Dr. Allan C.: Great. So today I'm going to be speaking about our work in environmental modeling, and how we're able to leverage satellite-based data sets to improve exposure modeling and the implications for environmental epidemiology. The premise of this is that getting exposures right matters. That when we're trying to reconstruct environmental exposures, we have to remember that the temperature outside, the quality of the air that we breathe, it changes dynamically. It can change over short distances and it can change very rapidly. And I think some areas saw that quite dramatically at the start of the pandemic with really big shifts in behaviors and commercial traffic and things like that, and not everyone. And also we've made many assumptions when we're started the field of understanding the impacts of extreme temperature. A lot of work has been based on data sets that are really grounded in where the local airports are because that's where the weather stations were. And most of us are not thankfully living at the airport. So you can have substantial variation in temperature. We've heard of the phrase the urban heat island, but in fact there's much more structure than that. There's an urban heat archipelago, a term that's also been used for several decades, substantial variation, even just thinking about block to block. And so how do we model that and then bring that into our studies to understand what people are really experiencing, where they live, go to school, where they sleep, et cetera. We need accurate estimates. And so my group builds geostatistical models using satellite data, and then we're investigating the implications of having improvements. What is the impact of reducing measurement error or thinking about having models that have less bias? When we're using satellite data, we use several different kinds of things that can be measured. Am I the ones that's supposed to [inaudible 00:02:15] okay, so we're using several different kinds of satellite data and then we're combining that with what may be, might be termed land use regression predictors. So from the satellite data, one parameter that we use a lot is a product called aerosol optical depth, which is the amount of light scattering related to particles in the entirety of the atmospheric column. And one of the limitations of this problem is that humans we're mostly doing our breathing in the bottom two meters of the atmospheric column. And if you have a lot of suspended particles, highly aloft, what you may end up with is a beautiful sunset, but if it doesn't mix down to the surface, then you're not necessarily going to breathe it. And so we have an imperfect proxy, right? Satellites are able to see that there's this light scattering that what the suns raised don't bounce directly back to the satellite, but they're not necessarily able to directly tell us what we're interested in, which is what am I going to breathe down at the ground level? Another parameter that we use from satellites is called land surface temperature. So thermal infrared remote sensing capabilities allow satellites to measure the skin temperature, the temperature, of the top few millimeters of whatever is the first thing that the sun's raise hit, whether that's bare soil or building rooftops or the tree canopy. But what we're experiencing when we walk around is an air temperature, which isn't exactly the same thing as surface temperature. The relation between those is complicated, right? Because it depends various thermal properties of radiant energy and whether there's wind and those kinds

of things. And then a third parameter that we're very interested in that satellites have done for a long time because of their applications in meteorology, is column water vapor, the amount of water that would precipitate out. And that uses principles of spectroscopy where these satellites have sensors that look for the absorption in very specific wavelengths in the electromagnetic spectrum that relate to water. And so because of the intense interest in meteorologic applications, satellites are very good at quantifying water vapor. But of course we see some opportunities to think about the relation of humidity and human health. Again, though, like the aerosol optical depth, the column water vapor is most often something that's for the whole atmospheric column. And we're living here just in the bottom two meters. So in all of these cases, we're getting these parameters from satellites, mostly NASA, NOAA, USGS, European Space Agency satellites. And we'll talk a little bit about some of the challenges and opportunities in satellite data specifically. So what we're doing is we're layering different kinds of information. So this is from some work that we did in Mexico City. Now, Mexico City had a legacy of a very, very bad air pollution. It's been improving over the years and a number of years ago, installed some air quality monitors. So they're measuring daily concentrations of fine particulate matter PM 2.5 with aerodynamic diameters less than 2.5 microns, but the, it's a pretty sparse network. You just have a few monitors throughout the city and there's a huge populace. And one of the characteristics that makes Mexico City a megacity is it's sort of sprawled out from the original administrative district. So on top of that ground monitoring network in a land use regression, you might layer on other kinds of information that's relevant particularly to the spatial distribution of sources of air pollution. So maybe you would put in information there about road networks or their utilization or traffic, some information about land use, population density, think about where the businesses are and things like that. Maybe you'd have some temporally varying features like meteorology to understand what are the conditions under which you have more secondary organic aerosol formation and things like that. And this is kind of how land use regression is often done. And now we're layering on top of that this spatio temporal layer from the satellite remote sensing. So a typical polar orbiting satellite might go overhead every day. You get a snapshot at the moment that that satellite goes overhead. There are also a set of satellites I'll talk about that are geostationary, so they're orbiting further up. And so they have to kind of zoom more if you think about the lens on the camera. But geostationary satellite will always be over the same longitude. And so it can take many snapshots over the course of the day. We'll talk about it. Pretty cool set of stationary satellites. So we're combining this kind of information and we use a geos statistical framework, which means that we're learning relationships where we have ground data so that we can try to infer using predictive modeling, what's happening where we do not have ground data. Some of the technical challenges are that these data sources include just staggering data volumes. NASA has this sort of planetary mission. They're producing data sets that go around the world. They're in sometimes esoteric formats, and there's just an enormous amount of it. There's an incredible amount of missing this. So you really have

to think about why is it that these data are missing aerosol optical depth, that light scattering? It can be missing for a number of reasons, but two of the most important ones are that it's nighttime and I continue to breathe at night and I hope that all of you do as well. And so that's a problem for us because we're mostly getting information during the time that things are going overhead. We also don't get it when there are clouds because we need light to be bouncing off the surface. We're not getting any information indoors. And indoors has incredibly important sources of air pollution. Now, ambient air pollution is, right now, what we're regulating in the us and we know that there's a strong relation between ambient concentrations and human health. It's still relevant, but boy, do I wish NASA would develop something that would peek indoors. They won't. Okay. I think one of the other challenges, and I'm not going to go into it so much today, but it's really almost like an earth science challenge, is that there's enormous measurement error. So those products that I mentioned, AOD, land surface temperature and column water vapor, they're generated through physics-based algorithms that are known as retrievals, which is using the properties of how light scatters or how thermal energy disperses. They're doing their best to infer what the truth is. But these are really indirect observations. And then we talked about how they could be for the entire column. So they're both measurement error, they're themselves noisy, and then they're not measuring directly the quantities that we're most interested in. So I have a big project that I'd be happy to talk about another time where we are using machine learning to clean up some of these retrievals with really, really fascinating results. We're able to really improve aerosol optical depth and we're developing some software that we're going to release that I think is going to ... Well, I have a lot of fun with that part, although that's upstream of the applications to human health. So one of the first projects that I worked on was we built the a spatio temporal model at a one kilometer resolution for Mexico City over a course of a couple years. And this model allowed us to leverage health data where we had information on cohorts of participants who had signed up for human health studies. We knew where they lived. And so the basic mechanism of assigning exposures of what we think people would breathe on average where they live, is based on a linkage from your residential address history. So the process is that if you're prospectively enrolling people in a study, you ask them about where they live, you stay up to date on whether they move, right? Or perhaps you're doing this a little bit retrospectively and you're asking them, where have you lived? When did you move? Now you don't necessarily, it's not so easy to remember all the addresses where you've ever lived. You go through a credit check, this is a challenge. But in general, if you can keep track of where people live, you start to link things in based on place, based on where they are. So this model, this kind of got me started in this field, and then we started doing things that were more complicated, more sophisticated, and building up bigger and bigger projects. So this is from a paper that was led by Daniel Carrion when he was a postdoc at Mount Sinai where we built a model of temperature that was hourly at a one kilometer resolution over the Northeastern and mid-Atlantic states of the US. And you see here on the right, a geostatistical model is predicated on

being able to train. You need some truth. So the ground truth that we used here was a network of data that's maintained by NOAA the National Oceanographic and Atmospheric Administration. So we have lots of these ground stations, and what you do is you train with that and you withhold some of them. And we use a process called cross-validation. In this case, we used a spatial cross-validation so we said, "How well would we do at this station if we knew nothing about what was going on around here, just inferring the concentration, the temperature at that location with the other information that we've trained into this model?" And here you see a list of some of the inputs that we use on the left, the kind of predictor space to kind of reconstruct what's happening at this NOAA set of stations. And then we compared a bunch of different predictive modeling approaches. So we said, how well can you do with sort of a simplistic geostatistical interpolation called inverse distance waiting? What about if we augment that with ordinarily squares regression? What if we start doing mixed effects because we say the relationship between the satellite based measurement and what's happening on the ground, that can vary day to day depending on meteorology and other things. Or what if we start using generalized additive models and we have smoothers so we can borrow information across space or across space time? And finally, what we found works really well in our hands is gradient boosting, which is a form of tree based models where you are really able to accommodate complex relationships between these predictors, that there's certain circumstances where if you've got a lot of inbound radiant energy and you've got a lot of surface pavement, then there there's going to be these kinds of interactions that we would think of interactions, that are occurring in that predictor space that help better explain this dependent variable. So we built this hourly model, and here's kind of a zoom in on what New York City looked like. This is midnight, so at nighttime, nighttime's a very important time for us because people need the physiologic respite of being able to cool themselves at night. And nighttime's also very important when you're using place-based exposure because that's when people are home and they give you their home address. So you've got this confluence where people are in their residence, they're home, but they're not necessarily able to get relief, particularly in these urban heat islands where you have dense surface materials and they hold onto that heat during the day and then they just don't cool down as much at night. So this is kind of zooming in and you see the island of Manhattan here in the middle of my slide. And then we overlaid on these larger grid cells, this rectangle, these represent the grid cells of the North American Land Data Simulation system, NLDAS2 which is a popular NASA product. It's what you would get when you were, if you go to the CDC's Heat and Health website, the environmental public health tracking website to look at climate related heat indicators. So the NLDAS2 it's a great product. It covers the US, it gives us lots of information about this historic exposure, but it's pretty coarse. So these larger rectangles here are like 11 by 14 kilometers. And what you see is that you're basically lumping, most people on the island of Manhattan are falling into one grid cell. And when I lived in Washington Heights kind of close to the middle of that, I wished on summer days that I were in the middle of the Hudson, but I wasn't.

It's kind of different there. So you lose out on that granularity, that spatial resolution. And what Daniel showed in his paper is that we were able to get much better predictive accuracy. So here we use the root means squared error as a metric of how well we're doing in our cross validation performance. And if you want to compare it proportionally, you should use the mean squared error. So we've got a third of the mean squared error of that NLDAS2, but I'll talk about the implication also of this sort of finer spatial resolution of being able to drill down on the neighborhoods where people live. I think anytime anyone tries to add to the zoom, I have to click on the screen again. So we built that and now that's on the left. That's the temperature model. We also had a daily Northeast PM 2.5 model where we're starting to cover these broader regions. And once you start to cover broader regions, it also lets you link not just to these cohorts like the one in Mexico City where people have specifically enrolled in, you've got a few hundred or a few thousand people, you can start to link into registries, population-based data sets. You can start to link into demographic information, thinking about pulling social variables from sources like the American Community Survey, ACS or the Census Bureau, others data sets like that. So on the right, it's a daily model because the EPA is mandating under the national ambient air quality standards that state and local agencies measure 24 concentrations of PM 2.5. So we're using that as our dependent variable, sort of midnight to midnight. And again, one of the challenges there of how informative can satellite data be, they're not telling you anything about what's going on at night, which is again, a time when I like to continue breathing. So we've got these predictive models, I gave you some examples, Mexico and then the northeast, we can do this reconstruction for the temperature model. We built it hourly, which is giving us information also about how much the temperature can vary within a day. And you start to see that you know can have asymmetry, it can get hot and stay hot, or it can change very rapidly. There are these dynamics and that NLDAS2 it is also hourly, but they're interpolating from a underlying North American regional reanalysis that really is a three hour time step. So maybe part of what we're doing is that we're actually finer resolution in both time and space, and we have these flexible machine learning models. We're using XG boost, but you also have to be really careful to avoid overfitting because all your data is coming from the same set of ground stations. And so different subsets of those data are not independent. Once you know that a station is in a neighborhood that runs hot or has higher air pollution, that that's kind of a attribute. And so you have to be careful because there's this important structure and non-independence in the training data, and you have to think about how to reflect that in your predictive modeling. Your observations are not independent identically distributed. So we have this paper from 2020 that was really about some of the challenges here. And then another one is in evaluating your model, we tend to put our meteorologic stations and our air pollution monitoring. They're much more prevalent in urban areas where a lot of the people live, of course. But it means that when you evaluate your model, some models kind of just borrow information across town and it's harder to know is this model appropriate everywhere? And we really were concerned

particularly about building a model that would do well where you don't have data from stations. So we've thought a lot about ways to use waiting and other approaches to really evaluate how well do we do when you're not near a station, how well do you do when you're in a suburb or you're in a rural area, right? Much of the United States lacks adequate ground monitoring. That's why we're building these models. So since then with some support from NIH's Echo program, they had an opportunity and infrastructure fund, we started building out national models. And so we built a national model for PM 2.5, and then I'll show you one for temperature. So on the left here is averaging the daily PM 2.5 across Meteorologic seasons in 2019. And this is a national model. And on the right, this is an example of zooming into the New York City area on the day that had the median air pollution concentrations for 2021 in this region. So we wanted this to be somewhat typical and this model, you'll start to notice when you look on the right, but I'll show you a little bit later, we actually switched some of our modality so that instead of focusing on this kind of rasterized or grided data, we started to build in both grided information and continuous fields, and now we're making point-based predictions. So we're able to really take advantage if you have someone's residential address, if you geocode it, which is the process of converting an address into its latitude and longitude, we're able to then make a prediction that we think is specific to that location. And that's really unique among large spatio temporal models, they tend to be grided. And so this is what it looks like for mean temperature on one day from our national temperature model. And this is sort of zooming in on New York City on that same day. But just note that the two color scales are reset to be sort of region specific on the span of colors that you can see. And if you're familiar with Manhattan, you can see even in the middle of the city there that we see that there's a tiny little bit of relief around Central Park where you have dense vegetation that's somewhat rectilinear there. So we'll talk a little bit more about resolution and its implications, but I want to move on from the models in a second. So these models are highly performant and I'll give you some examples of how they compare to other models. They're recent and updateable. So far they're through the end of 2021, but we're able to continue to add time to them. That's been a challenge for other models because for example, many PM 2.5 models rely on the EPA's national emissions inventory where you have a large industrial point source, industrial actors have to report their emissions, and then those all get pulled together into this emissions inventory, which is always released with a several year lag. So it's very hard to be able to predict using that kind of information if you want to be up to date. And we're not using that information. That lets us bring our models up to date so that we can be responsive to changing conditions. It's particularly useful when people have these kinds of cohorts or ongoing studies. They say, I need to be able to show the results of this study in less than five years. I need to be able to apply this. Yeah? [inaudible 00:22:56] So the question was how far back does it go? We started our models in the beginning of 2003, and that's sort of an era that that's maybe a golden age for this satellite remote sensing. We rely on the sensors that are known as MOTUS, which are very popular spectrometers that

are aboard to NASA satellites, Terra and Aqua, which went up in roughly 2000 and late 2002. So we went back to January of 2003. We've done models back to 2000, but 2003 is really a nice sweet spot because there's lots of ground data which is really relevant if you're going to do a geostatistical model and we have this sort of modern satellite era. There are newer satellites, which I'll talk about at the end, and they unlock new capabilities, but they don't go back quite so far in time. So there's sometimes they get questions about the 1990s, of course there was higher air pollution concentrations and more variability, which would improve statistical power and give you really relevant information if you're at a different part of the exposure response curve. But if we didn't have satellites, this particular approach, you'd have to go back to other things. So we have this daily model, it starts in. 2003 for now, it goes to the end of 2021, but we're continuing to update it. And the still improving part here is that we're able to integrate new ideas into this model. We've built it out as a completely automated pipeline data processing through machine learning. So that means that when new information is released, when the Environmental Protection Agency releases the air quality system information for another year, we're just able to put that into the pipeline and let the machine go. And I have some really nice servers and that helps. So I want to give some brief empirical comparisons, although I think it's more interesting to get to the implications of having better models. So one of the models that we compare with is the EPA's fused air quality surfaces with downscaling. So the EPA takes a 12 kilometer resolution chemical transport model called CMAC. And so this runs, they can run CMAC across the us. And the problem with a chemical transport model is they're trying to get at the physics of how things react and move, but sometimes the values that come out of it can have big differences from the ground truth. And so the fused air quality surfaces uses a basion approach to downscale that by tying it to the actual concentrations that were measured with these air quality system monitors. And one of the nice things about the EPA's product is that they release this, they have a tracked level. So census tracks are pretty small in terms of US population units. They have a tract level product that's national, that's daily, and they release this. So that created a really nice benchmarking opportunity for us because we're able to compare them, and also importantly, it's always hard to compare statistical models when you don't know what data someone already trained on because they already have the answer at that location. One of the great things about this EPA model is that they tell you explicitly which stations they trained on what data they used as an input. And so we were able to compare our model and the EPA's model at stations that they didn't train on, and we refit our model without using those stations at all. So we wouldn't learn anything about the bias. We didn't peek. We're sort of able to have a head-to-head validation. And in this case we're able to use in 2018, which was the most recent time that this FAQSD was available, we were able to look at 310 sites across the country. And many of those are reporting every day. Yes.

Speaker 2: So is it only PM 2.5 or are there any other [inaudible 00:26:50]

Dr. Allan C.: So this model is PM 2.5, and then I'll show daily minimum mean and maximum air temperature, and then I'll hint at some future directions. So when we make predictions from our model, because it can make point-based predictions, you make predictions wherever you want. When we make predictions, they're the same tracked centroids that the EPA is releasing, right? They're making predictions to the centroids. And then we have lower mean absolute error. The reason we use mean absolute error in instead of mean squared error here is that air pollution concentrations have a very long tail. They're never negative, but sometimes they're very, very high. And so we think that absolute error is a more appropriate metric. Here we have 16% lower mean absolute error when we average over all the years, and this is a metric I should say, that's also reweighted to account for the entirety of the country because if a bunch of these ended up being in New York City, you don't just want a metric that applies to New York City. So we sort of use an inverse waiting that's related to the amount of area that each monitor is the closest to, using what's called a veronoid diagram. And not only are we able to make the predictions the same tracks androids that they are, but we're able to make predictions anywhere we want, arbitrary points, because we have these continuous fields built into our model. And there's several advantages that one is that we're able to get at things like being near a roadway. And the model has learned that when it's nearer to a roadway, concentrations are higher. And even though it's not a dispersion model, there's no physics in it at all. It actually, we use some machine learning interpretation tools. We see that there's these strong gradients where the effect of being near a roadway dies off after just a few hundred meters, which is consistent with anyone who does monitoring on a transect around a roadway. And so we're able to get at those kind of local effects. And we're also, as I said, if you know where someone lives, we're able to make point-based predictions. So we also made predictions to the exact locations of those air quality monitors that we're testing on, and then we're able to get even lower mean absolute error. It's even more improved versus the FAQSD. So this is kind of a decomposition of how much better do we do when we try to play their game? And then what's the improvement of having that local information of being able to drill down within a centroid, sorry, within a tract, not just assuming that everyone is breathing the air at the cent, right? Some people will live on the edges. If you have to pick a representative place, the centroid isn't so bad. We also sometimes like the weighted population center that the census, decennial census produces. But I think this is a very cool kind of decomposition of the added benefit of not being as coarse of having something that's sort of at the subtract level. And when we compare our temperature model with three popular gridded models, PRISM, which produces and distributes a four kilometer grided product gridMET which brings in information from net NLDAS with the information from PRISM and Daymet, which is a one kilometer grided product from the Oak Ridge National Laboratory. What we see, so here we did, again, it's very hard to evaluate these because you don't know what people trained on. So we were able to get information from private weather stations from potentially nerdy folks who set up a wifi connected meteorologic station in their backyard. And so we use some

of that in our training, but we had leftover stations from the year 2014 onward. So we took a random draw of 10,000 of our extra stations that we hadn't trained on. We hadn't looked up them at all. And so we're able to compare these four models, ours, which is called the XG Boost, IDW Synthesis, Xis, an underappreciated Greek letter, and we're able to compare our model with these three grided models. And what we see is that we have a much lower mean squared error than each of these models. All right. I'm going to try to speed up. Just as a visualization, this is a very, very warm day nationally. This was one of the warmest days in 2021 and the New York area was a pretty warm day. And here on the left you see the one kilometer Daymet model. And on the right you see the kind of map that we're able to produce with Xis where we're able to make predictions on any arbitrary grid that we want, so this is kind of going at a finer grid here to look at the variation. So you have a combination of sort of effects that are related to higher accuracy, and then there can be sort of resolution effects that I'll talk about in a second. So what are the implications of having better models? So one of the things that we looked at, and Daniel really started this in our group, was looking at the relation between these exposures and the CDC's social vulnerability index, which is a metric that's scaled from zero to one, zero being the lowest vulnerability, one being the highest. And it's bringing together information across a number of different variables that the census has across a multitude of different domains. And so we were interested here, not in any kind of causal relationship, but we're just saying what is the relationship between the sort of cross-sectionally, how does temperature relate to social vulnerability? And we nested this analysis within counties because you need somewhat compact geographies so that you're not comparing sort of mountainous areas. So we fit a mixed effect model with county level intercepts and slopes where we're looking at the relationship between social vulnerability index as our main predictor and the air temperature as our outcome. What we found across all of the counties in that New England, this is using our one kilometer model first. What we found is that using our XG Boost based model here on the right, we saw a substantially higher difference in temperature on a heat wave day. So we picked sort of the hottest midnight, again, I'm going back to that nighttime temperature. Much, much higher than when you used that coarser, North American Land Data Simulation Systems because that's going to even within an urban area that's going to lump everyone together. So Mount Sinai right at the intersection of East Harlem and the Upper East Side. But if you say everyone from the Hudson River to the East River is experiencing the same thing, that that's just not going to be true. And what we saw when we zoom in on just a few specific counties looking within New York City or upstate, this is again from Daniel's paper, is that NLDAS on the left, it just doesn't have the same variation in temperature, so it's not able to accommodate, you have this essentially attenuation bias because you're assigning everyone the same exposure. When we look nationally, we've now done this with our Xis model, the new national temperature model. And when we take a very, very hot day, the minimum temperature on the hottest overall day in 2010, we pick that because we're going to pull some census data here. We do the same kind of analysis with

this mixed effect modeling. We see a difference of 0.69 centigrades Kelvin, and then these other grided products are just substantially attenuated. And in the case of these temperature models, they're not as coarse as the PM 2.5 example where the NLDAS is 11 kilometers by 14 kilometers, even Daymet, one kilometer. So some of this has to do with predictive accuracy and not just spatial resolution. When we do this analysis with PM 2.5, we calculated the 2018 average at every centrist track. So we construct daily predictions and then average them for the year 2018. And what we find with our model is that the fixed effect associated with vulnerability is essentially eight times greater than if you use the EPA's product. So we also, in collaboration, we've done some pilot analyses with a colleague who is able to link our data to large mortality data sets and then sees that this attenuation bias, if you use other sources of temperature data, you know, would underestimate the exposure response function that we see for extreme temperatures and mortality. So the theme there is that better exposure models reveal previously underestimated disparities and health impacts. And so this is going to help us advance climate and health epidemiology. And I'll give you a few examples briefly of the kind of work we've been doing. Some of the original support for going national was to link these exposures with the NIH's environmental influences on child health outcome study. And the hope would that be that using consistent exposure methodologies on these cohorts throughout the United States would help to better link in these significant environmental exposures with children's outcomes in a number of different health domains, from birth outcomes through neurodevelopment and many things like that. So that was the original impetus for scaling up our model to cover this many years and this large geographic range. In New York City, Mount Sinai has a very large biorepository with participants who have been genotyped. And so we were able to do some really neat gene environment interaction studies where we were looking at risk allele in APOL1 that's related to kidney disease. And we saw this sort of super additive risk where having higher air pollution and this risk allele was worse than having higher air pollution or the risk allele alone. And in this case, we were looking at sort of longer term air pollution because we're looking at the incidence of this chronic kind of kidney disease, the development of these very severe renal outcomes. And a study that we just went live with last week is a new preprint in Central Mexico leveraging publicly available records there and linking them into our updated Mexico City models where we're able to assign people exposures based on these sub municipal areas where they live. And one of the focuses, one the points of focus of this study has been to think about cause specific mortality, which is somewhat understudied once you leave the cardiovascular domain for air pollution, so we see these associations, we have this very large dataset with 1.5 million deaths from 2004 to 2019. We see these associations with hypertensive disease or with strokes or with chronic respiratory disease. And this has been well studied, but we're also able to look at these less well-studied outcomes where we're seeing associations with influenza, pneumonia, disease of the liver, renal failure, and of course doing the popular distributed lag, non-linear modeling. So we're starting to try to tease apart what is the relative timing in the relationship between air pollution and

mortality in this dataset. This is work led by Dr. Yvonne Gutierrez, a postdoc at Mount Sinai. And then Daniel led some work because there was a question whether we could take these case crossover designs that are so powerful in environmental epidemiology and apply them to preterm birth, where the challenge there is that the risk of delivering preterm changes a lot as you get later and later into your pregnancy. The likelihood of delivering changes a lot even over one month. And so Daniel led a simulation study of case crossover methods and whether you could appropriately reconstruct realistic effect estimates in relation to temperature when you have this rapidly changing outcome. What he found was that basically all the models were relatively unbiased. This is a short report in epidemiology that just came out last year and one of the fun things about this, we want to apply this case crossover methods with preterm birth, and I'll get to that, is that this analysis was done using sort of reproducible tools. One of them is the targets package and R, which is kind of an amazing tool for having this kind of reproducibility where you're able to get your exact same results back if you run it on a different computer or you start over from scratch. So that was a really fun project. Now we're linking to other large registries. So I mentioned in Mexico. So the focus of my New Ones proposal is to connect to Sparks, which is a statewide hospitalization registry. And so we're going to be using our models to connect with temperature, humidity, air pollution, and looking at spontaneous preterm births. So thinking about premature rupture of membranes and preterm early labor leading to these preterm births where we're able to go back in time. You wouldn't be able to do this with a prospective study because if you want to be able to go back in time, you know, you'd just need to enroll way too many people for an outcome like this to be possible. So leveraging large spatio temporal models and existing registries and health data sets is a really powerful approach to figuring out what happened. In future directions, I'm very excited about some of the new satellites that are coming along. This was a launch from last spring that put up one of the satellites that's monitoring the west coast. It's now called GOES 18. And so what's going to tell us about wildfires. It's going to tell us about some of the weird weather patterns that are being seen in the west and some of that wildfire, you have long range transport of lofted aerosols that can come east and they can down mix. They can influence our air quality here on the east coast as well. So one of our projects is reconstructing humidity. So we're using column water vapor, which is one of the remotely sensed metrics that I told you about. And so this is still work in progress. We don't have our final humidity model yet, but we're really interested in the interplay of humidity and temperature. Many of the meteorological indices that combine different elements, they were really developed maybe with occupational cohorts or predominantly older white male kind of folks. And we don't know whether that's the right way to think about the physiologic impacts of extreme weather and who is susceptible. So we're really interested in the temperature and humidity as a mixture. In this plot the color scale sort of reversed. The yellow is hotter here. This is using a newer product that's a higher resolution land surface temperature zoomed in on Mount Sinai if you're not familiar with the location. We're on the Upper East Side right

next to Central Park. And here you can see this incredible hotspot in the right is it's an MTA bus depot. So it's an asphalt covered building where they're just pumping heat out through the roof, right, because they have to have the ventilation for the buses and it's just lights up. And on the left we're able to distinguish the coolness of the Central Park reservoir from the thermal signature of grassy fields versus forest canopy. So starting to drill down, not from the one kilometer resolution, but really do I live in a building that is subjected to more heat? And that's going to have really profound impacts when you think about people's energy costs and the equity of cooling and the distribution of where we have tall trees and green space in our cities is horribly inequitable because of the legacy of residential segregation and structural racism. So being able to unpack that will help us to better assign exposures where people actually live. We're going to add gaseous pollutants, I'm super excited about the TEMPO satellite. I've been TEMPO Instrument, which is going up on a satellite in April. I've been working with them for five years. So you wait a long time before the actual launch. And so TEMPO is going to scan North America and include even down to most of central and southern Mexico hourly. And it's going to focus on gaseous pollutants, but also get aerosols. So this is going to be the first geostationary instrument. And geostationary means that it can take many scans per day. Everything is a mixture. We can measure many things from satellites. This is going to have profound implications when we think about the human health impacts and where we go from here in terms of human health adaptation. And of course, I want to thank my lab members, my current and former postdoc mentees, and I want to wrap up. Thank you so much.

Speaker 3: [inaudible 00:44:55] write algorithm so we don't have [inaudible 00:45:00]

Dr. Allan C.: And I'll repeat them.

Speaker 4: So the students have all [inaudible 00:45:25]

Dr. Allan C.: Yeah, that's great. So many of us are carrying around devices that are tracking our time activity patterns. And so you could combine that particularly with something like the hourly model where we start to say, "Where were you during the heat of the day?" In New York City, the population swells during the day. Or if you think about someone who works in the outdoors, a laborer, the conditions of where they work could be very different than where they live. So I think there are many opportunities to leverage that kind of very fine grain, time activity data using things like GPS devices. There are also some information on average commute kind of patterns and where people go the flows and fluxes between communities. But I think you're right that there's a lot of opportunities to leverage that with individual level information in health studies that have that, or some of us have opted into letting large corporations track us. Yep.

Speaker 3: Okay. Is there any [inaudible 00:46:33] question.

Dr. Allan C.: And my email's on the lower right if you have any questions. So

thank you. Thank you for your time.

Speaker 5: [inaudible 00:46:42] Thanks to our online audience for joining. And I do have a final announcement regarding Examiner. So [inaudible 00:46:55] we will add for the students, you will have an opportunity to meet to with the speaker and [inaudible 00:47:05] so because many of you have a classroom, so the current plan is, we have five slots open. So please do email me for [inaudible 00:47:16] and we can see how that works and then we can modify that later. And thank you. Thank you everyone.

Dr. Allan C.: Thanks so much.