going on. There's a lot of pivots that are driven by bright technology. And I believe, but I can't speak for the whole of the US nor the world. But I believe the technology is ahead of the science in many places. I think the science is much behind with the technology. And I think it's a real issue when it comes to data collection, because there's a ton of data, but the scientific framework or understanding, or even validation of the data is behind. So that is my response to the land users want it, but there's a lot of land users that collect a ton

of data themselves, and then have some sort of a black box algorithm that adjusts pivots to the irrigation needs.

CHUCK:
Well, I guess that's part of the question. Can we get that data from different sources to get a more robust database or there was a question on the chat Slack that while suggests that we need to be measuring carbon fractions of as a measurement, and that came up in the previous discussion as well. I'm just kind of sharing that. Kathryn, did you have?

KATHRYN ELMES:
Yeah, I would chime in, on picking up on a couple of threads in terms of, you know, whether it's thinking through the theory and what might drive theory or help support that or thinking through what is actually the end use of the data collected and who are the users. If I can come in with just a pragmatic example for those of you who don't know me, I'm with a company called Indigo Ag, we're deeply interested in, on farm greenhouse gas emissions, and a lot of that really stems from soils. And what we want to understand, if I can put it in broad strokes is, we want to understand soil activity. So, Alfred, I liked your initial overview of those two different items that we want to understand soil properties. And we want to understand management practices. I'd say we want to understand soil activity, and we want to understand on farm-greenhouse gas fluxes, and there's definitely an interaction between those and management practices.

KATHRYN ELMES:
So obviously we do want to understand the impacts of specific management practices in specific geographies, in specific climates, on specific crop types, et cetera. And I can go on and go on, but there's so much variability that we would love to be able to capture and really understand. So, I'd say, if I can loop back to Chuck's initial question. Of what are the data points that we think are of utmost importance to achieve this top of our list really are as mentioned soil organic carbon and bulk density. Yes, ideally, perfect world scenario. We could get down to meet her multi-meter depths and high frequency of measurement, both temporarily and spatially., But we'd also be interested in that top tier in texture and pH and wet aggregate stability. And then beyond that, if I had to talk about a next tier, it would be things like soil moisture and temperature, since they heavily influenced soil activity and nitrous oxide emissions.

KATHRYN ELMES:
And then of course, outside of soil, it would also be great to have, more sources of general emissions data on carbon dioxide, nitrous oxide, and methane from ag lands to, you know, better understand, for example, if we're feeding into models. You know, how are they working and where can they be better calibrated? So, I throw that out there just as more, if it's helpful to have a pragmatic example of what is an end use case. And, you know, we are just one end user, and I imagine additional end users who would potentially benefit from this would be the agronomist who could then provide insights to farmers on what are those appropriate management practices for their very specific field sites and situations.

CHUCK:
Thanks, Kathryn. Kristen, you are up next. Got a hand. You're muted.

KRISTEN HOFMOCKEL:
Alright. Thanks. Yeah. This is a great conversation. And I would agree with a lot of that list that you just put out there, Kathryn. One thing that I think is interesting to think about is, what bulk density, how critically important it is and how it can be really tricky to measure. So, I think that goes back to maybe some of the conversations this morning about having really standard methods for that and making sure that we're getting the best numbers that we can. But then thinking about also, so data that might not be as hard fought for. And if we, with new technologies, could get things like sensors in the ground where we could be getting real time and concentrations and trying to better understand what's driving those fluxes. And I think that there's a lot on the on the horizon in terms of some sensors that we could maybe leverage to have better data streams for helping us to understand how management is really influencing some of these things like greenhouse gas emissions.

KRISTEN HOFMOCKEL:
The other thing I was curious about to hear people's perspective on that maybe hasn't been mentioned, but relates to some of these other things is soil mineralogy. This is a hard fought for data, but it's, you know, texture only gets us so far and we can have soils that are very different in texture, but still have, you know, aren't changing in carbon very much. And it could be because of that mineralogy. And I think the mineralogy also plays into micronutrient availability and plant productivity and a lot of other aspects. And so, I was, you know, I agree with that top list. I was curious about people's thoughts on mineralogy and if that was a priority or not. And if it would help us to maybe understand some of these other bulk measurements like SOC, you know, I think that's really important to have a baseline so we can detect change, but if we want to understand why that change is happening in one place versus another, I'm not sure texture is gonna get us there.

CHUCK:
Yeah, good point.

CATHERINE:
Can I answer to that?

CHUCK:
Let me go, sorry. Can I move around. So, Cheryl?

CHERYL PORTER:
Yeah, I'm going to simultaneously braid several threads and also go off on a tangent. So, there's so many ways that a dynamic will information system could go, and if you try to do everything you're bound to fail. So, it's, I think it's really important to start simple. And that would be with, you know, first of all, simple list of soil properties that are mostly already calculated, but maybe add some complication you know, going to depth and adding a few variables that aren't commonly collected currently. But I also think that, you know, we've got some, some real grand challenges that we're facing and climate change is the biggest one and soil is going to be a huge part of the solution. You know, if we look at soil carbon sequestration. And it's also part of the problem when we look at the breakdown of organic matter and greenhouse gas emissions. So, if we could maybe perhaps focus on this as a grand challenge and maybe

decide the variables that are really needed to go forward and solving this big grand challenge, it's gonna be a major focus of soil research in the coming decade.

I, believe so. I don't know.

It just, a thought,

CHUCK:
Yeah. That's good. Cesar are you listening. I see you're online. So, he's already there, maybe not.

CESAR IZAURRALDE:
No. Chuck, am here.

CHUCK:
OK. Cesar, you do a lot of land use, carbon nitrogen flux modeling, erosion modeling, I guess, from your perspective you know, if you had a robust soil information system, what do you need? Or would you like to have? That's not there. I think Phil asked the question, what are we not measuring, but what would be helpful in the modeling to integrate the things that you're doing?

CESAR IZAURRALDE:
Well, ideally you know, currently we're finishing…

CHUCK:
The project that funded by NASA on estimating the carbon fluxes in crop plants across the United States, and one of the things that we needed quite a bit was the validation sites, especially with flux of data respiration and uptake of carbon. Now, we also model at lateral fluxes as well, but those in the end, they have uncertainty and depends on how well we can model erosion especially if you have some slopes and gradient and some length of slopes. You have some, you know, we predict that as well. Also, we predict carbon leaching as well in the systems or corn or for soybeans and so on. But the main thing is also the other thing that we know we always lack very good information is that we do have these soil databases, but we don't know exactly the partition of the carboon pools in terms of how it could more precise parameterization.

There's always the uncertainly there, we can spin up the model to a certain level, but that's remains, you know, we start the simulations for a given year and so on and we get the numbers but still having knowledge of the current status of those pools, because one thing is to get the data, the carbon stocks or the current concentration from the databases, the other thing is for the real farms is, you know, what is the dynamic of that system for a particular farm? And we are doing high resolution modelling and it's very nice, but still, we can do it. It still remains that uncertainty that we have and how close we can get to characterize that pool of system in terms of the current dynamics.

I stop here, maybe you…

ALFRED:

Yeah, well, somebody just jumped on Slack and said that erosion validation is a huge, huge need.

ALFRED:
(CROSSTALK)

CHUCK:
Yeah, we did some studies at the Coshocton Watershed by validating the model that we use, the epic model with real data, soil sediments, that have been collected at Coshocton sub watersheds. And, you know, when you have the data especially that have been collected for decades, then, I think, you can approximate how much particulate carbon you're losing and also get an estimate of how much soil carbon you are losing as well. But those remain small fluxes, still, compared to the major vertical fluxes that you have.

ALFRED:
Stephen Ogle and Phil, do you have any comments on that? Since you kind of done some big scale ecosystem type analysis? What would you like to have?

CATHERINE:
That's a good question, Chuck.

CATHERINE:
(LAUGHTER)

CATHERINE:
I mean, I can tell you what we used, but I think some others like Kristen here, but maybe what the next generation of models might be using. You know, we're trying to work towards that with the memes model here, but we're not quite there yet. But, yeah, I mean, we need texture, but I agree mineralogy, I think that's going to be the next generation model thing out there. Ph was brought up by Alfred, I think that's important bulk density. I mean, Phil talked about that. That's really, really critical for our modeling. You know, we sold that information across the horizons, like somebody else mentioned earlier, is important. How does this vary down deeper in the soil and, of course, the carbon measurements, the soil carbon, and we don't look as much of the inorganic carbon but out here in the west where I live, it's more important and something hopefully in the future we have better theory around, as Alfred brought up and be able to model that better.

And any information about the nutrients in the soil, the cations can be, so, we're not using those currently in our model, but we know that that's important. So, right off the top of my head, those are the things that I would think of and then there's the ancillary data, that's another question here.

So, I don't know if you're...

ALFRED:
I was going to try to transfer to that.

CATHERINE:
Well, maybe since I'm talking to you, I just add in to this ancillary data. You know, we have pretty good weather data sets out there. I think they're sufficient for what we're trying to do for the most part. But the land use to manage in particular, the management data is really more difficult to come by and in the context around dynamic soils information system, we get a measurement on something but then don't know about the management of the site or as important, I would argue, is this history of the management of the site for the last maybe two, three or four decades. It's not as useful, definitely not as useful to not have that additional information there. So, those are some things (INAUDIBLE). I'll pass over to Phil now. He may have some things he'd like to add.

WOMAN:
Well, I'm looking around the screen to see if there's another Phil in the room. (CHUCKLES) I'm not a modeler and process modelers have a long list of things they would love to have and need to have in many cases then, you know, I know that mineralogy, for example, as Kirsten brought up, can be key for some modeling, some processes, especially if you're crossing into highly weather soils that have variable charge mineralogy. One thing we haven't mentioned yet, though, with respect to remodeling is we've been doing a fair amount of machine learning predictions, statistical modeling, really of trace (INAUDIBLE), which are very, very difficult to model at the process level but we're having surprising success at modeling, using machine learning approaches. Of course, you need a very dense data in order to develop a good training set, but when you have that, we found that we can improve the predictions for trace gases, for example, for nitrous oxide anyway, three-fold over what an untrained process level model or should I say (INAUDIBLE) process level model will do for insights.

So ,and it's, you know, what machine learning tells you in this case is that it will tell you the reduced set of variance that you need in order to make informed predictions and with a surprisingly few set of predictors like five or six, we were able to make this know, to predict with 50 % or more fidelity that fluxes coming out of sites that have not been used to train the model.

WOMAN:
So, it's you know, I think we're learning a lot about what properties are important from machine learning approaches that really have not, we haven't even started to apply them, really, to soil predictions or soil process predictions yet. And so, I think, you know, in five or six years when we've got a lot more experience at this level, we may have a very reduced set of properties that we think are key and there may be some surprises there that are especially valuable.

ALFRED:
So, that kind of lens into one of the other questions is the proximal remote sensing. You know, I've been on a couple of workshops looking at a microchip (INAUDIBLE) soil and can sense things every second – water, temperature, maybe even CO2 or whatever. But then there's the remote sensing large scale and some of you been involved in fires scaling up. How, I must say, how useful… What would you like to see out of that kind of endeavor, particularly the proximal sensing where microchips in the ground that can measure everything everywhere?

WOMAN:

Well, not looking too far, not trying to be too futuristic here. We don't know a lot of things that you would think we should know now, simple things like management history, which can affect processes. We cannot, to my knowledge, we cannot detect no-till, for example, remotely now with any with any level of veracity. And how simple is that? So, to be able to predict continuous no-till would be, I think, pretty important if you're trying to do a regional level analysis, a regional level prediction of what soil carbon is doing. And that's, to me, that's about as simple a management variant as you might be able to collect, but yet we can't yet do it.

STEVEN:
I would say that I think we are... I agree that there's a lot of advancement that's needed, but I think we're moving in the right direction and I have cause for optimism. I hope in that maybe we might not be able to do the no-till flawlessly but what we definitely can look into remotely is tillage. So, was there a tillage event? That is something that we might be able to remotely detect and validate the ground truth information provided by growers when they say this is the practice that we carried out and these are the dates we can use remote sensing to validate and whether it's for tillage or for things like cover crops, I'm really optimistic about remote sensing to be used for cover crop detection. We need to be able to, I mean, as much as I want to be optimistic about all of the management, history and data that we get, there is always bound to be potential for error with the management history data provided. So, if someone says, "Hey, we purchased a cover crop seeds here," (INAUDIBLE), that's great.

On top of that, we can go back and validate there's been a cover crop in the ground. So, I'm definitely optimistic, I think, about the remote sensing capabilities. I agree with you, Phil, that there definitely needs to be additional advancement there.

STEVEN:
And then on the proximal sensing check, I'm excited about that also. That's something where I don't recall who mentioned the monitoring of trace gas fluxes, but, you know, if we have these proximal sensors and we can very easily check in on those when we see something maybe like a pulse, it would be fascinating to see what's actually happening in the ground at the same time. So, I think there's a lot of discovery that can be done once these data sources are available.

ALFRED:
Thanks, Kathryn. I don't know where I'm in the last. Stephen and then Alfred.

CATHERINE:
Thanks, Chuck. I just I would mention here on the no-till and the cover crop side, there is a group at New Hampshire led by Bill (UNKNOWN) and Steve Hagan, and they actually do have a product now on cover crops and no-till that, I think, is quite promising and looks like a nice product for (INAUDIBLE). It is wall to wall. So, that might be something to check into, those who are interested in those particular practices.

CATHERINE:

You know, the other thing to think about here is in a dynamic soil information system is that maybe there's, you know, and I don't know how this is going to work out. I don't know if any of us here on the call do but, you know, it might be that there's a certain subset of sites or a certain subset of areas where we collect more detailed information like we're talking about with soil centers, et cetera, that have been brought up at least a couple of times now. Maybe that's not information you collect across the whole network of sites that are in your information system but there might be a subsample where you collect that additional information, you know, obviously in a strategic way, probably in some way using statistics to get a robust subsample that you would be doing on that. It's just a thought and it's not going to occur to me where this is going exactly but it might be something to think about.

ALFRED:
So, just to respond to that. I think part of the idea of this workshop and the potential product is there are different groups and specific projects are collecting information that I don't think would go, well, the primary goal isn't develop a complete holistic network, the idea is where how can we compile different systems into a network back then could be put into a database and then provide information to different end users, so like the NRCS is taking information, they're taking about every five years, they do their soils inventory. But then there's somebody like Kathryn that might be taking more specific, site specific information on carbon and carbon sequestration or something like (INAUDIBLE) or Phil's doing on nitrous oxide flux that would feed into. So, how do you, I don't think we'll ever get to the point where we're going to have a $100 million and say, “OK, we're going to have this complete network.” I mean, that would be nice. But I think it's how do you collect information and put it into a system that could be make the soil information more robust?

CATHERINE:
So, if I could just, that sounds… that's probably what I was thinking it would be too, Chuck. It might be good if there's some way to relate these data sets to each other. You know, I think of the NRCS data sets is kind of nationwide here in the US. Of course, there's the other data sets from ISRIC and others that are global and then kind of encased within that, you have sites where you have more detailed information and to think of it that way might help organize that soil information in the system. But anyway, I'll stop there. I think, Alfred wanted to say something.

ALFRED:
Alfred and Kirsten.

PHIL:
So, on data, I think, you know, we in the digital source mapping community know where to shop for data, whether that's USGS or whether that's the hydrology labs or NRCS or anything. So, there is a lot of data around and a lot of particularly when it comes to the covariates that we use, we know where to shop for data. Is it on all different platforms and different sites in different formats? Yes. Is that a problem? No. We've learned to deal with it. We've incorporated it. We've got the ways to translate it. Is it a problem? I think harmonizing it is maybe very noble, but the data is developing faster than you can harmonize it. That would be my observation. On the remote sensing. I think there is so much remote sensing and data, particularly the Sentinel, if you use the Sentinel data. I mean, that is wonderful data that anyone should use that works with soil moisture. On the proximal sensors, maybe Kristen, you

know, we just put a grand in for a handheld XRD…

CESAR IZAURRALDE:
I mean, those things used to be the size of a of a small room, right? Remember? But now (UNKNOWN) makes a hand-held (UNKNOWN). We've got free XRS and we got a handheld (UNKNOWN). There's lots of little instruments that we take to the field all the time. I think we collect more time. We collect more data within the little group that I have that then perhaps an NRCS unit 30 years ago. I'm quite sure about that. We collect more data. The point is, how do we have a state? I would love to give it all away. We give it away to everyone who wants it. But there is not like the physicists, they have a repository where you can dump all your data. We have minimum meta data. But anyway, my first point was, if you want to harmonize the data, that's gonna take longer than and people might not be interested in harmonizing the data before they deposit it.

SPEAKER 2:
OK, Kirsten, and then actually, that leads into another question, the question on the…You're muted again. There you go.

KIRSTEN:
Alfred, that's the great lead in to kind of the same thing that I was thinking is how do we make it easy for people to give the data, right? So for all of us researchers can go to our long-term ecological research sites and generate the data. But it's not really necessarily where we need the data coming from. It's the working farms, right? And so what if we leverage the technology? I mean, OK, we can sample substrate from Mars. We've got to be able to generate data on local farms, right? So if there's an app or something like just pH, for example, because it came up, we can do a rapid test. We do it in all our basic biology tests with the pH strips, right? If there are ways that farmers could rapidly generate data, and your handheld devices are a perfect example of ways that we might be able to empower farmers to give those data so that we can better understand how to increase soil fertility and all of the other benefits of the soil and carbon storage and all the good things that we want to do.

KIRSTEN:
So I really like the way you're thinking about this. I think that's possible. I think that's possible now if we put our energy towards that.

SPEAKER 2:
So so actually it leads…Oh, well, Catherine, you got a quick comment?

CATHERINE:
Just quickly. Yes, absolutely, that's possible and the number of data points that we're already getting and can get from sensors are that on farm right now that should be accessible is massive.

SPEAKER 2:
So that actually, it kind of leads into quickly next set of questions is there's new methods coming about, there's a standard method, and Luca (UNKNOWN) mentioned yesterday about the lack of, well, I don't wanna say compatibility, but they ended up using one lab because of the variability between labs. What

do we need to do to standardize the methods, should we, or is it more harmonization, as Alfred mentioned? And if you got different user groups, they're gonna have different methods as well. So how should we go about standardizing or harmonizing?

CESAR IZAURRALDE:
I think the first thing I want to say to that, I don't believe (UNKNOWN). I think that's a piece of sort of...I mean, in these times to say that pH units, whatever he said, I don't believe any of that. I think there was a gross exaggeration. There was a political move to have it all done in one layer. So I don't think I want to see that data before he says that. He said that many times.

CESAR IZAURRALDE:
I think it's kind kind of belittling. So there is variation between lips labs, certainly when the method is more refined. But in our standard methods, like it's pH to two and a half of one to one, I mean, there is within two 2% I'm quite sure of about So that's my first response response. The other to is think people are going gonna different things things, certainly us in academia, we're not going gonna according to standards. We're going to use standards and then we're going gonna them and change them and involve evolve because we think it's better if you use a little bit of that And that makes that things change and it makes that some of the data are not as easily exchangeable as perhaps as they should be. But as long as we we share the data data, what Kirsten which we should, if we share the data and we say, alright, we've adopted the method a little bit, bit or tweaked it a little bit bit, then can see how it compares to their data.

SPEAKER 2:
Any other experiences comments on that? Kirsten?

KIRSTEN:
I think that the standard data is really nice for interpretation, right? It takes away a lot of the variability. So I think somewhere it came up like soil aggregation or something. There's a lot of different ways that people measure that. And it matters, and it matters a lot, right, in terms of what you interpret it. And I think from the research side, it's hard to standardize because we are adapting the method to get maybe a specific mechanism or to target specific hypotheses. And I think that's different than the end user of we want data streams coming off farms in order to inform management decisions. And so I guess I see them as two different goals. And I do think there's a big value to saying, OK, there's a standard test. And maybe we could move into something more sophisticated than an early spring nitrate test, right, that seems like a pretty low bar of something that we could do and come up with a standard test that's more informative, right?

KIRSTEN:
But the on the flip side, for research purposes, I guess, you don't think that's going to help us to innovate and to move theory forward, going back to what Alfred was saying earlier.

SPEAKER 2:
OK, right. There's a proliferation of methods. I think the key point is how do you compare the methods or compare the data, that's gonna be really key and maybe there are some techniques, machine learning

or some other ways even to kind of...

CESAR IZAURRALDE:
You used to have a thing called (UNKNOWN) and (UNKNOWN) and all these different (UNKNOWN). And we've developed these simple, better transfer functions to compare them. I'm sure you can do probably in the genomics work, you can do something like that. There is probably if you're creative.

SPEAKER 2:
and the results were completely different. And so I think there is some issues as far as standardization or QA, QC.

CATHERINE:
And I do think some of those issues might proliferate as on farm devices and sensors and measurement opportunities increase. So I'm very optimistic about handheld spectroscopic sensors and the values that those could bring. But when you're talking about so many devices and having to not only deal with calibration regionally, but calibration transfer, I think that issue is only going to exacerbate. Not that it can't be overcome. As Alfred pointed out, it has been overcome in the past with other approaches. But it's something that we should be cognizant of.

SPEAKER 2:
So, OK, we've got a few minutes left and actually the question came in on Slack, and that relates to our last comment or question here is from Slack it says, physical archiving, the space is decreasing US and GS and NRCS offices are being forced to get rid of samples.

SPEAKER 2:
is true in academia. I have a 30 year social soil archive from 30 plus years. And I got administrators always asking, why do you need to keep those mason jars up in the storage room of dirt? And so I guess the question is. is are the hurdles and how do we archive those samples, samples? mentioned we need a bunch of minus 80 degree freezers. You What's the and opportunities there for archiving and and the physical archive and the samples, not the data, that's another break out group.

CESAR IZAURRALDE:
I think there's a giant story somewhere for plant seeds. I forgot where it was, it might be Greenland or they have all the...

SPEAKER 2:
That's in Norway.

CESAR IZAURRALDE:
In Norway, yeah, I think we should look into that model. There might be something in that model they found. They must have the finance for it. And if we can store potatoes, for example, we I'm sure we can store soil samples. It's just the reason the funders might not be convinced that it's a good idea. But these folks have convinced the funders that we need to store potatoes. So I believe it's a very applicable model that we should look into and then we make.

SPEAKER 2:
Samantha?

SAMANTHA:
Yeah, I think it's been interesting 'cause with NEON we do have a very large biological repository where very diverse types of environmental samples are going, including soils. And we've partnered with the museum at Arizona State University, and it's been interesting working with the museum because museums, they're used to archiving specimens, but they are from an organism. So it's definitely been, I think, a paradigm shift for them to take jars of soil, bottles of ground up leaf litter, tiny little archive vials with two grams of soil for future proteomics and RNA sequencing and other stuff and get these big liquid nitrogen cryo doer's. So we've had working with them, but it costs several million dollars a year. And it's not at all trivial to spin up such archives.

SAMANTHA:
But that's kind of the way that we have gone, is to say, well, museums are really good at this, at curating samples and having good databases and storing them and loaning them and distributing them. But it's a it's than what they're used to. But not to say that they can't help us do that for soil archives as well.

SPEAKER 3:
Well, I really like this idea of thinking big and being bold, and it might take a salt mine in order to create a small repository that would be capable of storing the soils that we want to have. But why not?

SPEAKER 2:
We got a huge salt mine in Kansas, Kansas City area that they store records. (LAUGHS)

SPEAKER 3:
Yeah, (UNKNOWN). You could fit an Astrodome inside one of them.

SPEAKER 2:
Yeah. Good, that's a lot. Let's see, I think, oh, Kirsten, yeah.

KIRSTEN:
Yeah, I think complementary to that is to think about what we want to use those for in addition to archiving it, right. So we all have these big archives and I think even now making them available to other people, right, I think that could be really powerful. So some of the experiments that I know we all have archives are long-term ecological experiments where there's lots of data that go along with those samples. And so I think concurrent with thinking about how to give our samples to a storage facility is thinking also about what goes along with that and how to encourage the community to use those samples.

SPEAKER 2:
Any other comments on archiving it? I guess I go back...

SAMANTHA:
Just to follow up really quickly on what Kirsten said, I think the NEON soil samples have been really popular. They immediately people have picked up and wanted to use them because of that contextual metadata. I think we could say we've already measured these 20 parameters.

SAMANTHA:
Oh, but you want to know where the carbon is in the fractions or you want to know about mineral associated carbon. Awesome. And so I think that's such a good point that obviously with the archive would need to be linkages to all that rich contextual metadata that makes those samples so very valuable.

SPEAKER 2:
And Samantha and and Phil, you mentioned this yesterday, but not only is it the physical archiving of soil samples, but do we need to kind of have a set of reference sites that as new methods or new information comes available, that we can go back to and calibrate methods or learn new information, the theories that Alfred was talking about so that we can link old and new methods are old information or new information? The NEON, I forgot what the time frame, is it 20 years or so that it's in place. But that's mostly on native sites. Is it robust enough that encompasses the the solar variability and the climate variability that would be useful for an information network? (UNKNOWN) you're shaking your head. Yeah.

SPEAKER 3:
Yeah, no, I would just second that. Of course, I think that we have two or three large national networks in the US of and, course there's in some Canada and Europe, but if we're just talking about the U.S., bringing just in those networks, the LTTE are (UNKNOWN) NEON, LTA LTAR are (UNKNOWN), right there you've got four networks that would cover probably most of the soils in the US.

SPEAKER 2:
Right, and I argued in the last session, the land-grant university Network, if they set aside one acre, that could be used as reference, (UNKNOWN) at least what, 80, 90 land-grant universities, that would encompass a lot of variability, in addition to the LGERs and LTARs. What would it take to just set up a site just for future resources? Of course, then somebody told me, I need to write a grant in the South.

SPEAKER 2:
So, yeah, alright, I think it's 1:29, I think we're supposed to break at, well, my time, 1:30. So, so we'll finish up, up. You'll the rest of the afternoon, evening, morning off, off. Some us have to do a synthesis debrief here, but we will reconvene in the morning and be sure to rejoin. And that I really appreciate your participation in discussion. And we'll talk to you tomorrow. Thank you, everyone.

RANVEER CHANDRA:
So welcome back and while more people are joining in, we had a very active discussion in the first session of this breakout session where we discussed multiple things about who the different stakeholders are. How do we get consistent with all the different stakeholders across academia, government, private sector, farmers, and how do we get everyone on board? And how do we incentivize everyone to start sharing data? We also talked about what level of data needs to be shared, whether it should be a distributed database, how do you annotate the data, which becomes very important. And also about what's the metadata that's needed in the database. So we could continue discussion on all of that.

RANVEER CHANDRA:
And we also talked about creating like a soil ontology, the eCit work that Kathy's been leading and, the value of that, of creating an ontology so that, all the data can be harmonized and shared in a good way. And towards the end, we started talking about data fidelity. That is once you get all of this data, how do you make sure that this data is correct? How do you make sure that this data is well calibrated? Who should be doing that? How much of it should be done? Should you be heavily mandating that every data be accurate, calibrated? Or should you disconnect a lot of the data and include this information in the metadata so that this information could be thrown away? So that is a summary of the last hour of the breakout session.

RANVEER CHANDRA:
But now we continue talking more about these questions that are a FAIR framework. So for new people who've joined, FAIR refers to Findable, Accessible, Interoperable and Reproducible of the data. So do we need to do more than what we are doing? What else do we need to do to make sure that this data falls within that FAIR framework? That is one of the things the second is about how do we deal with data privacy and ownership concerns. Who has the data? We discussed some of this and the fireside chat this morning with private sector and their concerns, but also we discussed a little bit in the previous breakout session. But that is another key aspect that we want to discuss.

RANVEER CHANDRA:
And also in the training gap that is, this is soil science. How do you bring computer data scientists to soil science and how you bring soil scientists more towards computer science; both in terms of data sharing in terms of cryptography and encryption and the tools that are available. But how can you make use of some of the latest AI tools such as federated learning or machine teaching to make more use of the data? And finally, how should this data be stored to do, to enable both real time access now, plus use cases that might come about in the future. So with that, just wanted to see if anyone else had anything else you wanted to add as we, as we start the discussion. Pretty good. Do you want to summarize anything else from your previous session?

RANVEER CHANDRA:
RODRIGO VARGAS: No. I think you did a great job doing that. I would say that if anyone wants to jump in with something, please feel free to do so. If not, I can post some comments I think.

RANVEER CHANDRA:
Skye, you have something to say?

SKYE WILLIS:
Yeah. It's interesting, cause we we're talking about this in group C last time the ontologies needed to run the models and machine learning. It's all connected, but one thing that we've been doing in soil survey with our new approaches, for things like our PSP hub is trying to make the ontologies work, not just for the soil properties themselves, but for the metadata including management information. Which has been somewhat harder especially when there are existing models with very specific terminologies. So, we've decided that instead of having a common ontology, we need a Rosetta Stone that translates between all the other ontologies. And the problem we're having is nobody really knows enough to do that right now, but we think in NRCS, it's something we need to do. So we're, working on it and thinking about it and that I all I really wanted to add right now.

RANVEER CHANDRA:
Well, that's an interesting viewpoint. That is great, that the other breakout session is discussing this as well, but there are things that will be synergistic. And I think it's great to capture those as well across the different breakout groups. Yeah.

RODRIGO VARGAS:
I want to pick up the discussion on where we left before the breakout, and it was on data fidelity. So just to recap a little bit about that there's probably like two polarized points of view. One would be, let's just create a repository or options so we can have as much data as possible now. That's like data that is shared is better than no data that is accessible at all. And with that, it creates a challenge that you can have a lot of trash now, in this approach. So data fidelity becomes a big issue, and can we trust the data? On the other hand, could be an approach that is we want the best quality of data as possible so we have data that is usable, reproducible with great metadata, and it can be used for many different purposes. But on that approach, the challenge is that you restrict that to certain data characteristics. So there is this issue.

RODRIGO VARGAS:
And finally, there is something we haven't talked about, is the possibility of citizen science. So, with citizen science now, where would citizen fail now in this spectrum. On one hand, the potential solution for all of these is maybe having a great description of the metadata. How was everything collected and how everything is documented in a way that the end user can make that selection could be a way to go. But I want to bring this to, to the participants that were not here before. I think something that we were talking about, that if you look again, but how do we get a folding this, a spectrum, and what are your perspectives of this in either your personal experience on how are you collecting the data? How are you sharing the data? How are you using the data? And the vision, your personal opinion about how data depositories should be; the pros and the cons of this approach.

RANVEER CHANDRA:
Just to add to what you said as well, especially with citizen science. In metadata, you could include the methodology used to collect the data, but how do you even trust the metadata? Like, that's the other

thing. Especially with citizen science, I guess you can trust researchers; you can trust labs. How do you trust in like individual people? But just to add to that, maybe just to kick off the conversation here, as well as maybe, some of the latest tools that are being developed, data tools using AI and such can help. Which can help see if that if the metadata is correct based on what people are claiming, or be able to learn certain anomalies to say, you know what, this process couldn't have been followed in the collection of this data. But that's the idea; I wanted to get a lot of other peoples' opinion as well around this process. So, for example (UNKNOWN) had mentioned about in Europe, for example, you have these really strict standards, and you're trying to enforce that, which is great to get people to comply with those standards.

And then Andrew just called, they had called in from Australia and he was talking about how there, they're just collecting all that data and Alfred was saying, sort of "just collect all the data and then try to filter it out." But when we are building this kind of information system, what would be the recommended practice?

So any thoughts

RODRIGO VARGAS:
I believe Sky has her hand raised.

SKYE WILLIS:
Oh, sorry. I think I forgot to put it down from last time. But since I'm talking, I will ask, when you collect all the data, is that problematic to then sift through it to find the stuff that you want?

RANVEER CHANDRA:
Yeah. If you train your models on data, that is not harmonized, data that might not be as good or as correct, your models could be inaccurate. So that's the flip side. Yeah. And Kris, you have your hand up too, so yeah.

KRISTOFER COVEY:
Yeah. I, I just, uh, getting, uh, hi Rodrigo. Hi Ranveer. Good to see you all with sort of worlds colliding super fun for me. In the last breakout session, this idea of data quality came up and someone mentioned a program that was, that gave I think it was remotely sensed data, with quality indicators, almost the way that the IPCC issues, a sort of, "how confident are we in this, in the science?" If the database was set up to bring in all kinds of data from this really detailed, depth rich data, built by professional soil scientists, which would have a very high rating for data quality. And then, you know, the more citizen science end of data could take points down for the fact that we're not quite sure how it was collected.

KRISTOFER COVEY:
But I think when we look at the scale of a national soil information, particularly a dynamic one. The number of samples you're going to have to take, the irregularity and where the variability is out in the world. It seems almost impossible to imagine that all of that data is to be collected by professional soil scientists; highly educated, highly paid folks. When we would try and do inventories out in the American

West, you know, you send people to the middle of nowhere, and then you say, "OK, I want to get to this spot on the ranch." And someone says, "Well, OK, that's a four hour drive to that gate. And let me try and remember what the gate code is." And the idea that that's going to be a postdoc, I think obviously it's not going to be that.

KRISTOFER COVEY:
And then just the logistics of moving people around the field. And of course that rancher was there two months ago, moving cattle. And so is there a way for us to call it citizen science, but maybe it's a distributed inventory where we have a set of tools that could be deployed while people are already out in the landscape. And then it's about what is the kind of minimum useful unit? What are the things that you can measure where the information isn't completely destroyed by the fact that someone you don't know, just grabbed that sample from the ground? And I think in that, thinking about, Luca's comments about lab variability in Europe, and the fact that they basically just went to a single lab. I mean, that seems like as big a challenge as who's taking the data. So I'll throw all of that on the table and then quietly slink away into the darkness.

JIM TIEDJE:
So I was going to comment relative to other experiences and one that's always stuck in my mind is gen bank. I got a message on the screen. I'm not sure what that was for. Anyway, with regard to Gen bank, because when they started that many of the framers of Gen bank were worried about the quality of the data and what if all of this bad sequence, data got out there and all the problems that would create. In the end, the decision was made to just put it up and that the greater value would be to have the data and that the quality would sort itself out later. And that's in fact, what happened. Hindsight is, that was a great value to science to go ahead and put the data up. What then happened is that there were some, boutique subsets of that data, which were pulled out and then created to a higher level. Other people began to, make sure that their data was better quality. So there were sort of self-correction after the fact. So, I think eventually, I mean, that's a different kind of data set.

And potentially some of that relates here too, that it's maybe better to have the data available and allow the quality and other aspects of it to be corrected later.

JIM TIEDJE:
RANVEER CHANDRA: Great point Jim.

RODRIGO VARGAS:
That's a, great point. And some of the discussions has been that the soil science community at lags behind some other communities. Either the physicists, or in the KWR, you are showing James the microbiology community. So, just to build a little bit more on what you said is, there is this fear that, my data will be abundant in these databases. So your point that you are saying is that this auto correction will, let the community of the scientists collecting data to have better quality of data. Either because he's better at describing metadata, or just how the actual measurements or information that is there. But what happened with the bad data? Just simply was flagged? Or the community knew about that? Or

it just stayed there forever. Can you expand a little bit on that?

JIM TIEDJE:
Some of the bad data's still there. But actually, with time then GenBank for example, had more staff that would actually go through and be able to flag some of the bad data as an example. Some things that were, inconsistent or out of line. So where it interfaces with soil science is the soil microbiology and the sequence information about soil microbes is there. And a lot of the annotation of that is automatic. And some of it may not be correct. And so you can't. But, people who are knowledgeable can recognize the existing problems that that would be there.

RANVEER CHANDRA:
Good point Jim, Richard.

RICHARD OSTLER:
Alright, thank you. I tend to agree as well, but I think there's also a case of motivation as a data provider. So why are you pushing the data out there in the first place? And I guess I work at Roth Amsted and we have a requirement to publish data and make it available to the wider scientific community. And for us, we actively want the data to be used. So we promote the data as much as we can, and we invest a lot in the stewardship of the data to make sure it is of good quality. So for us that there is this motivation to have the data reused. And so we put the time and the effort into curating it to quite a high standard. And hopefully that is reflected in the kind of metadata that we produce alongside it. And another reason for trying to put a lot of effort into the stewardship of the data is, it hopefully allows the researchers using that data more independence. Because we have a small curator team and we don't have time to be supporting people with every data request. It's just not feasible.

So we have to, to make the data as independent as possible so that the researchers can pick it up, read through the metadata and understand how to use it.

RANVEER CHANDRA:
That's a great point Richard. So it's interesting where these are again on the one hand Richard you are saying that too, that, "Hey, we need to curate it. We need to make sure that the data is good quality." And that's what it means. And Jim raised another point where even in cases; this was a big debate then. And even if the oral data was still helpful, that is even if some of that was people were able to figure out what was good, what was not. And just to complicate this discussion even further, I wanted to add this other thing, which I mentioned in the previous breakout session was, this relates to question two here, how do you share data by making sure that you don't compromise any privacy? So one of the new computer science tools, what it enables you to do is to do AI on encrypted data.

RANVEER CHANDRA:
So data will stay encrypted yet. Yet you'll be able to do some artificial intelligence on top of it. Well, that works when the data is good. If we upload all the data where the fidelity is not quite sure, the AI will get compromised as well. So that would mean that well, you should be curating the data before putting it

up. But if you curate it well, it would limit the amount of data that you can get, for example to the point that Chris was mentioning. Only scientists can only collect so much data. If you want a large scale data set, you need citizen science. You need everyone to start contributing to this, to this dataset, if you want to build a truly dynamic soil map. I wanted to get other thoughts on this. Yeah, Mark.

MARK SCHILDHAUER:
Hi, this is a little bit backing up, but you know, the first question about the FAIR framework, I think it's really important, from my reading of the FAIR paper and interacting a lot with data repositories, to have a common means of representation of the data holdings. So that really leads right into, you know, community ontologies. And I guess in the earlier session, you heard about some of the work going on with eCit and the soil oncology. But it's also relevant in terms of data credibility and data quality. And I posted in the chat there, that there is another W3C, recommendation that's called Pravo that's for describing the provenance of a digital resource. So for instance, a dataset you could learn, you know, whether it was the output of some sort of a model run, or whether it's a field sample and who did it. And by adhering to the provenance model, you get a lot of consistency in that representation.

MARK SCHILDHAUER:
So I'm hoping that people are at least aware of provenance. And furthermore, I think that if you get into this framework of representation, which is kind of a graph technology, you have the possibility of annotation as well. So that a digital resource that for instance, is at the end point of some sort of a URI, if you put it into a graph that you can have potentially registered or authoritative users able to comment on that digital resource. And this is related to provenance again, that I use these data to calibrate my model. Or I used these data, and I found that there were some anomalies in it. And so you can build up a whole annotation framework using graph technology and some of these standard recommended, languages and frameworks that W3C has developed; which are part of what has been called in the past semantic web.

MARK SCHILDHAUER:
And then it's also been called link data. And, most recently it's called knowledge graphs. So I'm hoping that the, you know, the soil community, I think can really benefit from that because from what I've heard, I mean, yesterday we heard, I mean, I couldn't even write all the URLs of valuable resources. They were coming out fast and furious, you know? And so how does a researcher then find the data that are relevant for their area of interest? They're just all over the place on the web. And again, a knowledge graph can help organize that in a very consistent way that adheres to the FAIR framework, at least in my opinion.

RANVEER CHANDRA:
Yeah. Good point. Good point. Yeah. They need to do that question. So who should be maintaining this graph? So one question that we had asked previously was, who maintains the database? I guess one of the things that the discussion was tending towards was that this database will stay distributed. It's probably not going to be centralized, but there has to be mechanisms in which you can interact across these databases. And Mark to your point, this could be a knowledge graph. This could be to enable discovery of data as well. The other question that, Rodrigo had asked previously was how do you, so who does a lot of the, who pays for the cloud resources, for example, who does a lot of the hard work in

building the knowledge graph? Is it private sectors, is it federal agencies? Is it academia who takes this on? So do people have thoughts? Richard? Do you have your hand raised? (INAUDIBLE) Yeah.

RICHARD OSTLER:
Yes. Just to address the first part, one of the things that we're looking at for linking discreet published data sets to other research outputs is the pit graph. So this is using data-wise and to generate using the graph that we can use for DOI citation to link data sets to other published resources. But we also want to try and extend this to sessions for samples as well. So we can link a physical sample to data generated from that sample to any, publications from that sample. That pit graph work, I think is coming from, the project fray, which is one of the page 2020 projects for me.

RANVEER CHANDRA:
That's great. So what would you recommend Richard though? Should this be who should be doing it? Should it be done by everyone who's publishing some piece of work should be referring to the right datasets?

RICHARD OSTLER:
So I think if you're publishing work that's using existing data sets, then it's a must that you should cite those data sets, on the DOI. Because if you do that, then it gives you an explicit link between a publication and that dataset and that's, that's something that you can then track through this pit graph.

RODRIGO VARGAS:
Yeah. So there are several examples of databases out there, or datasets that they have. DOIs that are either each one of the specific data sets under those databases has a specific DOI. And of course, diversioning within them. And that is very important depending on which area of research you are, because it's not the only find-able and trackable, but also it is important for reporting in terms of impact now, in terms of what Richard said and I'm assuming that he is from this big group, this big effort. And how it has the impact now, how it has the information from these, these datasets it's called the DOI has permeated into users now, in this case tracked by DOI of publications for, reporting. Now we're talking about incentives, incentives of why to do this. It is very important then on the use and the impact of your purpose. So that's some very good common returns.

RANVEER CHANDRA:
Alright. So Michelle, you were trying to say something as well.

MICHELLE:
This is sort of in the sharing and bridging the gap or training soil science. I was interested in interacting with people who were doing well firm's natural language, and they have sort of made, I think, some options for non-coders to very powerfully, both share data and, and, amass data together. And so I thought it was really interesting opportunity to lower the bar so that, I wouldn't have to make a GitHub site, for example, as a non-coding kind of person. And, so I was really optimistic about it. Just working with some of the Wolf from people and thought it was great. But then comments from a colleague who was doing AI that were a little negative. I think, because he felt like, I dunno, I guess I didn't really understand his con.

MICHELLE:
So I'm curious about what you would think about that or others, if you've looked at it. Because it, they seem like you could easily tag and you could possibly build in. I'm very intrigued by this, you know, semantic web tools and also very easy to make, tags and ontologies and just conceptually in a way to teach people about all of this, they naturally lead you into good behaviours. So I was thinking about doing courses with that just because I thought it was such a useful, you know, it lowered the bar.

RANVEER CHANDRA:
Yeah. Yeah. Michelle, related to that, I would also like to point out this, this tool called GPT-3. I don't know how many of you are following it. This is done by OpenAI, so this is like a huge number of cores, and I think they use a hundred, 175 billion parameters to train a neural network model. So what that does is you can ask it questions in natural language and it then interprets it and does queries on large data, like for example, and it can also, for example, you should see what all it can do. You should just search for GPT-3 by OpenAI the interest. The reason I bring it up is that this relates to question number three here, which is how do we bridge this training gap in soil sites. So rather than having soil scientists learn the SQL and other query languages, or even in some cases that are in the AI, we can start issuing queries in natural language. For example, you could just say, tell me what is the soil nitrogen value in this particular region?

RANVEER CHANDRA:
And then the results should just come back rather than you learning a language or writing training your real models to do that. So, there is a lot of advances happening in AI around it, especially in big AI, that is AI using huge amounts of data that can help bridge some of these gaps. And essentially this is more around using large amounts of soil data. How can you bring soil scientists closer to huge amounts of data? But I think the other question you asked is how do you build these ontologies? Can you use similar tools that have been used in natural languages to build this graph? I think

MICHELLE:
After we just learn a little bit about it, we felt it was going to be a great way to actually gather farmer management data and we could lower the bar because it's so laborious for them to input. You know, make it more intuitive and also an additive to collect. But, I think this might be useful if you think there are better things are for people to shorten the learning curve.

RANVEER CHANDRA:
Yeah. Yeah.

RODRIGO VARGAS:
So, yeah. So building on that then in terms of data accessibility, not because the data is not there, but more in terms of the end user perspective now. There's more data out there. We have a lot of questions that class from data that is there. And maybe the limitations arguably could be our imagination, but also our technical skills on how to access the data. So Ranveer is saying just one option of how technology is moving forward to facilitate knowledge discovery, but definitely a bottleneck right now, might be from the end user on how to access this information. So, one of the points of discussion is, the gap between

the, let's say the soil science community and users and the computer science community. And where are we, how do you feel, where you are? Or your stakeholders who are students that have access to this information? In terms of the database may exist. But they don't have access because it's too much data out of the data is not in a searchable way. What are the things that we need to move forward for each one of the stakeholders, are your opinions?

MICHELLE:
Well, I guess on that, you know, I feel like there's such a need to bring them together. And so that my question was related to that. I'm working with some engineers trying to do AI and, and I have been really startled by the challenge of using soil's information with people doing remote sensing. And so some of this is really intriguing, and I think our understanding and soils data that was taken 20 years ago is still very useful, right? So this five-year, it's not going to become, obsolete. Some of it's going to have additive power into the future, but people's lack of understanding about, and the scale issues are huge. So I do think we need domain experts in this area and soil science have a lot to contribute, but they're just being left out because they're not keeping up with some of what I see. So we do need to speak similar languages.

RODRIGO VARGAS:
Yeah. In my case, I can share my experience from, academic perspective and training students. I encourage my students, since they arrived working with me to work with open-access softwares and, avoid, looking at your data. Now, you don't need to see your whole data in your screen. I have noticed that for some students, they feel uncomfortable, when they cannot see the whole data set now in the spreadsheet. And, it's that cultural issue of letting yourself move away from the actual visualization of each one of the values to something that is out there, and you extract a subset of that, or an analysis to answer a question. That is something that we tried to work with the students. But I think that many of them come with idea that, they want to see each one of the values.

RODRIGO VARGAS:
And, and working away from that, it just opens a lot of opportunities for data exploration and knowledge discovery, and that what it takes is also to, be open of learning other tools. Not as a requirement in a way that I have to do that now more on the idea that I get to learn these new tools to access this information for knowledge discovery. And this is one way that we are working at least with my students to open the opportunity to access this. But I have to say it's a learning curve. And, there might be some students that are going to be more open for that. We'll have more flexibility and interest in going into programming than others. But also recognizing there are some tools as Ranveer was saying. I'm not familiar with that one, the GP-3.

RODRIGO VARGAS:
But there are the data cubes where there are some pre select queries in some databases where you can access those things. So building these tools from the computer science perspective, but also from the soil science perspective now and bringing those gaps in knowledge and use of tools to access information is something that maybe we need to be working stronger on that front. OK.

RANVEER CHANDRA:

And I think, you know, there's an also an opportunity to take problems to the computer science side of things on the computer science side of things. A lot of papers are being published on things like machine teaching, where for example, there are experts that's meant for someone like soil scientists, where you have experts who are in a particular field, they are not machine learning experts. They are not AI experts. Can they then train the models so that they can then start using the data? That's the entire thing of, machine teaching. They use simulations to start augmenting a humans knowledge so that you can then start training human models. People are starting to use this in a few areas, like in computational chemistry for example, but not in soil science yet. It'll be interesting to see if any of those tools can be used for soil science.

RANVEER CHANDRA:
But I think the other problem which computer scientists should learn from soil science here, I think more of the input here would be good status things like as far as data, we are only getting some spot samples of soil. From there, how do you interpolate to the rest of the region? You're not getting complete information, you're getting sparse information. Some of the latest trends in AI is around sparse AI. That's AI on sparse data, but I think some of the soil scientists could educate the computer science also on some of the challenges that exist in using AI on some of the soil state. I think this could lead to new innovations on the AI side as well, but wanting to learn more from other people who've tried to use AI and the pros and cons and the things that have worked, things that haven't on soil data.

MICHELLE:
Skye might have really useful information if she's still on. I think the interpolation and knowing more about the polygon versus point work that NRCS is doing. I think they've done some really great stuff. So like maybe in terms of the building capacity, I think like a workshop from them would be very useful to people.

RODRIGO VARGAS:
Yeah. So, my research group. OK Skye, go ahead please.

SKYE WILLIS:
Oh, I was just going to ask, like I'm not sure that I'm drawing the connection, Michelle. So you might have to point me into what you're thinking of. But we have done a lot of that. Polygon versus…

RANVEER CHANDRA:
.. Raster and we do a lot of machine learning. We do a lot of… But I don't know. Our issues are actually often in how to share that information in a way that's understandable to non-experts and experts. And so, I don't know that we've actually solved the problem so much as we've just done well enough for a few audiences and we actually have multiple product streams because of that, because there's just different users need different things.

SKYE WILLIS:
Can you give an example, Skye, of when (INAUDIBLE) the challenges that you've run into in making it understandable?

RANVEER CHANDRA:
Making it understandable? Sure. So, our polygon maps often have multiple components per map unit and so the way the NRCS uses it for planning and providing Farm Bill information is to use what's called dominant component. But in order to do that, you actually have to go into the structure of what the map unit is assigned, which includes components and properties inside of that components and depths per component and so it's really complicated and there's an equation written but when a user goes to a soil survey and pulls that out, they usually just look at whatever the name the soil is, and that's the only information they get from it. And so that information is kind of obscured so that it seems more simple. But really there's a lot more detail involved.

SKYE WILLIS:
That's interesting, so what you're saying is in addition to... like in third question, it's also about how do we visualize the data? How do we present the data to the end users? And the end users could be very different. It could be any of the stakeholders, but the visualization and representation of the data is also data and insights is an important problem.

RANVEER CHANDRA:
Sure. In our current users, I mean, we have traditional users. You think about farmers, agronomic consultant, people going to have cell survey, people that are doing training, but then you get to the people who want to run these very large maps and often they're interested in just raw property. They don't want to deal with named categorical attributes, they just want... So, they'll go to completely different products like SoilGrids or maybe the POLARIS or something like that, and so we've tried to fill those in, but then it's not what our government system is built for. So, then we end up with like two separate streams that are related to each other, but they require different processing.

SKYE WILLIS:
Yes, that's great feedback.

RANVEER CHANDRA:
And something that came up in the other breakout room C was talking about how do you communicate that uncertainty? And that's something with like new Raster products, because people really think if the soil is mapped in a (UNKNOWN), we know that it's not all the same, but a lot of time we use or treat it that way. But when you put out a grid cell that's 30m by 30m or whatever, people like really expect that to be true and we don't have a good way to tell people when it's certain and when it's not, and then when we produce the maps, but for uncertainty, but people just don't seem to care. So, that's a huge issue for us like, how do we visualize that? How do we train our own people? How do we treat... How do we train the public to use our data? It's a huge issue and I don't have any really good thoughts on how to do it.

RODRIGO VARGAS:
I think that that just builds on the discussion of data fidelity. Not only the data fidelity of the point data or the raw data that is out there, but also the fidelity of the value-added products. So, whether you have now curated time series or a developed map, wall to wall map, that is also a product from this information that's available now to an end user. But it may have some quality issues that need to be, as

Skye was saying, shared and I fully agree that we need to put that into the information of uncertainty. But how we share uncertainty with the public, as Skye was saying, is very complicated and there are many working groups discussing these issues in other communities, but it is something that as a soil science community may need to think about. And in terms of data fidelity of value-added products, I just want to add something that Ranveer and Skye touched a little bit.

But it's this issue of application of machine learning and AI where I don't think is, we have to be very careful with that, because on one hand, there are many tools that are so amazing, and they get results in a surprising and beautiful way, but we don't know why. It's very difficult to interpret. So, interpretability of what is behind there, it is extremely important to understand that we are discussing a lot about the broad data and the data is like metadata, how someone collect data, what is allowed to collect the data. But we have not talked too much about the value-added products from these where we might need to be as critical as we are from, let's say, the lab that is collecting the sample, but also critical on the technique that is being used to develop that value added product. Whether that is, again, a time series that has been cleaned for (INAUDIBLE) field or a map and an understanding, again, the interpretability of how was there and what was done, I think is something that within any data information system we need to be aware of.

SKYE WILLIS:
Great point, Rodrigo. Any other comments on this? The one other thing I would like to add is around like related to, Rodrigo, what you were saying around fairness in AI and interpretive is like this, this thing around responsibility that's a particular field that is taking out, which is around how do you make sure that your AI is also fair? It's explainable, that is, you should be able to know why you've got a certain result and you should make sure that there's no bias in the results. Now, these come up often in the context of images and image recognition. But I think even in soil data, we need to be clear as to if an AI is predicting something, it's why is it doing that? If it is, what is the data that is leading to certain decisions that your AI is making? So, eventually, it actually boils down to data and the model as well.

But a lot of it is on building this map would require both, I believe, this dynamic map would be how would you collect the data but also an AI and being able to reason about why did an AI model take a certain decision or come up with a certain result? With that, I wanted to get others to talk as well. Dave, do you have any thoughts on these questions? I would love for... I know, sorry for calling on you. Dave (INAUDIBLE), I wanted to get your thoughts.

KRISTOFER COVEY:
No, I'm listening to all of the possibilities. I am not the collection curation data expert by any stretch of the imagination within the division. So, there are several folks on board in our, on this breakout. You've heard from Skye. I think Suzann Kienast-Brown and Drew Kinney are also on the probable, they definitely have more knowledge of this than I do. So, I'd say call on them.

SKYE WILLIS:
Call on Suzann and Drew as well. Do you have anything to add to this or add to the discussion? Anything you encounter when you're collecting data, using AI on this data? And to others as well, they have lots of experts, Richard, and others, if you have thoughts. And it could be about any of these four questions.

Right now, we are just compiling a lot of the discussion and then we synthesize this together into a deck that you all can comment on and we present tomorrow.

JIM TIEDJE:
I think just adding on to the discussion of AI and trying to understand what those machine learning as well and trying to understand what those models do. We haven't done a whole lot with AI in terms of our soil survey work, but certainly machine learning is our sort of go to tool for those types of digital soil mapping applications and something that we really try to not only communicate through our training for soil scientists, but also really try to enforce it, as in practice when we're doing these projects is understanding the data that we're putting into the models from a soil scientist perspective. So, making sure that the covariates that we're choosing, we can explain why those are important in terms of soil forming factors and the interaction between the covariates and then, of course, understanding the point data and sort of what is it representing? Is it representing all of the variability? But I think it's hard for our human brains to wrap around what's happening in some of these very advanced algorithms.

JIM TIEDJE:
So, the things that we can control are the data that we're putting into them and understanding how that might be affecting the outcome and then using it all of the sort of built-In tools with some of these algorithms to avoid overfitting and some of these other parameters that we can control and at least explain what's happening in the modeling domain from those perspectives. And then I think, evaluating the outcome, like we always encourage just a qualitative assessment. Does it look right? Does it make sense for what you know about soil landscape relationships and trying to do some kind of independent validation on those results so that that you have some level of confidence? Not only does it sit well with what you know, but quantitatively it also is reliable.

SKYE WILLIS:
Thank you. Thanks, Suzann. Drew, you want to add something to that?

RICHARD OSTLER:
Just that, you know, our traditional product that we produced for the last 80, 85 years, which is what we put out on Web Soil Survey. It's been a collective of information that we've collected over that and built upon that information over that period of time and continue to build upon. So, when we start looking into using some of our machine learning processes, we're still kind of building upon taking that and building upon what we still already know. So, we do have an extreme knowledge base to which to build from and I think that's probably the true value of our soil product that we produce, is that, you know, we've been the tried and true. We've tried and failed with some things and we've tried and succeeded with those and so we're always trying to build and expand upon what we know and that's where that's what's led us into the machine learning area so that we can not only expand upon what we know, but what we can expand to deliver to our customers and our customers are becoming infinitely more varied.

RICHARD OSTLER:
They're coming from all points of the compass and asking a lot of questions about our data and how to deliver, how can we deliver that data to the public and that's something we wrestle with quite a bit and we're just now in the infancy of looking at AI to help us kind of predict some of those things that the

public are asking of us. So, it's an interesting dilemma (CHUCKLES) and you're right. I mean, that's what leads us to this, these types of get togethers is to try and help us understand or better ways to deliver our information and better ways to collect our data and so we're always interested in opening those doors. So, that's really kind of the area we're wrestling with.

SKYE WILLIS:
Thanks.

RODRIGO VARGAS:
Turning a little bit off topic but getting back into what we're discussing in terms of centralized repositories versus distributed repositories and discussing also some of the cultural aspects of how we feel about our data and where data should be stored. What are your thoughts about the perception of having the data stored in a place where people know where it is calling that a university or a server, USDA or something like that, versus something that is more, let's call it diffuse where it's in the cloud where people don't have that feeling that it is stored somewhere. And I bring this up because some communities have a problem with this in the international community where my data will be stored. It is going to be stored by the EU somewhere else or the United States somewhere else or is going to be somewhere more like to say diffused or in the cloud system that we don't know where really it is, because these bring issues about the feelings of proprietary data where it is no physical contact, not physical contact, but an idea of where the data would be. I think (CROSSTALK).

SKYE WILLIS:
And to add to what Rodrigo was saying like even with cloud vector, we have this thing called sovereign cloud where you can see that this is the data cell. So, this is a region where this would be. But your point, Rodrigo, whether it should be distributed like people do feel ownership of the data, that could be another reason why you want to do it like that. This is my data, I want to host that. And even if it's in their cloud, it could be in their cloud subscription as opposed to saying, "You know what? This could be just in a common subscription." So, that could be another way to interpret your question, too. Yeah, Drew?

RICHARD OSTLER:
Yeah, and a lot of cases like in our case with the USDA, a lot of our information is based on an authoritative source, and so you have to go to that authoritative source to get that information. So, and that's predicated by policy. So, if some of our other agencies (AUDIO DISTORTS) within USDA, they have to go to our data, our soils data for that information, because it is the official data source and there's some legal ramifications why they have to do that. But that has not precluded us from looking at putting our data out on a cloud service and letting people generate information from that. It's just that once they do that, we make no claim that we didn't create that information. We, that's their interpretation of our data. So, it's more or less the approach that from a federal government perspective that we've had to do.

SKYE WILLIS:
Yes, Richard?

MARK SCHILDHAUER:
I think as well, it helps if you're putting stuff in the cloud, at least to begin with, if people have control over how their data is being accessed. I know Agrimetrics was mentioned in the first (INAUDIBLE) chat session, so I don't know how familiar people are with Agrimetrics. It's a company that's been set up by the government and a few research institutes in the UK with the aim of integrating lots of different types of agricultural data together. So, they've got a couple of different platforms that they've put together. There's a standard, fairly standard data catalogue where you can find data sets in that. But they also have a much more interesting linked data model so that they take various data sets and they are putting a lot of curation effort into them, adding semantic annotations and generating a link data model onto that so that you can do a really nice spatial queries over the UK and it will pull together data from pesticide usage, nitrogen usage, cropping and soil data, which is coming from places like the soil grids as well.

But for the people who are providing data, there's a whole bunch of public domain data sets out there which are being used but there's also lots of industry data sets and government data sets which have got more restrictions on there and so there's a very fine-grained access control over there. So, people who are providing data can if that, if they're not sure about releasing it to the public or the kind of people who they want to use it, they can control that through various kinds of subscription models. So, there's ways of doing it. But I think, once people get more comfortable with putting data into a cloud-based service like that, then it's like crossing a threshold. It gets easier and people are more receptive to it if once they start seeing benefits.

SKYE WILLIS:
Thanks, Richard. And yeah, I know I'm from the tech sector, so I definitely support the use of cloud and more because I think this this allows there's so many benefits of that. It makes things more accessible. It makes it more… But even within the cloud, I guess the other question was whether it should be a public subscription or should it be like or is there a need, for example, confidential stuff, for example, for national security reasons that you want to keep it in a sovereign subscription, not make it publicly accessible? Maybe at the scale at which we are talking about in this, like the vision of the particular working group that we are in, is to create this dynamic soil database for throughout the world that like fine grained resolution. So, if you're talking about that skill of a finding distribution map, do we run into national security issues? Are there other concerns that we have to keep into account as well? I see multiple hands so, Mark?

MICHELLE:
Yeah, I just wanted to ask for clarification of the use of the term cloud as opposed to the Web. If you're using cloud in the context of highly scalable computation, highly scalable storage, you know that maybe differentiates it from just the more conventional distributed set of resources over the Web. But when my group so, I'm a (INAUDIBLE) on several major data repositories, including NSF's official arctic data repository. When we've looked into the financing on cloud, the big problem is the charging for downloads. Right? So, if you live in the cloud, it's fine. But if you need to find data and then bring it down for local analyses, it always seems to come out as incredibly expensive. So, if you could comment on that, I'd appreciate it. I'd also just like to expand on what Richard just said, because you had asked earlier, you know, "Who's going to do this?"

MICHELLE:
And I think that there are community efforts, such as the Research Data Alliance, which is an international group that involves both informatics people, computer scientists, as well as domain experts planning for how to facilitate greater fairness of research data. And there's also in the United States, there's ESIP, you know, the Earth Science Information Partners, where there is this community activity, they call it a cluster to develop a soil ontology and address some of the other informatics issues. So, I agree. I think also, you know, Michelle earlier was concerned about the involvement of domain experts, but those are at least two, you know, one international and one sort of US but there's also increasingly with ESIP, EU, and Australian involvement in community development of scientifically themed solutions for the whole data discovery and access challenges.

SKYE WILLIS:
Thanks, Mark. This (INAUDIBLE) point that if our financing charging models, I think that's also coming down, it's becoming more and more accessible to... The prices are coming down that they think will keep happening given the scale of the way the cloud is expanding.

SKYE WILLIS:
And also with respect to these things around open data sets where people are willing to host more and more of these open data sets, making it easier and somehow, even sometimes even being financed by the cloud companies like I'm with Microsoft, we do that when we're hosting a lot of these data sets ourselves - for the research, for the scientific community and even paying for cloud credits for the researchers. But that I think it's something that more and more tech providers would start doing, cloud providers would start doing to make it accessible because from the benefits of cloud where it could improve performance, you are seeing on all the tech costs of maintaining a server. There's so many benefits, all the ones that you pointed out.

MICHELLE:
I mean, it's great. The cloud is great for storage and preservation, especially preservation is a big problem for a lot of us who are living from grant to grant. But for downloads, that's always the killer. And we end up just buying our own, you know, disk farms.

SKYE WILLIS:
You should check out some of these incentives or some of these programs that some of these cloud companies have. But I think there's more, like, I can talk for Microsoft. We have these things called the (UNKNOWN), we're also launching thing called Open Data Campaign, but that's just ours and the other cloud providers have something similar. But that said, I think this is where I think all of us going back to the question that Rodrigo was asking, who should be paying for this? I think we need to have programs across agencies to be able to fund the development of this. Part of the funding could come from the private sector, part of it from the public sector to enable this kind of research. But I think that regarding the finances, I think people would have to come together to make this happen.

FENNY VAN EGMOND:
Maybe just a clarification question from my side. I thought this was for now mainly US centered

initiative. But since you also mentioned global, I thought it would be useful to chip in on a couple of the experiences that we've heard back from countries around the globe in the International Network of Soil Information Institutes, which is part of GSP. And actually, once we started thinking about putting up a global soil information system, GLOSIS, a couple of years ago and we're now trying to build it. We had quite some discussions on this with the countries and they indicated that it's for them just really not an option to share this to a centralized library. They're not allowed by law to do that. They are the mandated authorities to curate their own data, their national data they're appointed. So, it's for them not even an option to go centralized. It is an option to go in the cloud as long as they really maintain control, they can update, they can access.

They can also at some point take out the data from the public domain, if maybe they figure out there is an error with it, or something has gone wrong. They want to have the control to take that data out of the public domain and maybe replace it or update or whatever.

FENNY VAN EGMOND:
So, that is really an aspect that for us, made it mandatory to go to a distributed or a federated system. And the other thing to consider, if you're also moving beyond US borders, is that in quite some countries, coordinates are considered private data. So, for instance, in France, the National Soil Institute cannot share a lot of their data because it has coordinates. They can only share aggregated information or maps or derived products or anything like this. You also mentioned several technologies to handle this before (INAUDIBLE), they're not applying all of that yet, with the result is that they're not sharing. So, they're making a lot of maps and derive products just to allow the data to be out there. But this is a huge obstruction because of EU privacy laws in this case. So, yeah, who puts out data? I think a lot of institutions like Richard was also mentioning. I mean, Rothamsted has it like a mandate or an objective to really share the data. ISRIC, where I work for Global Soil Information, it has the same. It is our mandate to really help the world with soil data.

And there's other institutions like that, that just have it as a as a mandate, as a vision to really help the world with soil data and to share and curate wherever we can. But, I think, for others it really boils down to use. So, if we show a use case, then, it's possible. In the Netherlands, at some point, they said, “OK, we need to have this soil information system because it can save us humungous costs in infrastructure.” So, because now by law, everyone is required to use the national available soil data, they can cut down only a couple of percent on the risk cost that they need to reserve for infrastructure projects like building a bridge and that already paid for the whole system, the national system, in two years. So, within two years, this whole soil information system was already earned back by saving on risk cost. There's more examples out there, but it's really about what is the use? What is the benefit?

And only then can we really convince also countries and institutes and companies and everyone to share with, if you're looking at a worldwide perspective, there are some challenges we face, I would say.

SKYE WILLIS:
Great example of how you were able to justify this. I think, you know, (INAUDIBLE), it would pay for

itself through some of these other use cases. I mean, that's amazing. And you make a good case for why this is, this needs to be distributed. You can't really think of it as a centralized database (CROSSTALK).

FENNY VAN EGMOND:
One addition, in Ethiopia, you can actually get sentenced if you take data out of the country. If they find you, with the USB stuck on you when you leave on a plane, you can actually get a fine or go to jail or whatever. It's really, that's maybe the worst example but it's a famous one.

SKYE WILLIS:
Thanks, Fenny. Kris?

KRISTOFER COVEY:
I would mirror a lot of what Fenny said from the individual landowner perspective as well, in talking to landowners and trying to get them to share data, whether you're a farmer and you just don't want your info out there or so much of the land that's privately held is held by large companies. And for them, this is part of their competitive advantage. These soil surveys, their information about their yields and how it's related to soil. They see that as proprietary information and don't want to share. So, I think about what's been successful about, for example, the FIA data set is that the aggregated maps that have been produced have helped to make markets or create large scale information. And there's been sort of there's a promise back to people that this location data is protected. I wonder if there are ways to kind of encourage folks looking at this. How do we encourage people that share data sharing data is worth it? Well, I think one way is to look at how people are being rewarded or likely to be rewarded in the near future for data.

KRISTOFER COVEY:
We see a lot of ecosystem service markets being proposed around agriculture to the extent to which those require people to provide data. We should be heading these markets and saying, you know, you should stipulate that that has to be public data. If you're going to pay people for ecosystem services, they should have to contribute to our understanding of how those ecosystem services work. I think this point, Fenny made about risk is super important. We talk to a lot of folks who are really interested in crop insurance and agricultural risk, ag loan folks. So, to be able to say to programs like the USDA, if you're going to give people money for cover crops, can you make it that they have to contribute samples along with that? Or if, you know, (INAUDIBLE) is going to be paying people for carbon sequestration, those should be public samples. If Rabobank is going to write people alone and risk is related to soil properties, we can really demonstrate that stuff at scale then building this sort of virtuous cycle where we say, "OK, this is important to understanding risk.

If you want my money, you have to give me this data." And in order to understand that at scale, I actually need this team of soil. Scientists have access to that. So, then it becomes a question about who has access and what they're allowed to do with it. And so, do we have sort of tiered access products where when people are coming in and they're saying, "This is publicly available." Are they saying, "You can use this to build maps and models, but you can't publish the locations," and unfortunately, I have to teach a class at 2:30 so I'm the guy who just babbled and ran.

SKYE WILLIS:
Thanks, Kris. (INAUDIBLE).

JIM TIEDJE:
So, I wanted to make a comment about two other cases that are sort of restrictive in data. So, one, for the biological data, some of you may be familiar with the Nagoya Protocol of the Conventional Biodiversity. Now, they're debating whether or not a genetic sequence data falls under the Nagoya Protocol. And the issue is, is benefit sharing. So, if you have sequence that might be of some value, then the country where that sequence came from then is to receive some value. Now, many countries are opposed to including digital sequence information in this benefit sharing issue. But it is a problem then for the soil microbial sequence data. It's supposed to now be when it's submitted to any database tracked by country, but with the idea that if the Nagoya protocol extends to the sequence data and there could be some benefit sharing for people who might use that for value. The second thing I wanted to mention in the biological sense is, I think, one of the very important aspects of soil biological data is about pathogens. But of course, that's also sensitive as well to farmers and markets.

So, in the US, for example, the way it's set up now is that to understand distribution of pathogens, it's important to have some data on that, to see where the outbreaks are going and what is spreading.

JIM TIEDJE:
But that information now is just restricted to county, not more specific than county and particular organism and date on which that was obtained. I know Australia also has that kind of limitation as well, because they... It deals with the market. You have pathogens that might be in the products that you would sell from that that area. And so, the intent is not to make that unsaleable because there's too much information about pathogens.

JIM TIEDJE:
(CROSSTALK)

SKYE WILLIS:
So, Jim, for this then the reason this pathogen data, it's not that it's not collected, it's because of privacy reasons that it is not shared?

JIM TIEDJE:
Yes, that's right.

SKYE WILLIS:
Interesting. Thanks, Jim.

MARK SCHILDHAUER:
Yeah. I just wanted to follow up on what Kristofer was saying around privacy concerns and access to data. And I think a really good parallel example to look at is the medical world. So, in the UK, for example, we've got the NHS, huge amounts of patient data in there but there is a mechanism for researchers and industry to access that. That data are anonymized format. So, then, it's about having

good information governance around the data so that you can demonstrate that you are a good, safe pair of hands to manage to use that data and that you're not going to be exploiting that data and all these other things. And I wonder if it's a useful model for soil data and agricultural data to look at those sorts of information, governance protocols that we have in medicine as a way of opening up access or is it really just another layer of bureaucracy to get to the data and say it's I can never quite make up my mind on that.

SKYE WILLIS:
Yeah, that's a good point. The thing is, though, health and medicine, it's sort of stricter. I'm not sure if we go that route, but that might be something even looking as a baseline to build on top of.

SKYE WILLIS:
(CROSSTALK).

MARK SCHILDHAUER:
You're right, it is restrictive. But I think that because you've got those restrictions, people have developed the protocols for sharing good quality data that is sufficiently well redacted to not identify patients from that data. So, you could conceivably do the same with geo located soil data because you can do it with population demographics. There's certainly parallels there. And I don't think we look enough at what epidemiologists do for accessing that kind of patient level data for looking at on farm data, for example.

SKYE WILLIS:
So, this is this is great and so I know we are at the end of this session and people have to leave. Just wanted to mention, Melissa just mentioned that there is a (UNKNOWN) Smart Farm Initiative, which is looking at some of these challenges as well for people who are interested. So, thank you so much for joining this breakout session. We create, they're compiling all our and tomorrow there'll be a presentation. One of us will get it, but we'll combine the feedback from these two breakout sessions. Thank you so much and look forward to continuing the conversation.

BRUNO BASSO:
Hello everyone. And welcome to the breakout room C. We will be discussing data analysis and models. We have prepared some questions to cover throughout the duration of this breakout. And if you are also, before we go into the question, I want to take this opportunity to introduce another panel member Alison Marklein on the planning committee as well as in this breakout. And I think I can see everyone. I think Kathe Todd-Brown is also present and Rafael Martinez-Feria is a post-doc at Michigan State. He'll be taking notes for us that will be synthesized when I present them tomorrow. So welcome again, everyone. And I would like to kick off the first question and we framed it around machine learning and AI, what is the promises of the current machine learning methods in AI and where are we concerned they will fall short? How do we deal with sparse and diverse data streams in soils? In the previous session I made a small introduction that where my lab and Rafael actually leads to the component of the machine learning and the ensemble approach using both crop models and machine learning.

BRUNO BASSO:
So we kind of familiar of some of the issues and things that we face. But besides that I'd like to open the floor for different comments and question you like to raise in the context of the first question. Anyone? I can continue to share a little bit of the experience that again, working in these fields, we feel that to avoid some of the limitations in the black box kind of approach in the ML. We coming from the school model of process-based models we feel that the combination of the two is a very important way, especially using ensemble approach of both of the categories. So I see two hands raised, both from John Jonathan Sanderman, I'll have Jonathan and then later to KATHE TODD-BROWN, please.

JONATHAN SANDERMAN:
Thanks Bruno. So kind of thinking about moving from breakout room A into C now we talked a lot in breakout room A earlier about kind of the temporal resolution of different data streams. And so I think maybe some of the silence initially in response to you asking this question is like, OK, what question are we trying to answer with machine learning AI right now?

JONATHAN SANDERMAN:
And thinking about, I mean, we were talking about some dynamics, so properties that vary on daily timestamps versus seasonal versus yearly versus decadal. And obviously the use of AI is more or less developed depending on what your purpose is. So it's a little hard to I mean like, my initial thought, well, it's really hard to provide an answer. I mean, we're talking digital slowly...

BRUNO BASSO:
Yeah, yeah, yeah. Exactly, for sure.

JONATHAN SANDERMAN:
Yeah. And so there's a little bit, I mean, I guess just, and I mean, it seems it's fit for purpose. Is it the right tool for your research question?

KATHE TODD-BROWN:
I mean, machine learning and AI is supposed to give us all the answers, right?

JONATHAN SANDERMAN:
That's right.

KATHE TODD-BROWN:
Feed them the right data stream and they will tell us everything we want to know.

BRUNO BASSO:
Yeah. Whether he's right or wrong. (LAUGHTER)

KATHE TODD-BROWN:
Exactly. So I think when this question came together, it was like, what are we asking of these methods? What are we expecting them to serve up for us in the future? And then where are we worried about them coming up short?

BRUNO BASSO:
Yeah. Well, what came earlier in the previous discussion, which I think we can introduce back here is the fact that we just don't feel, we understand, you know the performances of these machine learning being much more on a kind of a black box side and very limited to the data that they're fed in. And so that was relevant. But I think what John has raised is even more important because it's about, well, let's see first, what kind of question because of the different spatial and temporal questions. Yes. Chris from Indigo Ag. Welcome. Thanks for joining.

CHRIS:
Hey, thanks Bruno. Just, you know, one of my concerns with AI and machine learning is just over-feeding too, have you talked about that here too as you keeping to discuss?

BRUNO BASSO:
Good point and we'll try to get to that. Let's see Vanessa.

VANESSA BAILEY:
Sorry.

BRUNO BASSO:
And I'm sorry, I can see hands raised. You'll be next please. Add new text...

VANESSA BAILEY:
I just want to make a comment. I'm an empiricist. I've worked with a lot of people in this room, so they know that I'm not doing any machine learning myself, but I really look at it as a powerful new tool for maybe hypothesis generation rather than necessarily knowledge discovery. And so I think there's a tendency to think this is gonna solve all of our problems, well I'm actually hoping it narrows our questions.

BRUNO BASSO:

I just want to make a comment. I'm an empiricist. I've worked with a lot of people in this room, so they know that I'm not doing any machine learning myself, but I really look at it as a powerful new tool for maybe hypothesis generation rather than necessarily knowledge discovery. And so I think there's a tendency to think this is gonna solve all of our problems, well I'm actually hoping it narrows our questions.

BRUNO BASSO:
Would you say though, Vanessa, that machine learning would be a better tool to formulate hypothesis than models, than crop simulation model, I mean, process-based model that are more designed to capture those feedbacks while machine learning is so much more dependent on whatever the nature of the data going in? So...

VANESSA BAILEY:
I think it's just another tool. I don't think it replaces anything. I think it augments, right? Like I wouldn't want to go to the physician and replace, you know, all of the things he does in his office with just MRI. I would want both of them going together. And I would also like if there's a brand new discovery to be able to bring that together. So I think it's just another tool in our toolbox. We need to figure out what we're going to do about those sparse data streams. But I do think it's got a lot more power to help us sort through the diverse data streams. So again, I don't think it's gonna give us any answers, but I think it might help us focus some of our questions.

BRUNO BASSO:
For sure. Keith Shepherd, please.

KEITH SHEPHERD:
Yeah. Thank you. I think one of the advantages of machine learning is the ability to pick up non-linear relationships and complex patterns in a way that other methods can't. But having said that, I mean, I would just, the over-feeding problem has been mentioned. That's a massive problem and people not using enough, you know independent holdout data sets. The expression of uncertainty has been a weak point, but I think there's a lot of catch up being made on ways of now expressing uncertainty on estimations. But there are also dangers of, you know, jumping off a cliff with machine learning. I mean, with linear approaches, you're usually not too bad in terms of sort of going a bit beyond your data, but with machine learning it can be highly dangerous.

KEITH SHEPHERD:
But a big gap I see which, I mean, I work a lot in decision analysis and Bayesian approaches and we see that consistently the big data approach falls down because there are important variables and causal associations which are left out simply because there isn't enough data on them. And we always see the best solutions when we combine causal reasoning expert knowledge with available data. And I think Bayesian and approaches linked to GIS are beginning to be taken up now, but it's a scenario that really hasn't been exploited sufficiently. So you've got two advantages, one that you're combining the best of expert knowledge and data and using causal reasoning, but at the same time, you're propagating your uncertainties through at the same time as well. So two big advantages there. Thank you.

BRUNO BASSO:
Very, very good point. Thanks Keith. For sure. I'll share it. Arun.

ARUN PERSAUD:
Hi. So I'm a physicist, not another soil scientist, so that helps with a bit different background. I've broken the project maybe look at the carbon soil that gives real-time data. And so since I'm coming back from earlier, we talked about data position.

ARUN PERSAUD:
I see machine learning as a tool to like tell you where to do the next measurement and perhaps get your, get data faster and like find areas for kind of the discouraging processes that tell you uncertainties, where to go next to get better data sets. And that's kind of a, for sure that falls into sort of like analyzing the big data sets. Yeah. But perhaps it's a great tool to get better data sets and figure out the areas, where to look where you need more data and where not.

BRUNO BASSO:
Correct. Thanks. Good point. Allison you made a point earlier maybe useful to share it here as well. Just.

ALISON MARKLEIN:
Yeah. There's a lot of different ways that we can use machine learning. And someone mentioned earlier it's really good for generating hypothesis. And so we can take the results of machine learning or artificial intelligence algorithms and use that to narrow down what hypothesis we're gonna test and what experiments may be most useful to do. So instead of having thousands of possible treatments for a given site or series of sites, we can narrow that down to ask, to like a reasonable number of experiments to do.

BRUNO BASSO:
Excellent.

ALISON MARKLEIN:
Keith, do you have something to add as well?

BRUNO BASSO:
Maybe from before maybe the race the hand is still up there.

ALISON MARKLEIN:
I wanted to also note that machine learning and process-based models can be really powerful when used in combination, because you can run a series of ensembles or different simulations and get generate tons and tons of data, which can then be mined to generate new insights as well.

BRUNO BASSO:
That's right. And that's the direction we have taken in my lab is using most people quite diverse type of models, of process-based model to as well as the machine learning and run different algorithms together where they can learn from each other to eventually build metamodels. And that seems to work

out. Rafael, you said you wanted to add some points either here or to doing some of that work.

RAFAEL MARTINEZ-FERIA:
Yeah, I mean, so, yeah, so I guess just kind of like give a little bit of background what Bruno was talking about. So for us sometimes when we want to run an ensemble of different process-based models, that's very computational expensive, especially if we want to run it in a much special domain. So we've used machine learning to accelerate predictions and help interpolate in between a bunch of different locations. And we might not need to run the ensemble of the models every time. So machine learning there help us kind of accelerate that processing of data and you know the machine learning can, the main point that I'm thinking about is that sometimes we think about machine learning as a way to help us, you know, interpolate between things that we know.

RAFAEL MARTINEZ-FERIA:
And so if you have a grid of like a multi-dimensional parameter space that you cannot in reasonable amount of time look at every single instance of it, machine learning can help you very accurately fill in the blanks. But we need to be careful when we take that machine learning outside of that parameter space that we had some sort of certainty because then machine learning might not behave in expected ways. We can still look at those, but we need to be more careful.

RAFAEL MARTINEZ-FERIA:
And I guess like the difference there is that process-based modelling, we have an even more certainty because we're capturing the complexity in the system, the behavior of the system to do extrapolation. So like, long-term in the previous session, we were talking about long-term versus short term applications of the model. So process-based models seem to be better at capturing something that might be outside of the domain that we have data for while machine learning is really good to like interpolate and look at things that we cannot fully explore, because we just don't have the time and computational power to do it.

ALICE MARKLEIN:
I have a couple of questions from Slack.

BRUNO BASSO:
Yes. I saw that. And I would like to answer that, but before let's prioritize a question here in the group, and then I'll go back to Janell. Yes. Next question, Stephen Roecker.

STEPHEN ROECKER:
I guess the thing I'm interested in looking at with machine learning next in regards to this session is I've come across some new research and methods that I haven't implemented yet, but it's basically this notion of taking these machine learning models and converting them to something simpler and hopefully more stable.

STEPHEN ROECKER:
So the things I'm thinking about is like kind of doing partial effects type of looking at the partial effects of the model, and then with each one of the individual variables, instead of having the sort of noisy

prediction along, you know, a continuous predictor. There's methods that I've seen, where you can essentially take that relationship and convert it to something more smooth, like a spline or, and then feed that into other similar type of models, like generalized linear models and whatnot. So I don't know if other people have had experience with that, but that's one area that I'm interested in looking at. So I guess someone, lots of people talk about using machine learning methods as either hypothesis testing or hypothesis generating type methods. But that all seems correct to me.

STEPHEN ROECKER:
So again, I think, like trying to take the machine learning models and convert them into something that's maybe more manageable and hopefully more stable, because when you tend to plot these machine learning models in sort of a multi-dimensional space, particularly like with the random force type one, you can tend to get a rather noisy prediction space. And, you know, there's areas of the model where you're lacking data and it's coarse, so that when you kind of look at that unlike a grid, you kind of see like these weird, like predictions that are showing up in kind of no man's land. So that's what I'm interested in looking at next.

BRUNO BASSO:
Yes. Not very good student. The point from Slack is from Janell, like what are the areas in process-based model with the highest probabilistic uncertainties that machine learning could help basically, supposed to help solve? And then following the struggle between ML, AI and how different these from process-based model optimization? Well, as I work much more on process-based model, I don't see anything close to because an ideal crop simulation models is based on inputs and it's deterministic and should run once the level of inputs are supposedly of a good quality. So once the inputs are not necessarily represented in the system that creates uncertainties and not you know, be able to have reliable prediction.

BRUNO BASSO:
What we calibrate crop models into I don't think there is, that there should be any optimization going in except for the genetics describing the plants, because those are proxies that we have to describe, the duration and you know, the flowering and all that, a lot of uniqueness in things that we don't have to, you know, possibility of measuring it. So I think they're very, very distinct. In fact, in the next point, I'll make an example of AgMIP models and how we parameterize that and so on. So that answers kind of the second, but Kathe do you have a point on that?

KATHE TODD-BROWN:
Yeah. I'm actually going to disagree with you there, Bruno that there's some optimization routines that can certainly be augmented or informed with machine learning. And that the parameterizations of these models have room for improvement. And I, of...

BRUNO BASSO:
Give me an example. What would you parameterize in a model beyond an input?

KATHE TODD-BROWN:
I saw a carpenter overtime. What's the decay rate of your different pools? Especially if your pools are

empirically defined, which is certainly where the field seems to be moving towards with moms and other models. So I think that thinking about parameterization in the context of these machine learning algorithms could be really valuable to the field. And we can also remember that the parameterizations themselves have a range in uncertainty, and I I'm willing to bet Mike Dietze is gonna talk a little bit more about this. But we could, when we run these models, keeping in mind that really what we should be doing is running a parameter space that there's uncertainties associated with these parameterizations. And that should show up in our model projections.

BRUNO BASSO:
Yes. But that will become a different type of crop model that we normally use. But I fully agreed for sure. I mean, you parameterize the pool sizes in some ways depending on the land use, and that's a completely arbitrary. So you can certainly do that. The problem is that what I'm trying to get this often, the parameterization in many complex and mechanistic models have such a broad space and numbers of parameters that your use versus mine are completely that's beyond the definition of a system capturing, it becomes a gas because the users makes a big difference, you know, in parameterized. And so you get models that you basically, they're not transferable, they become an exercise that you have to parameterize for every single field. And that's what we run into that. Yeah.

KATHE TODD-BROWN:
I'd be interested to hear what Dietze has to say on this.

MICHAEL DIETZE:
Oh, goodness. You set me up there Kathe. I would argue you should never be guessing your parameters. You should be calibrating them to date, and if you have to calibrate them from every damn site, then that's what you have to do until you understand what's driving the heterogeneity in your system. There's a lot of heterogeneity in these systems that we don't understand and just, you know, sweeping it under the rug and pretending it doesn't exist isn't really gonna solve the problem. I'm also gonna add that, you know, the question of where the uncertainty lies is going to depend a heck of a lot on what timescale you're thinking about. I suspect over, over definitely short and quite likely over medium build even long-term scales. The initial condition uncertainties are going to be absolutely enormously dominant over any reasonable spatial scales. The fact that our sales match are just making stuff up.

MICHAEL DIETZE:
And then you put that in as an initial condition in a model, you know, propagate that uncertainty, you know, the memory is just so large in these systems. And then longer on the structural insert it's gonna go through the roof because there's so much we don't understand about process still. So even if you had the perfect parameters in the models, which we don't agree with, Kathe, we don't I still think...

BRUNO BASSO:
Oh, I agree with you as well. I mean, don't get me wrong on that.

MICHAEL DIETZE:
And the heterogeneity in the initial conditions in the structure. And I think the point that I've made in many other contexts is until you understand, until you truly understand what the dominant

uncertainties are in a model that you're using to make predictions, you're likely to be barking up the wrong tree in terms of your effort to constrain them. You know, you could be spending decades trying to constrain parameters and then if, which won't get you anywhere, if you had uncertainties in your initial conditions.

BRUNO BASSO:
Yes. But I share that that's definitely a correct way of doing that. But you need to formulate what you're trying to do in the question. I mean, I work on reproducing spatial and temporal variability of crop yields with a minimum amount of factors such that I can reproduce yields across any million acres by having, you know, integration of tools. And if you go with your correct, 100% approach of, you know, capturing that uncertainty then you really limited to be able to scale things. And the reality is we're truly capable of reproducing yields at a pretty good level of confidence.

BRUNO BASSO:
So how, you know, it depends how the model is also, there is a model structure behind that it depends you know, how much each of the parameters affects the ultimate goal. So there is a component there as well that needs to be you know, you can parameterize in some of these models, some of the parameters done will make a difference. And so they're kind of, you know, general across sets of condition that if you get the inputs as close as possible, you will get. So reality you'll be able to simulate yields independently anywhere. That's the reality, that's difficult to argue against.

MICHAEL DIETZE:
And I guess I would argue that I don't think doing it right is nearly as computationally limiting as you're implying, we've done this at fairly large scales ourselves, you know, we've done a continental scale, well, corner scale, you know, carbon data simulation exercise that actually deals with the uncertainties. And I agree that not all parameters are gonna be created equally. You have to focus your attention on the dominant uncertainties parametrically. But again, that requires actually getting in and, and doing the uncertainty analysis to understand which ones those are.

BRUNO BASSO:
Sure, sure. I mean, that's I think we can go in a circle here, but the point was also the highly dependence on the data going into parameterize, you know, what do you know about you? You can run a completely independent source of parameterization, but there are lots of assumptions made that could get you obviously not where you want to go. So the uncertainty, this is a critical piece and that came in the conversation earlier. So do how, Michael how do you see that, you know, the role of AI in sparse data, how would you know, what's the danger of using that? Or what are some of the solution towards that?

MICHAEL DIETZE:
I mean, personally, I probably wouldn't use machine learning in sparse data. I mean, similar to what others have said, you know, I think there's a lot of potential for hybridizing machine learning and process-based models. We started using them in those contexts as well. But relying on the process model structure a lot particularly when your data is sparse. And I'm gonna agree with the point that's been brought up multiple times before, which is, you know, it's really hard to separate the algorithm from the user because there's so many applications of machining learning these days where people are

just overfeeding, overfeeding, overfeeding, you know, when you were talking about getting response curls that just look like noisy wiggles, you're just feeding the noise, not the data. And it's so hard to kind of decouple the potential of these algorithms from the way that they're getting abused.

BRUNO BASSO:
Spot on. Stephen. You had a, yes.

STEPHEN ROECKER:
(INAUDIBLE).

BRUNO BASSO:
Oh, OK. And I guess I didn't, we didn't tackle, what are some of the, how is it used? So one of the recent paper we published on estimating and N2 emissions from high resolution data, both of space and time, and the Kellogg Biological Station was actually used, you know, a very rather simple approach of trained the random forest but providing ancillary data that came from the crop model. So things that were measured and by coupling the two, we basically improve the prediction. So knowing the nitrates through the dynamics of the nitrates from the crop simulation models that were measured improved the estimation of the prediction of N2. Any other points on that first point of discussion?

KEITH SHEPHERD:
Yeah. I'm gonna have to drop off soon, but I had a point on the third one, but you've also got other people on the call who know all about as much as I do and more about this, like John Sanderman for example, but I think in terms of new data streams, yes mid infrared spectral data is now starting to become mainstream.

KEITH SHEPHERD:
We have this initiative with German in the global soil GLOSOLAN, the Global Soil Laboratory Network hosted by the Global Soil Partnership, which is leveraging the KSSL USA soil laboratory, the magnificent you know, spectral and reference library they built up now. And they're trying to now get by and to turn that into a global calibration and estimation service. So John now is sort of heading a project of soil spectroscopy for the global good, which is taking that challenge up. And there's a group of us working on this, and it also connects back to GLOSIS. But I think this is a new direction, which has tremendous prospect for collecting large amounts of new soil geo-referenced soil data at scale. Also extending into sort of feel quotable, visible near infrared instruments as well as those become lower cost and more reliable. So I think that's one new data stream that has, I mean, we're leveraging that extensively to you know, we've produced this new 30 meters soil map of Africa, so a properties map of Africa but it's largely leveraging soil spectral data that we've collected in Africa.

KEITH SHEPHERD:
You know, based on 95% of the samples on there analyzed using conventional chemical crop means, with chemical means. So it's been a key tool for us to be able to collect data at the kind of scales that can be, to kind of allow digital soil mapping. So I think that's gonna be a big one going forward.

BRUNO BASSO:

Excellent. Thanks.

JONATHAN SANDERMAN:
I'll just add to that. But thanks, Keith. And I think at least in the US context, one of the great uses of spectroscopy in my mind is infilling some of this sparse data because the Kellogg Soil Survey Lab has scanned so many samples. They're at something like 82,000 samples and counting, we could then use some of these predictive models to then predict properties that weren't routinely measured at those other sites as a way of expanding essentially the reach of the soil database. And so, like one example we just did this, just published a paper really recently, we've just focused on the Great Plains, but we were able to predict three fractions of organic matter across the 8,000 samples in the Great Plains that had geo-referenced coordinates.

JONATHAN SANDERMAN:
And so at using, taking advantage of the fact that they already scanned 8,000 samples, we built the model with a few hundred. We did the lab work on a few hundred samples and applied it to those 8,000 samples to really scale our domain of reference.

SPEAKER:
I mean, we have field scanners in the hands-off village, what we call village enterprise agents. So we're actually testing business models that are providing visible infrared scanner fertilizer recommendations to thousands of farmers. So, you know, all those points are geo-referenced and being collected. So that we're really trying to get that into the commercial domain now. So they're generating profit from providing those services. And there are several commercial soil testing labs now you know, there's at least one in Africa and Kenya here providing commercial spectral services as well. So, you know, it's beginning to take off in terms of being viable from a deep business perspective, and that can then really generate a large number of data points. If there's cooperation to get those data.

BRUNO BASSO:
I guess we could, trying to tackle the question number two. What are the challenges with integrating models with data? So we kind of discuss a little bit the apparently digitization validation and benchmarking were also critical piece in the discussion previously of like Phil Robinson suggested, you know, having a series of benchmark sites. We could you know, validate against you know, these models and yes, Vanessa.

VANESSA BAILEY:
This was the question that I was looking forward to the most for me it's scaling like, honestly, as an empiricist, the scales that I take my measurements at are almost never the same scales that the models are trying to make predictions out. So we do a lot of arm waving and we do a lot of replication. We do a lot of assessments of heterogeneities, but that's, for me, the biggest challenge is linking the scales we're measuring now with the scales we're trying to make predictions on. And then also trying to use those predictive models to inform models at different scales themselves.

BRUNO BASSO:
Right. Well, the scale, this is obviously a critical piece here in anything we're trying to do that drives the

both the tools and, you know, the parameterization for sure.

VANESSA BAILEY:
Well, I guess where the reason why this got me so excited was, you know, the whole point of this workshop is the dynamic information system for soils. If we want to have a collective global data set, that is gonna be something that's a resource for modelers, are we constraining that to some sort of identification of what scale did you make these measurements at? Are we gonna say, you know, because I know that this has been a struggle, you know, I've worked with Kathe and others where, you know, we are looking at color data, we are looking at flux tower data, right. Those are really different things. So if we start trying to pull things out, like including microbial data, things like that, could we extract it from, you know, grams of material or micrograms of material, and it becomes a really big challenge if you want to have a global data set that can be robustly used.

BRUNO BASSO:
Yes. Well, I guess the answer to the scale there it goes back to the type of questions because the variables have different scales, you know, behavior across the scales. And so that's different properties have to be measured at different scales. And there was quite a bit of discussion around that. So I don't think that's a very extremely well, beside the point. I don't know how we're gonna to move towards that.

KATHE TODD-BROWN:
So I think Melanie had her hand first.

BRUNO BASSO:
Yes. Melanie.

MELANIE MAYES:
Yeah. Thanks. The point I want to make is right now, a lot of models still use basically empirical collections of data. And so, you know, I think the key point is that this system needs to have data that is available to where it can be called and accumulated by models. So to eliminate that intermediate step that we are often doing right now to come up with parameters, but having said that it's also going to need to have enough flexibility to where say a modeler could down select or choose the kind of sites that they want or data that they want. So I think more automation may help is, is the point I'm trying to make here.

BRUNO BASSO:
Eileen.

SPEAKER:
Yes so, I'm going to make basically the same kind of point that I made in group a, which is really where I belong. But I'm very concerned as a, as a person who makes measurements in the field. Anytime I work, with modelers or see modelers that I'm not working with about the very obvious, but very important part about knowing things like site history. And I made this point in, in part a that farmers who are doing soil health practices, for example, well, how many years have they done no till? How many years have

they done cover crops? Have they done that? All of those gory details. Aren't always with the metadata, with the data sets that a modeler gets. And of course, I understand that some models don't, don't deal with those processes anyway. But that's really concerning to me that that folks are making predictions without having that really important background information that determines how a system responds. I don't have a solution to that. I just, that's something that always concerns me as a field measurement person.

SPEAKER:
Coraline.

COLLIN:
Hi, thanks. Yeah, I think there's like one real big opportunity right now in that we're generating fundamentally new types of soil data than we have in the past. Big things that come to mind is like the huge amount of soil fractionation data that is becoming really powerful for predicting things. So, a metagenome make and microbiome information, all of this stuff is really important, but we have a lot of models that have been developed historically without this information. And so, they, they really don't know how to ingest or use this information in a meaningful way. And so, I think, you know, maybe we need more interest and more effort and developing models that are designed to use that information rather than trying to transform back that information, like to somehow be something that is a theoretical pool and a model that we don't totally have confidence in.

SPEAKER:
There is a question from Oh, I guess he may respond Jonathan, on the question on, from Slack to Jonathan about recommending part of salt carbon or salt particulate for monitoring purposes.

SPEAKER:
Huh? That's a good question. I mean, I guess it's, what's the question we're asking about soil monitoring? I mean, obviously, I mean, there's this idea, I mean, I guess the question is getting that these like leading indicators that we know it's still, I mean, in terms of them, we know soil, carbon stocks and total change really slowly. And so, if we're looking for signs that we're moving in a positive direction, I mean, obviously something like particulate, Carbon, I mean, there's been a lot of these lay bio carbon proxies have been proposed over the last five or 10 years. Well, I mean, some of them date back 30 or 40 years, the analytical methods and they've kind of Rose in prominence again. So, I don't know what I don't feel like I'm pleased to say what's best, but yeah, I mean, there's definitely, and I mean, I would hope this dynamic system is really considering a lot of these functionally important components of organic matter that we can actually see change in one or two seasons.

SPEAKER:
Right. So, one of the before one of the speakers was actually asked me that you thought it was useful to describe a little bit the approach using Ag MIP. I don't know if anyone is familiar with the community of how your cultural and modeling into comparison and improvement. So, if you're interested, I can cover some of that, which there's an extensive group working on calibration, and I'm trying to avoid, you know, code feeding is always the problem. And so anyway, if you're interested, I'll, I'll go over some of the findings, what we learned about running ensembles and the fact that the median of the models of

the ensemble was the best predictor, but there was quite a bit of a story around that. But again, that's very much focused on crop simulation models, targeting and yields predicting yield. Vanessa. Yes.

VANESSA:
Thanks. I wanted to jump in because there was a couple of things that were just said first, I really love both comments that Eileen and Colin made. And so, I wanted to say that because in these virtual interactions, you don't get that feedback. And it's really tough. And I know there's bandwidth reasons for us turning our cameras off, but I wanted to kind of reflect back on a couple of things that they both said, first off, we're talking about a dynamic information system. So, I think site history is critical. Some of the work we've done has shown that a soil, two soils that are at the same water content, if one's coming at (INAUDIBLE) that water content by a drought getting wetter, and the other one's drawing out from a flood are going to have wildly different wildly different carbon cycle patterns. We've demonstrated that with that papers published and in review right now. So, I think that this dynamic information system, the site history is like even more dynamic than what Eileen was talking about, about land management decisions, which are clearly critical.

So, I'm just wondering when you use the word dynamic, it's a really weighted word.

VANESSA:
Can we be really outside the box and position this so that we can actually have like time series data from some sites rather than, you know, soil maps have been relatively static where it's like, here's what we found the day we went out in 1962 and it's not going to change until we get another project funding, a survey possibly never. So, I think that's a really important element because the time series of direction of change of soil is going in nature change that not tied to what Colin said about with new data sets coming in. It may not be sufficient to just tack on a new module or a new parameter into a model. We may need whole new models. I mean, I feel bad. I'm not a modeler, but the number of times I keep saying the empiricist in me wants to have an if then set of models where I could say, if a soil has experienced drought, then go to this module.

And I don't think that's completely unheard of, but I think some of the then modules that I'm looking for are going to be completely different from anything we've got frameworks for right now, for instance, one of the new data streams that has probably caught on and been democratized in really cool ways.

VANESSA:
So, it's going to be huge is high resolution mass spec. You know, we've been talking a lot about the FTI CR mass spec data, where we can get thousands of carbon molecules identified in a single soil sample. We're not going to just like, metagenomes put each one of those molecules in a model, but there's different ways we can turn that data around so we can get nominal oxidation state to get an analysis of what is the thermodynamic favorability of that you take NASCO in conjunction with whether a soil has experienced drought or flood. We go a whole different direction in how we predict what the microbial

process of that carbon is going to be. We don't have models right now that are going to do that, but we have the information now that can feed into models that I think will be really powerful. So, I just, I really, I wish the virtual, wasn't such a static interface where we could just jump in and wave our arms more freely, but I just love the comments Colin and Eileen, made, and I wanted to jump on them.

SPEAKER:
Excellent. Thanks very much. Maylene.

MAYLENE:
Apologies. I didn't pull my hand back down.

SPEAKER:
Yeah, I wasn't sure because you just pop down any other comments.

SPEAKER:
So, I think feeding a little bit off of what Vanessa was saying there about having an if then switch for the models. I would really love to see a higher, a nested hierarchy for models. So, we have a really high resolution, poor space simulation that we use to feed into a core trans fluid transfer model or something along those lines that we can then sort of sequentially upscale. And that might give us places to hook in and out for these high-resolution mass spec or genomic profiles for the microbial communities in really new and innovative ways. And I think that that can like AI and ML methods, like some I've seen, I've seen process models being included under AI sometime like anything to, anytime you use computers to augment human decision it's AI, so you can wave your hands and, and we can include process models under the sphere. But I think that would be a really exciting direction to see the field go to is to try to develop these linkages across complexity scales and have those reflect the measurements that we can now take.

SPEAKER:
Yeah. In relation to that what's happening in the morning, the crop simulation world, this, you know, using GTL and genetic information to be able to predict the coefficient used in and crop models. So that that's one direction that we're using, certainly not for soils or microbial components, but in genetics that's, you know, it's happening.

SPEAKER:
But you could imagine something like abundance of (INAUDIBLE) Genesis genes or some kind of specific pathway like that being tied to prediction of proportional methane production versus CO2, or like, you can start picturing how those might be incorporated into parameterizations now.

SPEAKER:
Yeah. One point I wanted to make in response to Eileen which obviously a share our thoughts is that one of the ways we're trying to capture this three of the (INAUDIBLE) when it's independent, because that's what I'm trying to do. I don't try to depend on every single thing in a place and trying to make it work for one place or rather trying to scale it. And the way it's working for us is the integration with remotely sensed imagery, historically, where, you know, over a very large areas of the Midwest we've learned

over the last 10 years or more that there are parts of the fields that they're constantly producing higher yields. You, seen that being referred to the main production remarks, but then also by Joel and Jerry. And so, to us, that's a really an exciting way of running the models on the different stability zones.

SPEAKER:
So, you have a high stable zone to that, and even the tournament imagery that I showed using tournament imagery is a proxy for solve that any of upper transpiration and be able to satisfy that demand. So, the permanent reedition there, have one of the biggest uncertainties that Michael, we can talk forever about how you're going to get soil depth to really do anything valuable. You have to use proxies to be able to do that and, and reproduce that. So, it is an inverse engineering, but through data that comes in as ancillary, so thermal that tells you how deep your soil is. And in the unstable zones, the work that Rafael and I did, we showed basically areas more vulnerable towards a deficit, you know, all the sloping lands and the summit versus the areas that there are in the depressions where they came to have low yields in wet years and the high yields in dry years. And so that's captured very well already, and we've been scared of being able to scale and validate the approach of using remote sensing with high resolution yield mapping.

SPEAKER:
And so that's how we parameterized. So, I made the point that if you only run using soils and you don't capture management, you just miss a significant amount of that yield and for counting. one of the critical inputs in running carbon models is to be able to get the yields correctly, to be able to fit in residues and routes that allow to, you know, the soil farm dynamics to be done properly. And so often models I, think in terms of criticism, misload within what they are. So, if you're running a carbon models, you can't ignore, you have to give the input of what the carbon going in this, which is very, very uncertain. And so how we tackle that, it's basically an integration approach of using all the tools. And so, history is one of this, have we captured through remote sensing and understanding variability within the field and asked the models differently across the space.

SPEAKER:
Are you able to include so remote sensing, theoretically is getting better at doing things like tillage system and cover crop. But I would say theoretically, because it's not at least I haven't seen it really good yet, but are you incorporating any of that?

SPEAKER:
Yes. Yes. There is. There is quite a bit of advancement in capturing. The spectral signature of tilled soil is quite different than residues, as well as be able to capture cover crops. You know, the number of days to feel this green of season from a corn. And so, we're, we're all going in that direction, very empirical approach, but be able to try and understand the known the price of eggs to the route to review, to use yields. And that's what what's actually what's demanded on my case and believe it or not the carbon dynamics here, we do have to adjust the pool sizes depending on where you are, but mainly through whether you are in Africa, where do you got the resistant pools being much larger versus the Midwest. But also, because the data to parameterize these pools are often limited, so we have to be creative and, and it's, it's an empirical approach for to be able to capture what happens in the field, which we're getting really close in doing that.

SPEAKER:
I was going to add onto your pallet about the high-resolution Crop data is really critical for getting, obviously if we're clicking soil, carbon modeling here, nuts, I mean, inputs are half the equation. It's getting the outputs, right. And that's where I think as a group, we've really struggled with coming up with where do we get the high spatial and temporal data streams to get the decomposition side and erosion side of the mass balance. Right. And so, I mean, I think that's a lot of the job of this working group obviously, is to come up with creative ways of trying to increase the resolution of the soil data that can help us improve those (INAUDIBLE).

SPEAKER:
That's right. I mean, you don't need to go back to this, but you've seen yesterday that measurements pretty numerous in number of measurements were enabled to capture, you know, the variation in the observation was quite large as was like, you know, between five or eight tons, depending of you use the proper bulk density, or if you had the right number of samples. So, so I mean, you have to be realistic that point must have come up in group A for anyone that was there. The variability in this sort of soil, for example, soil carbon is extremely large for sure, Steven.

STEVEN:
Have you or anyone else identified (INAUDIBLE)for the collected data (INAUDIBLE). And (INAUDIBLE)I'm wondering in regards to like, something like the soil engineering (INAUDIBLE)goes back, hold it back up from that time window and use that you know, (INAUDIBLE) capturing the measurements (INAUDIBLE) make those possible.

SPEAKER:
Yeah. Unfortunately, Steve, I had a hard time capturing what you said you've seen anyone else says was able to capture better.

SPEAKER:
Has anyone gone and tied on the ground measurements back to like a spectral signal of when they were collected? So, like with something with KSSL, you know, some of those measurements go back to the fifties or better. And so, you know, with Google earth engine and whatnot, we can go back in time with the imagery. Has anyone tried to Like extract the ground on the ground measurements from an overlapping, you know image?

SPEAKER:
Yeah. I mean, there are people working on part of a center that works on bioenergy and marginal lands, and we're trying to use heavily this approach to see abandoned land and things that have changed over time. So, but not necessarily the way you asking, but there's a lot of potential using historical imagery and changes over time.

SPEAKER:
Yeah. Just say, I mean, there is kind of some merging field in digital soil mapping doing space, time modeling, and that's where, and that's one of the key features is trying to match the remote sensing

data to the time the sample was collected. I mean, I think that field of spatial temporal mapping right now is really hindered by the fact that you don't have a balanced sample set in time. I mean, you rarely have it in space and you never have it in time. And so that this creates huge uncertainties, but I mean, there's several papers that have attempted this with varying levels of success. And I think as nations move to a more regimented, like repeated soil sampling campaigns, that type of information will be highly, and that that type of modeling will become more and more useful. Moving on into the future.

SPEAKER:
For sure.

SPEAKER:
Any other comments or thoughts, Michael, please.

MICHAEL:
Yeah. To kind of link the kind of 0.2 and 0.3 on the height on the spectral side, I think there's really untapped opportunities, to more mechanistically link the process-based models with remote sensing data, rather than relying on kind of, you know, using remote sensing data to make indirect maps of proxy things, and then trying to relate those proxy things to the mechanistic model. So, you know, some of the things that we've been doing are teaching land models to actually predict spectral signatures explicitly so that you can then assimilate the satellite data regardless of what spectral resolution the satellite happens to be spectral hyperspectral multispectral, whatever, you know, direct assimilate that data directly into the mechanistic models in a way that's kind of internally consistent and allows the remote sensing data to actually inform many more processes within the model, because you're now kind of fully you're really using all the information that's in the spectra.

SPEAKER:
Right. That's absolutely a very good way of doing it. And it's how, how much into it are you as a group? I mean, publishing.

MICHAEL:
Well, this, the grad student who did this graduated a couple of years ago is pushing chapters out, but he's now at NASA Goddard working on the NASA SPG mission. And I think there's no real need to, I was kind of disappointed. I haven't seen a good connection between that team and this discussion because, you know, they're the folks launching the hyperspectral satellite, they should really be involved in, you know, kind of some of this next generation dynamics. So, mapping efforts.

SPEAKER:
For sure. Colin.

SPEAKER:
COLLIN

SPEAKER:
Yeah. Given that Mike's sort jumped to 0.3, I'm going to take the opportunity as well. And I think, you

know, the new data streams for us that we're really excited about now are all of the molecular data streams, all the things we're learning about. So, biodiversity and you know, for a long time, we've had the model soil processes and measure them, not knowing who we're looking at. Like if we've tried to understand photosynthesis, but didn't know if we were looking at a Sequoia or a Sapna Moss who would probably gain a lot by generating that information. And that's what's happening right now. And so, I know those data are super complex and changing a lot, but there's also a lot of opportunity there. And I think, you know, as these data sets become huge, it's going to be really exciting.

SPEAKER:
Sure. Excellent. We have 10 minutes now at this stage, we can just I'll take any points across anyone's wants to share it. Mike, do you still have the hands or new question? OK.

SPEAKER:
So, I guess I'll jump in here. So, w what, what do we think we need from a dynamic soil information system in order to enable all of these sort of exciting model developments that we're talking about here? Like, how are, how networked are we asking it to be, and, and what, what are the barriers to getting the data that well network so that we can just pick and choose where we pipe into?

SPEAKER:
Seems like we need something on par with Lucas. I would agree $8 million a year, apparently, which isn't a big amount if you think, I mean, for it says Peanuts when I thought it was that much. It was ridiculous.

SPEAKER:
Yeah. I mean, I don't know what kind of challenges they face in terms of coordinating that effort and, you know, how they target their samples. I mean, I just know from doing field work myself in the past that you know, it's never as easy as it sounds in these types of meetings getting landowner agreements and, you know, coordinating that on a scale. So, I guess I'd be curious to, you know, talk to, you know, the Lucas folks and, you know, see how they make all that kind of work without even, you know, in the background.

SPEAKER:
Yeah. Well, one thing that Luca Montana, Ella made clear, it's interesting, they've done, and I think it could work here, but you know, the fact that you were looking at protecting its soil functions in Europe, it goes beyond properties. That's basically they can impose things and here you don't always have the level of freedom of legislating approaches to. So obviously the that, that context is in favor to Europe to even beyond, you know, collecting the data, but panelizing for a certain behavior and which we don't do here, here, that the farm said, you know, they own the land and they, still do what they want in a way. So that's easier from Europe to, to be moving in that direction freely across permissions. And it's pretty much imposed.

SPEAKER:
But I mean, everyone's kind of pointing to spectral as though that will be like the saving grace in terms of terms of like, just increasing the volume of data. I mean, still at the end, at the end of the day, someone's going to go out and collect that sample. So, either you get a huge participatory system or,

you know, people are, you know, willingly submitting their own data, which, you know, would have its own problems in terms of adding bias to the system, because, you know, probably some farmers are going to be more inclined to do that than others, or again, you know, you have to have you know, as we say, boots on the ground and have someone go out there and facilitate that.

SPEAKER:
Cathine.

CATHINE:
I was going to start throwing people under the bus and asking if a forest had anything to contribute from the benchmarking work that his groups done. Since I noticed that he's on call and then would also be interested in hearing from Melanie on the work that her group's done using accelerator enzyme kinetics, and trying to incorporate that into a process model and sort of what type of infrastructure would you wish for in a dynamic soil system to support both scales.

SPEAKER:
Thanks, Cathine. Can everybody hear me?

SPEAKER:
Yeah.

SPEAKER:
yeah, this is for us. So, I have to apologize that during the first part of the session, I was on another meeting with my DOE program managers. I thought that that was important, but so I did it but, but I did want to weigh in on, on a few things. So, in the previous session, now we talked a lot about file formats and making data accessible and understand what I think, you know, the investments that we make there are important. And I know that, we won't ever converge to a single format, right. But I think, it's important to have some convergence because, you know, the computer scientists and the people that work with these data can always write converters, but can we get it down to a handful of useful extensible formats that can then allow us to more easily work with the data? So, I think, I think that's worth doing, you know, one of the things that I really focused on is this idea of model benchmarking.

And so, trying to, you know, we're looking mostly at global scale and regional scale models over longer time periods, but we're building this no extensible benchmarking framework that we're even going to be using for, you know, runoff events and things like that. So, it's, it's sort of agnostic to the temporal and spatial scales. And so, you know, we're very interested in using the right data to evaluate the models in the right way for what we expect them to be doing and to inform us by providing more constraints. So, getting information, you know, being able to organize the soils data in a way that makes that kind of model data integration is really super important for us. And so, we're, we are in, in my project doing a few things. And one of those is with organized a soil carbon dynamic working group to focus on these issues. So, we're trying to invest also in, you know, the generation of the data and organizing it in a way that brings it to the models in a, in a useful format.

And we recognize that data are, as it has been brought out in this session, that data are measured in

different ways. There's a lot of kind of messy metadata associated with, you know, things like soil history and, and climate history that, that is basically stored in, in some kind of, you know, virtual memory in the soil. And so being able to find new ways of capturing that information and including it in models is really important. So, we, you know, through our working group and we're interested in other people who might want to join it we're interested in trying to make those data sets more accessible and synthesizing new datasets from multiple databases for the purpose of really confronting our admittedly poor models too, you know, with the goal of trying to make them better.

SPEAKER:
So, I'm hearing you say things like descriptions of land, use history, or talking about methods. And to me that says some kind of semantic tool that we can apply to describe data sets. So, you don't have to go and read the primary literature and it's associated with the dataset and you can scrape that from the directly from the metadata is that,

SPEAKER:
I think that would save everybody time and trying to use the data and apply it to various grades.

SPEAKER:
Well, we only have five minutes left and then we'll reconvene tomorrow and 11 any last-minute thoughts?

MELANIE MAYES:
This, this is Melanie. I can respond to Kathy's request if you, if there's time.

SPEAKER:
Yes, there is.

MELANIE MAYES:
OK. So yeah, so what we have been trying to do is to bring more microbial data into models and use it in a mechanistic way. So, when we first started this work, we were using microbial biomass from chloroform fumigation. That's effectively one group of microbes altogether. But we did have a couple of different sorts of enzymes because that's actually where the action has happening with, with grabbing different kinds of carbon. And so, since then, what we have done is which basically tells you about what genes are there in the microbial system. And particularly we focused on genes that gene or production of enzyme position processes are, this is however, as I said in the session earlier today, this is a lot of complexity to think about for a sole information data system. And so earlier what I recommended was effectively just a few things just to the, I thought the chloroform fumigation just of microbial biomass was something useful to have as is partitioning between fungi, bacteria, and archaea. So, to me, those seem like really standard things to do.

MELANIE MAYES:
Enzyme measurements are also really fairly standardized things that we can do. Looking towards the future you know, maybe, maybe Kathy is right, you know, we should potentially be thinking genes, for example, that might come out of a mitogenomic analysis. It's not clear to me how useful is 16 S analysis

is because that It takes you to different things. So, a focus on simply the genes, I think, is the most appropriate way to start thinking about including that kind of data. But as I think, as I mentioned earlier, I think, you know, the most important thing for this honestly, is that it is accessible in the way that safe, fair data is, so that any modeler should be able to basically go and call that information instead of trying to compile subsets of all of the information that is available.

So, that it can, can be there for them, because basically there's lots of data that's always being updated, but, you know, if you're always building a subset to come up with some kind of an empirical coefficient, you're sort of always getting farther away from what the data is really telling you, instead of closer to it, Kathy, I don't know if that answers your question of me or not, but there it is.

KATHY:
The sort of, so I, I would, if we've got, we've got another minute or so I'm taking it. So, if, if you could think about where, how, what, what are the current hurdles in your group between taking those, those data's? And I know there are actually other enzyme measurements and other microbial data that are now starting to come online. What, what are the challenges with moving from the registered dataset into your modeling space? Is it just that everything needs to be reprocessed and that disconnects it from the broader dataset is it that people aren't giving you robust enough descriptors for sort of how the data was collected, so it's not usable, like what, what are you generally running into when you're trying to use these data?

MELANIE MAYES:
So, in general, what I would say is that that it isn't Present out in terms of a data set that you could just simply call in your model. And I think that's how the, the w the topic of this workshop might be helpful. So, if, you know, if say, if all the genomic data from soils were out there somewhere, and I know that there are things like the earth microbiome project that kind of has that sort of information, but being able to connect that with a lot of the soils metadata and the soils information is what's needed. So, what I would see is that the, an information system needs can't be responsible for everything, but it needs to be able to call from different sources to find the data, to support the model. So really what you're looking for is a system where you can draw on multiple data sets to contextualize it, so that you've got the metagenomics from one study contextualized with the soil survey that happened next door so that you can examine it all in context. Is that what I'm hearing you say? Yeah, I think that would be fabulous. Yeah.

Awesome. All right. Bruno, I think we're at time, aren't we?

BRUNO:
Yeah, we are. We are indeed. Well, thanks everyone so much. And it was a very, very productive, engaging conversation. So, we'll see you tomorrow at 11. Please join us again for the last day and the report of the breakout sessions. So have a good afternoon. Thank you. Bye.

SPEAKER:
Thanks everybody.